BioSystems 141 (2016) 31–39

Contents lists available at ScienceDirect

BioSystems

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

Organization principles of biological networks: An explorative study

Havva Kohestani[a], Alessandro Giuliani[b,*]

[a] Department of Biology, Faculty of Natural Science, University of Tabriz, Tabriz 5166616471, Iran
[b] Environment and Health Dept., Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Roma, Italy

ARTICLE INFO

Article history:
Received 7 July 2015
Received in revised form 12 January 2016
Accepted 29 January 2016
Available online 1 February 2016

Keywords:
Graph theory
Systems biology
Topology
Biocomplexity
Principal component analysis

ABSTRACT

The definition of general topological principles allowing for graph characterization is an important prerequisite for investigating structure–function relationships in biological networks. Here we approached the problem by means of an explorative, data-driven strategy, building upon a size-balanced data set made of around 200 distinct biological networks from seven functional classes and simulated networks coming from three mathematical graph models.

A clear link between topological structure and biological function did emerge in terms of class membership prediction (average 67% of correct predictions, $p < 0.0001$) with a varying degree of 'peculiarity' across classes going from a very low (25%) recognition efficiency for neural and brain networks to the extremely high (80%) peculiarity of amino acid–amino acid interaction (AAI) networks.

We recognized four main dimensions (principal components) as main organization principles of biological networks. These components allowed for an efficient description of network architectures and for the identification of 'not-physiological' (in this case cancer metabolic networks acting as test set) wiring patterns.

We highlighted as well the need of developing new theoretical generative models for biological networks overcoming the limitations of present mathematical graph idealizations.

1. Introduction

In 1952 the Dutch electrical engineer Bernard Tellegen (Tellegen, 1952) developed a theorem whose general importance in science has largely been underestimated (Mikulecky, 2001). Tellegen's theorem gives a simple relation between magnitudes that satisfy Kirchhoff's laws of electrical circuit theory. The Tellegen theorem is applicable to a multitude of network systems. The basic assumptions for the systems are the conservation of flow of extensive quantities (Kirchhoff's current law, KCL) and the uniqueness of the potentials at the network nodes (Kirchhoff's voltage law, KVL). Tellegen's theorem is a conservation principle of both potential and flux. The flux does not need to be an electrical current and the same holds for the potential. A flux of matter (in terms of relative concentrations of reagents and products) traverses a metabolic network and the free energy of the relative reaction is the potential. Similar reasoning holds for protein contact networks where the molecular motion flows between neighboring residues (nodes) and an interaction potential (intramolecular bonds) establishes between them (Di Paola and Giuliani, 2015).

As aptly stressed by Mikulecky (2001) the theorem opens the way to a 'network thermodynamics' strictly dependent from wiring architecture while largely independent of the constitutive laws governing the single elements.

The most spectacular empirical proof of Tellegen's theorem was probably in 1956, when Günthard and Primas (1956) realized that the matrix used in the classical Huckel method for deriving the energy of π molecular orbitals is a simple function of the adjacency matrix of the molecular graph.

Renewed interest in the study of complex networks (Albert and Barabási, 2002; Barabási and Oltvai, 2004), of 'biological systems as networks' boosted out in the beginning of this century.

In the present work, inspired by Tellegen theorem, we look for 'general principles' of network wiring architecture in the realm of biological networks. We are fully aware of the fact that flux networks (e.g. metabolic networks) and interaction networks (e.g. gene regulation networks) are drastically different as for their operational properties (Huang, 2004). Strictly speaking Tellegen's theorem only applies to flux networks, notwithstanding that, at a coarse-grain level, we are convinced the recognition of organization principles discriminating different kinds of networks could be of general interest.

* Corresponding author.
E-mail addresses: ha_koohestani90@ms.tabrizu.ac.ir (H. Kohestani), alessandro.giuliani@iss.it (A. Giuliani).

http://dx.doi.org/10.1016/j.biosystems.2016.01.004
0303-2647/

The existence of properties shared by biological and non-biological networks was already faced (Girvan and Newman, 2002; Alm and Arkin, 2003; Palla et al., 2005). These studies, at least to our knowledge, adopted a theoretical attitude focusing on features like community structure (Girvan and Newman, 2002; Palla et al., 2005) or scaling (Barabási and Albert, 1999). Other scientists focused on detailed analysis of the changes induced in the network wiring by disease conditions (Ideker and Krogan, 2012).

Here we adopted a somewhat different perspective: instead of focusing on 'average' properties of network classes, we adopted a bottom-up classification approach somewhat similar to quantitative structure–activity relationships (QSAR, Cronin, 2010) and epidemiological exploratory studies (Price et al., 2006).

Here the focus is no longer on the characterization of an 'ideal model' but on the actual discrimination of networks pertaining to different categories. This discrimination is obtained through a bottom-up strategy relying upon the empirical correlations holding between the data set descriptors. This approach allows for the derivation of 'general principles' in the form of principal components of the among descriptors correlation matrix (Benigni and Giuliani, 1994; Preisendorfer, 1988) acting as order parameters of the data set.

