Mol Divers
DOI 10.1007/s11030-014-9517-7

COMPREHENSIVE REVIEW

Trends in information theory-based chemical structure codification

Stephen J. Barigye · Yovani Marrero-Ponce · Facundo Pérez-Giménez · Danail Bonchev

Received: 7 December 2013 / Accepted: 7 March 2014

Abstract This report offers a chronological review of the most relevant applications of information theory in the codification of chemical structure information, through the so-called information indices. Basically, these are derived from the analysis of the statistical patterns of molecular structure representations, which include primitive global chemical formulae, chemical graphs, or matrix representations. Finally, new approaches that attempt to go "back to the roots" of information theory, in order to integrate other information-theoretic measures in chemical structure coding are discussed.

Keywords Information theory · Chemical structure · Statistical pattern · Information indices · Shannon's entropy · Mutual · Conditional and joint entropies

S. J. Barigye (✉) · Y. Marrero-Ponce (✉)
Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy, Universidad Central "Martha Abreu" de Las Villas, 54830 Santa Clara, Villa Clara, Cuba
e-mail: sjbarigye@gmail.com; sjbarigye@yahoo.com

Y. Marrero-Ponce
e-mail: ymarrero77@yahoo.es; ymponce@gmail.com

Y. Marrero-Ponce
Facultad de Química Farmacéutica, Universidad de Cartagena, Cartagena de Indias, Bolívar, Colombia

Y. Marrero-Ponce · F. Pérez-Giménez
Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Valencia, Spain

D. Bonchev
Center for the Study of Biological Chemistry, Virginia Commonwealth University, P. O. Box 842030, Richmond, VA 23284-2030, USA

Introduction

Historical background on information theory

In the life of the legends of science, there is that *manūscrīptus* (manuscript), or more precisely, that theory, that perhaps without a peek into the full magnitude of its impact, forever immortalizes their contributions to science and engenders scientific revolutions that affect all walks of life. Such could be said of Claude Elwood Shannon's landmark paper published by *The Bell System Technical Journal* in 1948, entitled "A mathematical theory of communication" [1]. From quotidian applications like mobile phones, internet, music and video players to more sophisticated systems like deep-space probes, and in an impressively diverse range of disciplines like linguistics, sociology, taxonomy, psychology, molecular biology, economics, statistical physics, neurobiology, ecology, thermal physics, quantum computing, the list is endless; *information theory*, as denominated later, has proved to be of monumental importance [2–13]. Yet, in a way it is sad that when Shannon passed away in 2001, he was virtually unknown to many people, probably because the most remarkable benefits of information theory are reaped decades after its introduction.

Definitely, it is natural to wonder what really makes information theory so important or applicable to such diverse fields of science. A search for an overly complex mathematical explanation may turn out to be frustrating. It is rather the simplicity in Shannon's inferences that awards his theory the elegance and intuitive applicability. Shannon suggests that by determining the ultimate limits of optimal communication, it is actually possible to achieve asymptotically error-free communication schemes, influenced by three important factors: (a) statistical knowledge of the information source, (b) the effect of noise in a communication channel, and (c) the

Published online: 05 April 2014

nature of the final destination. These theoretical limits are the source entropy and channel capacity as the lower and upper bounds, respectively. Certainly, the choice of the term *entropy* is rather unanticipated. A legend relates that Shannon chose the word entropy as a suggestion from von Neumann that, "call it entropy. No one knows what entropy is, so if you call it that you will win any argument" [14]. Interestingly, the existence of a parallel relationship between entropy formulations in the statistical mechanics and information-theoretic sense, according to Boltzmann and Shannon, respectively, has been demonstrated [10,15]. More than elucidating the similarity between these expressions, it has been clarified that the most accurate interpretation of classical thermodynamics entropy should be in terms of Shannon's measure of information, rather than as a measure of disorder. While the former is true for all thermodynamic processes, there are processes where the order-disorder interpretation does not hold [16].

In applications of information theory in other disciplines, interest has been placed on the analysis of the statistical structure (pattern) of information sources, in generic terms [2–11]. The justification for this paradigm is that common principles underlie the universe as the overall source of information.

Information theory in molecular structure characterization

This article offers a review of perhaps one of the least celebrated applications of information theory: theoretical characterization of molecular structures (as an information source) in mathematical chemistry, through the so-called information indices [17–20]. These are a subset within the universe of molecular structure characterizing parameters, collectively denominated *molecular descriptors* (MDs). For a comprehensive treatise of MDs, see ref [21]. Several reports could be mentioned in the literature that attempt to summarize the most relevant aspects of the IFIs defined so far [17–20]. However, no attempt is made to go backward and forward through information theory and its naturally related ideas, so as to reap the interesting analogies applicable to molecular structures. As a result there is a temptation to the inclination for regarding IFIs as "*graph invariants that view the molecular graph as a source of different probability distributions to which information theory definitions can be applied*" [21]. Note that "*information theory definitions*" is in reference to Shannon's entropy formula (see Eq. 1) and mathematical variants derived thereof. Of course, this statement is not improper, but leaves some unanswered questions, for example, what the conceptual constraints of its application are, or even the implication of the obtained result in information theoretic terms. Such interrogatives require a brief passage to the "place" where it all begun, *i.e.*, digital communication.

In this report an attempt is made to go back to the roots, and guide the readers through this exciting journey of the application of information theory in mathematical chemistry, certainly placing emphasis only on the aspects that are crucial to this context. Before that we will give a brief chronological outline of the IFIs defined so far, stratified according to the nature of the considered information source. Finally, novel insights toward molecular structure codification, derived from a purely information-theoretic understanding, are discussed (Figure 1).

Statistical patterns in chemical structure representations

In Chemistry, several approaches are used in the representation of molecular structures, and different classification schemes may be adapted for these representations, with the most common one being the dimensionality. Molecular representations are approximations (alphanumeric, topological, geometric, etc.) that attempt to describe chemical reality, and range from simple chemical formula to more complex models like 4D structural representations [21,22].

The ensuing IFIs a result of the analysis of the statistical structure of these models, using a quantity that measures how much "choice" or uncertainty is involved in the random selection of an event in a given model. This quantity is known as *Shannon's entropy* or *entropy of information* and is defined by:

$$H = -\sum_{i=1}^{n} p_i \cdot log_2 p_i \tag{1}$$

For brevity, in what follows the term *entropy* will be used. We will now discuss the different models as sources of chemical information, quantified as *source entropy*.

Note that the analysis of nuclei structure models as chemical information sources is beyond the scope of this report and will not be covered, see refs [23–26] for discussions in this context. Information-theoretic measures have also been used in the interpretation of electronic structure phenomena in the Hirshfeld partitioning scheme. Discussions in this context could be found in refs [27–30].

Chemical formula (0D molecular structure representation)

This constitutes the simplest molecular structure representation comprised of alphanumeric codes with the letters representing the different atom types, each accompanied by a number (subscript) representing the incidence of the atoms in the molecule. In fact, the first information index, defined in 1953 by Dancoff and Quastler [31], was derived from the analysis

