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

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

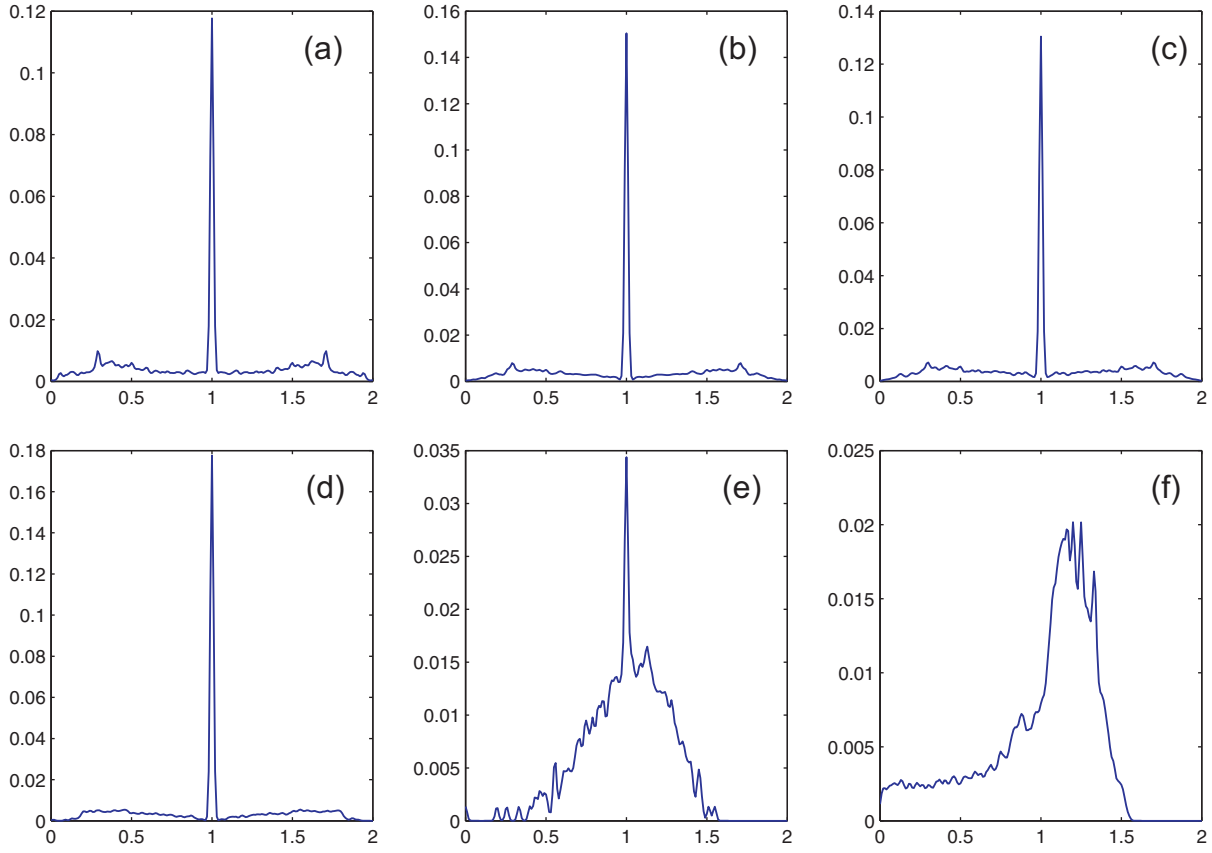

Fig. 1. Spectral plots of the metabolic networks of (a) *P. horikoshii*, (b) *E. coli*, and (c) *S. cerevisiae*. The sizes of the networks are 945, 2859 and 1812 respectively. The nodes represent substrates, enzymes and intermediate complexes. (d) Protein–protein interaction network of *H. Pylori*. Network size = 710. (e) Neuronal connectivity of *C. elegans*. Size of the network = 297. (f) Topology of the Western States power-grid of the United States. Network size = 4941. We plot the spectrum as a collection of the eigenvalues λ_i by convolving with a Gaussian kernel (with $\sigma = 0.01$), i.e. we plot $f(x) = \sum_{\lambda_i} 1/0.01\sqrt{2\pi}\exp(-|x-\lambda_i|^2/0.0002)$ along the vertical axis.

(Note that this operator has a similar but different spectrum like the operator investigated in Chung (1997) which is usually studied in the graph theoretical literature as the (algebraic) graph Laplacian (see Mohar, 1991 for this operator). A nonzero solution u of the equation $\Delta u - \lambda u = 0$ is called an eigenfunction for the eigenvalue λ. Δ has N eigenvalues, some of which may occur with higher multiplicity. The eigenvalues of this operator are real and non-negative. The smallest eigenvalue is always $\lambda_0 = 0$, since $\Delta u = 0$, for any constant function u.

