Information and Organic Molecules: Structure Considerations via Integer Statistics

Daniel J. Graham†

Department of Chemistry, Loyola University of Chicago, 6525 North Sheridan Road, Chicago, Illinois 60626

Received March 18, 2001

Information in relation to organic molecules was investigated in a previous work (Graham and Schacht, *J. Chem. Inf. Comput. Sci.* **2000**, *40,* 187). The topic is given further consideration here with the help of integer statistics. Discussed are the ramifications of an integer variable Ω_t which quantifies the total number of binding complexions for an organic molecule. Offered is a statistical view of the maximum allowed number of independent regions $\boldsymbol{D}$ expressed by the molecule, dependent on Ω_t. We illustrate the distribution properties of $\boldsymbol{D}$ along with upper limit estimates of the regioinformation $\boldsymbol{\mu}$, also dependent on Ω_t. Integer statistics based on elementary number theory establish the key distribution properties of $\boldsymbol{D}$ and $\boldsymbol{\mu}$. In so doing, the traits distinguishing high regioinformation molecules are enumerated. The statistical approach encompasses all possible molecules and conditions, not just those reported to date in chemical databases. The aim is to view the regioinformation expressed by molecules in an alternative and general way.

I. INTRODUCTION

What questions does a chemist pose about an organic molecule?[1] The rudimentary include the following: what atoms are represented in the molecule and what is the empirical formula, complete formula, and Lewis representation? More complicated queries are as follows: what is the functionality,[1,2] are there asymmetric centers, in what ways is the molecule unusual,[3] and how does the molecule interact/react with other materials?

The above questions pertain to information contained in the molecule. The following query should also be made: what is the total number of binding complexions allowed for the molecule by the experimental conditions?[4] The answer is an integer quantity, henceforth designated Ω_t, which gauges the total number of messages expressible by the molecule. Experiment and chemical structure theory offer time-honored insight into the messages, their action, and complexity.[1,2,5] We focus here solely on the message number.

Why is Ω_t significant? A cornerstone of chemistry is that organic molecules behave to a very good approximation as composite systems. Such forms the basis for regioselectivity: the complexions of one molecular region can operate independently of other regions.[2,6] As a consequence, Ω_t plays a pivotal if obscure role in virtually all chemical structure/function machinery. As will be discussed, Ω_t determines the maximum allowed number of independent regions and the maximum regioinformation. The number of possible molecules, conditions, and thus Ω_t-values is infinite. And integer statistics based on elementary number theory reveal several critical features of the regioinformation. The details are provided in the remaining sections. This paper follows an investigation of information and the base code for organic molecules.[7] It contributes to a growing literature connecting the Shannon information to molecular properties.[8−13] Our focus here is shifted to the base considerations of regiostructure.

II. MOLECULAR REGIONS AND INFORMATION CONTENT

Organic molecules by and large present distinct regions, each expressing its own number and flavor of binding complexions. This enables (among other things) reagents such as OH^- and Br_2 to target particular atom groups (e.g. $-COOH$, $CH_2{=}CH$-) of receptor species while the integrity of other groups is maintained.[1,2,6] Such is a type of information processing which involves complexion (message) sensing of individual regions. Lewis diagrams, Dreiding models, and computer graphics offer ready visualizations of the regions and complexions. The composite nature of molecules enables everything from the data compression exercised when communicating chemistry (e.g. drawing Lewis diagrams) to the robust, diverse chemistry itself.

To view molecular information more generally, we consider any composite species with D = 2 independent regions, each demonstrating Ω_1 and Ω_2 number of binding complexions. The system as a whole evinces $\Omega_1 \cdot \Omega_2 = \Omega_t$ complexions as depicted via the upper panel of Figure 1. Let all information processing, e.g. interaction with a reagent, involve random sensing of complexions. The probability that *a particular complexion* will be sensed is $1/\Omega_t$. By contrast, the probability p_1 that a sensed complexion will be allied with region 1 is

$$p_1 = \frac{\Omega_1}{\Omega_1 + \Omega_2} \tag{1}$$

with an analogous expression for p_2. Information theory asserts that the Shannon measure μ (the number of bits) associated with the two complexion groups is[14,15]

$$\begin{aligned}\mu &= -K \sum_{i=1}^{i=D} p_i \ln(p_i) \\ &= -K\,[p_1 \ln(p_1) + p_2 \ln(p_2)]\end{aligned} \tag{2}$$

[†] Corresponding author fax: (773)508-3086; e-mail: dgraha1@luc.edu.