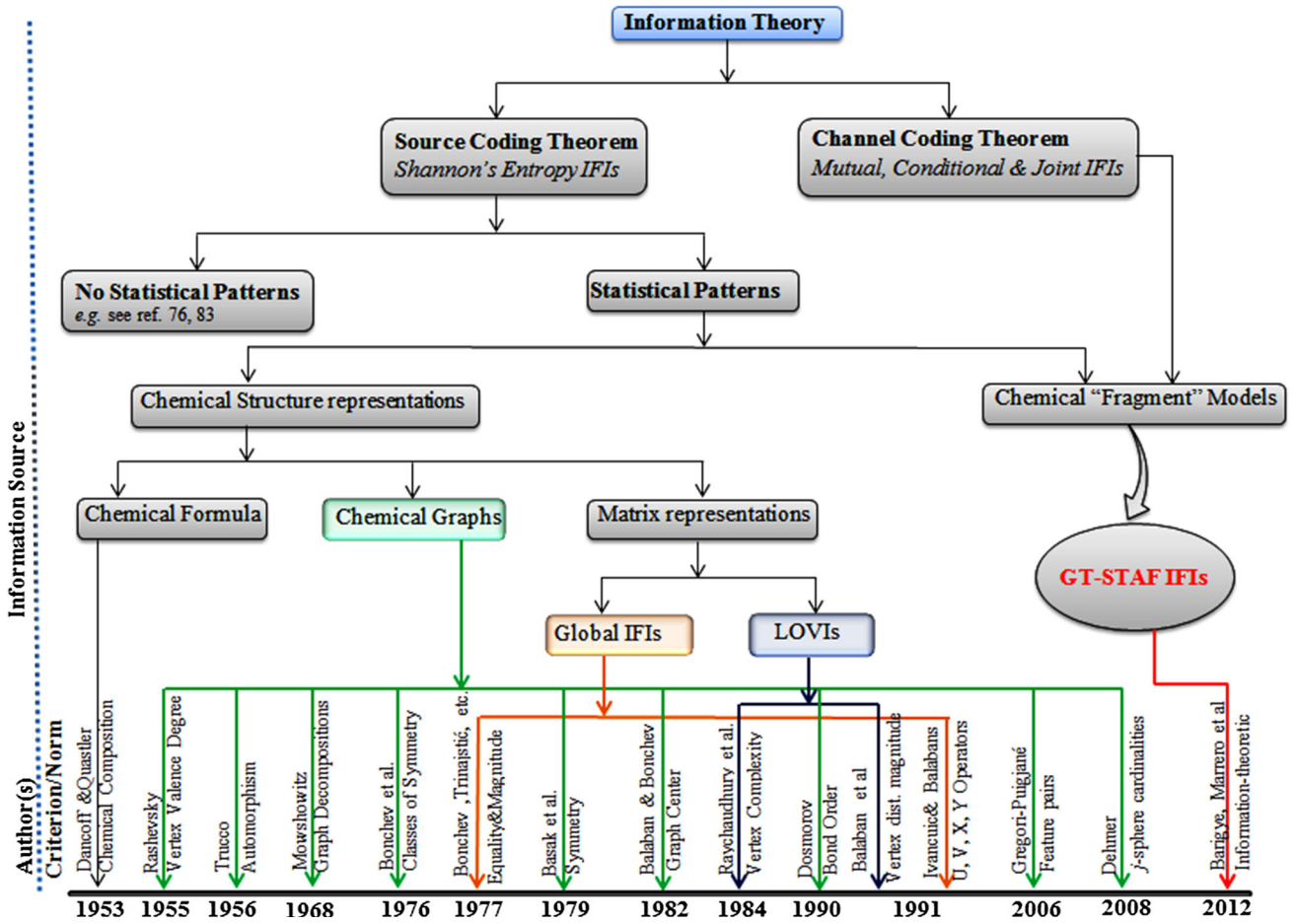

Fig. 1 Chronological description of the most relevant IFI definitions, structured according to the different information sources. Note that some IFI definitions involved various stages of development. In this case, only the pioneering definitions are included

of the statistical structure of this primitive representation as the information on the types of atoms in a molecule, thereby creating an ideal hierarchy in the chronological development of IFIs. Accordingly, atoms of the same chemical element are clustered together and their probabilities calculated, forming a **probability distribution function** (p.d.f). Applying Eq. 1 to this p.d.f gives the *information index on chemical composition*, which is a measure of the compound compositional diversity.

Chemical graphs as an information source

The use of graphs to represent molecular structures dates way back to the 19th century with the seminal work of Arthur Cayley in which he attempts to enumerate alkane isomers [32]. Other early contributions to chemical graph theory are attributed to James J. Sylvester and Crum-Brown, see ref [33] for a detailed treatise. A chemical graph has been defined as a model of the chemical system used to characterize the interactions between chemical objects (atoms, bonds, groups of atoms, molecules, ensembles of molecules, etc), see ref [20]. Evidently, graphs as approximations of chemical reality are expected to contain important structural information (see Fig. 1) [22,34–36].

Pioneering work to analyze the possible statistical patterns of molecular graphs in topological terms was performed by Rashevsky in 1955, using the vertex valence degree as criterion for topological homogeneity [37]. It follows that, two vertices are equivalent if at (graph) distance k from each vertex, where $1 \leq k \leq \eta$, there exists vertices of equal valence vertex degree (δ). Thus topologically equivalent vertices are members of equivalent class g_i for $i \in \{1, ..., m \,|m \leq n\}$ from which the graph entropy, denominated the *topological information content index*, is computed.

Trucco, a year later gives a more accurate definition for this index based on the notion of automorphism graphs and orbits [38,39], which he denominates the *vertex orbital information indices*. It follows that vertices belong to the same equivalence class if permutations on this class are structure preserving, *i.e.*, they belong to the same vertex orbit of the automorphism group. This notion was extended to consider edge orbits, defining the *edge orbital information indices*.

Yet more complex considerations in this line-treated subgraph orbits as a generalization of edge orbits. This is discussed in a little later so as to maintain a proper chronological order, see *Bertz Index* (see Fig. 1).

Graph complexity is further studied by Mowshowitz, formalizing mathematical definitions of relative complexity of graphs [40]. He discusses an entropy measure based on decompositions corresponding to a class of graph homomorphisms (information content of a structure relative to a system of symmetry transformations that preserve the system's invariance). In ref [41], he introduces an index for the *chromatic information content*, I_c ***(G)*** of a graph G defined as the minimum entropy over all finite probability schemes constructed from chromatic decompositions having rank equal to the chromatic number of G.

Toward the end of the 1960s, there is significant shift of interest among theoretical chemists from the direct analysis of structural graphs to the use of matrix theory and algebraic methods on graphs to characterize molecular structures. The use of matrix representations as a source of information will be discussed later (see Fig. 1).

However, a few years later, interest in molecular graphs as a source of information is revived. In ref [42], Bonchev et al. introduce the *molecular symmetry index* based on the distribution of atoms in different classes of symmetry in a molecule. Each class of symmetry includes atoms able to exchange position through operations of the symmetry point group to which the molecule belongs. This molecular symmetry index complements the orbital information index, in accounting for specific molecular geometry and conformations. Using the so-called orbits on the automorphism group of graph connections as criterion for equivalence, Bertz proposes a more complex definition of orbital information indices (IFIs) considering adjacent edges, multiple edges, and loops [43], denominated as the *connection orbital information content* or *Bertz index*. This index yields a more detailed description of the molecular structure than previous similar indices. In ref [44], mention is made on the possibility of using more complex subgraphs but without further comment (see Fig. 1).

Motivated by the notion of Hosoya's graph decomposition [45], Bonchev and Trinajstić [46] define mean and total information content, denoted by $\bar{I}_Z$ and I_Z, respectively, to analyze the statistical nature of k-matchings of a molecular graph, where the cardinality of the equivalence class c_k, is equal to the non-adjacent number $p(G, k)$ expressed as the number of ways of choosing k disjoint lines from a given graph G. Note that in the case of acyclic graphs, $\bar{I}_Z$ and I_Z are equivalent to the IFIs on polynomial coefficients.

In a series of reports with the first in 1979, Basak et al. "redefine" the index for orbital information content, to codify molecular complexity for hydrogen-filled multigraphs, through the so-called indices of neighborhood symmetry [47–52]. It follows that two vertices v_i and v_j of a multigraph MG are said to belong to equivalence set of kth order topological homogeneity if they satisfy the following conditions: (1) are of the same chemical element, (2) possess the equal vertex degrees, (3) the same conventional bond order, and (4) the same atomic neighborhood up to the kth order. Note that in the case of graph vertices of the same chemical element, the neighborhood information content of maximal order calculated for the H-depleted molecular graph coincides with orbital information content. Other related indices are: structural information content (SIC), bonding information content (BIC), complimentary information content (CIC), and redundant information content (RIC). This methodology has been widely used in QSAR and QSPR studies with relevant results (see Fig. 1) [47–49,51,52].

Information theoretic measures for the statistical distribution of vertices with respect to the graph center, collectively denominated *centric information indices*, were proposed by Balaban [53,54] and Bonchev [19,55,56], in order to quantify the "clustering" tendency of the vertices (or edges) about the graph center (or polycenter). An all-inclusive hierarchical algorithm for the identification the graph center was also proposed by these authors, achieving plausible discrimination of graph vertices (or edges), see ref [55].

The *information bond index* for a molecule, as proposed by Dosmorov [57], is computed as the total information content with respect to the conventional bond order, *i.e.*, single, double, triple, and aromatic bonds (see Fig. 1).

Recently Dehmer et al. [58–61] have proposed atom-based IFIs on the complete topological neighborhood of each vertex in G, using information functional approach on the j-sphere cardinalities of a graph. A measure for local entropy of G using the information functional has also been defined, see ref [18].

A peculiar information-theoretic measure derived directly from the topology of the molecular graph is proposed in ref [62] where Shannon's entropy is computed over the statistical distribution of atom-centered feature pairs at different paths lengths. A set of entropic values for the distribution of all the atom-centered feature pairs in a molecule forms the corresponding molecular profile, denominated the SHED profile (see Fig. 1).

Matrix representations as an information source

Matrix representations constitute probably the most important source of MDs, in general. Although the use of matrices in graph theory dates way back to 1900, with the seminal work of Poincaré on incidence matrices [63], it is not until 1970s that interest develops in the use matrix representations for chemical graphs. This is mainly attributed to the pioneering work by Harary [64] in 1969 and Hosoya [45] in 1971 on the use of the distance matrices in chemical graphs theory (CGT), which opened way to the definition of an

impressively wide collection of graph-theoretic matrix representations. The introduction of the distance matrix in CGT, permitted to give a matrix-based understanding of even the first known topological index, the Wiener index. A comprehensive monograph of matrices used in CGT is provided by Janezic et al., see ref [65].

