Information Content in Organic Molecules: Quantification and Statistical Structure via Brownian Processing

Daniel J. Graham,* Christopher Malarkey, and Matthew V. Schulmerich

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

Received February 16, 2004

Information and organic molecules were the subject of two previous works from this lab (Graham and Schacht, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 187; Graham, *J. Chem. Inf. Computer Sci.* **2002**, *42*, 215). We delve further in this paper by examining organic structure graphs as objects of Brownian information processing. In so doing, tools are introduced which quantify and correlate molecular information to several orders. When the results are combined with energy data, an enhanced informatic view of covalent bond networks is obtained. The information properties of select molecules and libraries are illustrated. Notably, Brownian processing accommodates all possible compounds and libraries, not just ones registered in chemical databases. This approach establishes important features of the statistical structure underlying carbon chemistry.

I. INTRODUCTION

An organic molecule transports mass, charge, energy, and angular momentum. As an electronic device, it also expresses information and processes that carried by neighboring molecules. Under routine storage conditions, a molecule's information fluctuates about an average value. In the presence of reagents and other instability sources, molecular information is transformed, increasing or decreasing along the way.

As is well-appreciated, there exists an infinite number of possible organic molecules, with only a fraction investigated thus far. Present-day archives maintain data for -20×10^6 compounds, yet the valence isomers of $C_{30}H_{62}$ alone pose -4×10^9 possibilities.[1] Clearly organic molecules, like other component-based systems (e.g. languages and transistor circuits), offer astonishing variety, complexity, and information capacity. This facet impacts several areas of chemical research, including combinatorial design, drug discovery, and molecular library organization.[2]

Information and organic molecules motivate ongoing investigations in this laboratory, with two works having been reported in this journal.[3,4] In the first, the information expressed by individual molecules was probed at a base level, that is, apart from details regarding the atom configurations.[3] In a follow-up study, molecular information was examined at a regio-level, this time taking structural attributes into account with the help of integer statistics. The capacity for an organic molecule to operate as a collection of independent, functional units was a central theme.[4]

We delve further into the subject in this and a companion paper. Departure is made from our previous studies by taking explicit account of atom (carbon, hydrogen, etc.) and bond configurations allowed by chemical structure theory. Further, a Brownian processing/random walk model is developed which is able to quantify and correlate molecular information to several structure orders. This information is a "read-only" type and is thus independent of specific laboratory conditions and chemical reactions. We apply this model to select molecules and libraries (sets of molecules) so as to view chemical information in a new and general way. As in our forerunner studies, statistical methods underpin all the key conceptuals. With the latter in mind, we also chart some of the informatic distribution properties of organic molecules. These are important to understanding the statistical structure underlying carbon chemistry.

The paper which follows examines several transformation issues surrounding molecular information; these pertain to all chemical reactions where information is enhanced or diminished.[5] Using Brownian methods, we will quantify several descriptors for organic reaction pathways. In so doing, useful geometrical devices will be introduced for characterizing carbon-based reactions from an informatic standpoint. Our research has included working with experimental (as opposed to computational only) probes of molecular information. Accordingly, one such probe will be featured in a forthcoming work so as to address the principal facets of Brownian information processing.[6]

Information and organic molecules compose broad, active fields. Our contributions aim at the probability framework surrounding this subject; accordingly, we seek bridges between chemical structure and information theory and the computational machinery of modern day. Our work augments a growing literature, for which an abbreviated bibliography is included in the reference section. We call particular attention to research by Schneider and by Bajorath and co-workers. Schneider has investigated the Shannon information associated with biopolymers and the thermodynamic ramifications.[7−9] In extensive contributions to this journal, Bajorath and co-workers have explored the application of Shannon information metrics to molecular descriptors and chemical databases; the findings have included important distinctions between natural product and synthetic compounds with clear applications to molecular design and virtual screening.[10−16] We note further the work of Morowitz for pioneering new statistical approaches to molecular libraries and the capacity offered by carbon chemistry.[17] Last, we note the classic

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

10.1021/ci0400213 CCC: $27.50 © 2004 American Chemical Society
Published on Web 07/22/2004

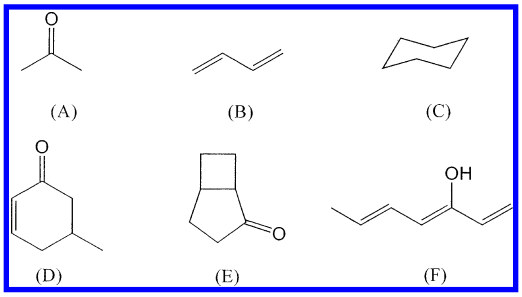

Figure 1. Structure graphs for organic molecules. Graphs A−E offer a sampling of molecular representations in a two-dimensional format.

research by Robinson[18] and the later studies by Corey and co-workers.[19] These works compellingly illustrate the importance of the systematic, topological analysis of molecular structure. Our own work is similar in spirit to those referenced in that it focuses on the structure relationships within and between organic molecules. Our particular approach is different, however, as it views molecular structure as a type of random variable source.

II. QUANTIFYING THE INFORMATION EXPRESSED BY ORGANIC MOLECULES: A BROWNIAN PROCESSING MODEL

How can the information expressed by an organic molecule, for example, any of the six represented in Figure 1 be quantified? The question is not a simple one, as "information" can be defined along disparate mathematical lines identified by Shannon, Fisher, and others.[20] The quantity depends as well on the nature of the measuring device. How does one proceed given all the complications?

We find an answer at the intersection of molecular recognition chemistry and applied probability: to quantify information, one examines precisely how a molecule functions as a random variable or code generating source during extended query events. We look to nature for a blueprint and working construct, with guidance provided by Bennett in his review of the thermodynamics of computation.[21]

As is well-known, an extensive apparatus has evolved which reads and translates molecular structure data encoded on polymers. We refer to cellular systems wherein, for example, the polymerase catalyzed synthesis of RNA molecules takes place; a schematic for this apparatus is offered in Figure 2.[22] With each data processing step, an enzyme systematically registers the identity of a monomeric site—traditionally abbreviated A, T, G, and C—and incorporates a complement species (U, A, C, G) into a growing RNA strand with the assistance of reservoir material. The rate and direction of the enzyme are controlled by the environmental conditions—temperature and the U, A, C, G-concentrations. Importantly, the recognition chemistry exhibits both parallel and serial modes: parallel as the composition of a given A,T,G,C-site is registered in a discrete step; serial as the sites are registered according to a strict nearest-neighbor order. With each processing event, the *uncorrelated* information acquired by the enzyme/growing polymer strand can be represented as $\log_2(4) = \log_2(2^2) = 2$ bits; correlated information involving multiple sites is, however, part and parcel to the reading and copying. Bennett and Hopfield have addressed independently the intriguing processing issues surrounding RNA polymerase.[21,23] Bennett refers to the

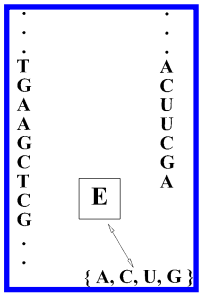

Figure 2. One example of Brownian processing. Schematic for a polymerase enzyme (E) which controls the synthesis of an RNA molecule (right-hand-side string) complementary to a DNA polymer (left-hand-side). The copy procedure features both serial and parallel information processing. The result is a polymer with information equivalent to the source. A, T, G, C, U stand, respectively, for adenine, thymine, guanine, cytosine, and uracil. The brackets denote reservoir material for the copying process. Details regarding carbohydrate-phosphate structures have been omitted.

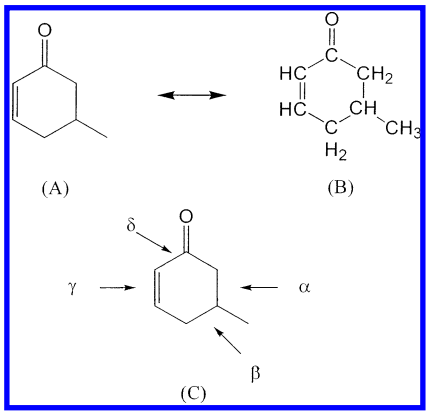

Figure 3. Structure graphs for 3-one-5-methyl-cyclohexene molecule. Graphs (A), (B), and (C) correspond to Figure 1D. (B) is an expanded form of (A), while (C) represents a molecule with several collision sites labeled arbitrarily.

