Chemical Physics Letters 570 (2013) 147–152

Contents lists available at SciVerse ScienceDirect

Chemical Physics Letters

journal homepage: www.elsevier.com/locate/cplett

Extended GT-STAF information indices based on Markov approximation models

Stephen J. Barigye [a,*], Yovani Marrero-Ponce [a,b,c], Vitalio Alfonso-Reguera [d], Facundo Pérez-Giménez [c]

[a] Unit of Computer-Aided Molecular 'Biosilico' Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy, Universidad Central 'Martha Abreu' de Las Villas, Santa Clara 54830, Villa Clara, Cuba
[b] Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, P.O. Box 22085, E-46071, València, Spain
[c] Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Spain
[d] Department of Telecommunications Engineering, Faculty of Electrical Engineering, Universidad Central 'Martha Abreu' de Las Villas, Santa Clara 54830, Villa Clara, Cuba

ARTICLE INFO

Article history:
Received 31 January 2013
In final form 19 March 2013
Available online 29 March 2013

ABSTRACT

A series of novel information theory-based molecular parameters derived from the insight of a molecular structure as a chemical communication system were recently presented and usefully employed in QSAR/QSPRs (J. Comp. Chem, 2013, 34, 259; SAR and QSAR in Environ. Res. 2013, 24). This approach permitted the application of Shannon's source and channel coding entropic measures to a chemical information source comprised of molecular 'fragments', using the zero-order Markov approximation model (atom-based approach). This report covers the theoretical aspects of the extensions of this approach to higher-order models, introducing the first, second and generalized-order Markov approximation models.
© 2013 Elsevier B.V. All rights reserved.

1. Introduction

The search of chemical models, which characterize better or at least distinctly, the intrinsic features of molecular structures, constitutes an area of sustained interest in theoretical chemistry. This is simply because there is no single model with which all physical, chemical and physicochemical phenomena could be effectively rationalized. This places the molecular structure in some sort of partially discovered 'mystery' requiring more research for its full elucidation. Several models, derived from a diverse range of theories such as physical chemistry, quantum chemistry, graph theory, information theory, among others, have been proposed, in the effort to achieve better approximations of chemical reality [1].

The debate on which model offers the best approximation would be subjective, as each model is characterized by particular strengths and limitations, evocative of the 'No free lunch' theorem [2]. What is however true, is the need to improve (or generalize) the existing models or define new ones that demonstrate greater sensitivity to progressive structural changes and considerable simplicity. This Letter deals with the former: the generalization of previously defined information-theoretic parameters, or simply information indices, to consider *n*-gram Markov approximation models [3,4] in chemical structure codification.

Recently, novel insights in chemical structure codification, based on Shannon's source coding and channel coding theorem paradigms were discussed, introducing a new family of information indices, collectively denominated the GT-STAF (acronym for Graph Theoretical Thermodynamic STAte Functions) information indices [5–8]. These models followed the analysis of statistical patterns of sets of molecular 'fragments', similar to words that comprise a natural text, as a chemical source. Various generalizations to consider higher dimensional communication system models, and diverse source originator algorithms were discussed [6,7]. In this report, we intend to dig deeper into the 'information-theoretic modeling reserves' to unveil paradigms extendable to chemical structural codification, with particular interest in Markov approximation models. First, we will briefly discuss a few theoretic aspects dealing with data statistical structure, placing emphasis on the features that will be posteriorly used in chemical structure codification.

2. Statistical structure of data and pattern substitution

Pattern substitution is a simple statistical encoding technique based on the replacement of frequently appearing patterns (or sequences) in an information source with super-characters [9]. These sequences are concatenations of characters, which are generally, stochastic in nature. In other words, in a natural information source, character concatenations are not a coincidence, that is, some sequences are more likely than others. For instance in the English language, digrams like TH, HE or AN are more frequent than let us say XP, KV, WZ, etc. Several Markov models, as approximations of an ordinary text message, could be analyzed. If we take as an example the zero-order Markov approximation models (comprised of independent and equally probable characters) [3,4] of an

* Corresponding author.
E-mail addresses: stephen@uclv.edu.cu, stvnjns.barigye@gmail.com (S.J. Barigye).