10.1021/ci0102923 CCC: $22.00 © 2002 American Chemical Society
Published on Web 01/24/2002

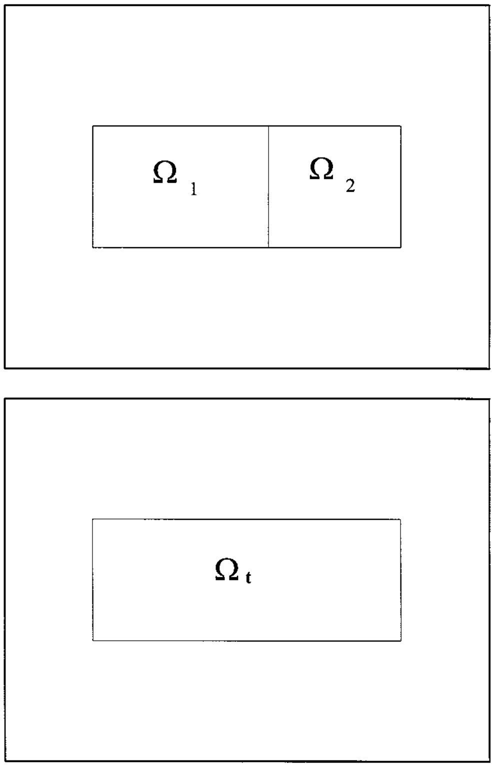


Figure 1. Schematic for a composite and single-region system (upper and lower panels, respectively).

where K = 1/ln(2) ≈ 1.44. Obviously, μ depends on the relative magnitudes of p_1 and p_2: the maximum value of μ in this example is 1 bit if $\Omega_1 = \Omega_2$. For this and all other cases, D sets the number of terms in the μ-summation. Broadly speaking, μ-values are enhanced when the Ω_i are nearly equal and are diminished when the Ω_i are markedly disparate. Note also that $\mu = 0$ bits for any species lacking in composite behavior, i.e., evincing only a single region (Figure 1, lower panel).

Models and chemical structure theory provide ways of estimating D and Ω_t. Integer statistics, however, offer insight regarding the maximum possible D and μ, given particular values of Ω_t. This insight charts an interesting and general perspective of regioinformation. It underscores which D- and μ-values are more likely than others, given a spectrum of molecules and conditions.

The fundamental theorem of arithmetic holds that any integer can be expressed as a unique product of prime numbers (2, 3, 5, 7, ...), e.g. 120 = 2•2•2•3•5.[16] A system expressing Ω_t complexions thus poses a maximum D equal to the number of prime factors of Ω_t. Note that we are taking all distinct regions of a molecule to be nontrivial, each asserting at least two complexions.

We will distinguish symbols for the maximum allowed number of independent regions and the maximum regioinformation using bold italics. Thus the quantity represented by $\boldsymbol{\mu}$ will follow as before from eq 2, yet it will depend exclusively on Ω_t having $\boldsymbol{D}$ number of prime factors, viz.

$$\boldsymbol{\mu}(\Omega_t) = -\mathrm{K}\,[p_1 \ln(p_1) + p_2 \ln(p_2) + ... + p_{\boldsymbol{D}} \ln(p_{\boldsymbol{D}})]$$
$$= -\sum_{i=1}^{\boldsymbol{D}} p_i \ln p_i \qquad (3)$$

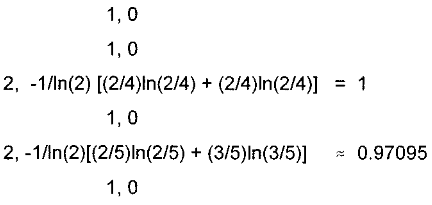
Table 1. ***D*** and $\boldsymbol{\mu}(\Omega_t)$ (Measured in Bits) for $\Omega_t \le 100$

Ω_t	D	$\mu(\Omega_t)$	Ω_t	D	$\mu(\Omega_t)$
2	1	0	51	2	0.6098
3	1	0	52	3	1.022
4	2	1.000	53	1	0
5	1	0	54	4	1.980
6	2	0.9709	55	2	0.8960
7	1	0	56	4	1.727
8	3	1.584	57	2	0.5746
9	2	1.000	58	2	0.3451
10	2	0.8631	59	1	0
11	1	0	60	4	1.887
12	3	1.556	61	1	0
13	1	0	62	2	0.3298
14	2	0.7642	63	3	1.457
15	2	0.9544	64	6	2.584
16	4	2.000	65	2	0.8524
17	1	0	66	3	1.199
18	3	1.561	67	1	0
19	1	0	68	3	0.8929
20	3	1.435	69	2	0.5159
21	2	0.8812	70	3	1.431
22	2	0.6193	71	1	0
23	1	0	72	5	2.292
24	4	1.974	73	1	0
25	2	1.000	74	2	0.2918
26	2	0.5665	75	3	1.548
27	3	1.584	76	3	0.8404
28	3	1.309	77	2	0.9640
29	1	0	78	3	1.122
30	3	1.485	79	1	0
31	1	0	80	5	2.192
32	5	2.321	81	4	2.000
33	2	0.7495	82	2	0.2713
34	2	0.4854	83	1	0
35	2	0.9798	84	4	1.778
36	4	1.970	85	2	0.7732
37	1	0	86	2	0.2623
38	2	0.4537	87	2	0.4488
39	2	0.6962	88	4	1.496
40	4	1.858	89	1	0
41	1	0	90	4	1.922
42	3	1.384	91	2	0.9340
43	1	0	92	3	0.7533
44	3	1.103	93	2	0.4305
45	3	1.539	94	2	0.2460
46	2	0.4021	95	2	0.7382
47	1	0	96	6	2.565
48	5	2.299	97	1	0
49	2	1.000	98	3	1.418
50	3	1.483	99	3	1.289
			100	4	1.863

with K again equivalent to 1/ln(2). By analogy with eq 1, p_i represents the ith prime factor f_i of Ω_t divided by the sum of all the prime factors:

$$p_i = \frac{f_i}{(f_1 + f_2 + ... + f_{\boldsymbol{D}})} \qquad (4)$$

The counting sequence Ω_t = 2, 3, 4, 5, 6, 7, ... leads to ***D***, $\boldsymbol{\mu}$ pairs as follows:


1, 0

1, 0

2, -1/ln(2) [(2/4)ln(2/4) + (2/4)ln(2/4)] = 1

1, 0

2, -1/ln(2)[(2/5)ln(2/5) + (3/5)ln(3/5)] ≈ 0.97095

1, 0

with a more extended list provided via Table 1. Clearly this application of the Shannon information hinges on the "mix"

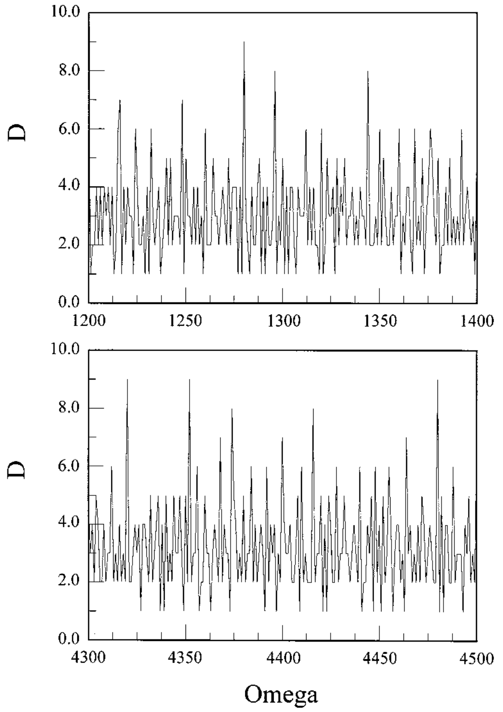


Figure 2. ***D***-versus-Ω_t computed over arbitrarily chosen intervals. Sequential points have been connected by lines.

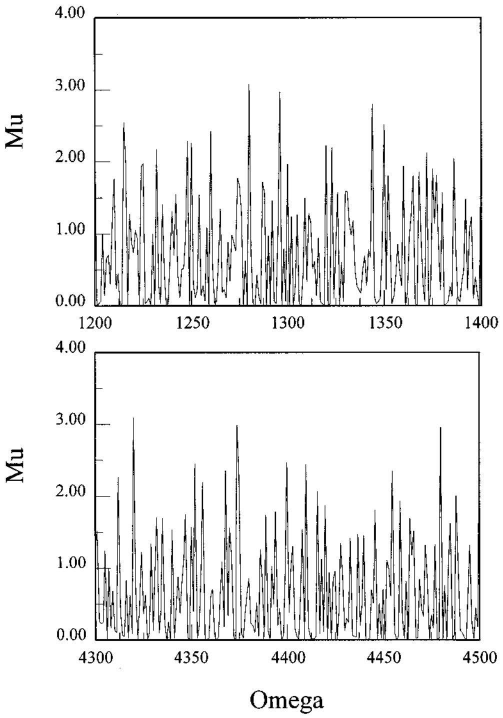


Figure 3. ***μ***-versus-Ω_t computed over the same intervals exercised in Figure 2. Sequential points have been connected by lines.

of prime factors for a given Ω_t. The Shannon information is an increasingly fascinating tool for assessing molecular structure properties.[8−13] To the author's knowledge, ***D*** and ***μ*** in relation to molecular information have not been considered in previous literature.[17]

D and ***μ*** offer interesting features. It can be shown that (1) $0 \leq \boldsymbol{\mu}(\Omega_t) < \mathrm{K}\ \ln(\Omega_t)$ bits for all possible Ω_t, (2) $\boldsymbol{\mu}(\Omega_t) = -\mathrm{K}\cdot 1\cdot \ln(1) = 0$ bits for all Ω_t prime ($\boldsymbol{D} = 1$), (3) $\boldsymbol{\mu}(\Omega_t) = \mathrm{K}\cdot\ln(\boldsymbol{D})$ bits for all $\Omega_t = \mathrm{f}^{\boldsymbol{D}}$, and (4) $\boldsymbol{\mu}(\Omega_t) < \mathrm{K}\cdot\ln(\boldsymbol{D})$ bits for all Ω_t in which ***D*** number of prime factors are not all equal to one another. $\boldsymbol{\mu}(\Omega_t)$ is enhanced for Ω_t having multiple, nearly equal prime factors. $\boldsymbol{\mu}(\Omega_t)$ approaches zero bits for Ω_t having a few widely disparate prime factors. The latter case is notably more prevalent since a large integer x is associated typically with ln(ln(x)) number of prime factors.[18] In other words, the typical ***D*** allied with Ω_t complexions is $\ln(\ln(\Omega_t))$, for example, $\ln(\ln(\Omega_t \approx 50 \times 10^6)) = 2.875.3$.

