Base Information Content in Organic Formulas

Daniel J. Graham* and David V. Schacht†

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

Received November 29, 1999

Three questions are addressed concerning organic formulas at their most primitive level: (1) What is the information per atomic symbol? (2) What is the level of system redundancy? (3) How are high-information formulas distinguished from low-information ones? The results are simple yet interesting. Carbon chemistry embodies a code which is low in base information and high in redundancy, irrespective of database size. Moreover, code units associated with halocarbons, proteins, and polynucleotides are especially high in information. Low-information units are more often associated with simple alkanes, aromatics, and common functional groups. Overall, the work for this paper quantifies the base information content in organic formulas; this contributes to research on symbolic language, chemical information, and molecular diversity.

INTRODUCTION

An interesting type of code analysis runs as follows. The symbols contained in a large body of material, e.g., a lexicon, are examined for their frequency of occurrence. The information content (measure) is then computed as

$$i_1 = -K\sum p_j \ln p_j \qquad (1)$$

where $K = 1/\ln(2) \approx 1.44$ and p_j is the occurrence probability of each jth symbol. The summation is taken over all symbols whence $\sum p_j = 1$. The result i_1 signifies the number of bits per symbol in the first order of analysis.[1]

This approach accommodates multiple correlations. Regarding pairs, for example, one can examine the probability $p_i(j)$ of observing a certain ith symbol in an aggregate, given that a particular jth symbol is already present. A formula similar to eq 1 applies:

$$i_2 = -\mathrm{K}\sum p(i,j) \ln p_i(j) \qquad (2)$$

where $K = 1/\ln(2)$ as before and $p(i,j) = p_i p_i(j)$. A double summation is carried out over i, j with the result i_2 based on conditional probabilities.

The i_n values depend on how the probabilities are defined. For example, one must choose whether or not to pay attention to symbol sequences. It is always true, however, that the maximum information i_0 is expressed when all the symbols occur with equal probability. That real codes embrace disparate probabilities means that the number of bits per symbol is less than i_0—often considerably so.[1,2] A true measure of information obtains from analyzing correlations at increasingly higher order—pairs, triples, quadruples, etc. It is a rule that $i_0 \geq i_1 \geq i_2 \geq ...$; with increasing correlations among symbols, a reduction occurs in the information.[1−5]

Note that "information" is being viewed only in a statistical sense and not in its usual context of "facts" and "data". The idea is closely related to statistical entropy[1,4−6] and is important for the following reasons. A code with information measure less than i_0 is redundant, and redundancy is integral to a code's operation and efficacy.[1−5]

Among other things, redundancy underpins the sparseness of viable symbol aggregates. English, for example, exercises 26 symbols—more if spaces and punctuation are counted. Yet a mere handful of aggregates is utilized compared with the number possible. In the case of six-letter arrangements, for example, there are $6^{26} \approx 1.7 \times 10^{20}$ possible, yet a fluent practitioner of English requires only ca. 10^4 words total, regardless of letter number.[7] Redundancy is not deleterious to code operations. Rather it reflects the need for error correction and noise suppression during communication. Redundancy can be defined formally as follows:

$$R = 1 - i_{\mathrm{lim}}/i_0 \qquad (3)$$

where i_{lim} is the limiting value of the information measure. The quantities i_{lim}, i_0, and R have been investigated for several systems, beginning with Shannon's analysis of codes and communication devices.[8,9] Excellent treatments of information theory and its applications are found in several places.[1−5]

Written languages offer one genre of codes. Chemical formulas offer another with striking analogies. Words of a language are constructed via characters of a particular type and quantity. Chemical formulas are also specified by enumerated symbols. Both language and atomic symbols belong to immutable sets. Words are combined to form sentences; the analogous holds for chemical formulas positioned in reaction sequences. In language, new words are introduced to reflect and affect culture. An analogous situation holds for chemical formulas, for example, in the synthesis of new compounds.