Figure 2 operation as "Brownian computation"; the terms are appropriate given the random walk nature of the enzyme, the directional bias set by local conditions. RNA polymerase by no means offers an isolated example of a Brownian-design processor. The apparatus for code translation during protein synthesis poses another showcase example.[22]

Enzymes and encoded polymers are concrete entities. They further inspire a model for quantifying molecular information at the covalent bond level. As in the Figure 2 schematic, the model features both parallel and serial modes and is brought about using random walk machinery. Via this model, both the uncorrelated and correlated information can be established for individual organic compounds, the principles grounded on information theory and the universal language of chemistry. We illustrate matters via one of the Figure 1 compounds.

Figure 1—and all organic texts and journals—present structure graphs based on elemental composition, electron distribution, and Angstrom-scale geometry. These properties are delineated via symbol composites of letters, intersecting lines, and integer numbers. The graph for 5-methyl-2-cyclohexene-1-one (**D**, Figure 1), for example, denotes an assembly of **C−H**, **C═C**, **C═O**, and **C−C** units. This pictorial is rendered more explicitly via the **(A)** and **(B)** assemblies of Figure 3. The integer valence of atoms plus covalent bond stability enable all of the abbreviations commonly practiced in this communication scheme.[24]

To quantify information, one models the sensing of the molecular code units in a way which mimics real Angstrom-scale recognition and prepares a detailed tape recording along the way. The rationale is as follows. Under real laboratory conditions, a reagent such as Br_2 can discriminate the different regions of 5-methyl-2-cyclohexene-1-one via extended strings of collisions, for example, in a liquid-phase environment. If a few of the contact sites were labeled (arbitrarily) as in the **(C)**-Assembly of Figure 3, a collision scenario might be encoded as $\alpha\beta\delta\gamma\alpha\gamma\alpha\alpha\beta\alpha\delta\alpha\delta\beta$..... The real-life details would depend of course on the electric charge interactions of *both* the contact sites and sensing reagent. Yet implicit in the scenario is a Brownian processing or random walk model by which any molecule can be probed via its structure graph for information content. This content can be viewed—like the structure graph itself—as stationary and intrinsic to the molecule, that is, independent of all dynamics and environmental factors. In simplest terms: to quantify molecular information, one simulates the reading and recording steps of a Brownian (i.e. random walk) processor; structure graphs such as in Figure 1 serve as read-only data files.

In the model, atom-bond-atom assemblies such as **C−H**, **C═O**, etc. serve as the fundamental code units; these form an alphabet for generating message strings. The expression of information depends on the source file organization (i.e., the atom/covalent bond network) and the processing mode by which the file is read. In Brownian *serial* processing, a molecular graph is sensed randomly, one code unit at a time subject to a strict nearest-neighbor selection bias; this is analogous to the sequential-site-workings of a polymerase enzyme as discussed by Bennett.[21] In the case of 5-methyl-2-cyclohexene-1-one (Figure 3**(C)**), the allowed read-only and tape-recorded strings would thus include

(C-C)(C=O)(C-C)(C=C)(C-H)(C-C)...

(C=O)(C-C)(C-C)(C-H)(C-C)(C-H)...

(C-H)(C-C)(C=O)(C-C)(C-H)(C=H)...

where sequential units are always nearest-neighbors in the parent graph. In exercising the model, the strings are recorded in a Brownian fashion where the allowed code units are tagged randomly and registered as in Figure 4. "Forbidden" (i.e., never-observed) sequences would include ...(**C−H**)-(**C═O**)... such as depicted in the lower right-hand corner. Just as in random walk processing along a biopolymer, the overwhelmingly most probable event following recognition at one site involves a nearest neighbor. To summarize: in the serial version of Brownian processing, a molecular graph functions as a sequential data file which is queried and recorded by an error-free processor.

Brownian parallel processing is different in that nearest-neighbor code units are sensed randomly; however, they are registered up to several at a time. For 5-methyl-2-cyclohexene-1-one, viable two-unit strings would include

O═C−C, **H−C−H**, **H−C−C**, ...

while three-unit strings would include

H−C−C═O, **C−C−C**, **H−C−C−H**, ...

as indicated in Figure 5. The Brownian flavor derives from the random selection of strings, again so as to emulate real molecular recognition in a disordered phase. In all processing scenarios, the viable code unit assemblies are ones in which the letters (atomic symbols) share one or more lines (chemical bonds) in the structure graph. An example of a forbidden string is illustrated in the lower right-hand corner of Figure 5. The operating principle can be summarized as follows: in Brownian parallel processing, a molecular structure graph serves as a random access file which is queried and copied by an error-free processor.

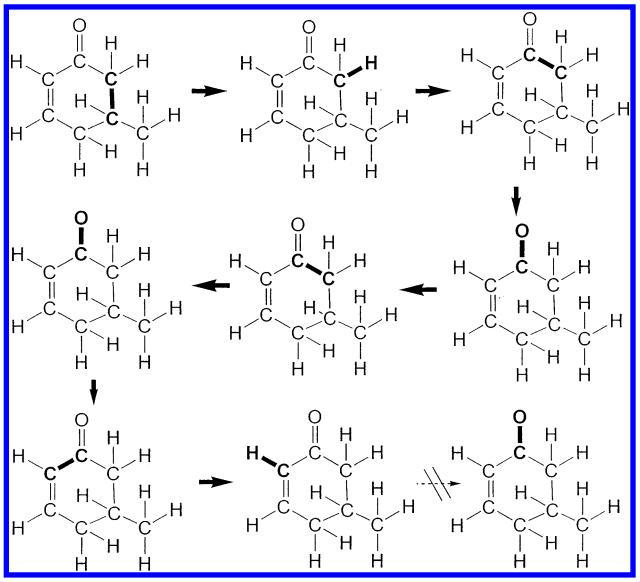

Figure 4. Molecular structure graphs subject to Brownian serial processing. The code units (states) accessed sequentially in a sample random walk allowing only nearest-neighbor transitions have been indicated using bold typeface. The two graphs of the lower right-hand corner illustrate one example of a forbidden transition.

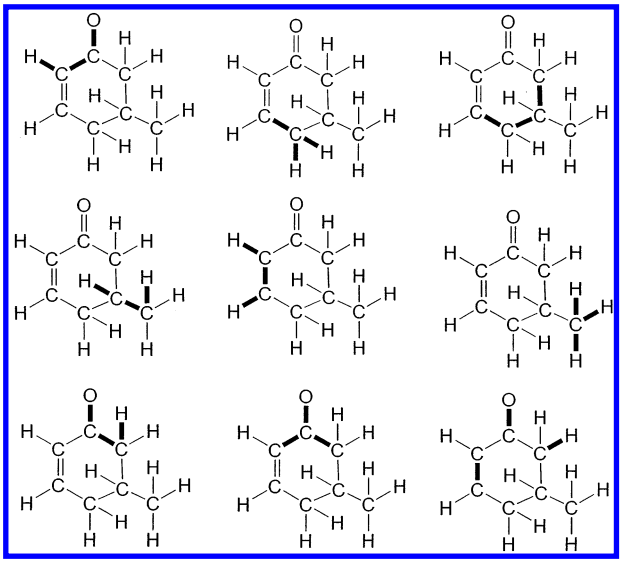

Figure 5. Structure graphs subject to Brownian parallel processing. Code unit triplets which are sensed randomly have been indicated using bold typeface. The selection rules require that all code units share at least one atom and bond symbol. The graph of the lower right-hand corner depicts an example of a forbidden triplet. Analogous rules hold for code unit doublets, quartets, etc.

III. METHODOLOGY: GRAPHS, BROWNIAN ALGORITHMS, AND RECORDING TAPE ANALYSIS

In the Brownian/random walk model, extended tape recordings of code units are constructed for a given

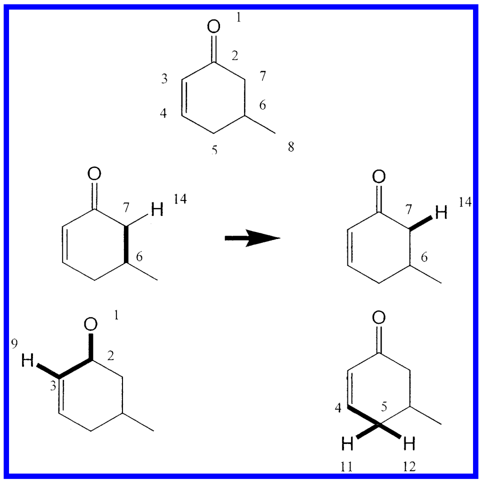

Figure 6. Molecular graphs and coding. The example of 3-one-5-methyl-cyclohexene molecule is illustrated. The numbering is arbitrary, and only the vertices and Roman letters need to be labeled initially. The middle graphs illustrate a sample transition in Brownian serial processing. The lower-most graphs describe a transition allowed in parallel processing. The array/matrix details for these transitions are given in the text.

compound. The unit sequence is significant in serial recording but is immaterial when parallel processing is exercised. These tapes are probed for information content in a manner akin to formal code or language analysis. The details are illustrated as follows, again for 5-methyl-2-cyclohexene-1-one.

