The Structural Information Content of Chemical Networks

Matthias Dehmer[a] and Frank Emmert-Streib[b]

[a] Institute of Discrete Mathematics and Geometry, Vienna University of Technology,
Wiedner Hauptstrasse 8 – 10, A-1040 Vienna, Austria
[b] Department of Biostatistics, Department of Genome Sciences, University of Washington,
1705 NE Pacific St, Box 355065, Seattle, WA 98195-5065, USA

Reprint requests to M. D.; E-mail: mdehmer@geometrie.tuwien.ac.at

Z. Naturforsch. **63a,** 155 – 158 (2008); received October 27, 2007

We present an information-theoretic method to measure the structural information content of networks and apply it to chemical graphs. As a result, we find that our entropy measure is more general than classical information indices known in mathematical and computational chemistry. Further, we demonstrate that our measure reflects the essence of molecular branching meaningfully by determining the structural information content of some chemical graphs numerically.

Key words: Structural Information Content; Graph Entropy; Information Theory; Chemical Graph Theory.

The structural analysis of complex networks is a current and ongoing research topic in, e. g., computational physics, computational biology and mathematical chemistry [1 – 3]. Besides the investigation of typical structural properties of given graphs, the development of methods for characterizing certain network classes topologically is also currently of particular interest. For example, in [4] an approach to predict the theoretical existence of different structural classes of complex networks by using the theory of graph spectra [5] has been recently presented. Such methods can be particularly used to investigate growth mechanisms of complex networks. Regarding the structural analysis of graph-based systems in the scientific areas mentioned above, especially in mathematical chemistry there exists a large number of methods to, e. g., analyze and quantify structural information, combinatorial complexity, and molecular branching of chemical structures [1, 6 – 13]. Generally, in mathematical and computational chemistry chemical systems are often represented as graphs where the vertices correspond to chemical components and the edges to the interactions between those components, respectively. A classical problem in mathematical chemistry is to measure the structural information content of chemical graphs by using so-called information indices [1, 8]. Classical methods from applied mathematics to determine the structural information content of graphs have been provided by, e. g., [14 – 16]. These methods are mostly based on the problem to find a partition of the underlying vertex set. Starting from such a partition, e. g., by utilizing vertex orbits, one can straightforward obtain a probability distribution and, hence, the structural information content of a graph defined as its Shannon's entropy [8]. In this paper we present an information-theoretic method to determine the structural information content of undirected and connected graphs and apply it to chemical graphs to quantify their molecular branching. This method was recently introduced in [17]. We will see that the application of this mathematical framework to chemical graphs generalizes information indices frequently used [8, 9]. However, at this point we want to emphasize that in this case our structured objects are comparable in a numerical sense. That means, whenever our entropy measure is applied to arbitrary undirected and connected graphs, we always obtain a numerical value for quantifying and, hence, comparing the structural information content of such graphs. In this paper, this will be mainly done to detect molecular branching between chemical graphs. This assumption is important to mention because there are methods which turn out that not all considered structured objects are comparable regarding their branching [18]. More precisely, in [18] an algebraic characterization of branching was presented that is based on deriving certain inequalities to compare branching of given graphs [18]. Finally, a major result of [18] was that there exist graphs which are not

0932–0784 / 08 / 0300–0155 $ 06.00 © 2008 Verlag der Zeitschrift für Naturforschung, Tübingen · http://znaturforsch.com

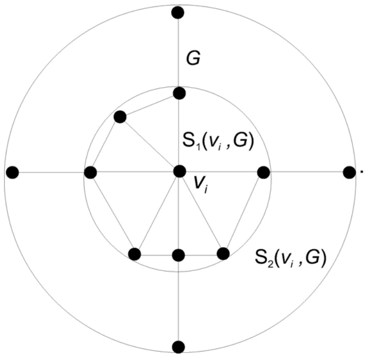

Fig. 1. Visualization of j-spheres. For the shown graph we obtain $|S_1(v_i,G)| = 7$ and $|S_2(v_i,G)| = 4$.

comparable by using this algebraic criterion. Therefore we see that for expressing a novel method to detect, e. g., branching between graphs, it is necessary to highlight the underlying principle for analyzing and, hence, for comparing the graphs (e. g. with respect to branching) under consideration.