Matrix representations, viewed as an information source, have permitted defining numerous IFIs, through the analysis of matrix statistical patterns. The IFIs on the vertex distance matrix, proposed by Bonchev and Trinajstić in 1977, were the first IFIs derived from the analysis of the statistical structure of graph theoretic matrices [46]. Other graph theoretic matrices that were later used as an important source of IFIs are: vertex- and edge- adjacency matrices, edge-distance matrices, and vertex- and edge-cycle incidence matrices (see Fig. 1).

Whole molecule information indices

Generally, the computation of these IFIs constitutes the analysis of the statistical pattern of the matrix elements. Taking into consideration that practically the same formalism is followed for all matrix representations in the definition of matrix-based IFIs, we will only take as an example the distance matrix, following the original definitions proposed by Bonchev and Trinajstić [46].

Let $\mu(g_i)$, denote the cardinality of set g of homogenous elements i (equivalence class (g_i) where $1 \leq i \leq \rho(G)$, $\rho(G)$ is the diameter of G. Two criteria have been followed in the definition of the equivalence classes, that is, equality and magnitude.

Equality criterion

Corollary *Matrix entries belong to an equivalence class if their values are equal. Using the equality criterion, two information measures are defined, total information content on the distance equality and mean information content on the distance equality, expressed by Eqs. 2 and 3, respectively:*

$$I_D^E = \frac{N(N-1)}{2}\log_2\frac{N(N-1)}{2} - \sum_{i=1}^{\rho(G)}\mu(g_i)\log_2\mu(g_i) \quad (2)$$

$$\bar{I}_D^E = -\sum_{i=1}^{\rho(G)}\frac{2\mu(g_i)}{N(N-1)}\log_2\frac{2\mu(g_i)}{N(N-1)}. \quad (3)$$

Let G denote a connected graph with a finite set of vertices V. The topological distance d_{ij} the length of the shortest path that connects vertices v_i and v_j of G [66]. Taking the graph in Fig. 2a as an example, Matrix D is the distance

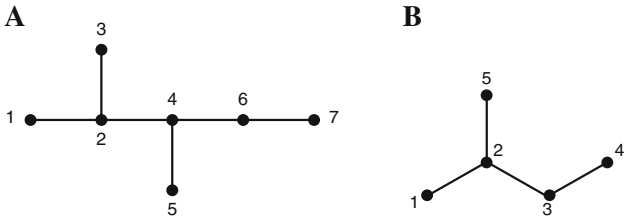

Fig. 2 **a** Branched tree graph, **b** The labeled chemical graph of the molecule of Isopentane (the numbers correspond to the labels that are assigned to the non-hydrogen atoms (vertices) in the molecular structure)

matrix corresponding to this G (see Fig. 1).

$$D = \begin{bmatrix} 0 & 1 & 2 & 2 & 3 & 3 & 4 \\ 1 & 0 & 1 & 1 & 2 & 2 & 3 \\ 2 & 1 & 0 & 2 & 3 & 3 & 4 \\ 2 & 1 & 2 & 0 & 1 & 1 & 2 \\ 3 & 2 & 3 & 1 & 0 & 2 & 3 \\ 3 & 2 & 3 & 1 & 2 & 0 & 1 \\ 4 & 3 & 4 & 2 & 3 & 1 & 0 \end{bmatrix}$$

For simplicity, only elements of the upper triangle submatrix are considered in the computation of the IFIs of distances, thanks to the symmetrical nature of D. Note also that the elements in the principal diagonal (zeros) are not considered since they do not offer any structural information.

The matrix entries are partitioned (distributed) in equivalence classes and their respective cardinalities obtained:

$$\mu(g_1) = 6; \quad \mu(g_2) = 7; \quad \mu(g_3) = 6; \quad \mu(g_4) = 2.$$

Now let's apply Eqs. 2 and 3: $I_D^E = 39.568\,bits$, $\bar{I}_D^E = 1.884\,bits\ per\ element$

Magnitude criterion

The high degeneracy of the equivalence-class-based IFIs, as they were denominated by Bonchev and Trinajstić [46], prompted them to devise means of reformulating these IFIs to increase their discriminating power, and thus applicability in QSAR studies. This incentive yielded a new class of IFIs, the magnitude-based IFIs (see Fig. 1) [20,46].

Corollary *An element is considered as an equivalence class whose cardinality is equal to the magnitude of the element* [19,21].

As in the case of the equivalence criterion, the magnitude criterion yields two information measures : *total information content on the distance magnitude* and *mean information content on the distance magnitude*, expressed by Eqs. 4 and 5:

$$I_D^W = W \log_2 W - \sum_{i=1}^{\Delta(G)} \mu(t_i) * i \log_2 i \tag{4}$$

$$\bar{I}_D^W = - \sum_{i=1}^{\Delta(G)} \mu(t_i) \frac{i}{W} \log_2 \frac{i}{W}, \tag{5}$$

where W is the Wiener's number (the total sum of vertex–vertex distances in the graph or half-sum of all the elements d_{ij} of the distance matrix) [21,67].

Further applications of information theory on the vertex distance matrix dealt with the analysis of the statistical pattern of vertex distance degrees σ_i according to the equality and magnitude criteria, respectively, see paper by Skorobogatov et al. (see Fig. 1) [68].

A special kind of distance matrix is the molecular influence (leverage) matrix, derived from the spatial coordinates of atoms in a given molecular conformation. Using this matrix representation, Consonni et al. proposed the GETAWAY IFIs [69,70].

Information indices as local vertex invariants (LOVIs)

Based on the rationale that some chemical and biological properties do not depend on the whole molecular skeleton but rather specific features (*e.g.*, functional groups) within a molecule, the late seventies are characterizing by increasing interest placed on obtaining MDs defined at atomic or substructure level, in some sort of "bottom-up" approach, with the most successful locally defined indices being the electrotopological indices proposed by Kier and Hall [71]. The IFIs would definitely not lurk behind.

A series of information-theoretic local vertex invariants (LOVIs) and their respective molecular graph invariants on the distance matrix are proposed through a concerted effort of various authors, with important contributions from Raychaudhury et al., Klopman, Balaban, Ivanciuc, and most recently, Kostantinova et al. [50,72–78]. Therefore, instead of partitioning the matrix elements, or the vertex degrees in equivalence sets according to their homogeneity, an analysis of the statistical arrangement of vertices with respect to a reference point, *i.e.*, vertex $v_i \in V$ is performed. Bearing in mind that some of these measures are related, only the key notions will be explained, and the corresponding variants will be simply mentioned (see Fig. 1).

Using the equality criterion, the *vertex complexity index* (v_i^c) was defined as:

$$v_i^c = - \sum_{m=0}^{\eta_i} \frac{{}^m\mu(g_i)}{N} \log_2 \frac{{}^m\mu(g_i)}{N}, \tag{6}$$

where ${}^m\mu(g_i)$ is the cardinality of the equivalence set of distances from vertex v_i equal to m, η is the atom eccentricity, and N is the number of graph vertices [50].

The magnitude criterion analog of the vertex complexity index, denominated the *vertex distance complexity*, $\tilde{v}_i^d$, or *mean local information on distances*, u_i was also defined:

$$\begin{aligned} \tilde{v}_i^d \equiv u_i \equiv H_{(D)i} &= - \sum_{m=1}^{\eta_i} {}^m\mu(g_i) \frac{m}{\sigma_i} \log_2 \frac{m}{\sigma_i} \\ &= - \sum_{j=1}^{A} \frac{d_{ij}}{\sigma_i} \log_2 \frac{d_{ij}}{\sigma_i}, \end{aligned} \tag{7}$$

where ${}^m\mu(g_i)$ is the cardinality of the equivalence set of distances from vertex v_i equal to m, σ_i is ith vertex distance degree. Derivations of the vertex distance complexity index include: normalized vertex distance complexity and the relative vertex distance complexity ($\boldsymbol{v_i}$), see refs [50,72–75]. Other closely related measures are the information content on vertex distance magnitudes, unfortunately denominated the "mean" extended local information on distances ($\boldsymbol{\gamma_i}$) and the extended local information on distances ($\boldsymbol{x_i}$), computed as the total information content on vertex distances from vertex $v_i \in V$ [74,75].