0009-2614/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cplett.2013.03.057

English text message, these do not yield logical texts while higher order models produce greater approximations to comprehensible texts. In practice, the trigram (third-order Markov) word model is generally used, which estimates the probability of the next word given the preceding two words. The natural idea is to assign simple characters to frequent character patterns, yielding some sort of super-alphabet.

Once a super-alphabet is defined, the information source adapted to a Markov approximation model may be subjected to entropy coding tree algorithms, to achieve greater optimality, such as Shannon–Fano encoding, Huffman encoding and the Lempel-Ziv-Welch encoding, among others [4,10–12]. These entropy coding algorithms work on the principle of ascribing shorter codewords to the most frequent *n*-grams while longer codewords are allocated to the least frequently used ones. This is also known as block coding [10]. A block code is defined as *any code that manipulates groups of codewords, either by concatenation or by the attribution of new codewords for specific groups of source symbols* [10]. In this sense, the codewords for the *n*-concatenations are considered as block codes.

Block codes result in considerably longer codewords and the codeword dictionary is also significantly greater. The key advantage is that the mean bit/codeword is considerably reduced, which enables much more data to be packed/transmitted [10]. These codes permit squeezing a piece of text message by means of a super-alphabet (data compression). By block coding one can reach the theoretical limit of 100% coding efficiency with arbitrary accuracy, but at the price of using an extended dictionary of codewords.

Most generally, block codes can be attributed different 'sub-block' or fields. For instance, a given field can be reserved for the payload (the sequence of codewords to be transmitted) and another field to the overhead (the information describing how to handle and decompress the payload). This constitution enables the use of block codes in error correction schemes [4,10].

Other than contributing to efficient telecommunication and information technology systems, these algorithms (or closely related procedures) provide powerful tools, whose applications could be extended to other fields, for example in Ref. [13], amino-acid sequences of natural antimicrobial peptides are treated as a formal language, reminiscent of phrases in a natural language, and a set of regular grammars built to describe this language. These regular grammars could in essence be considered as Markov approximation models. The ensuing set of grammars is posteriorly used to create new, unnatural AmP sequences. For similar applications in DNA sequences, see Ref. [14]. We will now discuss the extrapolation of Markov approximation models to chemical structural codification.

3. Chemical structure codification

3.1. Theoretical background: zero-order Markov approximations

Before we proceed, let us have a brief recapitulation of the definitions presented in previous reports [5–8].

Consider as a chemical information source *S* a set of 'molecular fragments' for a given molecular structure. The set *S* is generated according to predetermined criteria, denominated the event, which could be graph-theoretical, chemical, or physicochemical [7].

Let us take as a simple example the molecular structure of Isopentane (see Figure 1), where the numbers correspond to the labels that are assigned to the carbon atoms (vertices) in the molecular structure. Recently, several events as source originator algorithms (subgraph generators) have been proposed [7].

In this illustration only one event, i.e. connected subgraphs, will be used, for details concerning other source originator algorithms

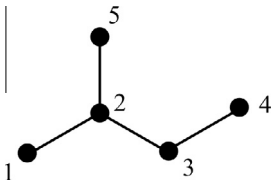

Figure 1. The labeled chemical graph of the molecule of Isopentane (the numbers correspond to the labels that are assigned to the non-hydrogen atoms (vertices) in the molecular structure).

see Ref. [7]. The connected subgraphs algorithm is based on the graph-theoretical concept of subgraph orders and types, according Kier–Hall nomenclature [1,15].