For chemical formulas, we were intrigued by questions normally directed at coding systems: (1) What is the information per symbol? (2) What is the level of system redundancy? (3) What distinguishes high-information units from low-information ones? The purpose here is to address these questions as confined to carbon-based chemistry. This work contributes to a growing body of research on symbolic language, chemical information, and molecular diversity.[10−18]

* To whom correspondence should be addressed. Fax: (773) 508-3086. E-mail: dgraha1@luc.edu.

† Present address: Northwestern University, Evanston, IL 60208.

10.1021/ci990182k CCC: $19.00
Published on Web 06/21/2000

Table 1. Code Symbols Employed by the CRC Organic Lexicon, First-Order Occurrence Probabilities, and Eq 5 Tally[a]

symbol	occurrence probability	eq 5 tally	total % i_1
H	0.487	0.506	30.5
C	0.384	1.04	62.4
O	0.0785	1.32	79.8
N	0.0262	1.46	88.0
Cl	0.0106	1.53	92.2
Br	4.52×10^{-3}	1.57	94.4
S	3.70×10^{-3}	1.60	96.2
F	2.68×10^{-3}	1.62	97.5
I	1.61×10^{-3}	1.63	98.3
Si	5.29×10^{-4}	1.64	98.6
P	3.85×10^{-4}	1.64	98.9
As	2.82×10^{-4}	1.65	99.1
B	1.69×10^{-4}	1.65	99.2
Se	7.19×10^{-5}	1.65	99.3
Na	3.16×10^{-5}	1.65	99.4
K, Ca, Ge, Sn, Sb, Hg, Te, Pb, Fe	$<3.00 \times 10^{-5}$	1.66	100

[a] Atomic symbols based on double Roman letters are treated as single characters.

The results are simple yet interesting. At their most primitive level, organic formulas offer a system low in information and high in redundancy, irrespective of database size. High-information formulas are allied with (among other things) halocarbons and the building blocks of proteins and polynucleotides. Low-information formulas are more often associated with simple alkanes, aromatics, and common functional groups. Importantly, work for this paper leads to quantification of the base information content.

ANALYSIS

The contents of several representative databases were utilized. The principal ones were assembled from the Table of Physical Properties of Organic Compounds published in the *Handbook of Chemistry and Physics*, 52nd ed., by the Chemical Rubber Co. (CRC).[19] This source lists data for 13 600 organic compounds of 1−216 carbons: "those of wide application in teaching, industry, medicine, and research". Each molecular formula was encoded in a digital format. Multiple abridged versions (90%, 75%, 50%, etc.) were compiled from randomly selected entries of the parent files.

At its most primitive (base) level, a chemical formula signifies an aggregate state of atoms of a particular number and identity. Our analysis was aimed only at this level; hence, no account was taken of the atom connectivity and relative spatial coordinates. For example, the formula for the methyl benzoate molecule was encoded simply as $C_8H_8O_2$, with the same entry registered for *o*-acetylphenol. By not distinguishing structural isomers, the study was akin to making no discriminations among anagrams, for example, the English words "eat", "tea", and "ate". This is not a small point obviously as the isomers exceeded 50 for many listings, e.g., $C_8H_8O_3$.

In conducting the zero-order analysis, the database symbols C, H, O, N, F, ... were first identified (Table 1); i_0 followed simply from imagining each to occur with equal probability. For the 24 symbols employed by a chemical database—or any other coding system—one has

$$\begin{aligned} i_0 &= -K\sum p_j \ln p_j \\ &= -[1/\ln(2)]\sum(1/24)\ln(1/24) \\ &= 4.59 \end{aligned} \quad (4)$$

For the first-order analyses, the occurrence probabilities were examined in detail. Sample results based on 100% of the CRC lexicon are presented in Table 1, listed in descending order of the symbol probabilities. In the case of Cl, for example, one arrives at a probability of 0.0106 by dividing the number of Cl occurrences in the CRC organic lexicon (3679) by the total number of symbol occurrences (347 779). For the coding system H, C, O, ..., i_1 then follows from eq 1:

$$\begin{aligned} i_1 &= K\sum p_j \ln p_j \\ &= [-1/\ln(2)][0.487\ln(0.487) + \\ &\qquad 0.384\ln(0.384) + \\ &\qquad [0.0785\ln(0.0785) + ...] \\ &= 1.66 \end{aligned} \quad (5)$$

Table 1 includes running tallies of the above summation and percentage of the limiting i_1 value. As would be anticipated, terms allied with H and C contribute most significantly. Interestingly, 12 symbols, H, C, ..., As, account for >99% of the first-order information. Note that, throughout the analysis, atomic symbols based on double Roman letters such as Cl, Br, and As were treated as single characters. The number of code symbols thus corresponded to the number of atoms represented in the database.