We now introduce a finite, undirected and connected graph [19] as $G = (V,E), |V| < \infty$, $E \subseteq \binom{V}{2}$. $\mathcal{G}_{UC}$ denotes the set of finite, undirected and connected graphs. Starting from $G \in \mathcal{G}_{UC}$, $\sigma(v) = \max_{u \in V} d(u,v)$ is called the eccentricity of $v \in V$, where $d(u,v)$ denotes the shortest distance between u and v. $\rho(G) = \max_{v \in V} \sigma(v)$ is called the diameter of G. We further define for $G = (V,E) \in \mathcal{G}_{UC}$ the vertex sets

$$S_j(v_i,G) := \{v \in V \,|\, d(v_i,v) = j,\ j \geq 1\}, \tag{1}$$

where $S_j(v_i,G)$ is called the j-sphere of v_i regarding G. Now, we define a so-called information functional as a monotonous function which quantifies the structural information of an underlying graph. In [17] we introduced for a vertex $v_i \in V$ a special form of an information functional:

$$\begin{gathered} f^V(v_i) := \alpha^{c_1|S_1(v_i,G)|+c_2|S_2(v_i,G)|+\cdots+c_\rho|S_\rho(v_i,G)|}, \\ c_k > 0, 1 \leq k \leq \rho, \alpha > 0. \end{gathered} \tag{2}$$

Here c_k are arbitrary real positive coefficients. We generally choose the c_k such that they are not all equal because it generally holds

$$\begin{aligned} &|S_1(v_1,G)| + |S_2(v_1,G)| + \cdots + |S_\rho(v_1,G)| \\ &= |S_1(v_2,G)| + |S_2(v_2,G)| + \cdots + |S_\rho(v_2,G)| \\ &\quad \cdots\cdots\cdots \\ &= |S_1(v_{|V|},G)| + |S_2(v_{|V|},G)| + \cdots + |S_\rho(v_{|V|},G)|. \end{aligned}$$

According to (1), $S_j(v_i,G)$ denotes the j-sphere of v_i regarding G and $|S_j(v_i,G)|$ its cardinality. Figure 1 shows the procedure to compute the vertex spheres for a graph G exemplarily. We remark that the j-spheres can be computed efficiently by using Dijkstra's algorithm [20]. In order to define a probability value for each vertex by using our information functional f^V, we define the quantities [17]:

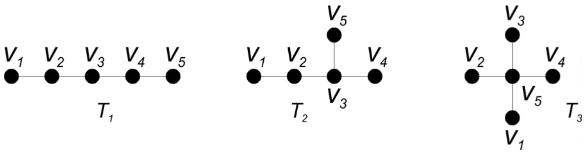

Fig. 2. Three chemical graphs (trees) with 5 vertices. The trees are arbitrarily labeled.

$$p^V(v_i) := \frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)}, \tag{3}$$

for which holds

$$p^V(v_1) + p^V(v_2) + \ldots + p^V(v_{|V|}) = 1. \tag{4}$$

From this we define the structural information content of G as its corresponding entropy [17]:

$$I_{f^V}(G) := -\sum_{i=1}^{|V|} \frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)} \log\left(\frac{f^V(v_i)}{\sum_{j=1}^{|V|} f^V(v_j)}\right). \tag{5}$$

Now we determine the structural information content (5) for some chemical graphs shown in Figure 2. The graphs in Fig. 2 are trees because they are connected and cycle-free. These trees often occur in chemical systems where the trees represent chemical molecules or components thereof [1, 8, 9, 12]. By using the information functional (2), we immediately obtain