Several mathematical operators have been applied to obtain the respective global (molecular) graph invariants. Konstantinova and Paleev in ref [76] use the summation operator on a vector of LOVIs $H = [H_{(D)i} | 1 \le i \le N]$, where N is the number of vertices in the G, to obtain *the information distance index* (H_{D}). Note that the summation operator was originally proposed by Raychaudhury et al. [50] on v_i^c and $\tilde{v}_i^d$, yielding the graph vertex complexity and graph distance complexity indices, respectively. Their difference with respect to H_{D} lies in the use of normalizing factors in their computation (see Fig. 1).

Later on Ivancuic and the Balabans [74,75,77], propose probably the most relevant global information-theoretic operators, namely: U, V, X, and Y indices, for the corresponding locally defined measures on the vertex distance matrix, *i.e.*, u_i, $v_{i,}$, x_i, and γ_i, respectively. Expressions of these operators on LOVIs derived from vertex distances are given by Eqs. 8–11:

$$U = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (u_i \cdot u_j)^{-1/2} \tag{8}$$

$$V = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (v_i \cdot v_j)^{-1/2} \tag{9}$$

$$X = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (x_i \cdot x_j)^{-1/2} \tag{10}$$

$$Y = \frac{B}{C+1} \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} a_{ij} \cdot (\gamma_i \cdot \gamma_j)^{-1/2}, \tag{11}$$

where a_{ij} are the elements of the adjacency matrix, B the number of edges, and C the cyclomatic number. Examples of QSPR evaluations successfully performed with these indices could be found in refs [77,79,80]. Note that these information theoretic operators have been generalized to attain applicability for all graph-theoretic matrices, see refs [81,82].

There exist other entropic measures, such as *electronic delocalization entropy* proposed in ref [83], though strictly speaking, they are not information-theoretic measures. It is thus not surprising that these descriptors are not considered as IFIs in ref [21] (see Fig. 1).

Back to information theory: new trends in information indices

In this section, we attempt to trace the origins of information theory, discussing its fundamental concepts (or results) in a simplified and understandable manner, with the aim of comprehending the allure of this theory and discover possible analogies applicable to chemical structures, as communication systems.

Statistical patterns and source coding

Consider, as an entropy source, a natural English text retrieved from the *NewYork Times* (see Fig. 3a). This text is comprised of a set of words which belong to the universal collection of words referred to as a *dictionary*. On average, irrespective of the context of the text, there are alphabetical characters that tend to appear more frequently than others. In other words, in an English text, the characters are not completely random, but rather there exists a statistical structure, in that the probability of appearance of some characters is higher than others.

If we need to store this text, or to transmit this message to a recipient, then it is desirable that we are able to encode this text in the smallest possible bits or size (*i.e.,* data compression) so that it occupies smaller space on a storage disk, or to enable faster and reliable transmission, and when decoded (decompression) the original text is retrieved without loss of information [14,84,85].

This inference brings us to *Shannon's source coding theorem (SCT).* In source coding, it is known that through appropriate choice of *variable-length codes* [14,84,85], with highly probable symbols (or words) assigned short codewords and low probability ones longer codewords, low weighted average codeword length could be achieved (see illustration in Fig. 3b and 3c). Note that this selection of the codes should be such that these are uniquely decodable, that is, the codewords should be prefix-free. However, one logical question arises: "What is the smallest average codeword length ($L_{\min}$), achievable for a given symbol originator (input source)?" The SCT establishes that this theoretical lower bound is in fact given by the *source entropy*, H, obtained directly from the statistical structure of the message.

Given a source A describing particular event with a set of symbols $\{a_1 \ldots a_n\}$ and a p.d.f. $p(a) = \{p_1 \ldots p_n\}$, the entropy for this source, $H(A)$, is given by Eq. (1).

On the other hand, the expected codeword length l_i of the coded source symbols, $L(A)$, is defined as:

$$L(A) = -\sum_{i=1}^{n} p_i \cdot l_i, \tag{12}$$

where p_i represents the probability associated to symbol a_i.

It follows that $H(A) \leq L(A)$. In the case where the expected code word length matches the source entropy, it is said that the code is optimal [14,84,85].

In the illustration in Fig. 2, the source entropy, $H(A)$ and the weighted average codeword length $L(A)$ are 4.031 and 4.304 bits, respectively; thus $H(A) < L(A)$. Evidently, adapted coding scheme in this example is not necessarily the most optimal one. There exist coding algorithms that yield closer approximations to the theoretical bound of the

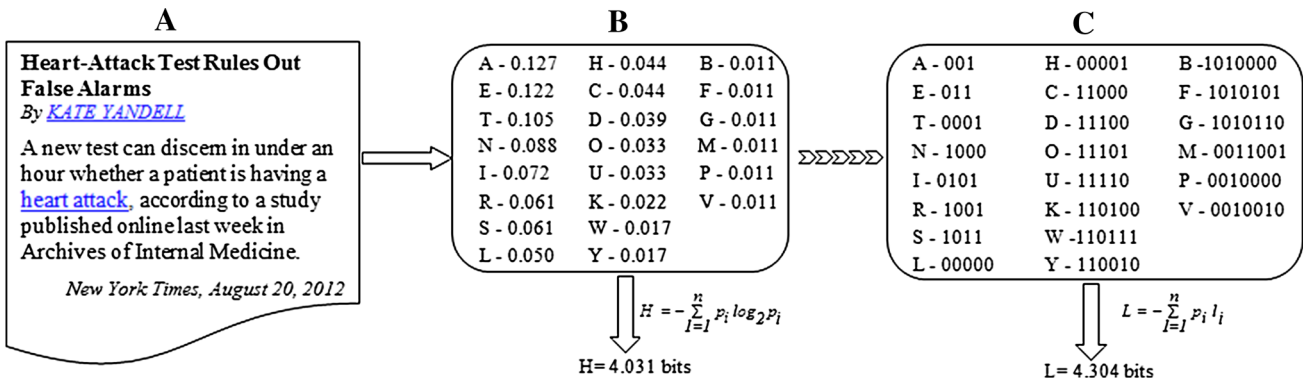

Fig. 3 Illustration of computation of source entropy and average codeword length using variable-length codes

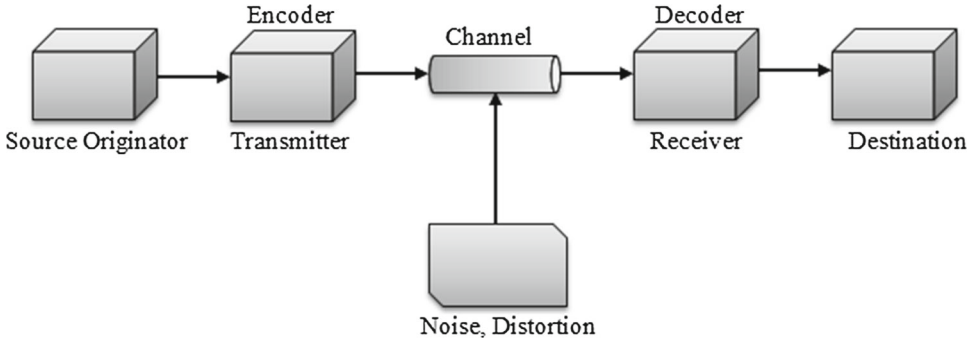

Fig. 4 Schematic diagram of a communication system with a noisy channel

expected codeword length, for discussions about optimal source coding, see refs [14,84,85].

In the illustration above, the words were considered to be comprised independent alphabetical characters. However in a normal text, character concatenations are not a sheer coincidence, that is, some sequences are more probable than others. For example in the English language, diagrams like TH, HE, or AN are more frequent than let's say XP, KZ, WZ, etc. This analysis could be extended to consider more complex concatenations such as trigrams, words, or even phrases. Here the point is to assign the shortest codewords to the most frequent n-grams and the longest codewords to the less frequently used ones. This inference stirred interest in use of the so-called block codes to achieve better data compressions (*i.e.*, reductions in the mean bit/codeword), and thus faster transmissions [86,87]. Two probability schemes could be explored to describe the statistical structure of an information source of this nature: (1) The transition (or conditional) probabilities for n-gram sequences, and (2) The joint probabilities for n-gram structures, *i.e.,* considering relative frequencies of character concatenations.

Channel-coding theorem

We will now discuss some aspects of channel coding placing emphasis on the ones that will be crucial in posterior stages. Consider that when this text, that we shall for convenience call, text X is transmitted along a communication channel, and text Y is received at the destination. The fundamental question is if the message that was sent from the source is identical to the one received at the other end. Two extreme cases do exist:

(1) Text X identical to text Y (noiseless channel): This means that text X did not undergo any distortions, along the communication channel in the sense that no part of this information was lost or altered. This means that $p(x, y) = p(x)$, and thus the entropy of Y, $H(Y)$ is a deterministic function of $H(X)$.
(2) Text Y is independent of text X(useless channel): In this case, it does not matter what text is sent from the source, text Y will be indifferent. This means that $p\,(x,y) = p(x)p(y)$.