One initiates matters by identifying and numbering the graph vertices (representing carbon atoms) and other Roman letters (heteroatoms), as shown in Figure 6. Only the skeleton (topmost graph) needs to be labeled initially as a simple algorithm based on atom valence rules can expand the data to include all relevant H-symbols. The result is a combination array and integer matrix pair:

array entries	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	(matrix column indices)
O 1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
C 2	2	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
C 3	0	1	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
C 4	0	0	2	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	
C 5	0	0	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	
C 6	0	0	0	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	
C 7	0	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	
C 8	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	
H 9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 11	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 12	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 13	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
H 14	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
H 15	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
H 16	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
H 17	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
H 18	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	

This procedure is not new but rather has been employed extensively in molecular structure studies based on adjacency matrices.[25,26] Owing to symmetry, only the upper or lower triangular matrix half is really necessary for applying the Brownian model. For clarity, we make use of complete matrix versions in this section.

The array and integer matrix encode a lattice upon which a random walk can be executed. The middle graphs of Figure 6 show an example of a walk-step allowed in serial processing: ...(**C**−**C**)(**C**−**H**).... Thus the (**C**−**C**)-unit of the left-hand-side graph is marked in the array/matrix format using bold-face type:

array entries	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	(matrix column indices)
O 1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
C 2	2	0	1	0	0	0	*1*	0	0	0	0	0	0	0	0	0	0	0	
C 3	0	1	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
C 4	0	0	2	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	
C 5	0	0	0	1	0	*1*	0	0	0	0	1	1	0	0	0	0	0	0	
C 6	0	0	0	0	*1*	0	**1**	*1*	0	0	0	0	*1*	0	0	0	0	0	
C 7	0	*1*	0	0	0	**1**	0	0	0	0	0	0	0	*1*	*1*	0	0	0	
C 8	0	0	0	0	0	*1*	0	0	0	0	0	0	0	0	0	1	1	1	
H 9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 11	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 12	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 13	0	0	0	0	0	*1*	0	0	0	0	0	0	0	0	0	0	0	0	
H 14	0	0	0	0	0	0	*1*	0	0	0	0	0	0	0	0	0	0	0	
H 15	0	0	0	0	0	0	*1*	0	0	0	0	0	0	0	0	0	0	0	
H 16	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
H 17	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
H 18	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	

The nearest-neighbor selection rule constrains the possible transitions to six; these have been indicated above using italic typeface. In the model, every allowed transition is taken to be equally probable; one of the transitions is selected randomly thus posing a new tagged, atom-bond-atom state, such as represented by the middle, right-hand-side graph of Figure 6. The code unit (**C**−**H**) affiliated with this new state and its possible transitions are then registered accordingly:

array entries	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	(matrix column indices)
O 1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
C 2	2	0	1	0	0	0	*1*	0	0	0	0	0	0	0	0	0	0	0	
C 3	0	1	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
C 4	0	0	2	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	
C 5	0	0	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	
C 6	0	0	0	0	1	0	*1*	1	0	0	0	0	1	0	0	0	0	0	
C 7	0	*1*	0	0	0	*1*	0	0	0	0	0	0	0	**1**	*1*	0	0	0	
C 8	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	
H 9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 11	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 12	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 13	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
H 14	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	
H 15	0	0	0	0	0	0	*1*	0	0	0	0	0	0	0	0	0	0	0	
H 16	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	

Array/matrix pairs offer constructions which are isomorphous to molecular graphs. A recording tape is made as a lengthy random walk proceeds over the constructions so as to note every code unit encountered. The extended strings for 5-methyl-2-cyclohexene-1-one may look as follows:

```
...(C-H)(C-C)(C=C)(C-H)(C=C)(C-H)(C=C)(C-H)(C=C)(C-C)(C-C)(C-C)(C=C)(C-C)(C-C)(C-H)(C-C)(C-C)(C-C)(C=C)(C-C)(C=O)(C-C
)(C-C)(C-C)(C-H)(C-C)(C-C)(C-C)(C-C)(C-C)(C-C)(C-H)(C-C)(C-C)(C=O)(C-C)(C-H)(C-H)(C-C)(C-C)(C=C)(C-C)(C-H)(C-C)(C=C)(C-
H)(C=C)(C-H)(C-C)(C-H)(C-C)(C=O)(C-C)(C-C)(C-H)(C-C)(C-C)(C-C)(C-H)(C-H)(C-C)(C-C)(C-H)(C-C)(C-H)(C-H)(C-H)(C-H)(C-C)(
C-C)(C-H)(C-C)(C-H)(C=C)(C-C)(C-C)(C-C)(C=O)(C-C)(C=C)(C-H)(C=C)(C-C)(C-H)(C-C)(C=C)(C-C)(C-C)(C-H)(C-C)(C-C)(C-C)(C-
H)(C-H)(C-H)(C-H)(C-C)(C-H)(C-C)(C-C)(C-C)(C-C)(C-H)(C-H)(C-C)(C-H)(C=C)(C-C)(C=O)(C-C)(C-H)(C-C)(C-C)(C-H)(C=C)(C-H)(
C-C)(C-C)(C-H)(C-C)(C-H)(C-H)(C-H)(C-H)(C-H)(C-C)(C-H)(C-H)(C-H)(C-H)(C-H)(C-C)(C-H)(C-C)(C-C)(C=O)(C-C)(C-H)(C-C)(C-C)
(C-C)(C-H)(C=C)(C-C)(C-H)(C-C)(C=C)(C-C)(C-H)(C-H)(C-C)(C-C)(C=C)(C-H)(C-C)(C-C)(C-C)(C-H)(C-H)(C-C)(C-H)(C-H)(C-C)(C-
H)(C-H)(C-H)(C-C)(C=O)(C-C)(C=C)(C-H)(C=C)(C-C)(C-C)(C-C)(C-C)(C-C)(C-C)(C-C)(C-H)(C-C)(C-C)(C-C)(C=C)(C-C)(C-C)(C-H)(
C-C)(C-C)(C-C)(C-C)(C=C)(C-C)(C=C)(C-C)(C-C)(C-C)(C=C)(C-H)(C-C)(C-H)(C-C)(C-H)(C-C)(C-H)(C-C)(C=C)(C-C)(C-H)(C=C)(C-
H)(C=C)(C-C)(C-C)(C=O)(C-C)(C-C)(C=O)(C-C)(C=C)(C-H)(C-C)(C=O)(C-C)(C=O)(C-C)(C=C)(C-C)(C-H)(C=C)(C-C)(C-H)(C-C)(C=
O)(C-C)(C=C)(C-C)(C-C)(C-H)(C-C)(C-H)(C-H)(C-C)(C-C)(C-H)(C-C)(C-C)(C=O)(C-C)(C=C)(C-C)(C-H)(C-C)(C=C)(C-H)(C-C)(C-H)(
C=C)(C-C)(C-H)(C-C)(C-C)(C-H)(C-C)(C-H)(C-C)(C-C)(C-H)(C-C)(C-C)(C-C)(C-C)(C-C)(C-H)(C-C)(C=O)(C-C)(C=O)(C-C)(C-H)(C-
C)(C=O)(C-C)(C-C)(C-H)(C-H)(C-C)(C-H)(C-C)(C-C)(C-C)(C=C)(C-H)(C-C)(C-H)(C-C)(C-H)(C-C)(C-C)(C-H)(C-H)(C-H)(C-H)(C-H)(
C-C)(C-C)(C-H)(C-H)(C-H)(C-C)(C-C)(C=C)(C-H)(C-C)(C-H)(C-H)(C-H)(C-C)(C-C)(C-H)(C-H)(C-H)(C-H)(C-H)(C-H)(C-H)(C-H)(C-C)
(C-H)(C-C)(C-C)(C-H)(C-H)(C-C)(C-H)(C-C)(C-C)(C-C)(C-H)(C-H)(C-C)(C=O)(C-C)(C-H)(C-C)(C-C)(C-C)(C=C)(C-H)(C=C)(C-C)(C-
C)(C-H)(C-C)(C-H)(C-H)(C-H)....
```

In studying a given molecule, Brownian serial tapes are assembled which are typically several thousand code units in length. These enable information probes to several orders of structure correlation.