$$\begin{aligned} I_{f^V}(T_1) = -\Bigg[& 2\frac{\alpha^{c_1+c_2+c_3+c_4}}{D^{T_1}} \log\left(\frac{\alpha^{c_1+c_2+c_3+c_4}}{D^{T_1}}\right) \\ & + 2\frac{\alpha^{2c_1+c_2+c_3}}{D^{T_1}} \log\left(\frac{\alpha^{2c_1+c_2+c_3}}{D^{T_1}}\right) \\ & + \frac{\alpha^{2c_1+2c_2}}{D^{T_1}} \log\left(\frac{\alpha^{2c_1+2c_2}}{D^{T_1}}\right) \Bigg], \end{aligned} \tag{6}$$

$$\begin{aligned} I_{f^V}(T_2) = -\Bigg[& \frac{\alpha^{c_1+c_2+2c_3}}{D^{T_2}} \log\left(\frac{\alpha^{c_1+c_2+2c_3}}{D^{T_2}}\right) \\ & + \frac{\alpha^{2c_1+2c_2}}{D^{T_2}} \log\left(\frac{\alpha^{2c_1+2c_2}}{D^{T_2}}\right) \\ & + \frac{\alpha^{3c_1+c_2}}{D^{T_2}} \log\left(\frac{\alpha^{3c_1+c_2}}{D^{T_2}}\right) \\ & + 2\frac{\alpha^{c_1+2c_2+c_3}}{D^{T_2}} \log\left(\frac{\alpha^{c_1+2c_2+c_3}}{D^{T_2}}\right) \Bigg], \end{aligned} \tag{7}$$

and

$$I_{f^V}(T_3) = -\left[4\frac{\alpha^{c_1+3c_2}}{D^{T_3}}\log\left(\frac{\alpha^{c_1+3c_2}}{D^{T_3}}\right) + \frac{\alpha^{4c_1}}{D^{T_3}}\log\left(\frac{\alpha^{4c_1}}{D^{T_3}}\right)\right], \tag{8}$$

where

$$D^{T_1} = 2\alpha^{c_1+c_2+c_3+c_4} + 2\alpha^{2c_1+c_2+c_3} + \alpha^{2c_1+2c_2}, \tag{9}$$

$$D^{T_2} = \alpha^{c_1+c_2+2c_3} + \alpha^{2c_1+2c_2} + \alpha^{3c_1+c_2} + 2\alpha^{c_1+2c_2+c_3}, \tag{10}$$

$$D^{T_3} = 4\alpha^{c_1+3c_2} + \alpha^{4c_1}. \tag{11}$$

We clearly see that we get function families to express the information content of chemical graphs. That means that the resulting entropy measures are now functions depending on the free parameter α and the coefficients c_k. Choosing different values for α leads to different entropy functionals. We want to remark that the coefficients give us the possibility to weigh structural properties of the graphs. In contrast to this, classical information-theoretic indices [8, 9] characterize molecular structures based on graph-theoretical characteristics, e. g., degree distributions or vertex orbits, are mostly described by

$$I(G) = |V|\log(|V|) - \sum_{i=1}^{n}|V_i|\log(|V_i|) \tag{12}$$

or

$$I_m(G) = -\sum_{i=1}^{|V|} P_i \log(P_i), \tag{13}$$

where $|V|$ is the number of vertices of G, n the number of different sets of vertices (regarding a certain property), $|V_i|$ the number of elements in the i-th set of vertices, and $P_i = \frac{|V_i|}{|V|}$. We see that in (12) and (13) there are no free parameters or coefficients because, e. g., the quantities P_i are completely determined by the chosen partitioning. Hence, our measure in (5) is more general than the indices mentioned above because we have now the possibility to modify an information functional f and to use different values for α and c_k.

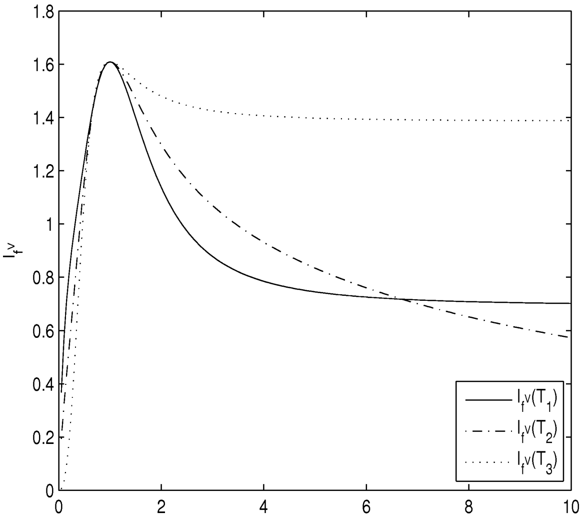

Fig. 3. Structural information content of the chemical graphs shown in Figure 2. The $I_{f^V}(T_i)$ are plotted in dependence on α.