Greater attention to aggregate structure was given in second-order studies and beyond. Thus, symbols composing each database string were next examined pairwise to see whether given a C, there was another C in a string, given a C there was also an H, given a C there was also a Br, and so on. This procedure was carried out for all pairs CC, CH, CBr, ..., leading to conditional probabilities and i_2 (eq 2).

The higher order analyses were extended versions of second order. In fourth order, for example, each formula was probed to see whether given a C, and two H's, there was also a Br. This was extended to all quadruples, and i_4 was computed using a modified version of eq 2. Ultimately, the correlations were investigated up to the 10th order. This obtained answers to questions 1 and 2 regarding the base information content and system redundancy.

Question 3 concerned the relative information for different formulas. To address this, a variation of the base analysis was performed. Diverse formulas of interest were expanded and appended with the (arbitrarily chosen) symbol "@", e.g., $C_8H_8O_2$ was converted to CCCCCCCCHHHHHHHHOO@. The new string was then scrambled randomly and repeatedly, e.g., CCCCCCCCHHHHHHHHOO@ → CHCHOHC@HCOHHHCHCC, HCCCOHHCCCHH@OHHCHC, HHH@HHCHCCCCCOOCHHC, etc.

Each scrambled arrangement was considered according to sequence. In the case of CCH···C@H···CC, for example, a given database was first examined to see which entries contained at least one C, second, at least one C and another C, third, at least one C, another C, and an H, and so on. With each symbol examined, entries which failed to meet

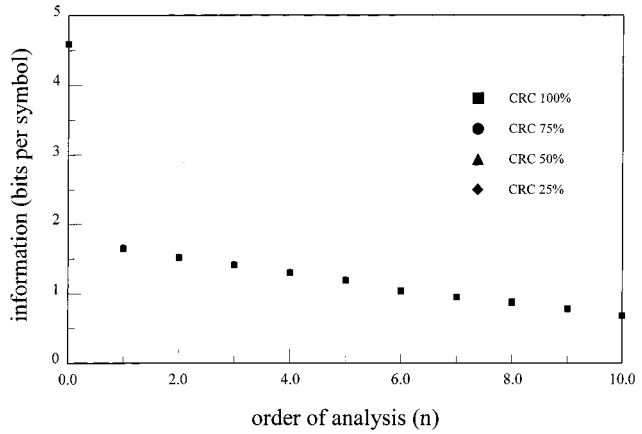

Figure 1. Information measure versus analysis order. Data marked by squares are based on the CRC lexicon of 13 600 formulas. Data marked by circles, triangles, and diamonds are based, respectively, on randomly abridged versions of 75%, 50%, and 25%. The degree of symbol overlap shows the results to be insensitive to database size.

the matching criteria were eliminated from the data pool. When @ appeared for consideration, all database listings not composed of $8 + 8 + 2 = 18$ symbols—equivalent to the number of atoms represented in the original string $C_8H_8O_2$—were eliminated. Thus, the formula state for the dichloromethane molecule CH_2Cl_2 would survive the first symbol round for CCHCHOHC@HCOHHHCHCC, but not the second. For the alternative arrangement HCHCHCCHCOHC@CHCHHO, CH_2Cl_2 would survive the first three rounds (H, C, and H) but would be eliminated at the fourth. Representing a molecule containing $1 + 2 + 2 \neq 18$ atoms, CH_2Cl_2 would be eliminated during any round in which @ appeared for consideration. The idea at work is that, with each symbol examined, more information about the base formula state is provided. Such information was probed while asserting no bias or particular strategy, hence the repeated random scramblings of the character sequences.