2.2. Compute the spectral density

We convolve the spectrum of a network with a kernel $g(x, \lambda)$ and get the function

$$f(x) = \int g(x,\lambda)\sum_k \delta(\lambda,\lambda_k)d\lambda = \sum_k g(x,\lambda_k) \tag{2}$$

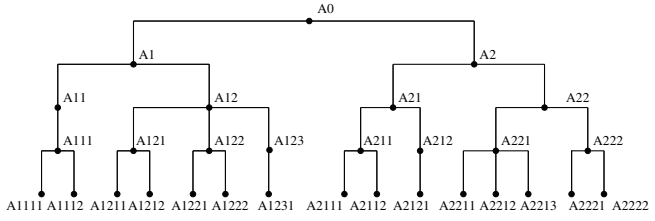

Fig. 2. Evolution of a graph A0 along a definite tree: A1 and A2 have been produced independently in the 2nd generation with a certain evolutionary process from A0. In the same way, A11 and A12 have been produced from A1 and A21, A22 from A2 and so on. Continuing in the same fashion, we end up with the graphs A1111,. . .,A2222 in the 5th generation. The tree which is shown here is our true tree. Any evolutionary process can be applied on A0 to produce A1111,. . .,A2222 along this tree.

Clearly

$$0 < \int f(x)dx < \infty \tag{3}$$

Here we use the Gaussian kernel $1/\sqrt{2\pi\sigma^2}\exp(-(x-m_x)^2/2\sigma^2)$ with $\sigma = .01$ for all computation purposes. Choosing other types of kernels does not change the result significantly since a kernel does not change the distribution unless the value of the parameter is profoundly different. Hence the choice of the parameter value is important (Banerjee and Jost, 2007b, 2009a).

We compute the spectral density f^* by normalizing f as:

$$f^*(x) = \frac{f(x)}{\int f(y)dy} \tag{4}$$

Table 1
Distance table between metabolic networks of *P. horikoshii* (Γ_{Ph}), *E. coli* (Γ_{Ec}), *S. cerevisiae* (Γ_{Sc}); protein–protein interaction network of *H. pylori* (Γ_{Hp}); neuronal connectivity network of *C. elegans* (Γ_{Ce}) and US power-grid network (Γ_{PG}). All distances are computed using the metric $D(\Gamma_1, \Gamma_2)$.

Network	Γ_{Ph}	Γ_{Ec}	Γ_{Sc}	Γ_{Hp}	Γ_{Ce}	Γ_{PG}
Γ_{Ph}	0.0000	0.0904	0.0661	0.1694	0.4704	0.4704
Γ_{Ec}	0.0904	0.0000	0.0641	0.1036	0.4902	0.5074
Γ_{Sc}	0.0661	0.0641	0.0000	0.1340	0.4574	0.4738
Γ_{Hp}	0.1694	0.1036	0.1340	0.0000	0.5086	0.5380
Γ_{Ce}	0.4704	0.4902	0.4574	0.5086	0.0000	0.2429
Γ_{PG}	0.4780	0.5074	0.4738	0.5380	0.2429	0.0000