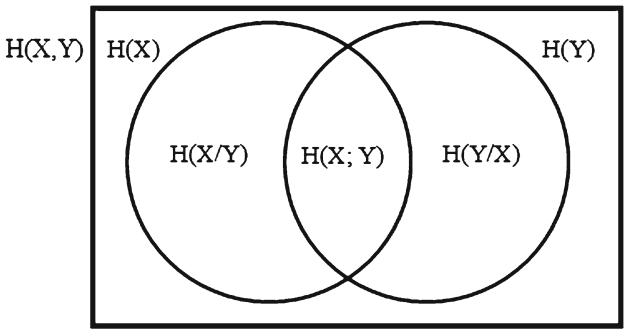

Fig. 5 Venn diagram illustration for the different entropy measures calculable for a noisy channel

In between these extremums, there are messages that will be transmitted with some degree of distortion probably due to noise, or some kind of physical distortion. Such transmissions are characteristic of noisy channels. Figure 4 is a schematic diagram of a communication system containing a noisy channel.

Here, two statistical processes are involved: the source and the noise. Several entropy measures could be calculated: the source (input) entropy, $H(X)$, the entropy of the received message (output), $H(Y)$, the joint entropy of input and output, $H(X, Y)$, and finally two conditional entropies $H\,(Y/X)$ or $H\,(X/Y)$, the entropy of the output when the input is known or conversely, see Fig. 5 for Venn diagram illustration.

The entropic measures $H(X, Y)$ and $H\,(Y/X)$ are defined according to:

$$H(y/x) = -\sum_{x}\sum_{y} p(x, y)\log p(y/x) \tag{13}$$

$$H(x, y) = -\sum_{x}\sum_{y} p(x, y)\log p(x, y). \tag{14}$$

Note that,

$$H\,(x, y) \leq H\,(x) + H\,(y) \tag{15}$$

with equality only when $p\ (x,y) = p(x)p(y)$, *i.e.*, in the case of independent sources.

It is easily verified from set's theorem (see Fig. 5) that,

$$H\ (x,\ y) = H\ (x) + H\ (y/x)\ . \tag{16}$$

From Eqs. 15 and 16, it is evident that $H(y) \geq H(y/x)$. This means that conditioning never increases the entropy of an event.

The key interest in this case is whether we are able to transmit a message over a noisy channel with a vanishing error probability, achieving satisfactory approximations to the original message. A measure of the correlation between X and Y, denominated *mutual information (MI)*, helps determine the amount of "additional" information to be introduced to correct the message errors. It is logical, however, that to determine MI, knowledge about the uncertainty (entropy) introduced by the channel noise, conveniently called *equivocation* is necessary. This equivocation is the conditional entropy $H(Y/X)$, that is, the uncertainty of Ygiven the knowledge of X. Subtracting the conditional entropy from the output entropy $H(Y)$, yields the MI, denoted by $H(X;\ Y)$. Shannon's channel coding theorem, establishes the upper limit of "trustworthy" communication along a noisy channel as the maximum mutual entropy, max $H(X;\ Y)$, also known as the channel capacity [14,84,85]. Further information-theoretic inferences related to transmission along a noisy channel are beyond the scope of this MS.

A "metric" understanding of MI is proposed by Kullback and Leibler [88] as a special case of a more general quantity-denominated *relative entropy* or *Kullback-Leibler divergence*. The relative entropy, denoted by D(p||q), is the "distance" between two probability distributions $p(x)$ and $q(x)$. It could also be understood as a measure of the additional bits of information necessary to correct the error in assuming that the probability distribution of a source is $q(x)$ when in reality is $p(x)$. The D(p||q) is given by the formula:

$$D\ (p\ \|q\ ) = \sum_{x\in X} p(x) log \frac{p(x)}{q(x)}. \tag{17}$$

In preceding paragraphs, we discussed two extremums for channel coding corresponding to the noiseless and useless channel, respectively. In the case of a noisy channel, an approximation to error-free communication requires a tradeoff between these extremums, in that $p(x,\ y)\ >\ p(x)p(y)$. This tradeoff is equivalent to the relative entropy for the two probability distributions $p(x,\ y)$ and $p(x)p(y)$. If $p(x,\ y)\ >>>\ p(x)p(y)$, it means that x and y are highly correlated, while if $p(x,\ y)\ -\ p(x)p(y)\ \rightarrow\ 0$, x and y are weakly correlated. This means that relative entropy is in fact directly related to MI. Thus MI is the measure of the inefficiency in assuming the channel probability distribution of $p(x)p(y)$ when in reality it is $p(x,\ y)$. From Eq. 18, we can thus express MI as:

$$D\ (p(x,\ y)\ \|\ p(x)q(x)\ ) = H(X;\ Y) = \sum_{x}\sum_{y} p(x,\ y)\ \log\ \frac{p(x,\ y)}{p(x)q(x)}. \tag{18}$$

Note that when $p(x,\ y)\ =\ p(x)p(y)$, $H(X;\ Y)\ =\ 0$ and when $p(x,\ y)\ =\ p(x)$, $H(X;\ Y)\ =\ H(X)$

New insights in information theoretic chemical structure codification

Source coding theorem in information index computations

Having introduced these concepts and procedure in the preceding section, ground is provided for their application to chemical structure codification. Consider as an information source S a set of subgraphs that describe a molecular structure. According to ref [21], a *molecular subgraph* is a subset of atoms and related bonds, which is in itself a valid graph usually representing molecular fragments and functional groups. The set S is formed following predefined models, known as events, based on graph-theoretic, physicochemical, or chemical considerations [89]. These events constitute the semantic context of the information source (or message).

From the illustration in Fig. 3, logical analogies could be traced. The subgraphs are equivalent to the words in the input message, with these formed by concatenations of alphabetical characters, which are the vertices in the former. The idea in this case is to determine of the chemical source entropy as a measure of the structural diversity. We will consider as an illustration the chemical graph of Isopentane (see Fig. 2b). In a recent publication, various events were proposed as criteria for generating sets of subgraphs [89].

In this example, the connected subgraphs algorithm will be exclusively considered. This model is based on the exploration of subgraphs of different orders in a chemical graph. A treatise on other source originator models can be found in ref [89].

Accordingly, for G in Fig. 2b, the connected subgraphs obtained for different orders based on the atomic relations are:

Order 0: C_1, C_2, C_3, C_4, C_5
Order 1: C_1–C_2, C_2–C_3, C_3–C_4, C_2–C_5
Order 2: C_1–C_2–C_3, C_1–C_2–C_5, C_2–C_3–C_4, C_2–C_3–C_5
Order 3: C_1–C_2–C_3–C_4, C_2–C_3–C_4–C_5, C_1–C_2–C_3–C_5
Order 4: C_1–C_2–C_3–C_4–C_5

These subgraphs will constitute the information source. Subsequent source entropy computation is easy (see illustration in Fig. 6a).

Other molecular structure entropy measures like: negentropy and standardized Shannon's entropy could be applied

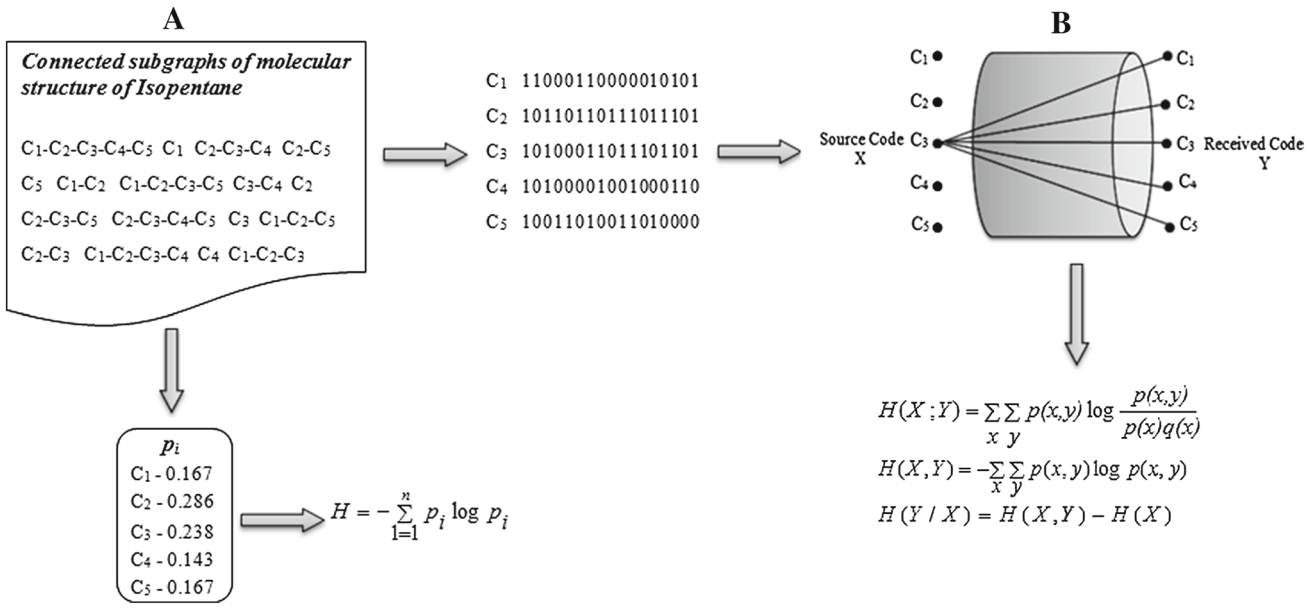

Fig. 6 **a** Calculation of chemical source entropy for the molecular graph of Isopentane. The chemical source is a set of connected subgraphs (for details see refs [89–91]); **b** Graphic illustration of communication along a noisy channel. During the **MI**, **CE**, and **JE** computations, each input code is matched with the output codes for comparisons on the degree of similarity among these (see below)

to the p.d.f obtained for the chemical source in 6a, but these will not be discussed here as are not classical information theory measures [21].

