Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry 32 (2008) 131–138
Computational Biology and Chemistry
www.elsevier.com/locate/compbiolchem

Brief Communication

Structural information content of networks: Graph entropy based on local vertex functionals

Matthias Dehmer [a,*], Frank Emmert-Streib [b,c]

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

Received 16 June 2007; received in revised form 11 September 2007; accepted 14 September 2007

Abstract

In this paper we define the structural information content of graphs as their corresponding graph entropy. This definition is based on local vertex functionals obtained by calculating *j*-spheres via the algorithm of Dijkstra. We prove that the graph entropy and, hence, the local vertex functionals can be computed with polynomial time complexity enabling the application of our measure for large graphs. In this paper we present numerical results for the graph entropy of chemical graphs and discuss resulting properties.

Keywords: Structural information content; Graph entropy; Information theory; Gene networks; Chemical graph theory

1. Introduction

In computational chemistry (Cramer, 2004) complex systems are often represented as graphs (Harary, 1969) where the vertices correspond to the components and the edges to the interactions, respectively. Therefore, graph-theoretical measures and techniques are extremely useful to analyze and quantify the complexity of chemical structures or reaction networks (see, e.g., Bonchev, 1983, 2003; Caporossi et al., 2003; Gutman and Polansky, 1986; Minoli, 1975; Randić, 1975). For example, a well-known contribution in this direction has been made by Randić (1975) who developed a topological index for measuring the extent of branching of the carbon-atom skeleton of organic molecules (Caporossi et al., 2003). A problem that is strongly related to the problem of measuring the complexity of chemical structures or systems is to determine the *structural information content* of such systems representing graphs (Bonchev, 1983; Mowshowitz, 1968b; Rashewsky, 1955; Trucco, 1956). In this context *structural information content* is defined as the entropy of the underlying network topology (Bonchev, 1983; Mowshowitz, 1968a, b). Classical methods to measure the entropy of graphs are often based on the problem to find a partition of the underlying vertex set. Starting from such a partition, one can straightforward obtain a probability distribution. A first attempt in this direction has been given by Rashewsky (1955). Rashewsky (1955) defined the entropy of directed/undirected and unweighted graphs by partitioning the vertices in sets of indistinguishable vertices according to their dependence on local and non-local degree-dependencies. He finally applied his concept to graphs of metabolic reactions and living organisms. Further algebraical approaches to determine the entropy of graphs have been stated by Trucco (1956) and Mowshowitz (1968b). These approaches are mainly based on obtaining certain vertex partitions by using pure algebraical principles and tools, e.g., determining groups of a graph. Hence, the obtained vertex partitions lead us directly to an entropy of a graph. Particularly, Mowshowitz (1968b) examined properties of graph entropies by determining graph automorphisms (Harary, 1969). Then, in order to define the entropy of a graph, it is necessary to determine the automorphism group of the corresponding graph algorithmically. However, it is well known that this procedure is for arbitrary graphs computationally extensive (McKay, 1981). Now, we finish our

* Corresponding author.
E-mail addresses: mdehmer@geometrie.tuwien.ac.at (M. Dehmer), fes99@u.washington.edu (F. Emmert-Streib).

1476-9271/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2007.09.007

brief review on techniques to measure the entropy of graphs by stating a listing of indices for characterizing chemical graph structures:

- Indices based on topological characteristics, e.g., vertex orbits (Bonchev, 1983; Mowshowitz, 1968b; Rashewsky, 1955; Trucco, 1956).
- Indices based on chromatic decompositions (Bonchev, 1983; Harary, 1969).
- Indices based on adjacency matrices (Bonchev, 1983).
- Indices based on vertex degrees (Bonchev, 1983).
- Indices based on the incidence of a graph (Bonchev, 1983).
- Indices based on cycle matrices of a graph (Bonchev, 1983).
- Indices based on distance matrices of graphs (Bonchev, 1983).
- Indices based on graph decompositions, e.g., the HOSOYA-graph decomposition (Bonchev, 1983; Hosoya, 1971).
- Centric indices and connectivity indices (Bonchev, 1983).