Accordingly, for the molecular structure in Figure 1, the connected subgraphs obtained for different orders based on the atomic relations are:

Order 1: C1–C2, C2–C3, C3–C4, C2–C5
Order 2: C1–C2–C3, C1–C2–C5, C2–C3–C4, C2–C3–C5
Order 3: C1–C2–C3–C4, C2–C3–C4–C5, C1–C2–(C5)–C3
Order 4: C1–C2–(C5)–C3–C4

These subgraphs will constitute the information source. Just like for a natural English language text, where some letters are more frequent than others, for the chemical source above, the vertices (letters) that form the 'molecular fragments' possess a statistical structure. The analysis of the statistical patterns of this information source gives the degree of uncertainty (or lack of homogeneity). This quantity is known as *Shannon's entropy* or *entropy of information* and is defined by:

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

where p_i is the probability associated to vertex v_i and n is the number of vertices that constitute molecular graph G.

Figure 2A, illustrates the computation of Shannon's entropy (SE) for the chemical information source obtained for the molecular structure of Isopentane using the connected subgraphs algorithm.

If we consider the logarithm to base two, the entropy for the chemical information source in Figure 2A, H (S) = 2.258 bits. The H (S) is the number of bits on average required to describe the source as a random variable. From a source coding point of view, H (S) delimits the smallest mean codeword length achievable for a given source (minimum data compression). This is a fundamental result of Shannon's source coding theorem. Several algorithms have been proposed, aimed at achieving maximum approximation to this lower bound, for details see [4,10–12].

Once the lower limit is established, the next step would naturally be the application of a coding scheme that offers plausible approximation to this bound. We will however for posterior practical implications consider a fixed-length binary coding tree scheme, based on the incidence of vertices (letters) in the subgraphs (words), forming an *n*-length binary codewords, where *n* is the subgraph number (fixed codeword size). In other words, the interest here is not code optimality but rather dissimilarity of the vertex codewords, if applicable.

Given a set of subgraphs, $S = \{s_g | \, 1 \leqslant g \leqslant n\}$, generated according to a predefined criterion, the codeword for v_i is sequentially assigned:

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

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

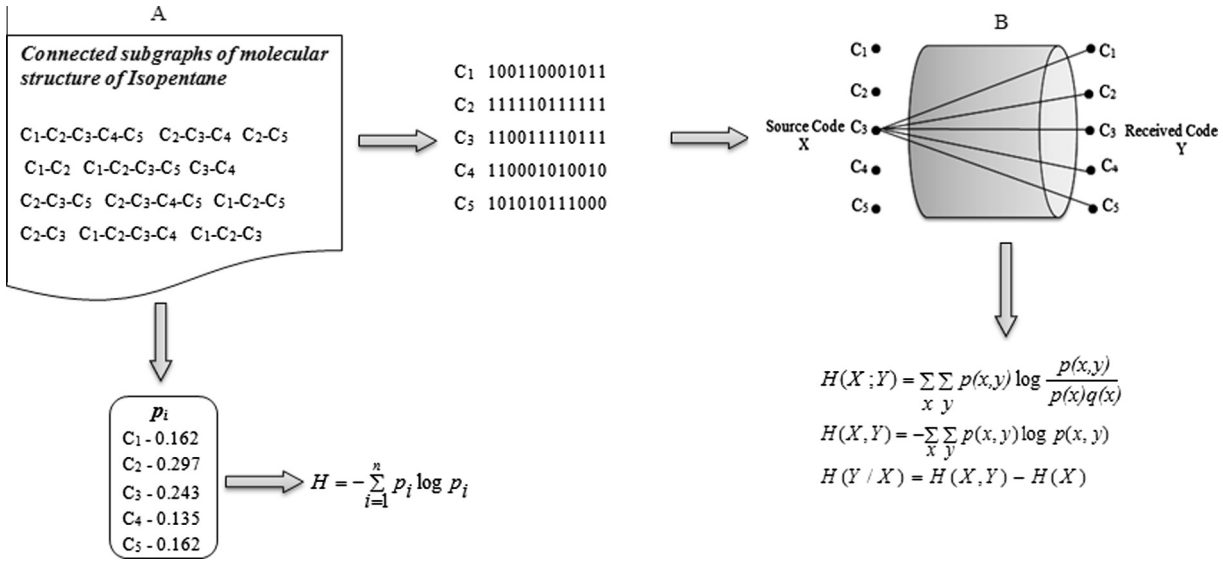

Figure 2. (A) First-order Markov approximation model for the chemical source entropy computation for the molecular structure of Isopentane. The chemical source is comprised of connected subgraphs of orders 1–4. (B) Schematic representation of the relations between inputs and outputs in a noisy channel. Note that the MI, JE and CE analysis is carried out for each of the input codes with respect to the output codes (see below).