The nature of ***D*** and ***μ*** is apparent via a few computations and the resulting graphs. Figures 2 and 3 illustrate the ***D***- and ***μ***-functionality computed over arbitrarily chosen intervals. Here, sequential points have been connected by lines whereby a random variable dependence on Ω_t is indicated. By contrast, scatter plots (no connecting lines) such as in Figures 4 and 5 reveal multiple loci. The "curves" indicated in the lower portion of Figure 5, for example, derive from $\boldsymbol{\mu}(\Omega_t)$ with $\Omega_t = 2\cdot\mathrm{f}, 3\cdot\mathrm{f}, 5\cdot\mathrm{f}, ...$ The "edge" suggested in Figure 5 near $\boldsymbol{\mu}(\Omega_t) = \mathrm{K}\ \ln(2) = 1$ bit derives from $\Omega_t = f^2$ and $\Omega_t = \mathrm{f}_1\cdot\mathrm{f}_2$, $\mathrm{f}_1 \approx \mathrm{f}_2$. More subtle "edges" are manifest near $\boldsymbol{\mu}(\Omega_t) = \mathrm{K}\ \ln(3) \approx 1.585$ bits for $\Omega_t \approx f^3$; near K ln(4) ≈ 2.000 bits for $\Omega_t \approx f^4$, and so forth.

$\boldsymbol{\mu}(\Omega_t) = \mathrm{K}\cdot 1\cdot\ln(1) = 0$ bits for all Ω_t prime. It follows that the total number of ***μ***-values = K ln(1) bits over the Ω_t-domain can be approximated by a function $\pi(\Omega_t)$ well-known in number theory/integer statistics:[16−18]

$$\pi(\Omega_t) \approx \Omega_t/\ln(\Omega_t) \tag{5}$$

It can then be shown that the fraction ρ of ***μ***-values = K ln(1) as a function of Ω_t is approximately

$$\begin{aligned}\rho &= \frac{d}{\mathrm{d}\Omega_t}\frac{\Omega_t}{\ln(\Omega_t)}\\ &= \frac{\mathit{1}}{\ln(\Omega_t)} - \frac{\mathit{1}}{[\ln(\Omega_t)]^2}\end{aligned} \tag{6}$$

One can inquire as to the ***D***- and ***μ***-distributions for different neighborhoods of Ω_t. These also obtain from straightforward computations. The boxes in Figures 6 and 7, for example, mark the fraction of ***D***- and ***μ***-values ≤ d and x bits, respectively, for Ω_t of order 10^2. Circles, triangles, etc. mark analogous data for Ω_t of order 10^3, 10^4, etc. These data can be modeled in the usual ways via polynomials and other arbitrarily selected functions. For example, the boxes in the lower panel of Figure 7 mark the fraction of ***μ***-values ≤ x bits computed explicitly for $\Omega_t \approx 5 \times 10^7$; the continuous curve derives from the best-fit application of the following:

$$\boldsymbol{S}(\mathrm{x},\Omega_t) = [1 - \exp(\text{-x}/\tau)^\gamma] + 1/\ln(\Omega_t) \tag{7}$$

With τ,γ as adjustable parameters, a "stretched exponential" applies well to ***μ***-distributions computed over multiple decades of Ω_t.

Tables 2 and 3 list quantiles for the distributions which can be interpreted numerous ways.[19] For example, for Ω_t of order 10^2, 10^3, ..., 10^7, 5×10^7, the median value of ***μ*** is 0.693, 0.453, ..., 0.0866, 0.0633 bits (Table 3); the corresponding median ***D*** are 2, 3, ..., 4, 4 as reported in Table 2. Only about 5% of Ω_t-values in the neighborhood of 10^4 bespeak ***D*** > 6, ***μ*** > 2.04 bits; for $\Omega_t \approx 50 \times 10^6$,

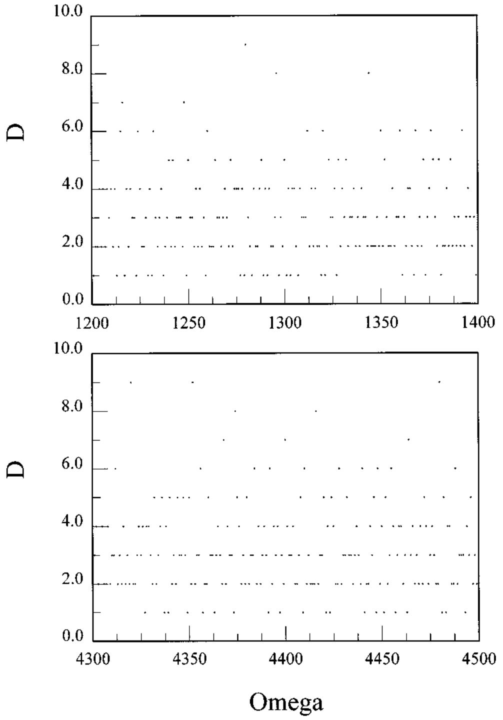


Figure 4. ***D***-versus-Ω_t: points (sans connecting lines) over intervals of Figures 2 and 3.