These indices are also based on the main idea to create certain partitions for defining the entropy of chemical graphs, where the elements of such a partition can be described by graph-theoretical quantities (Bonchev, 1983). Finally, classical graph entropy definitions from computer science have been introduced by Körner (1973); Simonyi (2001) and Csiszár et al. (1990). Körner (1973) and Simonyi (2001) originally focused on the problem to determine the performance of a best possible encoding of information emitted by a certain information source. Finally, Csiszár et al. (1990) defined the entropy of a graph with respect to a certain probability distribution P, where P has been defined by using the so called vertex packing polytope of a graph (Csiszár et al., 1990). Further definitions and techniques to determine a graph entropy can be found in (e.g., Bonchev, 1983; Simonyi, 2001).

Some words of precaution: In the present paper we provide a definition for the structural information content of graphs as its corresponding entropy. We want to emphasize that we are dealing exclusively with discrete objects, namely graphs, and, hence, all resulting entities derived from these discrete objects can be interpreted unambiguously as SHANNON-like entities. For this reason there can be no confusion about words like *information* because in this paper *information* is always used in a SHANNON-like sense.

Our graph entropy definition is based on a functional that is based on graph-theoretical quantities related to certain vertex sets. Starting from this definition, we assign a probability value to each vertex in the graph by using metrical properties of graphs (Skorobogatov and Dobrynin, 1988). Hence, by defining such a functional, we avoid the problem to determine vertex partitions with pure algebraical principles because such procedures are often computationally expensive. We show that our procedure to determine the structural information content of a graph requires polynomial time complexity and, hence, can be computed even for large graphs as, e.g., gene networks, consisting of hundreds or even thousands of vertices. Finally, we present some numerical results to demonstrate how to calculate our measure.

2. Information Theoretic Functionals for Graphs

In this section, we define a functional to measure the entropy of graphs that is based on metrical graph properties (Skorobogatov and Dobrynin, 1988). In order to define this functional, we start with some basic notations (Cover and Thomas, 2006; Harary, 1969; Skorobogatov and Dobrynin, 1988). We want to mention that in this paper we deal throughout with undirected, connected graphs without loops and multiple edges.

$G = (V, E), |V| < \infty$ denotes a finite undirected graph, where $E \subseteq (< MML : MFRACLINETHICKNESS = \text{"0"} > V2 < /MML : MFRAC >)$. $G = (V, E)$ is called connected if for arbitrary vertices v_i and v_j there exists an undirected path from v_i to v_j. Otherwise, we call G unconnected. In the following, $\mathcal{G}_{\mathrm{UC}}$ denotes the set of finite, undirected and connected graphs. The degree of a vertex $v \in V$ is denoted by $\delta(v)$ and equals the number of edges $e \in E$ which are incident with v. Starting from $G = (V, E) \in \mathcal{G}_{\mathrm{UC}}$, $\sigma(v) = \max_{u \in V} d(u, v)$ is called the eccentricity of $v \in V$, where $d(u, v)$ denotes the shortest distance between u and v. We notice that $d(u, v)$ is a metric. $\rho(G) = \max_{v \in V} \sigma(v)$ is called the diameter of G. Now, we define for $G = (V, E) \in \mathcal{G}_{\mathrm{UC}}$ the following vertex sets:

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

is called the j-sphere of v_i regarding G. Fig. (1) shows the process to determine j-spheres. In order to explain this process exemplarily, we choose an arbitrary vertex $v_i \in V$. As a first step, we determine all vertices $v \in V$ such that $d(v_i, v) = 1$ where the number of such vertices forms the cardinality of the 1-sphere $S_1(v_i, G)$, i.e., $|S_1(v_i, G)|$. As a second step, we determine all vertices $v \in V$ such that $d(v_i, v) = 2$ and, hence, we get $|S_2(v_i, G)|$. We continue this step until $j = \rho$ holds. By using, e.g., the well-known DIJKSTRA-algorithm (Cormen et al., 1990; Dijkstra, 1959), we repeat this process for each vertex $v_i \in V, 1 \leq i \leq |V|$. Now, we observe that