C_1 100110001011
C_2 111110111111
C_3 110011110111
C_4 110001010010
C_5 101010111000

Suppose this source code (message) is transmitted along a noisy channel [4,10], such that the codeword sequence for vertex v_x is received instead of the one corresponding to v_y. This means that the received message is not necessarily the same as the one sent out by the transmitter (see Figure 2B for illustration). The mutual information (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, defined as:

$$H(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)q(x)} \tag{2}$$

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 probability distribution function P(X, Y) is formed, where $P(X,Y) = \{p(x,y) : p(x,y) = f(x,y)/f_T | x \neq y \wedge f_T = \sum_{x=1}^{n} \sum_{y=1}^{n} f_{(x,y)},\ x = y\}$

Applying Eq. (2) gives the molecular *MI index*. Other entropic measures like *joint and conditional entropies*, denoted by *JE* and *CE*, respectively, could also be computed with Eqs. (3) and (4), obtaining corresponding molecular information indices.

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y) \tag{3}$$

$$H(Y/X) = H(X,Y) - H(X) \tag{4}$$

Note that in the computation of the joint probability distribution function, an assumption is made that zero components in the vertex codewords are information-less, i.e. they do not lie in the context of the source originator algorithm, and thus their mutual frequencies are left out.

Using the same example of the molecular structure of Isopentane, let us illustrate the computation of the MI, JE and CE information indices. For operational convenience, mutual frequencies and corresponding joint probabilities are represented by frequency and joint probability matrices, denoted by F and P, respectively, as shown below:

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

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

Applying Eqs. (2) and (3) to matrix P, yields:

$MI(X, Y) = 3.001$ *bits* $\quad JE(X, Y) = 6.566$ *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. (4) and by the chain rule, CE(Y/X) = 4.294 bits.

This approach was extended to consider information coding for communication systems with three and four source dimensions and corresponding applications to molecular structure codifications were discussed, see Ref. [6].

Up to this point, we considered zero-order Markov approximations, i.e. the symbols that constitute the chemical source were considered as separate independent entities. However, higher order Markov approximation models could be considered as well. This will be the primary focus of this report.

3.2. First-order Markov approximations

In this case, as opposed to the analysis of statistical patterns of the chemical source symbols as separate entities, pair-wise concatenations of chemical symbols are considered. In other words, for a given chemical source, an exploration of valid digram blocks (adjacent vertex–vertex pairs or edges) is performed and these are assigned 'super-characters'. Consequently, it is to the super-alphabet that the proposed entropy coding scheme is applied. Likewise, the different information-theoretic entropy computations for

the noisy communication system model are performed. We will now illustrate the application of the first-order Markov approximation to a chemical information source.

Let A be the source alphabet comprised of a series of valid binary concatenations, whose sequences form the 'molecular fragments' that comprise the chemical source. If we take as an example the chemical source introduced in Section 3.1, the set of valid super-characters, A = {C_1C_2, C_2C_3, C_3C_4, C_2C_5}. Therefore, in place of analyzing the statistical patterns in terms of singular symbols (vertices) that comprise the 'molecular fragments', the distribution of digram blocks in the chemical source is explored. Figure 3A shows the computation of SE for the defined chemical source, based on the application of Shannon's fundamental formula (Eq. (1)) to a probability distribution function (p.d.f) derived from the digram (vertex pair) participation frequencies in the set of molecular 'fragments'.

The SE for the chemical source generated for the molecular structure of Isopentane, H (S) = 1.979 bits. Additionally, the Shannon's channel coding entropic measures for communication along a noisy channel could be applied to the codeword sequences for the digram characters (see Figure 3B), yielding the mutual information, conditional and joint entropy-based molecular parameters. For the example in Figure 3B, applying Eqs. (2)–(4), H(X; Y) = 0.965 bits, H(X, Y) = 3.813 bits and H(Y/X) = 1.834 bits, respectively.