Likewise, we may be interested in designating codewords to source symbols (vertices) using a coding-tree scheme based on the incidence of vertices (letters) in the subgraphs (words), forming n-length binary codewords, where n is the subgraph number (fixed codeword size). Let us consider as *payload* the vertex incidences and the subgraphs as *overhead information (*describing how to handle payload). In this algorithm, codewords are assigned to vertices according to successive choices between 0 and 1 at each branch in the coding-tree scheme. Given a set of subgraphs, $S = \{s_g | 1 < g < s_n\}$, generated according to a predefined criterion, the codeword for v_i is sequentially assigned:

1, if v_i is included in s_g, where $1 < g < s_n$
0, otherwise

For the chemical source in Fig. 2, the corresponding fixed length (17 bit) codewords for the vertices would therefore be:

C_1 11000110000010101
C_2 10110110111011101
C_3 10100011011101101
C_4 10100001001000110
C_5 10011010011010000

Certainly, the interest here is not code optimality but rather dissimilarity of the different vertex codewords, if applicable. Indistinctive (not uniquely decodable) codewords are not "penalized" but rather considered informational about the topological similarity of the compared vertices.

Channel-coding theorem in information index derivations

In a noisy channel, due to signal distortions and/or additive noise, transmitted data are usually susceptible to errors, in that if a codeword sequence for vertex v_x is sent from a source originator, then it may not necessarily be the same obtained at the receiver's end, but rather one for vertex v_y (see Fig. 6b for illustration). The MI for vertex codewords for v_x and v_y, $H(X;Y)$ gives a measure of the true information content at the receiver's end.

For vertex codeword pairs $({}^c v_x, {}^c v_y)$, mutual frequencies and subsequent joint probabilities, $p(x, y)$ for 1 bit length "sequences" are computed. Thus a joint p.d.f $P(X, Y)$ is formed, where

$$P(X,Y) = \{p(x,y) : p(x,y) = f(x,y)/f_T | x \not\leq y \wedge f_T = \sum_{x=1}^{n}\sum_{y=1}^{n} f_{(x,y)}, x = y\}.$$

When Eq. 19 is applied to the joint p.d.f $P(X, Y)$, the molecular *MI index* is obtained. Likewise, Eqs. 14 and 16 are used to calculate the *joint and conditional entropy* IFIs, designated by *JE* and *CE*, respectively.

This approach could be extended to consider digram or trigram joint probabilities. We, however, limited ourselves to 1 bit sequences for simplicity.

It should be remarked that since the zeros in the vertex codewords do not lie in the context of the source originator

algorithms (are indicative of nonparticipation of vertices in subgraphs rather than participation), their mutual frequencies are not considered.

We will now demonstrate the calculation of the *MI*, *JE*, and CE indices, using as an example the molecular graph of Isopentane just as in the "Source coding theorem in information index computations" Section. Given the computational advantage offered by matrix-based operations, the mutual frequencies and their respective joint probabilities are condensed in frequency and joint probability matrices, designated by F and P, respectively:

$$\mathrm{F} = \begin{bmatrix} 7 & 6 & 4 & 2 & 3 \\ 6 & 12 & 8 & 4 & 6 \\ 4 & 8 & 10 & 5 & 4 \\ 2 & 4 & 5 & 6 & 2 \\ 3 & 6 & 4 & 2 & 7 \end{bmatrix}$$

$$\mathrm{P} = \begin{bmatrix} 0.167 & 0.143 & 0.095 & 0.048 & 0.071 \\ 0.143 & 0.286 & 0.190 & 0.095 & 0.143 \\ 0.095 & 0.190 & 0.238 & 0.119 & 0.095 \\ 0.048 & 0.095 & 0.119 & 0.143 & 0.048 \\ 0.071 & 0.143 & 0.095 & 0.048 & 0.167 \end{bmatrix}$$

Applying Eqs. 14, and 18 to matrix **P**, yields: $JE(X, Y)$ = 8.838 bits $MI(X, Y)$ = 5.273 bits

The $CE(Y/X)$ for G is obtained by substituting the values for $H(X) = JE(X, X)$ and $JE(X, Y)$ in Eq. 16 and by the chain rule, $CE(Y/X)$ = 6.087 bits

A generalization of this approach to higher dimensions based on information coding paradigms for three and four source communication systems was presented, see ref [91].

In the codification of molecular structure information, it is desirable that the indices used permit to discriminate isomeric structures. Consequently, in ref [90], schemes for codification of heteroatoms and unsaturated bonds were discussed to award greater applicability to the proposed approach in the characterization of molecular diversity. Also generalizations of the summation operator as the global characterization of the vertex codeword entropies were discussed [90].

The limits and constraints of information theory

Before we culminate this report, we believe it would be profitable to remind us of the conditions that need to be fulfilled to rationalize the use of information-theoretic parameters as uncertainty measures. Right from the years that immediately preceded Shannon's coining of information theory, it was evident that its axioms were not to be confined to digital communication. A quite similar challenge, though to a lesser magnitude, existed in other fields as well, *i.e.*, how to measure the information (or uncertainty) for a source or outcome of event. Shannon suggested that an ideal measure, which he denominated entropy (H), had to necessarily be a function of the statistical distribution of the elements that comprised an information source. Three important requirements were set down [1,14,92]: (1) H should be continuous in p_i, with its maximum value achieved for equally likely events, (2) H should to be a function of a random variable distribution function and should not depend on a set of concrete values of the observed phenomenon, (3) If an event is split into two consecutive events, then the initial H is given by the weighted sum of the H value for each event. So following the intuition that other than digital signals, information sources in general did possess statistical patterns, it did not take long before Shannon's entropy would be applied to other fields of science and of course mathematical chemistry was not left out! Therefore the partitioning of molecular structure representations into equivalence (distribution) sets on the basis of predetermined homogeneity criteria opens way to the use of information theory in the characterization of molecular information. Generally, molecular IFIs have been defined using the same scheme, following the computation of the total information content and mean information content, using Eqs. 19 and 20:

$$I(G, \omega) = N \log N - \sum_{g=1}^{G} n_g \log n_g \tag{19}$$

$$\bar{I}(G, \omega) = -\sum_{g=1}^{G} \frac{n_g}{N} \log \frac{n_g}{N}, \tag{20}$$

where ω is the homogeneity criterion, and n_g is the cardinality of equivalence set g. Note that other information-theoretic measures for characterizing molecular structures have been proposed but are essentially variants of Eqs. 19 and 20. In other words, for almost 60 years attention has been basically placed on the search of alternative ways of defining homogeneity criteria. However, this has dangerously given way to the tendency to overlook the very principles that justify the application information-theoretic concepts. For example, attempts have been made to use information-theoretic measures as tools to describe information sources at semantic level, probably due to the feeling of natural proximity of the term *information*. Although it is clear that once syntactic data are transmitted, semantic analysis is necessary in order for the recipient to profit from its acquisition, attempts to import information theory tools to describe any information source should only be at syntactic level. The substantial meaning of data should be ignored and terms like choice or uncertainty should be clearly traceable in the adapted algorithm. These simple rules are usually violated when ratios of physical data are mistaken for probabilities. A simple illustration in chemical structure codification would be, when no equivalence criterion of any kind is taken into account and ratios of an atomic property for each vertex in a G with respect to the corresponding molecular property (*e.g.,* atomic mass/molecular

mass) are considered as "probabilities" to which Eq. 1 is be applied. Much as such an index would unquestionably codify some sort of chemical information, and possibly be reasonably applicable to QSPR studies, it would be a fallacy to consider such a MD as an information index. We are certainly not detractors to the use logarithmic functions in the definition of MDs, but we believe that the "boundaries" of information theory are clearly delimited, and conceptual correctness should not be ignored. Rather than expressing criticism, we seek to challenge us to be mindful of the attribute of nature, as an information source, that justifies the use of information theory in general: the statistical distribution of elements.