The third-row graphs of Figure 6 illustrate sample triplets registered in Brownian parallel processing: (**H**−**C**−**C**═**O**) and (**C**−$\mathbf{CH_2}$) each obtain from the random selection of a nonzero matrix entry followed by a probe of interacting

matrix entries. These two examples are indicated as follows using bold-face type:

array entries	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	(matrix column indices)
O 1	0	**2**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
C 2	**2**	0	**1**	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
C 3	0	**1**	0	2	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	
C 4	0	0	2	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	
C 5	0	0	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	
C 6	0	0	0	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	
C 7	0	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	
C 8	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	
H 9	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 11	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 12	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 13	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
H 14	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
H 15	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
H 16	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
H 17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
H 18	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	

array entries	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	(matrix column indices)
O 1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
C 2	2	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
C 3	0	1	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
C 4	0	0	2	0	**1**	0	0	0	0	1	0	0	0	0	0	0	0	0	
C 5	0	0	0	**1**	0	1	0	0	0	0	**1**	**1**	0	0	0	0	0	0	
C 6	0	0	0	0	1	0	1	1	0	0	0	0	1	0	0	0	0	0	
C 7	0	1	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	
C 8	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	
H 9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 11	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 12	0	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	
H 13	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
H 14	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
H 15	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	
H 16	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
H 17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	

In parallel processing, all viable code aggregates are established by random selection. A separate recording is made for each type of aggregate: doublet, triplet, quartet, etc. For a given molecule, Brownian parallel tapes offer lengthy records featuring several thousand aggregates. These constructs also allow information probes to several correlation orders.

In examining tapes for information content, two landmark ideas are put into play. The first is that on average, the optimum number of bits H_I required to encode independent draws of a discrete random variable I (here an atom-bond-atom unit) expressing a frequency distribution p(i) over i-labeled states is given by the Shannon formula[27]

$$H_I = -\Sigma_i p(i) \log_2 p(i) \\ -(1/\log_e 2)\Sigma_i p(i) \log_e p(i) \qquad (1)$$

Eq 1 asserts that optimal coding—and thus the information quantity H_I—always follow from precise knowledge of the i-states and p(i). At a root level, H_I measures the size of the message imbedded in p(i) quite apart from any state correlations that may exist.

A second idea is needed to address cases of less-than-optimal coding. If during message analysis, one were to employ an alternative distribution, say q(i) in place of p(i), an *excess* number of bits would be required for coding. This excess can be quantified using a formula similar to eq 1[28]

$$K_I = +\Sigma_i p(i) \log_2 \{p(i)/q(i)\} \\ = +(1/\log_e 2)\Sigma_i p(i) \log_e \{p(i)/q(i)\} \qquad (2)$$

whereby K_I approaches zero bits as p(i)/q(i) approaches unity. K_I gauges the overlap disparity of two frequency distributions; it addresses the question: how many bits would be required for coding if one were to assume, however mistakenly, that the frequency distribution was q(i) instead of p(i)?

In Brownian information processing, molecular structure graphs and recording tapes furnish all of the state data and frequency distributions. For 5-methyl-2-cyclohexene-1-one, for example, expressing a total of 18 code-units of four different state categories: p(**C**−**H**) = 10/18; p(**C**=**O**) = 1/18 = p(**C**=**C**); p(**C**−**C**) = 6/18; these frequency values are independent of the processing mode. Then by eq 1

$$H_I = -(1/\log_e 2)\{(10/18)\log_e(10/18) + \\ (1/18)\log_e(1/18) + (1/18)\log_e(1/18) + \\ (6/18)\log_e(6/18)\} \approx 1.463 \text{ bits}$$

H_I is the uncorrelated (i.e., first order) message size which depends on the number of different code units and their occurrence frequency. H_I increases with increasing unit types and is further enhanced if the code units appear with near equal frequency. By this model, H_I would equal 0.00 bits for methane and carbon tetrachloride molecules, each expressing only a single type of atom-bond-atom state. By contrast, H_I would equate with 1.00 bit for dichloromethane molecule expressing two different code units, **C**−**Cl** and **C**−**H**, with equal frequency.

An eq 1 application examines the code units independent of neighbor properties. Variations of eq 2 enlighten in all probes of the correlated information (CI) for an organic molecule. For example, in a second-order analysis, a joint frequency distribution p(i,j) is constructed on the basis of molecular code unit pairs. This distribution is compared with q(i,j) = p(i) · p(j), the distribution that would be in effect if the i and j code units were expressed on a recording tape independently of one another, i.e., if their occurrence frequencies were completely uncorrelated. Following eq 2, one establishes a bit-value for CI_{IJ}, the second-order correlated information via

$$CI_{IJ} = +\Sigma_i p(i,j) \log_2\{p(i,j)/q(i,j)\} \\ = +(1/\log_e 2)\Sigma_i p(i,j) \log_e \{p(i,j)/p(i) \cdot p(j)\} \quad (3)$$

Note that by this strategy, CI_{IJ} equals zero bits if for all states p(i,j)/p(i) · p(j) proves equal to unity. CI_{IJ} answers the critically important question: to what extent does a structure graph predicate correlated signals, that is, a departure from noise.

The eq 3 theme is readily extended to higher CI-orders in both serial and parallel processing. In third order, code unit triples are examined via p(i,j,k) constructed from Brownian tapes, and compared with q(i,j,k) = p(i) · p(j) · p(k). A modified eq 3 yields a quantity CI_{IJK}, a measure of departure of p(i,j,k) from independent i,j,k-distributions. In our studies, we have typically analyzed Brownian tapes, serial and parallel, to a maximum of eight structure orders. For notation ease and clarity, we will make reference to the information values H_I using the symbol I_1, the subscript emphasizing the single summation, first-order nature of eq 1. Along the same lines, CI_{IJ}, CI_{IJK}, ... will be represented using the symbols $CI_n^{(S)}$ and $CI_n^{(P)}$. Here the subscript n conveys the structure order equivalent to the number of atom-bond-atom states

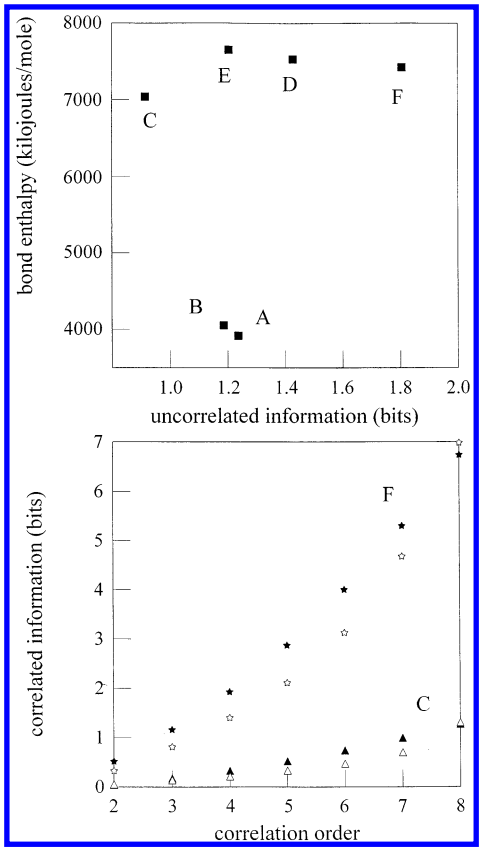

Figure 7. Uncorrelated and correlated information for Figure 1 compounds. Upper panel illustrates total covalent bond enthalpy versus uncorrelated information. The lower panel illustrate CI-values versus structure order for serial (filled symbols) and parallel Brownian processing (open symbols) for cyclohexane (C) and 1,3,5-heptatriene-3-ol (F).

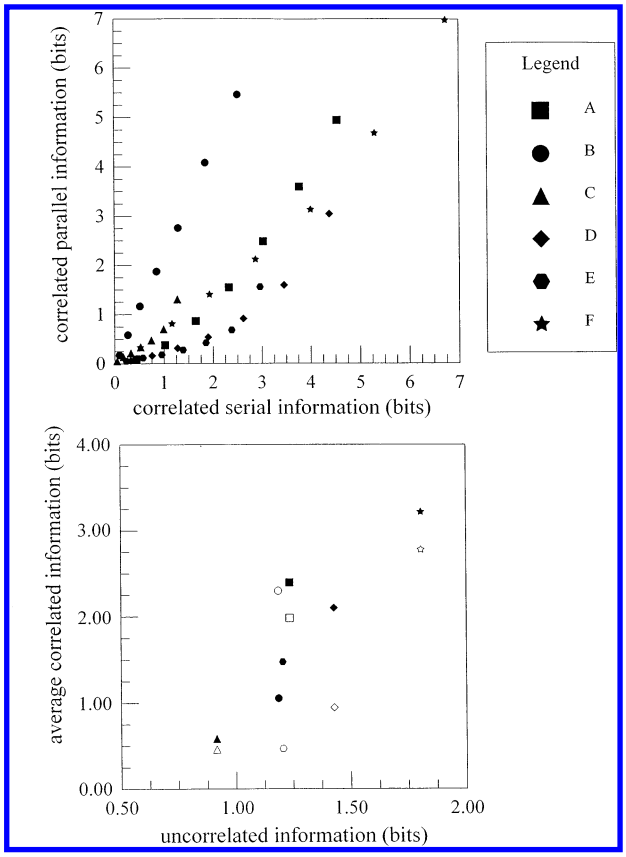

Figure 8. Additional informatic perspective for Figure 1 compounds. Upper panel shows $CI_n^{(P)}$-versus-$CI_n^{(S)}$ for the six compounds. Lower panel illustrates $\langle CI_n^{(P)} \rangle$ (open symbols) and $\langle CI_n^{(S)} \rangle$ filled symbols versus I_1.

correlated; the superscript specifies the mode of processing, serial or parallel.