The entropic measures derived from the first-order Markov approximations could be regarded as equivalents of the edge (bond)-based MDs, although the latter are not obtained from a similar chemical information source. The edge based MDs have been successfully used in various chemoinformatic studies, yielding in some cases better correlations for molecular properties than their vertex(atom)-based MD analogs [16–26]. In fact, edge-based analogs for almost all well-known vertex-based MDs have been defined, for example the Wiener index [27], the molecular connectivity indices [23], Schultz molecular topological index [28], Harary index [29], vertex orbital and centric information indices [30–33], etc.

3.3. Second order Markov approximations

This model represents stochastic processes where the choice of a letter (or symbol) depends on the preceding two symbols. A set of trigram frequencies would therefore be required to analyze the statistical nature of the information source. It is expected that better stochastic approximation to a logical source message should be achieved with this model and thus improved data compression [3,4]. Note that since the chemical information source is derived from undirected molecular graphs, the transition probabilities $p(i/j, k)$, $p(j/i, k)$ and $p(k/i, j)$ are indistinctive. As in the case of the second-order stochastic model, the different trigram blocks are substituted with super-characters and the respective trigram frequencies determined. Subsequent entropic computations for a chemical communication system are straight forward. Figure 4 illustrates the considerations for entropic computations using a third-order letter (vertex label) model for a noisy chemical communication system.

Applying Eqs. (1)–(4) to a joint p.d.f of trigram blocks in the chemical information source, yields H(X) = 1.990 bits, H(X; Y) = 1.939 bits, H(X, Y) = 4.384 bits and H(Y/X) = 2.394 bits, respectively.

Note that this procedure remains for a binary communication system model (comprised of a source originator and a recipient) and should not be confused with three (or four)-dimensional chemical communication system paradigms [i.e. for 3 (or 4) source communication systems], as these deal with cases in which codewords are sent from two (or three) sources and received by the same recipient or one source but received by two (or three) recipients, see [6]. These higher dimensional (triple and quadruple) communication system models could as well be applied to the information sources adapted to Markov approximation prototypes.

3.4. Generalized Markov approximations (n-order stochastic model)

Finally, we will consider a more generalized scheme for *n*-concatenations. Here the super-characters represent more general structures such as: the most frequent molecular fragments in medicinal chemistry, for example chemical functional groups for example: the carboxyl group, conjugated unsaturated bonds, hydroxyl groups, etc. The notion in this case is to assign super-characters to these predefined chemical patterns which in addition to the 'normal alphabet' (in this case 'normal alphabet' refers to the set of all the vertex labels that form the molecular 'fragments' of the information source, considered as independent entities) collectively constitute a chemical super-alphabet. There exists a series of structural keysets (fingerprints) in the literature such as BCI [34], MDL [35], Extended E-State [36], etc. assembled on the basis of structural, chemical, physicochemical and conformational considerations. Using a large and diverse dataset, a query with various