For different formulas, the results were examined over several hundred to a few thousand string randomizations to yield $N_D(n)$. This quantity is defined as the average number of applicable (i.e., surviving) formula states in a database as of the nth random examination round. The information scaled as the natural logarithm of the number of surviving states, with the results for different formulas compared using reduced variables. Thus, the quantities $\Xi = -\ln[N_D(\upsilon)/N_D(0)]$ versus $\upsilon = n/n_t$ were considered: $N_D(\upsilon=0)$ signifies the initial number of formula states in the data pool, while n_t represents the total number of atomic symbols in a formula string of interest. Throughout the analysis, careful attention was paid to the statistical properties of all $N_D(n)$ and related quantities.

RESULTS AND DISCUSSION

Figure 1 illustrates typical i_n as a function of analysis order n. Included are results based on an entire CRC formula lexicon (13 600 states) and randomly abridged versions of 75%, 50%, and 25%. In all cases, i_0 demonstrated a value of 4.59 bits/symbol, with a steep decline of 1.66−0.682 bits/symbol registered over the first to tenth order. The bits per symbol are reduced at increasingly higher order; this means that redundancy is a significant attribute of the coding system.

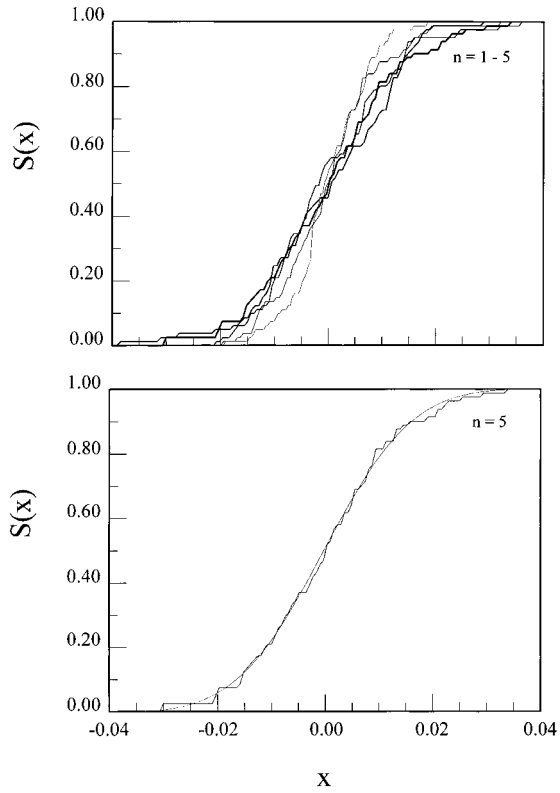

Figure 2. Cumulative probability distribution functions $S_n(x)$ for the random variable $X = (N_D(n) - \langle N_D(n)\rangle)/N_D(0)$. The upper panel illustrates typical results for $n = 1-5$ based on 2500 random scramblings and database probes for CCCCCCCCCHHHHHHHHOO@; for storage efficiency, results were collected using 85 bins. The plots for different n values are distinguished by use of a line-blackening scale proportional to n: darkest for $n = 5$ and lightest for $n = 1$. The lower panel illustrates the cumulative distribution observed for $n = 5$ plus the best-fit normal distribution (thinner line).

The degree of symbol overlap moreover shows the results to be insensitive to the database size.

For strings such as CCCCCCCCCHHHHHHHHOO@, there exist 19!/8!8!2!1! = 37 413 090 possible symbol arrangements; for expanded $C_6H_{14}N_2O_3$@ (hydroxylysine and isomers), there are 5.35×10^{11} distinct arrangements. Fortunately for the Ξ studies, $N_D(n)$ and related quantities demonstrated all the hallmarks of a normal random variable.[24] Figure 2 attests to this as compiled from 2500 randomizations/database probes for $C_8H_8O_2$@. Illustrated in the upper panel are the cumulative probability distribution functions $S_n(x)$ observed for the random variable X:

$$X = \frac{N_D(n) - \langle N_D(n)\rangle}{N_D(0)} \tag{6}$$

$S_n(x)$ gauges the likelihood of observing all possible values of $X \leq x$; typical results for the $n = 1-5$ (first to fifth database examination rounds) are illustrated. One finds the width of each distribution to depend not only on the symbol grouping particulars, but also on the value of n. All of the cumulative distributions—a total of 19 derived from the random scramblings and database probes of CCCCCCCCCHHHHHHHHOO@—adhere closely to the normal distribution. The normalcy property is conveyed in the lower panel for the $n = 5$ distribution.

