BioSystems 107 (2012) 186–196

Contents lists available at SciVerse ScienceDirect

BioSystems

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

Structural distance and evolutionary relationship of networks

Anirban Banerjee*

Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany

ARTICLE INFO

Article history:
Received 23 February 2011
Received in revised form 4 November 2011
Accepted 6 November 2011

Keywords:
Metabolic networks
Graph spectra
Graph Laplacian
Normalized graph Laplacian
Structural difference of networks
Graph evolution
Evolutionary relationship of networks

ABSTRACT

Exploring common features and universal qualities shared by a particular class of networks in biological and other domains is one of the important aspects of evolutionary study. In an evolving system, evolutionary mechanism can cause functional changes that forces the system to adapt to new configurations of interaction pattern between the components of that system (e.g. gene duplication and mutation play a vital role for changing the connectivity structure in many biological networks. The evolutionary relation between two systems can be retraced by their structural differences). The eigenvalues of the normalized graph Laplacian not only capture the global properties of a network, but also local structures that are produced by graph evolutions (like motif duplication or joining). The spectrum of this operator carries many qualitative aspects of a graph. Given two networks of different sizes, we propose a method to quantify the topological distance between them based on the contrasting spectrum of normalized graph Laplacian.

We find that network architectures are more similar within the same class compared to between classes. We also show that the evolutionary relationships can be retraced by the structural differences using our method. We analyze 43 metabolic networks from different species and mark the prominent separation of three groups: Bacteria, Archaea and Eukarya. This phenomenon is well captured in our findings that support the other cladistic results based on gene content and ribosomal RNA sequences. Our measure to quantify the structural distance between two networks is useful to elucidate evolutionary relationships.

1. Introduction

In evolving systems, some dynamics play a role to organize the connections between the components of that system. In a broad sense, due to the interplay between the structure and dynamics, biological and other networks evolve with different evolutionary dynamics are expected to have different structures while the networks constructed from the same evolutionary process have structural similarities. It is important to find the prominent structural difference between different types of networks, e.g., metabolic, protein–protein interaction, power grid, co-authorship or neural networks. Studies of common features and universal qualities shared by a particular class of a biological network is one of the most important aspects of evolutionary studies. In that regard, one can think about the differences between the networks within a same class (for instance among all metabolic networks), and also pose a question: are two evolutionary metabolic networks from two different species more similar than others?

In the last few years different notions of graph theory have been applied and new heuristic parameters have been introduced to analyze different aspects of network topology such as degree distribution, average path length, diameter, betweenness centrality, transitivity or clustering coefficient, etc. (see Newman, 2003 for details). These quantities can capture some specific but not all qualitative aspects of a graph. With these parameters, it is not always easy to distinguish or compare the topology of different real networks and to predict their source of formation. A popular trend is to categorize networks according to their degree distribution which is the distribution of k_n, the number of vertices that have degree n. It has been observed that most of the real networks have power-law degree distribution (Albert et al., 1999; Barabási and Albert, 1999; Guimera et al., 2005; Jeong et al., 2000, 2001; Redner, 1998) which is a very general network quality. Graphs with same degree sequences can have a very different synchronizability (Atay et al., 2006a,b). The invariants like average path length or diameter of a graph can vary widely depending on the details of the preferential attachment rule chosen (Jost and Joy, 2002b). Thus the power-law degree distribution fails to distinguish networks from different systems. The relative frequencies of small motifs help to categorize real networks into some superfamilies (Milo et al., 2002, 2004) but it cannot distinguish the networks very well within a superfamily. Hence focusing on specific features and qualities is

* Current address: Department of Mathematics and Statistics, Department of Biological Sciences, Indian Institute of Science Education and Research-Kolkata, Mohanpur 741252, India.
E-mail addresses: banerjee@molgen.mpg.de, anirban.banerjee@iiserkol.ac.in

0303-2647/$ – see front matter © 2011 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2011.11.004

not enough to reveal the structural complexity in biological and other networks.