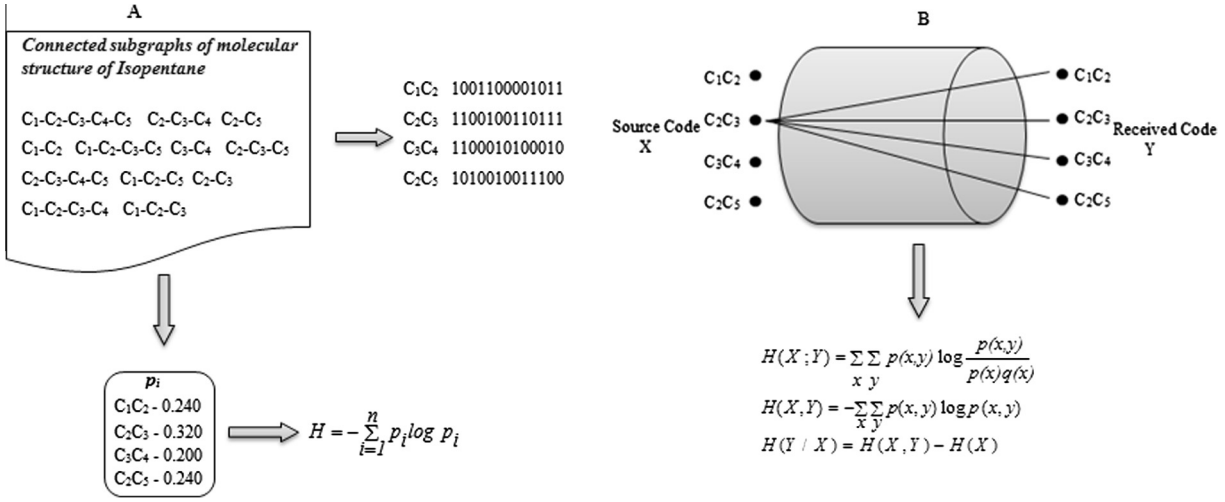

Figure 3. (A) Second-order Markov approximation model for chemical source entropy computation for the molecular structure of Isopentane. (B) Schematic representation of the relations between inputs and outputs in a noisy channel.

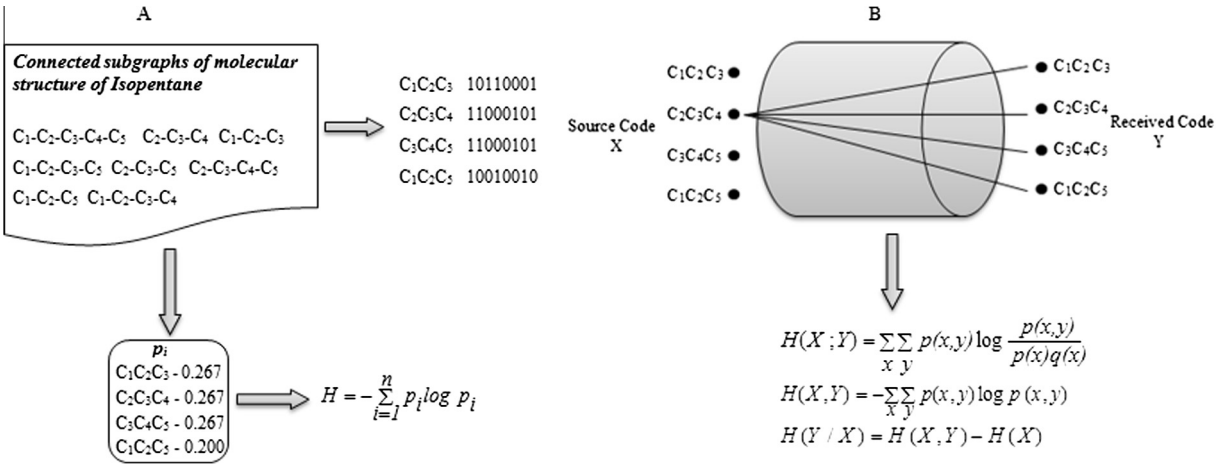

Figure 4. (A) Third-order Markov approximation model for chemical source entropy computation for the molecular structure of Isopentane. Note that for simplicity the chemical source was limited to connected subgraphs of orders 2–4. (B) Schematic representation of the relations between inputs and outputs in a noisy channel.

keysets is performed and the most prevalent 'molecular fragments' are identified and assigned super-characters (*n*-gram blocks). As an example, Table 1 illustrates a selection of the most frequent 'molecular fragments' according to the PubChem and Substructure keysets [37], obtained using an Otava diversity dataset (15 000 compounds). The designation of super-characters would be in terms of such molecular 'fragments', which together with the set of ordinary vertex labels would form a generalized super-alphabet.