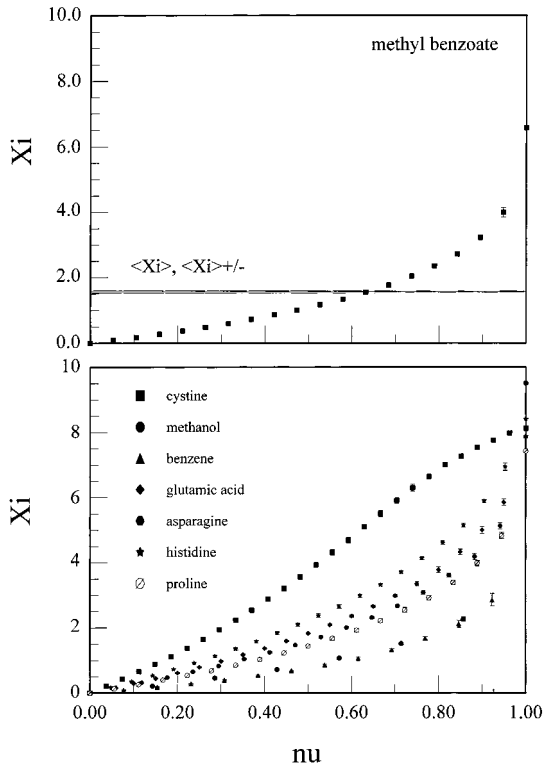

Figure 3. Ξ versus υ for diverse formulas. The upper panel shows results for $C_8H_8O_2$ (methyl benzoate and structural isomers), with the vertical lines indicating plus/minus the estimated standard deviation. The horizontal lines at the ordinate scale values 1.56, 1.64, and 1.48 mark, respectively, $\langle\Xi\rangle$, $\langle\Xi\rangle^+$, and $\langle\Xi\rangle^-$. The lower panel contains results for diverse formulas, with the names of prominent structural isomers used for convenience. Corresponding $\langle\Xi\rangle$ and windows bracketed by $\langle\Xi\rangle^+$ and $\langle\Xi\rangle^-$ appear in Table 2.

Because of results such as in Figure 2, one is able to estimate very accurately the means allied with the n value using small (500−3000 symbol arrangements typical) randomly-selected portions of sample space. In each case, significance testing (Student's t test, sign tests, and Mann−Whitney rank-based test) demonstrated no significant differences at the ≥95% confidence level between the means of different, randomly-selected portions of the sample spaces.[25]

Figure 3 presents Figure 2-type information in a different way. Plotted in the upper panel is Ξ versus υ observed for $C_8H_8O_2$@ (methyl benzoate and structural isomers). The vertical lines allied with each data point (filled boxes, one for each examination round) indicate ±1 standard deviation about the average. The horizontal line at the ordinate value 1.56 indicates $\langle\Xi\rangle$, or the average of all the estimated $-\ln[N_D(\upsilon)/N_D(0)]$ over the interval $0 \le \upsilon \le 1$. The thinner neighboring lines at 1.63 and 1.48 indicate $\langle\Xi\rangle^+$ and $\langle\Xi\rangle^-$; these represent the average of estimated $-\ln[N_D(\upsilon)/N_D(0)]$ plus/minus the estimated standard deviation for each string population.