In this article, we propose a method to quantify the structural differences between two networks. The basic tool we employed to characterize the qualitative topological properties of a network is the normalized graph Laplacian (in short Laplacian) spectra (Jost and Joy, 2002a). The multiplicity of the smallest eigenvalue λ_0 is equal to the number of components in the graph. The distance of the highest eigenvalue λ_{N-1} from 2 reflects how far the graph is away from the bipertiteness. Another property of the spectra of a bipartite graph is if λ is an eigenvalue, $2-\lambda$ is also an eigenvalue of that graph and hence the spectral plot will be symmetric about 1. The first nontrivial eigenvalue λ_1 (for connected graph) tells us how easily one graph can be cut into two different components. For the complete connected graph with N vertices, all nontrivial eigenvalues are equal to $N/(N-1)$ (see Chung, 1997; Jost, 2007 for the details). Not only the global properties of a graph structure are reflected by the Laplacian spectrum, local structures produced by certain evolutionary processes like motif joining or duplication are also well captured by the eigenvalues of this operator (Banerjee and Jost, 2007a, 2008a, 2009a). For instance, a single vertex (the simplest motif) duplication produces eigenvalue 1, which can be found with a very high multiplicity in many biological networks. Duplication of an edge (motif of size two) that connects the vertices i_1 and i_2 generates the eigenvalues $\lambda_{\pm} = 1 \pm (1/\sqrt{n_{i_1} n_{i_2}})$, and the duplication of a chain $(i_1 - i_2 - i_3)$ of length 3 produces the eigenvalues $\lambda = 1, 1 \pm \sqrt{1/n_{i_2}((1/n_{i_1}) + (1/n_{i_3}))}$ (where n_i is the degree of the vertex i). The duplication of these two motifs create eigenvalues which are close to 1 and symmetric about 1. For certain degrees of vertices, the duplication of these motifs can generate specific eigenvalues 1 ± 0.5 and $1 \pm \sqrt{0.5}$ which are also mostly observed in the spectrum of real networks. If we join a motif Σ (with an eigenvalue λ) with an eigenfunction that vanishes at a vertex $i \in \Sigma$ by identifying the vertex i with any vertex of a graph Γ, the new graph will also have the same eigenvalue λ. As an example, if we join a triangle that itself has an eigenvalue 1.5 to any graph, it contributes the same eigenvalue to the new graph produced by the joining process (for more details see Banerjee and Jost, 2007a,b, 2008a, 2009a,b). See Jost and Joy (2002a), Rangarajan and Ding (2002) and Atay et al. (2004) for how the spectra can influence dynamical properties like synchronization. Thus the various local structures of a graph can leave significant traces in the spectrum which is a good characteristic. The distribution of the spectrum has been considered as a qualitative representation of the structure of a graph (Banerjee and Jost, 2007b). In other way around, with the good algorithms one can reconstruct a graph from its spectrum (up to isospectrality) (Ipsen and Mikhailov, 2002). Comparative studies on real networks are difficult because of their complicatedness, irregular structure and different sizes. Graphs of similar sizes can be aligned on each other to compare the structural similarities. For any graph, all eigenvalues of the graph Laplacian operator are bounded within a specific range (0–2). This is an added advantage when comparing spectral plots of graphs with different sizes.

Spectral plots that can distinguish networks of different origins have been widely used to classify real networks from different sources (Banerjee and Jost, 2008b). Since networks constructed from the same evolutionary process produce very similar spectral plots, the distance between spectral distributions can be considered as a measure of structural differences. Hence it can be used to study the evolutionary relation between networks. In this paper we quantify this distance with the help of a divergence measure (Jensen–Shannon divergence) between two distributions. We consider this as a quantitative distance measure of those two structures and show that the evolutionary relationships between the networks can be derived from their topological similarities captured by this quantification.

To find the efficiency of this method, we apply it on the simulated networks constructed from the artificial evolutionary processes. The method successfully shows that the evolutionary relations between the networks can be retraced by their structural differences. Afterwards we apply this method to the metabolic networks of 43 species and show that the phylogenic evidences can be traced from the measurement of their structural distances.

1.1. Previous work

In the last few years, different methods such as elementary mode analysis (Schuster et al., 2000), method of singular value decomposition (SVD) of extreme pathways (Price et al., 2002), comparison of extreme pathways and elementary mode (Papin et al., 2004), etc. have been applied to characterize and compare metabolic pathways and networks.

Different graph theoretical approaches like comparison of the network indices, degree distribution and motif profile (Zhu and Qin, 2005) have been explored to compare metabolic network structures. For the evolving system, a general graph alignment method has been considered for the cross-species analysis of interaction networks (Berg and Lässig, 2006).

Several other methods such as multivariate analysis on the enzyme and substrate ranking (Poldani et al., 2001), comparison of network similarity by obtaining the similarity score between the vertices (Heymans and Singh, 2003), enzyme, reaction, and gene contents comparison (Ma and Zeng, 2004) have also been applied to reconstruct the phylogeny comparing the metabolic networks. Different operations from the set algebra have been used on the network to trace the phylogeny (Forst et al., 2006). Metabolic network structures have been compared by using graph kernel to reconstruct the phylogenetic tree (Oh et al., 2006). Mazurie et al. (2008) has predicted cross species phylogenetic distance by computing the distances between the vectors with the components of several network-descriptors which are estimated on the NIP (network of interacting pathways). Borenstein (2008) has predicted the phylogenetic tree by comparing the *seed* compound content.

In this paper, we implemented a method that is based on the graph spectrum and which carries many qualitative aspects of a graph to compare different network structures. This is a very general graph theoretical method and can be applied to any kind of networks without having any prior knowledge about their source. Our aim is not to reconstruct the phylogenetic tree, but rather to find the evolutionary closeness between the networks from the same evolving system. In the same context, Erten et al. (2009) performed a phylogenetic analysis of protein–protein interaction networks based on the conservation and divergence of modular components, and Mano et al. (2010) attempted to find the co-evolutionary relationships between metabolic pathways by comparing them to the evolutionary relationship between different organisms based on the combined similarities of all of their metabolic pathways.

2. Methods

2.1. Spectrum of graph Laplacian

The normalized graph Laplacian operator (Δ) is represented on an undirected and unweighted graph Γ that represents a network with a vertex set $V = \{i : i = 1, \ldots, N\}$. The vertices i and j are called neighbors if they are connected by an edge. The degree n_i of a vertex i is the number of neighbors of i. The graph Laplacian (Banerjee and Jost, 2008a; Jost, 2007; Jost and Joy, 2002a) has been defined as the $N \times N$ matrix $\Delta = (\Delta)_{ij}$, $i, j = 1, \ldots, N$ where

$$(\Delta)_{ij} := \begin{cases} 1 & \text{if } i = j \\ -\frac{1}{n_i} & \text{if } i \text{ and } j \text{ are neighbors} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$