Conclusions

In this report, an effort was made to review of the most significant strategies in information theory-based chemical structure codification. A typical characteristic of practically all these strategies (*i.e.,* IFIs) proposed in the literature is that they only use the expression for Shannon's entropy or its variants and thus are limited to the perspective of the chemical structure as an information source rather than system. In this sense, other information-theoretic measures, recently introduced in the definition of IFIs are discussed, with the hope that these will spur new insights toward the chemical structure.

Acknowledgments The authors are grateful to Abbe Mowshowitz and Elena Konstantinova for sending us their manuscripts. The authors acknowledge the partial financial support from Spanish Ministry of Science and Innovation (MICINN, project reference: SAF2009-10399). Marrero-Ponce thanks to the program '*International Professor*' for a fellowship to work at Cartagena University in 2013–2014.

References

1. Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423. doi:10.1002/j.1538-7305.1948.tb01338.x
2. Mandelbrot BB (1968) Information theory and psycholinguistics: a theory of word frequencies. In: Lazarsfeld PF, Henry NW (eds) Readings in mathematical social science. The MIT press, Cambridge
3. McMillan B (1997) Scientific impact of the work of C. E. Shannon. Paper presented at the Proceedings of the Norbert Wiener centenary congress on Norbert Wiener centenary congress, East Lansing, Michigan, 1997
4. Ebling W, Jiminez-Montano MA (1980) On grammars, complexity and information measures of biological macromolecules. Math Biosci 52:53–71. doi:10.1016/0025-5564(80)90004-8
5. Cosmi C, Cuomo V, Ragosta M, Macchiato MF (1990) Characterization of nucleotidic sequences using maximum entropy techniques. J Theor Biol 147:423–432. doi:10.1016/S0022-5193(05)80497-7
6. Schneider TD, Mastronarde DV (1996) Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method. Discrete Appl Math 71:259–268. doi:10.1016/S0166-218X(96)00068-6
7. Theil H (1967) Econ Inf Theory. North Holland Publishing Company, Amsterdam
8. Maasoumi E (1993) A compendium to information theory in economics and econometrics. Econ Rev 12:137–181. doi:10.1080/07474939308800260
9. Dimitrov AG, Lazar AA, Victor JD (2011) Information theory in neuroscience. J Comput Neurosci 30:1–5. doi:10.1007/s10827-011-0314-3
10. Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106:620. doi:10.1103/PhysRev.106.620
11. Ulanowicz RE (2011) The central role of information theory in ecology towards an information theory of complex networks. In: Dehmer M, Emmert-Streib F, Mehler A (eds.) Birkhäuser Boston, pp 153–167. doi:10.1007/978-0-8176-4904-3_7
12. Bernaola-Galvan P, Roman-Roldan R, Oliver J (1996) Compositional segmentation and long-range fractal correlations in DNA sequences. Phys Rev E 53:5181–5189. doi:10.1103/PhysRevE.53.5181
13. Bonchev D (2009) Information theoretic measures of complexity. In: Meyers R (ed) Encyclopedia of complexity and system science, vol 5. Springer, Heidelberg, Germany, pp 4820–4838. doi:10.1007/978-0-387-30440-3_285
14. Desurvire E (2009) Classical and quantum information theory an introduction for the telecom scientist. Cambridge University Press, New York
15. Jaynes ET (1957) Information theory and statistical mechanics II. Phys Rev 108:171–190. doi:10.1103/PhysRev.108.171
16. Ben-Naim A (2011) Entropy: order or information. J Chem Educ 88:594–596. doi:10.1021/ed100922x
17. Balaban AT, Ivanciuc O (1999) Histological development of topological indices. In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach Science Publishers, The Netherlands, pp 32–39
18. Dehmer M, Mowshowitz A (2011) A history of graph entropy measures. Inf Sci 181:57–78. doi:10.1016/j.ins.2010.08.041
19. Bonchev D (1983) Information theoretic indices for characterization of chemical structures. Research Studies Press, Chichester, UK
20. Bonchev D (2005) My life-long journey in mathematical chemistry. Int Electron J Mol Des 4:434–490
21. Todeschini R, Consonni V (2009) Molecular descriptors for chemoinformatics, 1st edn. Wiley-VCH, Weinheim
22. García-Domenech R, Gálvez J, de Julián-Ortiz JV, Pogliani L (2008) Some new trends in chemical graph theory. Chem Rev 108:1127–1169. doi:10.1021/cr0780006
23. Bonchev D, Tashkova C, Ljuzkanova R (1975) On the correlation between enthalpy of formation, atomic number, and information content of alkali halides. Dokl BAN 28:225–228
24. Bonchev D, Kamenska V, Kamenski D (1977) Informationsgehalt chemischer elemente. Monatsh Chem 108:477–487. doi:10.1007/BF00902003
25. Bonchev D, Kamenska V (1978) Informationscharacteristiken der perioden und unterperioden im periodensystem. Monatsh Chem 109:551–556
26. Bonchev D, Kamenska V (1978) Information theory in describing the electronic structure of atoms. Croat Chem Acta 51:19–27
27. Nalewajski RF, Parr RG (2001) Information theory thermodynamics of molecules and their hirshfeld fragments. J Phys Chem A 105:7391–7400. doi:10.1021/jp004414q
28. Nalewajski RF (2002) Applications of the information theory to problems of molecular electronic structure and chemical reactivity. Int J Mol Sci 3:237–259. doi:10.3390/i3040237
29. Nalewajski RF, Broniatowska E (2003) Entropy displacement and information distance analysis of electron distributions in molecules and their hirshfeld atoms. J Phys Chem A 107:6270–6280. doi:10.1021/jp030208h