The lower panel contains analogous results for diverse formulas identified (for convenience) by the names of prominent structural isomers. The corresponding $\langle\Xi\rangle$ values appear in Table 2 along with error windows bracketed by $\langle\Xi\rangle^+$ and $\langle\Xi\rangle^-$. Specifically one finds Ξ for C_6H_6 and CH_4O (benzene, methanol, and isomers) to demonstrate minimal ascent with υ; $C_6H_{12}N_2O_4S_2$ and $C_6H_9N_3O_2$ (cystine and histidine) show much more pronounced inclines. Clearly with each symbol examined in strings composed randomly from $C_6H_{12}N_2O_4S_2$, significantly greater information results compared with that for C_6H_6 and CH_4O. As with i_n values, the Ξ results also proved insensitive to the database size. The Ξ functionality rather hinged on the formula state attributes, not surprisingly the mix of heteroatom (N, O, S, F, Cl, ...) symbols.

Results for additional formulas appear in Table 2 in order of descending $\langle\Xi\rangle$. Figure 4 offers the same results pictorially: the abscissa and ordinate denote, respectively, the Table 2 formula index and $\langle\Xi\rangle$ values. For each species, vertical bars locate $\langle\Xi\rangle^+$ and $\langle\Xi\rangle^-$. Overall, one finds $\langle\Xi\rangle$ for N- and halogen-containing units to exceed values computed for

Table 2. String Variables and $\langle\Xi\rangle$ Computed for Diverse Formulas[a]

index	string	prominent isomer	$\langle\Xi\rangle$	$\langle\Xi\rangle^+ - \langle\Xi\rangle^-$
1	CHBrClF	CHBrClF	4.69	0.110
2	$C_6H_{12}N_2O_4S_2$	cystine	3.96	0.107
3	$C_5H_5N_5O$	guanine	3.54	0.0758
4	$C_5H_5N_5$	adenine	3.44	0.103
5	$C_6H_{14}N_4O_2$	arginine	3.36	0.0942
6	$C_8H_{10}N_4O_2$	caffeine	3.22	0.0886
7	$C_6H_{14}N_2O_3$	hydroxylysine	2.92	0.0958
8	$C_6H_4Br_2$	dibromobenzene	2.91	0.145
9	$C_{17}H_{21}NO_4$	cocaine	2.88	0.0764
10	$C_3H_7NO_2S$	cysteine	2.84	0.106
11	$C_5H_{11}NO_2S$	methionine	2.84	0.140
12	CH_3	methyl radical	2.80	0.0726
13	$C_{18}H_{21}NO_3$	codeine	2.78	0.0766
14	$C_4H_5N_3O$	cytosine	2.77	0.150
15	$C_6H_9N_3O_2$	histidine	2.77	0.0800
16	$C_{17}H_{19}NO_3$	morphine	2.73	0.0732
17	$C_6H_4Cl_2$	dichlorobenzene	2.66	0.124
18	$C_6H_{14}N_2O_2$	lysine	2.60	0.0830
19	$C_5H_{10}N_2O_3$	glutamine	2.55	0.0954
20	$C_{11}H_{12}N_2O_2$	tryptophan	2.52	0.0838
21	$C_4H_8NO_3$	asparagine	2.47	0.0652
22	$C_4H_4O_2N_2$	uracil	2.44	0.0926
23	$C_5H_9NO_4$	glutamic acid	2.37	0.0748
24	$C_5H_6N_2O_2$	thymine	2.35	0.0108
25	$C_4H_7NO_4$	aspartic acid	2.33	0.0760
26	$C_9H_{11}NO_3$	tyrosine	2.21	0.0828
27	$C_4H_9NO_3$	threonine	2.15	0.0774
28	$C_3H_7NO_3$	serine	2.15	0.0660
29	$C_{20}H_{14}$	tryptycene	2.09	0.107
30	$C_6H_{13}NO_2$	leucine	2.08	0.0588
31	$C_6H_{13}NO_2$	isoleucine	2.08	0.0868
32	$C_5H_{11}NO_2$	valine	1.98	0.0782
33	$C_5H_9NO_2$	proline	1.97	0.0848
34	CH_4O	methanol	1.97	0.0410
35	$C_3H_7NO_2$	alanine	1.97	0.0794
36	C_2H_5N	glycine	1.96	0.110
37	$C_9H_{11}NO_2$	phenylalanine	1.95	0.0824
38	C_2H_4	ethylene	1.83	0.0662
39	C_2H_6	ethane	1.68	0.0600
40	$C_7H_6O_2$	benzoic acid	1.60	0.0816
41	$C_9H_{10}O_2$	ethyl benzoate	1.58	0.0728
42	C_3H_6O	acetone	1.57	0.0714
43	$C_8H_8O_2$	methyl benzoate	1.56	0.0784
44	C_4H_{10}	*n*-butane	1.54	0.0816
45	C_6H_{14}	*n*-hexane	1.52	0.0622
46	$C_{10}H_8$	naphthalene	1.50	0.0860
47	C_4H_6	butadiene	1.44	0.0686
48	C_6H_6	benzene	1.43	0.0778
49	C_7H_8	toluene	1.40	0.0710
50	C_6H_{10}	cyclohexene	1.36	0.0790