In closing this section we remark that organic structure graphs are time-honored for conveying subtleties of atom/bond configurations.[24] These would include the number and location of chiral and ionic centers, and the resonance forms for a given molecule. These attributes of electric charge can be accommodated by array/matrix pairs and Brownian processing. In studies of optically active compounds, for example, we have distinguished chiral and achiral carbon atoms using symbols "c" and "C", respectively. We have further denoted ionic centers such as in oxalate anion, using the conventional symbol "O^-". These distinctions do not alter the Brownian, tape recording, and information probe algorithms. Yet the methodology is not without practical limits, for example, concerning bond lengths and geometry. In work thus far, the chair and boat forms of cyclohexane and the eclipsed and staggered forms of ethane have not been distinguished from one another.

IV. SAMPLE RESULTS

A molecule expresses electronic energy and information via atoms and a bond network. An elementary view of this expression is contained in the upper panel of Figure 7. Here total bond enthalpy values for the Figure 3 compounds are represented along the ordinate axis, each following from standard look-up tables and procedures.[29] The I_1-quantities derive from eq 1 applied to the structure graphs (Figure 3) and range from 0.916 bits for cyclohexane molecule **(C)** to 1.86 bits for 1,3,5-heptatriene-3-ol **(F)**. Of the six compounds, acetone **(A)** and butadiene **(B)** contain the fewest atoms and bonds; not surprisingly, these molecules demonstrate the lowest total enthalpy values. That their I_1-values fall midrange emphasizes that molecular information depends on an interplay between the bond network size and code-unit diversity.

The lower panels of Figure 7 shed additional light via the quantities $CI_n^{(S)}$ and $CI_n^{(P)}$ for cyclohexane and 1,3,5-heptatriene-3-ol—the highest and lowest I_1 -compounds of the upper panel. In general, for all the Figure 1 compounds and for both processing modes, CI increases with increasing structure order: as additional code units are sensed during a Brownian query, greater correlated information about the bond network is acquired. The lower-panel data show how the magnitudes and scaling are different for different organic compounds and depend to an extent on the mode of processing. Of the Figure 1 compounds, cyclohexane **(C)** and 1,3,5-heptatriene-3-ol **(F)** demonstrate the lowest and highest CI-values, respectively, at all orders for both serial and parallel modes. We remark that CI data such as in Figure 7 are consistent for encoded tapes acquired from different Brownian runs. Relative standard deviations of ca. 0.01−0.05 are typical in each order of correlation. This means that error bars associated with the Figure 7 data are of size less than the symbol widths.

The Figure 1 molecules also lead to the data in Figure 8. The upper panel illustrates $CI_n^{(S)}$-versus-$CI_n^{(P)}$. Note in the case of butadiene **(B)**, a linear relation is evident between

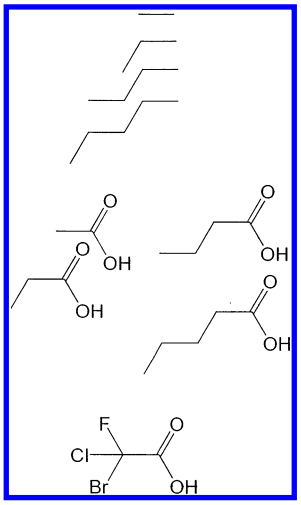

Figure 9. Structure graphs for three miniature libraries. The groupings are according to functionality: alkane, carboxylic acids, and a single perhaloacid.

the information quantities, whereas for the other compounds, the $CI_n^{(S)}$-$CI_n^{(P)}$ functionality is more complicated. The lower panel shows the dependence of the average values $\langle CI_n^{(S)} \rangle$ and $\langle CI_n^{(P)} \rangle$ on I_1. Such averages have been computed over structure orders 2−8 with equal weight coefficients applied; the serial and parallel data are distinguished via filled and open symbols, respectively. Here one learns that a four-carbon, relatively high-symmetry compound such as butadiene expresses about twice the CI-average during parallel sensing, compared with serial. By contrast, bicyclo[3.2.0]-heptane-2-one **(E)** expresses considerably greater CI in serial processing, compared with parallel. Details such as these reflect the natural complexities of molecular information content, especially where small, symmetric compounds are concerned. Not all informatic features, however, are case-specific. We have indeed found the CI-values for organic compounds lacking long-range symmetry and fused rings to be relatively insensitive to the mode processing details. Moreover, one observes a nominal linear dependence of CI-averages on I_1, as indicated in the lower panel of Figure 8.

Figure 9 adds perspective by illustrating three miniature libraries organized by functionality: (normal-) alkanes, carboxylic, and halo-substituted acids. Figure 10 follows by portraying $CI_n^{(S)}$ data for the libraries, the trends in parallel processing data being very similar. The upper panel pertains to the alkanes, while the lower panel shows data relevant to the acid libraries. Different symbols have been used to discriminate the number of carbon atoms for each library member; open squares mark results for the perhalo-compound (library of one molecule). Clearly the lowest $CI_n^{(S)}$ values are expressed by the saturated compounds of Figure 9—the molecules devoid of functional groups. Pentane molecule, for example, poses less than one bit of $CI^{(S)}$, even in the eighth structure order. By contrast, the carboxylic acid compounds express $CI^{(S)}$ which are an order of magnitude higher. Ethane, the smallest alkane, expresses the lowest $CI_n^{(S)}$ at all orders, compared with all other Figure 9 compounds. Note however, that the minimum $CI_n^{(S)}$ in the acid-library is registered by pentanoic acid, the carboxylic acid offering the longest saturated chain. Clearly a molecule's information does not depend in a trivial way on the number of carbon atoms of a given compound but rather on the

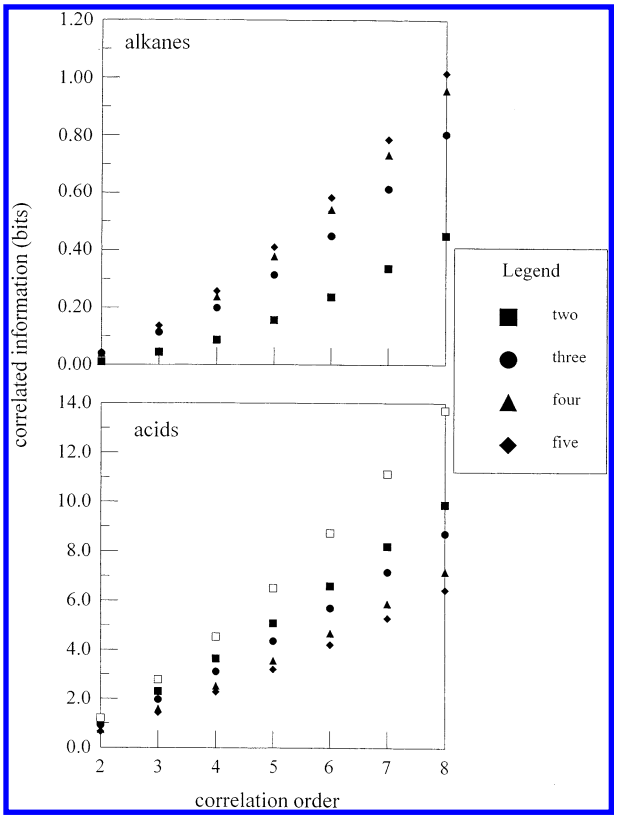

Figure 10. Correlated information for Figure 9 libraries. Upper and lower panels illustrate $CI_n^{(S)}$-versus-n for the three sets of molecules. The legend connects symbols with the number of carbon atoms in the molecules. Results for the perhaloacid are illustrated using open squares.

structure intricacies of the backbone, heteroatoms, and functionality. It is difficult to intuit these details from structure graphs by sight. Rather the formal application of Brownian information processing and tape analysis is required.