Principal component analysis (PCA) of topological descriptors generated a very clear correlation structure indicating three different layers of network organization and an unexpected 'network scalar'.

These 'principles' allowed us to both discriminate different network classes (so pointing to peculiar organization principles for each functional class) and to predict a 'pathological network character' of cancer regulation network. It is worth stressing that while components are each other orthogonal by construction on the entire data set, they have a rich local (within functional classes) correlation structure: any network class is a peculiar relational model in the topological component space.

2. Materials and methods

2.1. Data set

Eleven topological network descriptors and one derived feature defined as transmission load were computed for 222 networks pertaining to 10 classes in turn subdivided into 7 functional classes (the network are classified by their biological function) and 3 theoretical classes solely based on their wiring architecture. We also investigated 29 cancer metabolic networks pertaining to different cancer types that were used as 'test set' to check the ability of the extracted component space derived from the 222 networks (training set) to get rid of (in a totally data-driven way) the peculiar features of cancer networks.

2.1.1. Topological descriptors

Number of nodes: Nodes are the key building blocks of every network, also known as vertices in graph theory. A network is a system of nodes connected by links (or edges) (Strogatz, 2001). The number of nodes is a measure of the size of the network.

Shortest path number: The shortest path between two nodes is the path connecting them with the minimum number of links to traverse and corresponds to the shortest distance between two specified nodes in a network. The total number of shortest paths in a network is thus dependent on both network size and wiring architecture.

Average shortest path (ASP): or characteristic length is the average of shortest path lengths over all pairs of nodes in the network. ASP participates in the emergence of diverse large-scale behaviors of network systems. This is particularly evident in AAI (Amino Acid Interaction Networks), those networks having amino acid residues as nodes and their contact in 3D space as edges (Zhou et al., 2014; Di Paola et al., 2012). In AAI, ASP minimization is demonstrated to be at the basis of an efficient allosteric behavior of protein molecules (Tasdighian et al., 2013; Del Sol et al., 2006).

Average degree: The degree of a node is the number of nodes connected directly to that node also known as the number of direct connections of a node to other nodes via edges. For an undirected network including N nodes and L links, $\langle k \rangle = 2L/N$ denotes the average degree of network.

Network density: The ratio of actual connections between nodes of a network to the maximum potential connections is defined as the density of that network. For a network by N nodes potential connections is given by $N^*(N-1)/2$. This measure acquires values in the domain of [0,1].

Clustering coefficient: This descriptor has to do with the tendency of the nodes to cluster together constructing dense neighborhoods. This index builds upon the frequency of 'triads' satisfying the relation 'if A is directly connected to B and B directly connected to C, even A is directly connected with C'. For an undirected network, clustering coefficient of a node n is defined as $C_n = 2e_n/(k_n(k_n-1))$, where k_n is the number of all neighbors of node n and e_n is the number of connected pairs between all neighbors of node (Shannon et al., 2003). We considered the average of clustering coefficient over all the nodes of each network.

Network centrality: Centrality indices have different definitions depending upon the metrics adopted to define the 'importance' of a node. The simplest way (closeness centrality) to define the centrality of a node i is the inverse of the distances (in terms of edges to be traverse) from node i to all the other nodes of the network. Here we adopted the eigen vector centrality (EVC) that corresponds to the eigenvalue of the principal eigenvector of the adjacency matrix of a network. It is a continuous measure that depends on more than just the node itself, but also on the surrounding neighborhood (as opposed to degree centrality that is purely local). EVC is a measure of network well-connectedness (Canright and Engø-Monsen, 2006; Carreras et al., 2007). EVC is a *proxy* of the spreading power of a single node, or a collection of nodes, by the most central node or by the average EVC of all the nodes in the collection.

Network diameter: The longest distance between two nodes corresponds to network diameter in terms of number of edges to be traversed (Shannon et al., 2003). It is a measure of the topological width of the network.

Network heterogeneity: Network heterogeneity corresponds to node degree variance. Although this descriptor can be computed for all network types, heterogeneity is one of the most important features of scale-free networks. Scale-free networks have few highly connected nodes (i.e. 'hubs') and a great majority of low-degree nodes. This feature is a result of power law degree distribution in scale-free networks.

Modularity: Modularity corresponds to the number of links between residues of the same module minus the expected number links in the same module if the links of that network were distributed randomly. Modularity has values in the range of $[-1/2, 1]$ (Newman, 2006).

Transmission load: Transmission load is the ratio of average shortest path length to number of nodes in network TL = ASP/N. It builds upon the recognized correlation between number of nodes and ASP so giving a normalized measure of ASP linearly independent of the size of the network in terms of number of nodes. The higher the transmission load, the less optimal is the wiring of network in terms of transmission efficiency.

2.1.2. Network classes

We used Cytoscape v_2.8.1 and Cytoscape v_3.1.1 (Shannon et al., 2003) to perform network parametric analysis and importing