[a] Included are names of prominent structural isomers and the difference between $\langle\Xi\rangle^+$ and $\langle\Xi\rangle^-$. Data are listed in order of decreasing $\langle\Xi\rangle$. Figure 4 offers an illustrated version of the data. Results do not change significantly (95% confidence level or better) upon examining different, randomly selected portions of sample space.

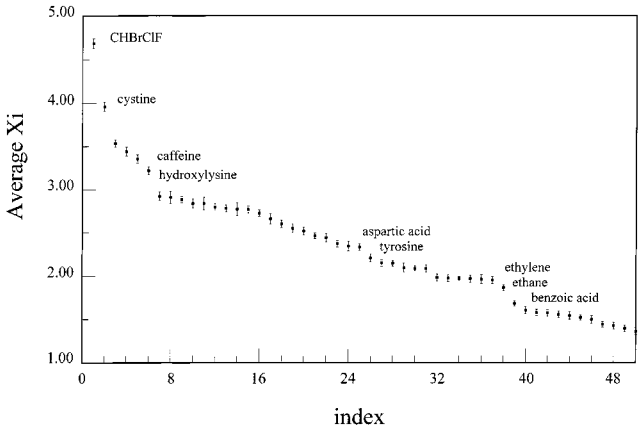

Figure 4. Table 2 data in pictorial form. The abcissa denotes the Table 2 formula indices while the ordinate scale refers to $\langle\Xi\rangle$ values; vertical bars locate $\langle\Xi\rangle^+$ and $\langle\Xi\rangle^-$.

hydrocarbons and common functionalities. The results span a range of diverse natural product, aromatic, and solvent formulas, with windows provided by $\langle\Xi\rangle^+$ and $\langle\Xi\rangle^-$ gauging the overlap for different species.

As is well known, organic formulas are most often specified using a mix of symbols and graphs; they are interpreted in terms of electronic interactions and other physical properties. The path adopted here is different, however, in that graphs, interactions, and physical properties have been ignored altogether. While off the beaten trail, this focus and its results are interesting (in our opinion) given the parallels between chemical formulas and coding systems in general. Information theory offers a vehicle to examine chemical structure apart from functionality, in this case at the most primitive level of carbon atom aggregation.

The most significant results concern the base system information and redundancy. For written languages (such as used for this paper), i_0 and $i_{\lim}$ are typically 4.7 and 1.9 bits/symbol; a typical system redundancy is thus ca. 0.60.[1,2,20] Organic formulas offer a markedly different case, expressing at their most primitive level ca. 0.68 bit/symbol and a redundancy of $1-0.68/4.59 \approx 0.85$. One notes these values to be upper and lower limits, respectively, as the order of the analysis was finite.

Qualitatively, the results do not surprise. The equivalence of C_6H_5- and Ph-, Me-, CH_3-, etc. arises only because certain aggregate states are expressed with great frequency. Inference and data compression are thus feasible. Quantitatively, the information and redundancy values intrigue. All of the formulas examined for this work are allied with real molecules. It is interesting that the base code tied to the molecules and their impact is among the most redundant. Redundancy is the antidote for noise and error in communication. The redundant nature of the chemical code is consistent with the effective transmission of molecular messages.