Carbon chemistry typically allows more than one stable configuration of a given atom composition. For example, the graphs **(D)**, **(E)**, and **(F)** of Figure 1 represent valence isomers with formula $C_7H_{10}O$. Figure 11 elaborates on this theme by showing enthalpy totals versus $\langle CI_n^{(S)} \rangle$ for a library of 58 $C_7H_{10}O$ isomers. The data have again been restricted to serial processing, as the trends are reflected as well in parallel. Via Figure 11, one finds the enthalpy values to demonstrate significant pooling over several levels; this trait is mirrored as well in more sophisticated molecular calculations.[30] By contrast, the information quantities display a more-or-less even spread about a mean value. This type of library statistical structure is elaborated upon in the lower panel. Here one compares $\langle CI_n^{(S)} \rangle$-distribution for the $C_7H_{10}O$ library with the standard normal distribution with mean zero and unity standard deviation. We have found these results typical for valence isomer libraries supported by carbon chemistry. Accordingly, an informatic classification of each compound of a given library proves useful according to measured CI. Figure 12 offers examples of such classifications for $C_7H_{10}O$. The left- and right-most compounds express the library minimum and maximum CI, respectively; the middle compound expresses information approximately equal to the mean value. Interestingly, the *Aldrich Catalog* lists 16 molecules with formula $C_7H_{10}O$, for example,

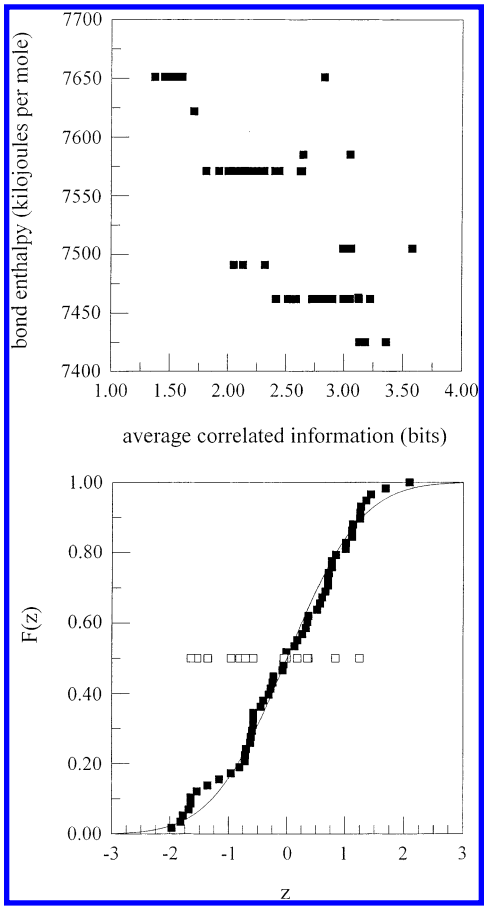

Figure 11. Informatic perspective for $C_7H_{10}O$ isomers. Upper panel shows bond enthalpy-versus-$\langle CI_n^{(S)}\rangle$ for a library of 58 valence isomers. Lower panel illustrates library distribution function; the solid line marks the standard normal distribution. In constructing the lower plot, $\langle CI_n^{(S)}\rangle$-values have been scaled relative to the library mean in units of the standard deviation, i.e., $Z = [\langle CI_n^{(S)}\rangle - \langle CI_n^{(S)}\rangle_{avg}]/\sigma$ with $\langle CI_n^{(S)}\rangle_{avg} = 2.44$ bits and $\sigma = 0.545$ bits. F(z) quantifies the fraction of compounds with $z \leq Z$. Open squares in lower panel mark $C_7H_{10}O$ compounds available from Aldrich Chemical.

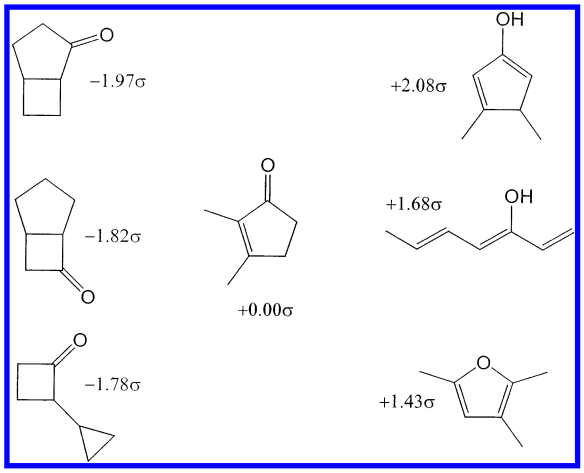

Figure 12. Further informatic perspective for $C_7H_{10}O$ isomers. The left-most graphs identify compounds expressing minimum CI; Z-values range from -1.78σ to -1.97σ. The middle graph refers to a compound expressing the mean library information. The right-most graphs show compounds expressing the largest CI with Z-values $+1.43\sigma - 2.08\sigma$.

norcamphor.[31] The informatic positions of these have been indicated in the lower panel of Figure 11 using open squares.

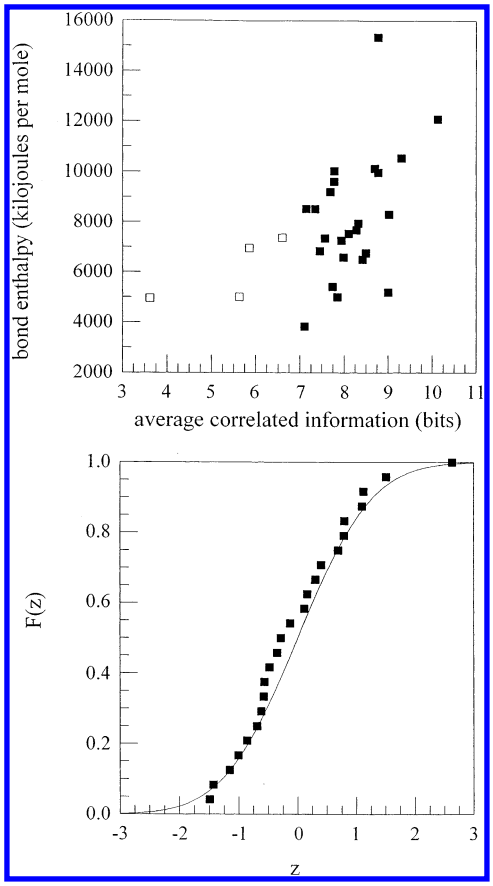

Figure 13. Informatic perspective for the naturally occurring amino acids and genetic code bases. Upper panel shows bond enthalpy-versus-$\langle CI_n^{(S)}\rangle$ for the organic acids and bases, represented by filled and open squares, respectively. Lower panel illustrates the cumulative CI-distribution for the amino acids; the solid line marks results for the standard normal distribution. The $\langle CI_n^{(S)}\rangle$-values have been scaled in the same manner as in Figure 10, and the data are reported numerically in Table 1. All calculations pertain to the nonionic forms of the molecules.

For this case, one finds the commercial products to span (more-or-less) the information measure permitted by the $C_7H_{10}O$ formula.

Figure 13 presents data for two critically important libraries by way of the naturally occurring amino acids and the organic bases of the genetic code.[32] As in Figure 11, the data have been restricted to serial processing; the data are reported numerically as well in Table 1. One finds that for the amino acids, tryptophan and glycine express the highest and lowest $\langle CI_n^{(S)}\rangle$, respectively. For the bases, guanine and uracil express the highest and lowest $\langle CI_n^{(S)}\rangle$, respectively. Note that all four bases express CI less than the values found for the amino acids. There is a nominal linear correlation between bond enthalpy and $\langle CI_n^{(S)}\rangle$. As with molecular libraries organized around common structure/function motifs (amines, aliphatics, etc.), the data invite comparison with normal law statistics. This comparison is exercised for the amino acids in the lower panel of Figure 13 where the solid line marks the standard normal distribution. These results lead each acid of Table 1 to be affiliated with a Z-value gauged in units of the library standard deviation.