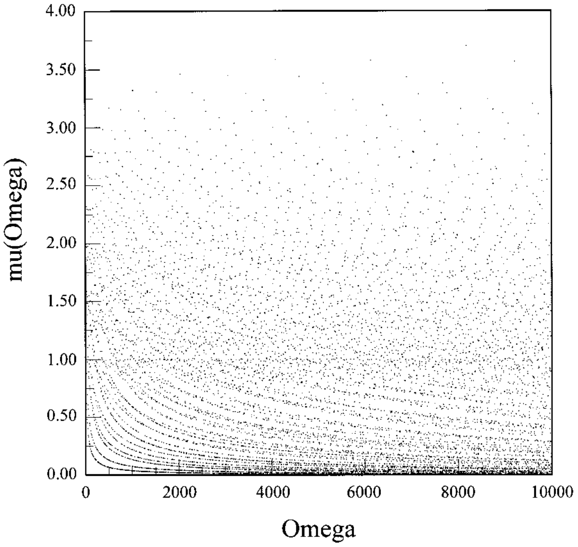


Figure 5. $\boldsymbol{\mu}$-versus-Ω_t computed over four decades.

approximately 1% of the possibilities exceed this $\boldsymbol{\mu}$-value. For Ω_t of order 50×10^6, ***D***- and $\boldsymbol{\mu}$-values ≥ 10, 2.75 bits are allied with the quantiles exceeding 0.995. There appear few changes in the ***D***-quantiles for $\Omega_t \approx 10^6 - 5 \times 10^7$.

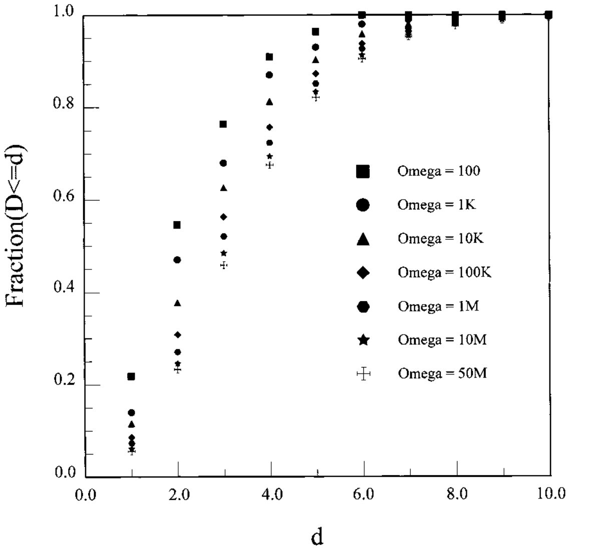


Figure 6. ***D***-distribution: the fraction of ***D***-values ≤ d for Ω_t of order 10^2, 10^3, ... 5×10^7.

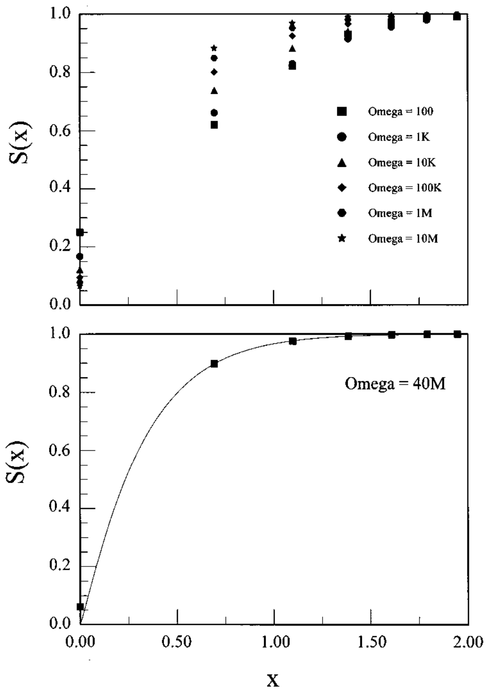


Figure 7. Upper panel: $\boldsymbol{\mu}$-distribution: the fraction of $\boldsymbol{\mu}$-values ≤ x bits for Ω_t of order 10^2, 10^3, ... 10^7. Lower panel: boxes mark the $\boldsymbol{\mu}$-distribution for $\Omega_t \approx 5 \times 10^7$; the continuous curve represents the best-fit application of eq 7.

Table 2. Quantiles for the Distribution of ***D***-Values

p	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99	0.995
Ω_t										
100	2	2	2	3	3	4	4	5	6	6
1K	2	2	3	3	4	4	5	6	7	7
10K	2	3	3	3	4	5	5	6	7	8
100K	2	3	3	4	4	5	6	7	8	9
1M	3	3	3	4	4	5	6	7	9	10
10M	3	3	4	4	5	5	6	7	9	10
50M	3	3	4	4	5	5	6	7	9	10

III. DISCUSSION

The number of carbon containing molecules registered in databases exceeds 16 million.[20] Each has the capacity to express information of its own and process that of other molecules. For organic compounds, there appear to be no limitations in size and diversity. There exist, for example, $20^{100} \approx 1.3 \times 10^{130}$ possible 100-moiety proteins constructed from 20 naturally occurring amino acids. 20^{100} far exceeds the current number of chemical database entries. Yet the possible proteins of this size are significantly more numerous if all the different stereoisomers are considered and/or additional amino acids are incorporated as building blocks. As is well-known, size is not everything when assessing molecular information. Rather it is the details of compart-