Another result concerns $\langle\Xi\rangle$ for diverse species. Notably, formulas allied with halocarbons, amino acids, and heterocyclic bases (and their structural isomers) demonstrated greater $\langle\Xi\rangle$ compared with hydrocarbons with common functionalities. One learns, for example, that the genetic code is sustained by formula units of relatively high information.[21] The analysis showed i_n, R, and $\langle\Xi\rangle$ to be insensitive to the database size. As with the genetic and other language codes, the root essentials of organic formulas can be expressed by databases of very limited size.

In summary, organic formulas were examined in a framework of simple code analysis. The results quantify a system at its base level, one which can be represented by much abridged databases. The authors were drawn to this study by the analogies involving symbolic language and the chemistry of molecular recognition and diversity.[10−18]

ACKNOWLEDGMENT

Support by the Department of Chemistry, Loyola University of Chicago, is appreciated. D.V.S. is grateful for the support of the REU program administered by the National Science Foundation. We appreciate helpful discussions with Professors David Crumrine, Jon Zerkowski, and Derek Nelson. We are grateful to two anonymous reviewers for comments and criticism.

REFERENCES AND NOTES

(1) Brillouin, L. *Science and Information Theory*, 2nd ed.; Academic: New York, 1962.
(2) Herdan, G. *The Advance Theory of Language as Choice and Chance*; Springer-Verlag: Berlin, 1966.
(3) Goldie, C. M.; Pinch, R. G. E. *Communication Theory*; Cambridge University Press: Cambridge, 1991.
(4) Morowitz, H. J. *Energy Flow in Biology*; Ox Bow Press: Woodbridge, CT, 1979.
(5) Feynman, R. P. In *Feynman Lectures on Computation*; Hey, A. J. G., Allen, R. W., Eds.; Addison-Wesley: Reading, MA, 1996.
(6) Lin, S. K. Molecular diversity assessment: logarithmic relations of information and species diversity and logarithmic relations of entropy and indistinguishability after rejection of Gibbs paradox of entropy. *Molecules* **1996**, 57−67.
(7) Marsh, G. P. *Lectures on the English Language*; Scribner, New York, 1862.
(8) Shannon, C. E. Prediction and entropy of printed English. *Bell Syst. Tech. J.* **1951**, *30*, 50−64.
(9) Shannon C. E.; Weaver, W. *The Mathematical theory of Communication*; University of Illinois Press: Urbana, IL, 1949.
(10) Gordon, J. E.; Brockwell, J. C. Chemical inference 1. Formalization of the language of organic chemistry: generic structural formulas. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 117−134.
(11) Gordon, J. E.; Brockwell, J. C. Chemical inference 2. Formalization of the language of organic chemistry: generic systematic nomenclature. *J. Chem. Inf. Comput. Sci.* **1984**, *23*, 81−92.
(12) Johnson, M., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
(13) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(14) *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1976.
(15) Turner, D. B.; Tyrrell, S. M.; Willette, P. W. Rapid quantification of molecular diversity and selective database. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 18−22.
(16) Agrafiotis, D. K. Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841−851.
(17) Wiswesser, W. J. Historic development of chemical notation. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 258−263.
(18) Read, R. C. A new system for the designation of chemical compounds 1. Theoretical preliminaries and the coding of acyclic compounds. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 135−149.
(19) *Handbook of Chemistry and Physics*, Weast, R. C., Ed., Chemical Rubber Co.: Cleveland, 1972.
(20) Zipf, G. K. *Human Behavior and the Principle of Least Effort*; Addison-Wesley: Cambridge, MA, 1949.
(21) Calvin, M. *Chemical Evolution*; Oxford University Press: Oxford, 1969.
(22) See, for example: *Principles of Molecular Recognition*; Buckingham, A. D., Legon, A. C., Roberts, S. M., Eds.; Blackie Academic and Professional: London, 1993.
(23) 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.
(24) Freund, J. E. *Mathematical Statistics*; Prentice-Hall: Englewood Cliffs, NJ, 1962.
(25) Conover, W. J. *Practical Nonparametric Statistics*, 3rd ed.; Wiley: New York, 1999.

CI990182K