Amino acids form the building blocks of peptides and proteins. Figure 14 offers data for a sample library of pentamers such as can be obtained via combinatorial synthesis. For all the peptide structures, the C-terminal

Table 1. Information Properties of Naturally Occurring Amino Acid and Genetic Code Bases[a]

molecule	I_1 (bits)	$\langle CI_n^{(S)} \rangle$ (bits)	Z-value (σ-units)
	Acids		
tryptophan	3.097	10.12	+2.63
tyrosine	2.986	9.303	+1.51
histidine	3.056	9.020	+1.12
cysteine	2.994	9.003	+1.10
cystine	2.993	8.779	+0.793
thyroxine	2.881	8.775	+0.787
phenylalanine	2.624	8.705	+0.691
asparagine	2.895	8.495	+0.405
aspartic acid	2.883	8.424	+0.308
glutamine	2.892	8.321	+0.168
glutamic acid	3.001	8.282	+0.115
methionine	2.926	8.104	−0.128
threonine	2.747	7.988	−0.286
hydroxyproline	2.829	7.942	−0.349
alanine	2.621	7.848	−0.476
arginine	2.799	7.783	−0.565
8-hydroxylysine	2.776	7.776	−0.575
serine	2.700	7.744	−0.618
lysine	2.556	7.694	−0.687
valine	2.535	7.571	−0.855
proline	2.778	7.459	−1.01
leucine	2.489	7.351	−1.15
isoleucine	2.315	7.148	−1.43
glycine	2.636	7.105	−1.49
	Bases		
guanine	2.867	6.597	N/A
adenine	2.765	5.854	N/A
cytosine	2.552	5.630	N/A
uracil	2.228	3.616	N/A

[a] Molecules are listed in order of descreasing $\langle CI_n^{(S)} \rangle$ and Z-value.

residue was held (arbitrarily) invariant at valine, while all the sequence possibilities were explored using valine, serine, histidine, arginine, and threonine as synthetic units (also chosen arbitrarily). Such a sample library offered $1 \cdot 5^4 = 625$ compounds, the $\langle CI_n^{(S)} \rangle$ observed to follow a normal distribution, as illustrated in the lower panel of Figure 14. Table 2 highlights the average and extremum data for this sample peptide library; the primary structures of pentamers expressing the highest, and lowest CI, and information closest to the library-mean are specified.

V. DISCUSSION

Via mass and charge, an organic molecule expresses information over the various degrees of freedom—translational, rotational, vibrational, etc. Via collisions and other electronic interactions, molecules transmit and sense this information and exercise a spectrum of outcomes, including reading, writing, copying, and transformation. The information posed by a given compound is closely related to its entropy. At a fundamental level, this latter quantity can be measured via the molecular partition function or Boltzmann sum over states.[33]

For a molecule, energy and information are distributed over several degrees of freedom. We have confined attention, however, to the covalent bond networks, for these are the centerpieces in molecular recognition.[34] Structure graphs do not equate with the real electric charge packages stored in lab bottles. However, informatic quantifications based on

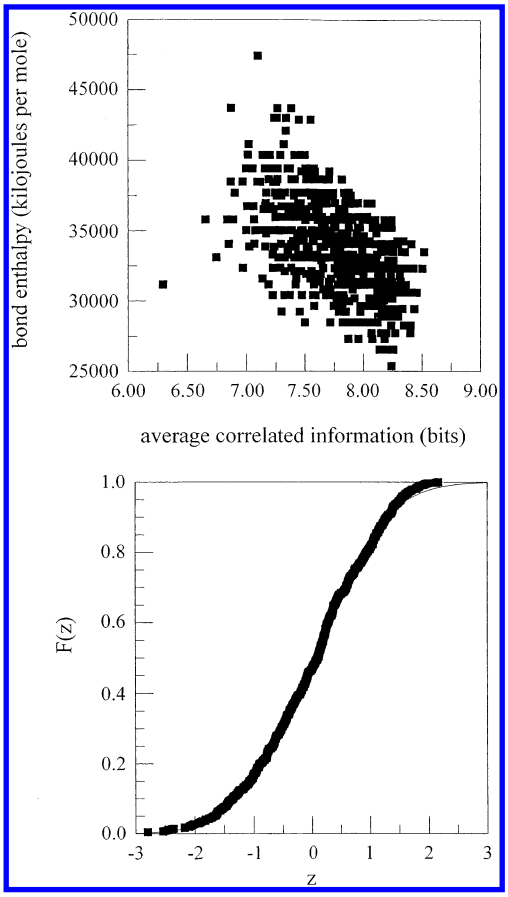

Figure 14. Informatic perspective for a sample peptide library. Upper panel shows bond enthalpy-versus-$\langle CI_n^{(S)} \rangle$ for a sample library of pentamers. The C-terminal end has been fixed arbitrarily at valine, while the other four residues feature possible combinations of valine, serine, histidine, arginine, and threonine. The lower panel illustrates the cumulative distribution for the 625-member library; the solid line marks results for the standard normal distribution. The $\langle CI_n^{(S)} \rangle$-values have been scaled in the same manner as in Figures 10 and 13. Table 2 identifies library members of maximum, minimum, and mean CI.

Table 2. Information Properties of NH_2-X_5-X_4-X_3-X_2-Valine-CO_2H Peptide Library[a]

molecule	$\langle CI_n^{(S)} \rangle$ (bits)	Z-value
NH_2-THR−HIS-THR−HIS−Val-CO_2H	8.517	+2.16
NH_2−HIS-THR−SER−HIS−Val-CO_2H	8.503	+2.12
NH_2-THR-THR−SER−HIS−Val-CO_2H	8.453	+1.98
NH_2−SER−HIS-THR-THR−Val-CO_2H	8.442	+1.95
NH_2−HIS-THR−SER−SER−Val-CO_2H	8.435	+1.93
NH_2−SER-THR-THR−VAL-Val-CO_2H	7.745	+0.00
NH_2−VAL-VAL-VAL-HIS−Val-CO_2H	6.851	-2.50
NH_2-ARG-VAL-VAL-VAL-Val-CO_2H	6.837	-2.54
NH_2−VAL-VAL-VAL-VAL-Val-CO_2H	6.745	-2.79
NH_2−VAL-VAL-ARG-VAL-Val-CO_2H	6.654	-3.05
NH_2−VAL-VAL-VAL-SER−Val-CO_2H	6.292	-4.06

[a] X_i = valine, serine, histidine, arginine, and threonine.

graphs are simple to carry out with new insights obtained in the process; importantly these insights are based on the universal icons used for communicating organic molecular structure and transformations.[35] As in all quantification procedures, "information" is not synonymous with "facts and data". Rather, the concept is statistical in nature, in section IV deriving from atom-bond-atom distributions logged during Brownian queries. Intriguingly, molecular information offers

two distinct realizations, parallel and serial, dependent on the details of code transmission and reception.

The primary aim of this paper was to present information quantification procedures applicable to all organic molecules. Accordingly, section III delineated new methodology for probing the subtleties of carbon chemistry. This methodology was realized using tools of both applied probability and chemical structure theory. Yet a second aim was also important, namely to examine the informatic framework of sample libraries through the lens of well-known error laws. The results are significant, because such laws transcend the structure/function details of any particular systems. In addition, they help connect the infinite possibilities posed by carbon chemistry with the mathematical statistics realm.

A molecule's energy scales with its capacity to perform meaningful work. In turn, a molecule's correlated information can be viewed as a measure of the capacity for controlling work, either its own or that in conjunction with another molecule. A two-carbon entity such as ethane (Figures 9 and 10) expressing minimal CI—less than 0.50 bits in the eighth order—exerts a control capacity barely exceeding that of white noise. By contrast, acetic acid—an ethane-type backbone plus a carboxy functional group—offers over nine bits of CI in the eighth order. Clearly the latter compound asserts significantly greater capacity for governing work execution. Note that molecular information is operative in settings very different from everyday paradigms, for example, the Dell device used to write this paper. In the latter case, information is stored and accessed synchronously in power-of-two increments only. Organic molecules follow a radically different and indeed more complicated playbook. They express and access information in noninteger denominations in a highly erratic fashion. Moreover, collections of molecules exhibit broad work control spectra; this facet is also very different from the algorithms offered by everyday information devices.

The informatics of real molecules depend to an extent on environmental factors such as solvent morphology. This dependence undoubtedly supports a mix of serial and parallel processing modes during real intermolecular interactions. In Brownian operations, serial modes offer greater code transmission redundancy and thus error correction facility. By contrast, the parallel operating modes are geared more toward enhanced diversity of the electronic messages expressed by molecules. This paper does not address molecular informatics for specific lab conditions, with solvent issues taken into account. Rather the emphasis is on the intrinsic information of molecules subject to the two extremes of serial and parallel processing. However idealized, three essential principles emerge from the section IV data assortment.

The first principle is that Brownian processing is able to affiliate every possible organic molecule with a set of correlated information values at the atom and covalent bond level. Thus for a nontrivial compound expressing > 1 type of code unit, there exists a bit-value in the set for every order of structure correlation, irrespective of the molecular functionality. A library of molecules is associated with a collection of CI-sets, irrespective of the structure/function theme. These were the central ideas of Figures 7−10. We note that correlation orders > 8 can readily be investigated for a molecule; however, in our experience we have found the scaling attributes for most compounds to be established by the sixth−eighth order; consistent CI-measures are thus readily obtained in these lower orders via both serial and parallel Brownian modes.