Table 3. Quantiles for the Distribution of μ-Values (Measured in Bits)

p	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99	0.995
Ω_t										
100	0.213	0.453	0.693	0.966	1.26	1.63	2.07	2.34	2.59	2.63
1K	0.146	0.279	0.453	0.679	0.959	1.34	1.89	2.26	2.65	2.72
10K	0.0933	0.180	0.306	0.479	0.713	1.06	1.61	2.04	2.51	2.59
100K	0.0533	0.113	0.199	0.326	0.519	0.819	1.35	1.81	2.36	2.45
1M	0.0266	0.0666	0.126	0.227	0.379	0.639	1.15	1.61	2.24	2.35
10M	0.0200	0.0400	0.0866	0.160	0.287	0.513	0.987	1.45	2.16	2.27
50M	0.0133	0.0333	0.0633	0.126	0.233	0.433	0.879	1.35	2.08	2.23

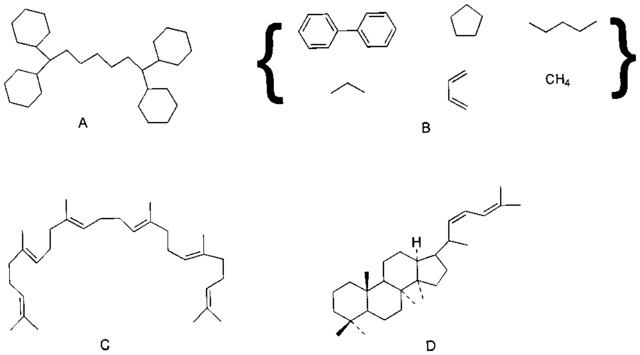


Figure 8. Lewis diagrams for arbitrarily selected isomers of $C_{30}H_{50}$. The diagrams contained in brackets collectively signify a van der Waals isomer.

mentalization which appear most significant in carbon-based chemistry. These ideas are implicit in Figure 8. Only carbon and hydrogen atoms are represented in selected Lewis diagrams of $C_{30}H_{50}$. Yet the internal organization is radically disparate: significantly different regioinformation is associated with the squalene (**C**) and steroid compound (**D**), compared with the alkane (**A**) and van der Waals cluster (**B**).

Regioinformation issues are usually addressed via diagrams such as in Figure 8. These two-dimensional graphs, their diagrammatic classes (e.g. terpenes, alkanes, etc.,), and local topological characteristics (single bonds, aromatic rings, etc.) have formed the basis of many quantifications of combinatorial library information[21−23] and the Shannon information for organic libraries[8−10] and biopolymers.[11−13] In a combinatorial peptide library, for example, where each product consists of eight amino acids with eight different possibilities for each, a typical compound might be labeled in binary form as 101 001 110 110 010 111 010 100; the regiodiversity can thus be quantified at 24 bits. In a topological approach, descriptors such as the number of single bonds have demonstrated Shannon information values of a few to several bits, depending on the library nature. Interestingly, natural product libraries express higher information contents for select descriptors, compared with synthetic libraries.[8−10]

In this paper, we have elected a different pathway by examining the impact of simple integer properties. We have viewed regioinformation as integral to molecular behavior, whereby the maximum allowed internal partitioning is tied to an integer variable. There appears no reason for a single value of Ω_t to be associated with a given molecule due to experimental variations (temperature, pressure, solvent, etc.). A statistical approach thus points only to what is likely and unlikely, yet also what can be exceedingly special about carbon-based systems.

In previous work, the base code for organic molecules was found to be low in information and high in redundancy.[7] Here we report that for possible molecules and conditions, the maximum allowed partitioning and regioinformation are also typically low. $\boldsymbol{D}$ and $\boldsymbol{\mu}$ pose maximum case scenarios. One finds the median values of $\boldsymbol{D}$ and $\boldsymbol{\mu}$ (Tables 2 and 3) to be < 5 and < 1 bit, respectively; such is the case for the nine orders of Ω_t examined for this paper. This is an important result. It demonstrates that distinction may be conferred for any species expressing more than five independent regions and a regioinformation exceeding one bit. For example, one can estimate via bond angle and excluded volume considerations the number of binding complexions of squalene molecule (**C**, Figure 8) to be in the range 10^4-10^5.[24] Under conditions where this species operates as a composite of six units (e.g. isoprenes), such is affiliated with the 0.90 quantile, expressing a regioinformation >2 bits. The statistical perspective offers that squalene is a very atypical organic molecule. One notes the rareness of cases where $\boldsymbol{D} = 1$, $\boldsymbol{\mu} = 0$ bits. For $\Omega_t \approx 50 \times 10^6$, for example, eq 6 offers that the percentage of $\boldsymbol{D} = 1$, $\boldsymbol{\mu} = 0$ possibilities is −5%. The statistical perspective asserts that at least nominal amounts of internal partitioning and regioinformation are likely for organic molecules and conditions.