Accordingly, the analysis of the statistical patterns is based on the generalized super-alphabet, similar to the first-order word Markov approximation model, and this way yielding a p.d.f for the super-alphabet. The preceding entropic computations come naturally, that is, SE is calculated in terms of the chemical characters from the super-alphabet, which comprise the chemical structure, as well as the MI, CE and JE computations based on the assignation of codewords to the chemical characters from the super-alphabet that constitute the molecular 'fragments' (or substructures). The key differences of this approach from the classical fingerprint paradigm is that for a given molecular structure, we are interested in the statistical distribution of the super-alphabet in the chemical source comprised of molecular 'fragments' while the classical fingerprint query follows the exploration of the existence (or not) of a set of predefined keysets in a molecular structure and features that do not fall in the defined keyset are ignored. Note also that although the number of maximum super-characters is predefined, the size of the super-alphabet is variable depending on the considered source, in the sense that only the super-characters with representativity in a generated chemical source are incorporated to the 'normal alphabet'.

It is important to clarify that although the notion of binary concatenations (digrams) has been applied, whether consciously or unconsciously, in the definition of IFIs (specifically the Bertz index), as well as the Markov chains in MARCH-INSIDE MDs, the formalism proposed in this report follows entirely different considerations, as a corollary of the application of digital communication paradigms in chemical structure codification. This approach represents an important methodological contribution in the sense that a practical structure is provided for the generalization of MDs defined at the atom-level (atoms) and vertices (chemical bonds), using higher order Markov approximation models.

Table 1
Selection of the most common 'molecular fragments' according to the PubChem and Substructure keysets.

Fragment[a]	Frequency	Fragment	Frequency
C ONS bond	87 423	O–C–O–C–C	14 386
Rotatable bond	67 677	C(–H)(=N)	14 361
Conjugated double bond	31 883	[#1]–C–O–[#1]	14 334
C(–C)(=N)	14 634	C:C–O–C	14 308
Cl–C:C–O–C	14 534	O–C–C–C=C	14 274
[As]-C:C-[#1]	14 524	N–C:C–C–C	14 271
ESSSR hetero-aromatic ring	14 490	C(–C)(–H)(=N)	14 265
C–N:C-[#1]	14 482	Cl–C:C–C=O	14 251
N:C–O-[#1]	14 482	N–N–C–N-[#1]	14 251
C–C–C–C–C–C	14 386	Carbonyl group	2288

[a] ESSR, Extended Smallest Set of Smallest Rings (ring which does not share three consecutive atoms with any other ring); ':' denotes bond aromaticity; '-' denotes single bond; '=' denotes double bond.

4. Conclusions

The theoretical aspects for the first, second and generalized-order Markov approximation models for the GT-STAF information indices are presented, offering a generalized scheme for these indices, previously defined for only the zero-order Markov approximation model. It is observed that the first-order Markov approximation model permits obtaining information indices analogous to the edge-based molecular descriptors. Posterior reports will be dedicated to the analysis of the structural information captured by these generalized models, with particular interest in the comparison of the variability, orthogonality and correlation capacity with molecular properties of the obtained information indices with respect to their vertex-based analog (zero-order Markov approximation model), in order to comprehend the real contribution of this extension scheme, if applicable.

References

[1] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, first edn., vol. 1, Weinheim, Ger., WILEY-VCH, 2009. p 667.
[2] D.H. Wolpert, W.G. Macready, IEEE Trans. Evol. Comput. 1 (1997) 67.
[3] C.E. Shannon, Bell Syst. Tech. J. 27 (379–423) (1948) 623.
[4] T.M. Cover, J.A. Thomas, Elements of Information Theory, second edn., John Wiley & Sons, Hoboken, New Jersey, 2006.
[5] S.J. Barigye, Y. Marrero-Ponce, O. Martínez-Santiago, Y. Martínez-López, F. Torrens, Shannon's, mutual, conditional and joint entropy-based information indices. Generalization of molecular descriptors defined from LOVIs, Curr. Comput.-Aided Drug Des., in press.
[6] S.J. Barigye, Y. Marrero-Ponce, Y. Martínez-López, F. Torrens, L.M. Artiles-Martínez, R.W. Pino-Urias, O. Martínez-Santiago, J. Comp. Chem. 34 (2013) 259.
[7] S.J. Barigye, Y. Marrero-Ponce, Y.M. López, O.M. Santiago, F. Torrens, R.G. Domenech, J. Galvez, SAR QSAR Environ. Res. 24 (2013) 3.