30. Parr RG, Ayers PW, Nalewajski RF (2005) What is an atom in a molecule? J Phys Chem A 109:3957–3959. doi:10.1021/jp0404596
31. Dancoff SM, Quastler H (1953) The information content and error rate of living things. In: Quastler H (ed) Essays on the use of information theory in biology. University of Illinois Press, Urbana, pp 263–273
32. Cayley A (1875) Ueber die analytischen figuren, welche in der mathematik bäume genannt werden und ihre anwendung auf die theorie chemischer verbindungen. Ber deutsch chem Ges 8:1056–1059. doi:10.1002/cber.18750080252
33. Rouvray DH (1989) The pioneering contributions of Cayley and Sylvester to the mathematical description of chemical structure. J Mol Struct (Theochem) 185:1–14. doi:10.1016/0166-1280(89)85003-1
34. Pogliani L (2000) From molecular connectivity indices to semiempirical connectivity terms: recent trends in graph theoretical descriptors. Chem Rev 100:3827–3858. doi:10.1021/cr0004456
35. Randić M (2003) Aromaticity of polycyclic conjugated hydrocarbons. Chem Rev 103:3449–3606. doi:10.1021/cr9903656
36. Randić M, Zupan J, Balaban AT, Vikic-Topic D, Plavšic D (2011) Graphical representation of proteins. Chem Rev 111:790–862. doi:10.1021/cr800198j
37. Rashewsky N (1955) Life, information theory, and topology. Bull Math Biophys 17:229–235. doi:10.1007/BF02477860
38. Trucco E (1956) A note on the information content of graphs. Bull Math Biophys 18:129–135. doi:10.1007/BF02477836
39. Trucco E (1956) On the information content of graphs: compound symbols; different states for each point. Bull Math Biophys 18:237–253. doi:10.1007/BF02481859
40. Mowshowitz A (1968) Entropy and the complexity of the graphs I: an index of the relative complexity of a graph. Bull Math Biophys 30:175–204
41. Mowshowitz A (1968) Entropy and the complexity of graphs IV: entropy measures and graphical structure. Bull Math Biophys 30:533–546. doi:10.1007/BF02476673
42. Bonchev D, Kamenski Kamenska V (1976) Symmetry and information content of chemical structures. Bull Math Biol 38:119–133. doi:10.1007/BF02471752
43. Bertz SH (1981) The first general index of molecular complexity. J Am Chem Soc 103:3599–3601. doi:10.1021/ja00402a071
44. Bonchev D (2003) Shannon's information and complexity. In: Bonchev D, Rouvray DH (eds) Complexity in chemistry, vol 7., Mathematical chemistry SeriesTaylor & Francis, London, UK, pp 155–187
45. Hosoya H (1971) Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. Bull Chem Soc Jpn 44:2332–2339. doi:10.1246/bcsj.44.2332
46. Bonchev D, Trinajstic N (1977) Information theory, distance matrix, and molecular branching. J Chem Phys 38:4517–4533. doi:10.1063/1.434593
47. Basak SC, Roy AB, Ghosh JJ (1979) Study of the structure-function relationship of pharmacological and toxicological agents using information theory. In: Avula XJR, Bellman R, Luke YL, Riegler AK (eds) Proceedings of 2nd international conference on mathematical modelling, University of Missouri, Rolla, pp 851–856
48. Basak SC, Raychaudhury C, Roy AB, Ghosh JJ (1981) Quantitative structure-activity relationships (QSAR) studies of bioactive agents using structural information indices. Ind J Pharmacol 13:112–116
49. Basak SC, Magnuson VR (1983) Molecular topology and narcosis. A quantitative structure-activity relationship (QSAR) study of alcohols using complementary information content (CIC). Arzneim-Forsch/Drug Res 33:501–503
50. Raychaudhury C, Ray SK, Roy AB, Ghosh JJ, Basak SC (1984) Discrimination of isomeric structures using information theoretic topological indices. J Comput Chem 5:581–588. doi:10.1002/jcc.540050612
51. Basak SC (1987) Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach. Med Sci Res 15:605–609
52. Basak SC (1999) Information theoretic indices of neighborhood complexity and their applications. In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Reading, UK, pp 563–593
53. Balaban AT (1979) Chemical graphs. XXXIV. Five new topological indices for the branching of tree-like graphs. Theor Chim Acta 53:355–375. doi:10.1007/BF00555695
54. Balaban AT, Bertelsen S, Basak SC (1994) New centric topological indexes for acyclic molecules (trees) and substituents (rooted trees), and coding of rooted trees. MATCH Commun Math Comput Chem 30:55–72
55. Bonchev D, Balaban AT, Mekenyan O (1980) Generalization of the graph center concept, and derived topological indexes. J Chem Inf Comput Sci 20:106–113. doi:10.1021/ci60022a011
56. Bonchev D (1989) The concept for the centre of a chemical structure and its applications. J Mol Struct (Theochem) 185:155–168. doi:10.1016/0166-1280(89)85011-0
57. Dosmorov SV (1982) Generation of homogeneous reaction mechanism. Kinetics and Catalysis
58. Dehmer M, Varmuza K, Borgert S, Emmert-Streib F (2009) On entropy-based molecular descriptors: statistical analysis of real and synthetic chemical structures. J Chem Inf Model 49:1655–1663. doi:10.1021/ci900060x
59. Dehmer M, Grabner M, Varmuza K (2012) Information indices with high discriminative power for graphs. PLoS ONE 7(2):e31214. doi:10.1371/journal.pone.0031214
60. Dehmer M, Borgert S, Emmert-Streib F (2008) Entropy bounds for molecular hierarchical networks. PLoS ONE 3(8):e3079. doi:10.1371/journal.pone.0031214
61. Dehmer M, Emmert-Streib F (2008) Structural information content of networks: graph entropy based on local vertex functionals. Comp Biol Chem 32:131–138. doi:10.1016/j.compbiolchem.2007.09.007
62. Gregori-Puigjané E, Mestres J (2006) SHED: Shannon entropy descriptors from topological feature distributions. J Chem Inf Model 46:1615–1622. doi:10.1021/ci0600509
63. Poincaré H (1900) Second complément à l'Analysis situs. Proc London Math Soc 32:277–308. doi:10.1112/plms/s1-32.1.277
64. Harary F (1969) Graph theory. Addison-Wesley, Reading, MA
65. Janežič D, Miličević A, Nikolić S, Trinajstić N (2007) Graph theoretical matrices in chemistry., Mathematical chemistry monographsUniversity of Kragujevac & Faculty of Science Kragujevac, Kragujevac
66. Ivanciuc O, Balaban AT (1996) Design of topological indices. Part 3. New identification numbers of chemical structures: MINID and MINSID. Croat Chem Acta 69:9–16
67. Wiener H (1947) Structural determination of paraffin boiling points. J Am Chem Soc 69:17–20. doi:10.1021/ja01193a005
68. Skorobogatov VA, Konstantinova EV, Nekrasov YS, Sukharev YN, Tepfer EE (1991) On the correlation between the molecular information topological and mass spectra indices of organometallic compounds. MATCH Commun Math Comput Chem 26:215–228
69. Consonni V, Todeschini R, Pavan M (2002) Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. Part 1. Theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci 42:682–692. doi:10.1021/ci015504a

70. Consonni V, Todeschini R, Pavan M, Gramatica P (2002) Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. Part 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies. J Chem Inf Comput Sci 42:693–705. doi:10.1021/ci0155053
71. Hall LH, Kier LB (1995) Electrotopological state indices for atom types: a novel combination of electronic, topological, and Valence state information. J Chem Inf Comput Sci 35:1039–1045. doi:10.1021/ci00028a014
72. Klopman G, Raychaudhury C (1988) A novel approach to the use of graph theory in structure-activity relationship studies. Application to the qualitative evaluation of mutagenicity in a series of nonfused ring aromatic compounds. J Comput Chem 9:232–243. doi:10.1002/jcc.540090307
73. Klopman G, Raychaudhury C, Henderson RV (1988) A new approach to structure-activity using distance information content of graph vertices: a study with phenylalkylamines. Math Comput Modelling 11:635–640. doi:10.1016/0895-7177(88)90570-5
74. Balaban AT, Balaban TS (1991) New vertex invariants and topological indices of chemical graphs on information on distances. J Math Chem 8:383–397. doi:10.1007/BF01166951
75. Ivanciuc O, Balaban TS, Balaban AT (1993) Chemical graphs with degenerate topological indices based on information on distances. J Math Chem 14:21–33. doi:10.1007/BF01164452
76. Konstantinova EV, Paleev AA (1990) Sensitivity of topological indices of polycyclic graphs. Vychisl Sistemy 136:38–48
77. Ivanciuc O (2002) Building-block computation of the Ivanciuc-Balaban indices for the virtual screening of combinatorial libraries. Int Electron J Mol Des 1:1–9
78. Mekenyan O, Bonchev D, Balaban AT (1988) Topological indices for molecular fragment and new graph invariants. J Math Chem 2:347–375. doi:10.1007/BF01166300
79. Balaban AT, Feroiu V (1990) Correlations between structure and critical data or vapor pressures of alkanes by means of topological indices. Rep Mol Theory 1:133–139
80. Ivanciuc O, Ivanciuc T, Cabrol-Bass D, Balaban AT (2000) Evaluation in quantitative structure-property relationship models of structural descriptors derived from information theory operators. J Chem Inf Comput Sci 40:631–643. doi:10.1021/ci9900884
81. Ivanciuc O, Ivanciuc T, Balaban AT (1999) Vertex- and edge-weighted molecular graphs and derived structural descriptors. In: Devillers J, Balaban AT (eds) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach Science Publishers, Amsterdam, The Netherlands, pp 169–220
82. Ivanciuc O, Balaban AT (1999) Design of topological indices. Part 20. Molecular structure descriptors computed with information on distance operators. Rev Roum Chim 44:479–489
83. Ramos de Armas R, González Díaz H (2004) Markovian backbone negentropies: molecular descriptors for protein research. I. Predicting protein stability in arc repressor mutants. Protein struct Funct Bioinform 56:715–723. doi:10.1002/prot.20159
84. Hamming RW (1986) Coding and information theory, 2nd edn. Prentice-Hall, Englewood Cliffs
85. Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. Wiley, Hoboken, New Jersey
86. Lin S, Costello DJ Jr (1983) Error control coding: fundamentals and applications. Prentice-Hall, Englewood Cliffs, NJ
87. Blahut RE (1983) Theory and practice of error control codes. Addison-Wesley, Reading, MA
88. Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22:79–86. doi:10.1214/aoms/1177729694
89. Barigye SJ, Marrero-Ponce Y, López YM, Santiago OM, Torrens F, Domenech RG, Galvez J (2012) Event-based criteria in GT-STAF information indices: theory, exploratory diversity analysis and QSPR applications. SAR & QSAR Environ Res 24:3–34. doi:10.1080/1062936X.2012
90. Barigye SJ, Marrero-Ponce Y, Santiago OM, López YM, Torrens F (2013) Shannon's, mutual, conditional and joint entropy-based information indices. Generalization of global indices defined from local vertex invariants. Curr Comput-Aided Drug Des 9:164–183
91. Barigye SJ, Marrero-Ponce Y, Martínez-López Y, Torrens F, Artiles-Martínez LM, Pino-Urias RW, Martínez-Santiago O (2013) Relations frequency hypermatrices in mutual, conditional and joint entropy-based information indices. J Comp Chem 34:259–274. doi:10.1002/jcc.23123
92. Dmítriev VI (1989) Applied information theory. Mir Publishers, Moscow