This provides us with the additional possibility to emphasize certain structural characteristics of a graph. To demonstrate that our entropy measures can meaningfully reflect molecular branching, we plotted in Fig. 3 the entropies (6), (7) and (8) in dependence on α. Here, we used $c_1 := 1, c_2 := 2, c_3 := 3, c_4 := 4$. From Fig. 3 we find that $I_{f^V}(T_2)$ is almost everywhere larger than $I_{f^V}(T_1)$. Further, $I_{f^V}(T_3)$ is almost everywhere larger than $I_{f^V}(T_2)$ and $I_{f^V}(T_1)$. Indeed, this result corresponds to our intuition that the branching increases starting from T_1, through T_2, to T_3. Starting from the information functional (2), this can be explained mathematically by the fact that a stronger branching of a graph leads to larger values of the cardinalities of the corresponding j-spheres.

In further studies we will aim to investigate the differences between the structural information content of networks and their functional information content [21]. Especially, with regard to gene networks we expect that there are significant differences.

Acknowledgements

We would like to thank the referees of this paper and Danail Bonchev for fruitful discussions. Matthias Dehmer has been supported by the European FP6-NEST-Adventure Programme, contract no. 028875.

[1] D. Bonchev, Complexity in Chemistry. Introduction and Fundamentals, Taylor and Francis, New York 2003.

[2] M. Fischermann, I. Gutman, A. Hoffmann, D. Rautenbach, D. Vidović, and L. Volkmann, Z. Naturforsch. **57a**, 49 (2002).

[3] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).

[4] E. Estrada, Phys. Rev. E **75**, 0161031 (2007).

[5] D. M. Cvetcovic, M. Doob, and H. Sachs, Spectra od Graphs. Theory and Application, Academic Press, New York 1997.

[6] S. H. Bertz, J. Am. Chem. Soc. **103**, 3241.

[7] S. H. Bertz, Bull. Math. Biol. **45**, 849 (1983).

[8] D. Bonchev, Information Theoretic Indices for Characterization of Chemical Structures, Research Studies Press, Chichester 1983.

[9] D. Bonchev and N. Trinajstivć, J. Chem. Phys. **67**, 4517 (1977).

[10] G. Caparossi, I. Gutmann, P. Hansen, and L. Pavlovi, Comput. Biol. Chem. **27**, 85 (2003).

[11] M. V. Diudea, I. Gutman, and L. Jäntschi, Molecular Topology, Nova Publishing, Huntington, New York 2001.

[12] I. Gutman and O. E. Polansky, Mathematical Concepts in Organic Chemistry, Springer, Berlin 1986.

[13] D. Minoli, Atti. Acad. Naz. Lincei, VIII. Ser., Rend., Cl. Sci. Fis. Mat. Nat. **59**, 651 (1975).

[14] A. Mshowitz, Bull. Math. Biophys. **30**, 175 (1968).

[15] N. Rashewsky, Bull. Math. Biophys. **17**, 229 (1955).

[16] E. Trucco, Bull. Math. Biol. **18**, 129 (1956).

[17] M. Dehmer, Cybernetics and Systems, in press.

[18] I. Gutman and M. Randić, Chem. Phys. Lett. **47**, 15 (1977).

[19] F. Harary, Graph Theory, Addison Wesley Publishing Company, Reading, MA 1969.

[20] E. W. Dijkstra, Numerische Math. **1**, 269 (1959).

[21] R. M. Hazen, P. L. Griffin, J. M. Carothers, and J. W. Szostak, Proc. Natl. Acad. Sci. **104**, 8574 (2007).