The results expressed via integer statistics make intuitive sense. As part of the investigation, several sets of 30−120 hundred organic molecules tabulated in the *Handbook of Chemistry and Physics* were chosen randomly.[25] $\boldsymbol{D}$-values were subsequently assigned for each species based on organic chemistry training, in particular lab synthetic experience. For example, a $\boldsymbol{D}$-value of 1 was assigned for cyclopentane molecule; $\boldsymbol{D}$ was taken to equal 2 for benzoic acid, 3 for salicylic acid, and so forth. To be sure, such assignments are qualitative and somewhat arbitrary—they vary depending on the chemist and the molecule. Yet the cumulative distributions obtained from such exercises proved qualitatively consistent with ones posed by integer statistics. In Figure 9, typical distribution results for estimated $\boldsymbol{D}$-values (CRC data, hollow boxes) and for $\Omega_t = 100$ (integer statistics, filled boxes) are plotted. In all cases, one observes the most significant fractions of organic molecules to assert two or three regions; the cumulative distributions express an exponential dependence.

Ω_t offers an intriguing descriptor for molecular similarity and diversity studies.[8−10,21,22] Two or more molecules are possibly similar (in the chemical sense) if their Ω_t express common prime factors and/or numbers of prime factors; the molecules are markedly dissimilar otherwise. Further, the diversity within a molecular library is maximized when the greatest number of prime factors are represented in the set of Ω_t-values. One observes such features graphically in

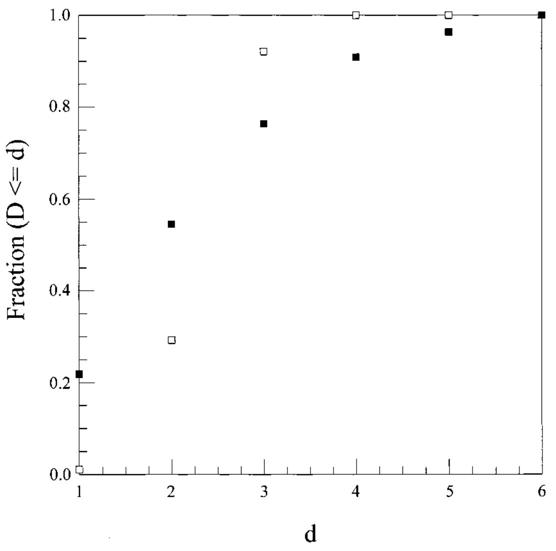


Figure 9. ***D***-distribution: the fraction of ***D***-values ≤ d for set of 120 organic molecules chosen randomly from the *Handbook of Chemistry and Physics* (hollow boxes). ***D***-values have been assessed based on principles of functional group chemistry and skeletel integrity. Results from integer statistics for $\Omega_t = 100$ (cf. Figure 6) are plotted using solid boxes.

Figures 4 and 5: Ω_t with common prime factors and/or numbers of prime factors are allied with significant clusters and edges of ***D***- and ***μ***-values and a wide spectrum of Ω_t expresses widely diverse ***D*** and ***μ***. It is unlikely, of course, that Ω_t can serve in lieu of the more usual robust descriptors such as fragment substructures, bond indices, and functional groups. The molecules *n*-butanol and *n*-butanoic acid demonstrate similar sets of Ω_t-, ***D***-, and ***μ***-values but markedly different responses to sodium hydroxide reagent. Ω_t offers its greatest utility as an initial screening device in similarity/diversity studies.

Two additional aspects are interesting. ***D*** and ***μ*** express hallmarks of random variables (Figures 3 and 4).[26] This demonstrates that small changes in expressed Ω_t can be allied with significant changes in the regioinformation. A molecule could thus demonstrate significant composite behavior only under conditions favoring Ω_t having multiple factors. It is unfortunate there is no procedure for converting diagrams such as in Figure 8 to exact Ω_t. This is all the more compelling reason for a statistical view of molecular regioinformation.

Multiple chiral centers ensure alliance of a molecule with the highest quantiles. Steroid compounds (e.g. **D**, Figure 8) typically evince several centers,[27] while other natural products and biopolymers can exhibit tens of asymmetric centers.[27,28] Molecular crystals of select space groups (e.g. $P2_12_12_1$) can evince of order 10^{23} chiral unit cells.[29] This points to an important feature of regioinformation, because for each asymmetric center, there automatically exists two binding complexions. This means that Ω_t for highly chiral molecules is associated with multiple factors, all based on the smallest of the prime numbers. In an exemplary case where Ω_t is of order 2^N, N being the number of chiral centers, $\boldsymbol{D} \approx$ N and $\boldsymbol{\mu} \approx$ K ln(N) bits. This means that for a protein with N $\approx$ 100, $\boldsymbol{\mu} \approx$ K ln(100) $\approx$ 6.644 bits. In terms of regioinformation statistics, such a molecule is allied with the $(1-10^{-30})$ quantile. The statistical perspective reinforces one's present conceptions of high information molecules, namely natural products, biopolymers, and crystals with high degrees of chirality. The lowest information molecules would be those lacking functional groups and asymmetric carbons.

Integer statistics offer a framework for molecular regioinformation. This framework quantifies the maximum case scenarios along with their distribution properties. The author was intrigued by the subject, first via the base code for organic molecules,[7] and later by the bridge with integer statistics. Current efforts are focused on the statistics of information processing in molecular recognition and chemical reaction events. The results will be reported in a forthcoming paper.