[8] S.J. Barigye, Y. Marrero Ponce, F. Pérez-Giménez, D. Bonchev, Trends in information theory based chemical structure codification. Chem. Rev., submitted for publication.
[9] M. Nelson, J. Gailly, The Data Compression Book, M&T Books, New York, 1995.
[10] E. Desurvire, Classical and Quantum Information Theory, Cambridge University Press, New York, 2009.
[11] R.W. Hamming, Coding and Information Theory, second edn., Prentice-Hall, Englewood Cliffs, NJ, 1986.
[12] J. Korner, Coding of an information source having ambiguous alphabet and the entropy of graphs, in: Transactions of Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Academia, Publishing House of the Czechoslovak Academy of Sciences, Prague, 1971; pp. 411–425.
[13] C. Loose, K. Jensen, I. Rigoutsos, G. Stephanopoulos, Nature (London) 443 (2006) 867.
[14] I. Rigoutsos, A. Floratos, L. Parida, Y. Gao, D. Platt, J. Metab. Eng. 2 (2000) 159.
[15] L.H. Hall, L.B. Kier, J. Pharm. Sci. 67 (1978) 1743.
[16] S. Nikolić, N. Trinajstić, I. Baučić, J. Chem. Inf. Comput. Sci. 38 (1998) 42.
[17] I. Lukovits, I. Gutman, MATCH Commun. Math. Comput. Chem. 31 (1994) 133.
[18] C. Cao, Acta Chim. Sin. 54 (1996) 533.
[19] E. Estrada, J. Chem. Inf. Comput. Sci. 35 (1995) 31.
[20] E. Estrada, J. Chem. Inf. Comput. Sci. 36 (1996) 844.
[21] E. Estrada, J. Chem. Inf. Comput. Sci. 39 (1999) 1042.
[22] E. Estrada, Chem. Phys. Lett. 336 (2001) 248.
[23] E. Estrada, N. Guevara, I. Gutman, J. Chem. Inf. Comput. Sci. 38 (1998) 428.
[24] E. Estrada, N. Guevara, I. Gutman, L. Rodriguez, SAR QSAR Environ. Res. 9 (1998) 229.
[25] E. Estrada, I. Gutman, J. Chem. Inf. Comput. Sci. 36 (1996) 850.
[26] E. Estrada, A. Ramirez, J. Chem. Inf. Comput. Sci. 36 (1996) 837.
[27] I. Gutman, E. Estrada, J. Chem. Inf. Comput. Sci. 36 (1996) 541.
[28] E. Estrada, L. Rodriguez, Comput. Chem. 35 (1997) 157.
[29] B. Lučić, A. Miličević, S. Nikolić, N. Trinajstić, Croat. Chim. Acta 75 (2002) 847.
[30] E. Trucco, Bull. Math. Biophys. 18 (1956) 129.
[31] E. Trucco, Bull. Math. Biophys. 18 (1956) 237.
[32] D. Bonchev, A.T. Balaban, O. Mekenyan, J. Chem. Inf. Comput. Sci. 20 (1980) 106.
[33] D. Bonchev, J. Mol. Struct. (Theochem) 185 (1989) 155.
[34] I. Baskin, A. Varnek, Fragment descriptors in SAR/QSAR/QSPR studies, molecular similarity analysis and in virtual screening, in: A. Varnek, A. Tropsha (Eds.), Chemoinformatics Approaches to Virtual Screening, The Royal Society of Chemistry, Cambridge, UK, 2008.
[35] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, J. Chem. Inf. Comput. Sci. 42 (2002) 1273.
[36] L.H. Hall, L.B. Kier, J. Chem. Inf. Comput. Sci. 40 (2000) 784.
[37] C.W. Yap, J. Comp. Chem. 32 (2010) 1466.