When CI-averages are combined with energy data, a type of informatic portrait is obtained based on the intricacies of structure. Examples of these portraits were presented in the upper panels of Figures 11 (valence isomers), 13 (amino acids), and 14 (peptides). Note that "energy" by way of bond enthalpy is arguably the most accessible ordinate quantity, given extensive look-up procedures and immediate contact with structure graphs. However, there are really no restrictions here; ionization energies, heats of formation, etc. would be equally valid in assembling informatic portraits. The pivotal idea is that energy/CI plots offer powerful tools for comparing the informatics afforded by organic molecule collections. Compounds affiliated with far-apart points in a portrait express markedly different electronic messages—and vice versa. Further, the scatter of points offers a measure of the message diversity within a given library. Organic molecules offer infinite possibilities; it is thus difficult to conceive of appreciable gaps in the energy/information plane imposed by chemical structure theory. Electronic message wise, carbon chemistry furnishes an exceedingly tunable, space-filling system.

The second principle is that in general, the greater diversity of code units, the more enhanced the correlated information is for a molecule. Bond network diversity—or the lack of it—determines an uncorrelated message size I_1 gauged by eq 1. The larger-size messages of natural products such as amino acids offer significantly greater potential for CI over and above noise-levels, as quantified by eq 3. It is not surprising that amino acids, peptides, and genetic code constituents express the relatively high CI-values that they do, given their control signal and operational demands in biological settings. By contrast, aliphatics and other minimally functional compounds such as solvent materials offer only sparse CI. The take-home point here is that the informatics of a molecule at the covalent bond level clearly tie directly to the depth and sophistication level of chemical action.

The third principle is that the CI-distributions for libraries organized around common structure/function themes adhere to simple error laws. This was the principal message in the lower panels of Figures 11−14. This is significant because it facilitates the use of sampling and inference methods when presented with unfamiliar chemistry situations. Extensive information calculations can readily be carried out for a library of several hundred molecules such as in Figure 14. Yet a pentamer library featuring 20 different moieties poses $20^5 \approx 3.2 \times 10^6$ possibilities; it is a ponderous task to survey all the members in such a library. Further, for an isomer library based on any nontrivial formula (e.g. $C_8H_{13}O_2Cl$), there is no facile algorithm for identifying all of the possible structures. Given these kinds of unfamiliar circumstances, it is straightforward to construct random samples of compounds in computer memory and apply the Brownian methods accordingly. By this strategy, one can readily assemble informatic statistical portraits and model distributions—not necessarily Gaussian—with unbiased CI-estimators having two favorable outcomes. First, the distribution statistics enable predictions of the practical limits of work control capacity afforded by the library structure/function

theme. Second, sampling methods enable identification of "lead" or extremum CI compounds whose information can be fine-tuned via derivatization studies.

VI. SUMMARY AND CLOSING

Methods for quantifying the information content of organic molecules and libraries at the covalent bond level have been presented. The procedures are grounded on Brownian/random walk machinery and chemical structure graphs according to two distinct processing modes. A molecule's correlated information combined with energy data offers a type of informatic portrait for comparing encoded messages. We are presently adapting Brownian methods to incorporate the perturbational effects of solvent molecules. Further, the transmission of messages over channels involves errors; the code correction facility of organic molecules and solvents is thus also being investigated with the results presented in a later paper. For now, the present study has focused on the informatics intrinsic to molecules. The paper which follows examines the informatics intrinsic to chemical reactions.

ACKNOWLEDGMENT

Support for one of the authors (M.V.S.) from the National Science Foundation via the Research Experiences for Undergraduates Program is greatly appreciated. Discussions with Professor James Babler are enlightening and much appreciated. The helpful criticism of anonymous referees is also appreciated.

REFERENCES AND NOTES

(1) Pimentel, G.; Spratley, R. D. *Understanding Chemistry*; Holden-Day: San Francisco, 1981; Chapter 19.
(2) See, for example, Terrett, N. K. *Combinatorial Chemistry*; Oxford University Press: Oxford, 1998. Bunin, B. A. *The Combinatorial Index*; Academic Press: San Diego, 1998.
(3) Graham, D. J.; Schacht, D. Base Information Content in Organic Molecular Formulae, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 942.
(4) Graham, D. J. Information Content in Organic Molecules: Structure Considerations Based on Integer Statistics. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 215.
(5) Graham, D. J.; Schulmerich, M. Information Content in Organic Molecules: Reaction Pathway Analysis Via Brownian Processing. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1612−1622.
(6) Graham, D. J.; Marlarkey, C.; Sevchuk, W. Information Content in Organic Molecules: Brownian Processing, Solvent Channels, and Noise Effects. Manuscript to be submitted.
(7) Schneider, T. D. Measuring Molecular Information. *J. Theor. Biol.* **1999**, *201*, 87−92.
(8) 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.* 125−137.
(9) Schneider, T. D. Information Content of Individual Genetic Sequences. *J. Theor. Biol.* **1997**, *189*, 427−441.
(10) 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.
(11) Stahura, F. L.; Godden, J. W.; Xue, L.; Bajorath, J. Distinguishing Between Natural Products and Synthetic Molecules by Descriptor Shannon Entropy Analysis and Binary QSAR Calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245.
(12) 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.* **2001**, *41*, 1060.
(13) Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233.
(14) Godden, J. W.; Bajorath, J. Chemical Descriptors of Distinct Levels of Information Content and Varying Sensitivity to Differences Between Selected Compound Databases Identified by SE-DSE Analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 87.
(15) Godden, J. W.; Xue, L.; Bajorath, J. Classification of Biologically Active Compounds by Median Partitioning. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1263.
(16) Godden, J. W.; Xue, L.; Kitchen, D. B.; Stahura, F. L.; Schermerhorn, E. J.; Bajorath, J. Median Partitioning: A Novel Method for the Selection of Representative Subsets from Large Compound Pools. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 885.
(17) Morowitz, H. J. *Energy Flow in Biology*; Ox Bow Press: Woodbridge, CT, 1979.
(18) An insightful commentary on Robinson's systematic approach to molecular structure is given by Fleming in *Selected Organic Syntheses*; Wiley: London, 1973.
(19) Corey, E. J. General Methods for the Construction of Complex Molecules. *Pure Appl. Chem.* **1967**, *14*, 19.
(20) Reza, F. M. *An Introduction to Information Theory*; Dover: New York, 1994; Chapter 3.
(21) Bennett, C. H. Thermodynamics of Computation − A Review. *Intl. J. Theo. Phys.* **1982**, *21*, 905.
(22) Lehninger, A. L. *Biochemistry*; Worth Publishers: New York, 1970; Part IV.
(23) Hopfield, J. J. Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity. *Proc. Natl. Acad. Sci.* **1974**, *71*, 4135.
(24) 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. 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**, *24*, 81−92.
(25) King, R. B. *Applications of Graph Theory and Topology in Inorganic Cluster and Coordination Chemistry*; CRC Press: Boca Raton, 1993; Chapter One.
(26) *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1976.
(27) Shannon C. E.; Weaver, W. *The Mathematical Theory of Communication;* U. of Illinois Press: Urbana, IL, 1949. See, also: Goldie, C. M.; Pinch, R. G. E. *Communication Theory*; Cambridge University Press: Cambridge, 1991.
(28) Kullback, S. *Information Theory and Statistics*; Dover: New York, 1997; Chapter One.
(29) Cox, J. D.; Pilcher, G. *Thermochemistry of Organic and Organometallic Compounds*; Academic Press: New York, 1970.
(30) The authors had access to a number of quantum molecular modeling routines based on Gaussian basis sets. In our experience, energy pooling over several levels is a general trait of valence isomer libraries, irrespective of the modeling approach.
(31) *Aldrich Chemical Catalog*; Sigma-Aldrich: Milwaukee, WI, 2001.
(32) White, A.; Handler, P.; Smith E. L. *Principles of Biochemistry*, 5th ed.; McGraw-Hill: New York, 1973; Chapter 14.
(33) Khinchin, A. I. *Mathematical Foundations of Statistical Mechanics*; Dover: New York, 1949. See, also: Khinchin, A. I. *Mathematical Foundations of Information Theory*; Dover: New York, 1957.
(34) See, for example: *Principles of Molecular Recognition*; Buckingham, A. D., Legon, A. C., Roberts, S. M., Eds.; Blackie Academic and Professional: London, 1993.
(35) See: Wiswesser, W. J. Historic development of chemical notation. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 258−263. See, also: *Essays on the History of Organic Chemistry*; Traynham, J. G., Ed.; Louisiana State University Press: Baton Rouge, 1987.

CI0400213