ACKNOWLEDGMENT

Support by the Department of Chemistry, Loyola University of Chicago is appreciated. The author is grateful for helpful discussions with Professors David Crumrine, Jon Zerkowski, Derek Nelson, and Gordon Ramsey and the technical assistance of Eric Gage. The questions and comments of an anonymous reviewer are also appreciated.

REFERENCES AND NOTES

(1) An excellent discussion of molecular structure queries can be found in the classic: Shriner, R. L.; Fuson, R. C.; Curtin, D. Y. *The Systematic Identification of Organic Compounds*, 5th ed.; Wiley: New York, 1964.

(2) le Noble, W. J. *Highlights of Organic Chemistry*; Dekker: New York, 1974.

(3) See, for example: Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular Diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1−10.

(4) The terms "binding complexions" are used in the most general sense, referring to the total number of binding states (electronic, conformational, etc.) expressed by a given molecule.

(5) See, for example: Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599−3601. Mil'man, B. L. A Complexity Measure for Chemical Compounds. *J. Struct. Chem.* **1988**, *29*, 957−960.

(6) House, H. O. *Modern Synthetic Reactions*, 2nd ed.; Benjamin: Phillipines, 1976.

(7) Graham, D. J.; Schacht, D. V. Base Information Content in Organic Formulas. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 187.

(8) Godden, J. W.; Stahura, F. L.; Bajorath, J. Variability of Molecular Descriptors in Compound Databases Revealed by Shannon Entropy Calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 796.

(9) Stahura, F. L.; Godden, J. W.; Bajorath, J. Distinguishing Between Natural Products and Synthetic Molecules by Shannon Descriptor Entropy Analysis and Binary QSAR Calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245.

(10) Godden, J. W.; Bajorath, J. Differential Shannon Entropy as a Sensitive Measure of Differences in Database Variability of Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *41*, 1060.

(11) Schneider, T. D. Measuring Molecular Information. *J. Theor. Biol.* **1999**, *201*, 87−92.

(12) Schneider, T. D. Theory of Molecular Machines. I. Channel Capacity of Molecular Machines. *J. Theor. Biol.* **1991**, *148*, 83−123. II. Energy Dissipation from Molecular Machines. *J. Theor. Biol.* **1991**, *148*, 125−137.

(13) Schneider, T. D. Information Content of Individual Genetic Sequences. *J. Theor. Biol.* **1997**, *189*, 427−441.

(14) Brillouin, L. *Science and Information Theory*, 2nd ed.; Academic: New York, 1962.

(15) Feynman, R. P. *Feynman Lectures on Computation*; Hey, A. J. G., Allen, R. W., Eds.; Addison-Wesley: Reading, MA, 1996.

(16) Hardy, G. H.; Wright, E. M. *An Introduction to the Theory of Numbers*, 4th ed.; Clarendon: Oxford, 1971.

(17) An extensive discussion of prime numbers and allied number theoretic functions is given by Riesel, H. *Prime numbers and computer methods for factorization*, 2nd ed.; Birkhauser: Berlin, 1994.

(18) Kac, M. *Statistical Independence in Probability, Analysis, and Number Theory*; Mathematical Society of America, distributed by Wiley: New York, 1959.

(19) See, for example: Conover, W. J. *Practical Nonparametric Statistics*, 3rd ed.; Wiley: New York, 1999.

(20) Brown, T. L.; LeMay, H. E., Jr.; Bursten, B. E. *Chemistry, the Central Science*; Prentice Hall: Upper Saddle River, NJ, 2000; Chapter 25.

(21) Willett, P. Using Computation Tools to Analyze Molecular Diversity. In *A Practical Guide to Combinatorial Chemistry*; Czarnik, A. W., DeWitt, S. H., Eds., American Chemical Society: Washington, DC, 1997.

(22) Terrett, N. K. *Combinatorial Chemistry*; Oxford Press: Oxford, 1998; Chapters 5 and 9.

(23) Zhao, P. L.; Zambias, R.; Bolognese, J. A.; Boulton, D.; Chapman, K. Sample Size Determination in Combinatorial Chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 10212.

(24) Flory, P. J. *Statistical Mechanics of Chain Molecules*; Oxford University Press: Oxford, 1989.

(25) *Handbook of Chemistry and Physics*; Weast, R. C., Ed.; Chemical Rubber Co.: Cleveland, 1972. This resource lists data for 13 600 organic compounds: "those of wide application in teaching, industry, medicine, and research."

(26) Freund, J. E. *Mathematical Statistics*, Prentice Hall: Englewood Cliffs, 1962.

(27) Roberts, J. D.; Caserio, M. C. *Basic Principles of Organic Chemistry*; W. A. Benjamin: New York, 1965; Chapter 30.

(28) See, for example: Suh, E. M.; Kishi, Y. *J. Am. Chem. Soc.* **1994**, *116*, 11205.

(29) See, for example: Kahr, B.; McBride, J. M. *Angew. Chem., Int. Ed. Engl.* **1992**, 31, 1. McBride, J. M.; Bertman, S. B. *Angew. Chem., Int. Ed. Engl.* **1989**, 28, 330. McBride, J. M. *Angew. Chem., Int. Ed. Engl.* **1989**, *28*, 377.

CI0102923