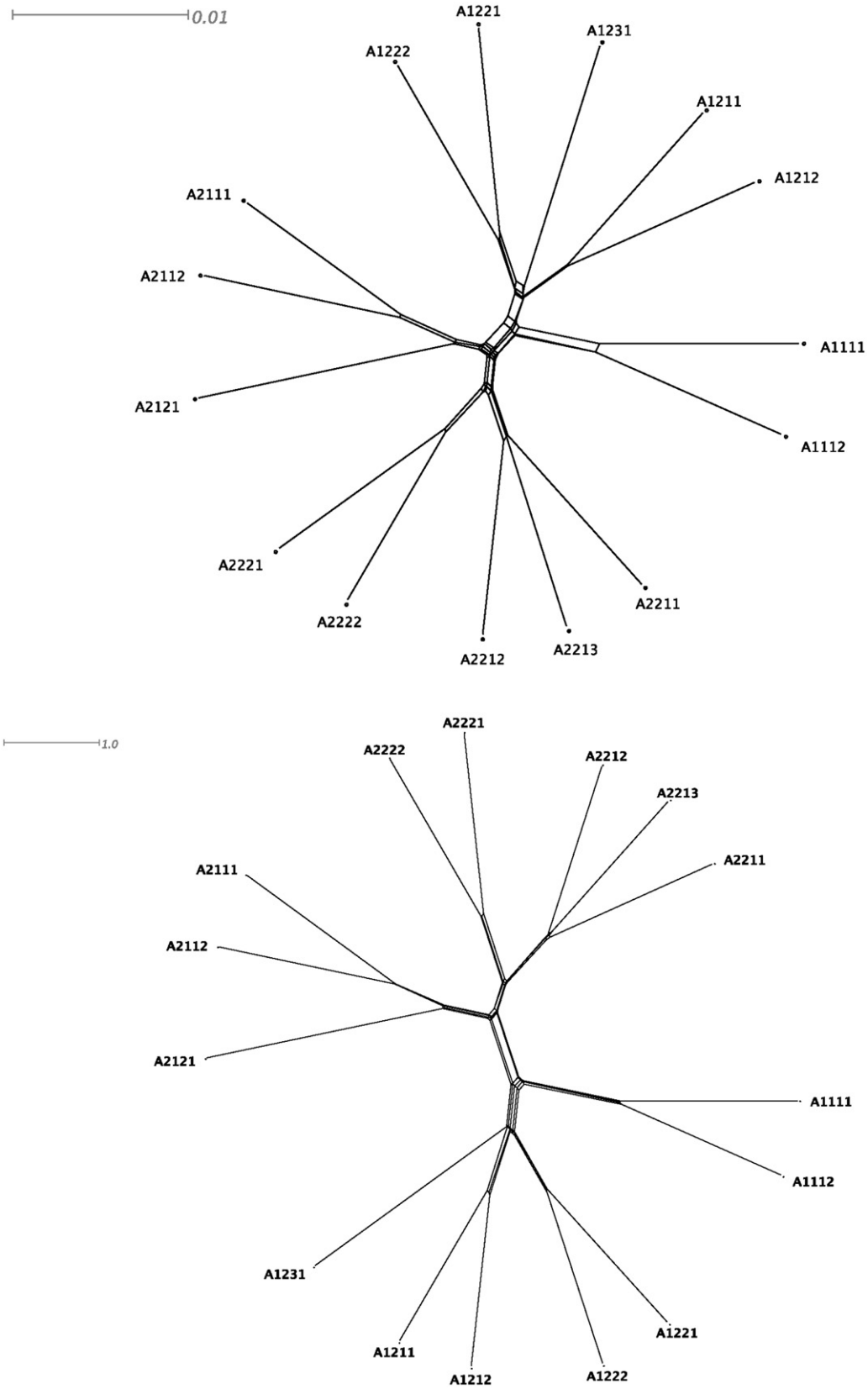

Fig. 3. Retracing the evolutionary relations, i.e. the leafs of a definite tree, between the networks (Fig. 2) constructed by the 'Edge-rewiring' evolutionary mechanism (upper figure) and 'Duplication–divergence' evolutionary mechanism (lower figure). Both the figures show the splits network for the structural distances (calculated by our proposed metric) of the graphs from the 5th generation. Each band of parallel edges indicate a split. For example, the parallel lines closer to A1111 and A1112 represent the split {A1111, A1112} versus the others. This tree-like splits network shows that the evolutionary relationships among those graphs is clearly captured by our distance measure. The figure has been produced by using Neighbor-Net (Bryant and Moulton, 2004).

2.3. Jensen–Shannon divergence as a measure for the structural distance

In a discrete system, Kullback–Leibler divergence measure (K–L) for two probability distributions p_1 and p_2 of a discrete random variable X is defined as

$$KL(p_1, p_2) = \sum_{x \in X} p_1(x) \log \frac{p_1(x)}{p_2(x)} \tag{5}$$

Note that the K–L divergence measure is not defined when $p_2 = 0$ and $p_1 \neq 0$ for any $x \in X$. This measure is not symmetric i.e. $KL(p_1, p_2) \neq KL(p_2, p_1)$ and does not satisfy the triangle inequality and hence cannot be considered as a *metric*.

Jensen–Shannon (J–S) divergence measure for two probability distributions p_1 and p_2 is defined as

$$JS(p_1, p_2) = \frac{1}{2} KL(p_1, p) + \frac{1}{2} KL(p_2, p); \text{ where } p = \frac{1}{2}(p_1 + p_2) \tag{6}$$

This measure is symmetric and unlike the K–L divergence measure, it is also defined when one of the probability measures (p_1 or p_2) is zero for some value of x (for more details see Lin, 1991). The square root of J–S divergence is a metric (for details Österreicher and Vajda, 2003).

For two different graphs Γ_1 and Γ_2 with spectral density (of graph Laplacian) f_1^* and f_2^* respectively, we define the structural distance $D(\Gamma_1, \Gamma_2)$ between two different graphs in terms of the J–S divergence measure between f_1^* and f_2^*:

$$D(\Gamma_1, \Gamma_2) = \sqrt{JS(f_1^*, f_2^*)} \tag{7}$$

Theoretically, there exists isospectral graphs but they are relatively rare in real networks and qualitatively quite similar in most respects (see Wilson and Zhu, 2008 for a systematic discussion). For example, all complete bipartite graphs of the form $K_{m,n}$ (with $m + n$ = constant) have the same spectrum. In this case distances between those structures will be zero which is one drawback of this measurement.

Since the eigenvalues of the normalized graph Laplacian are bound between [0, 2] the spectral distributions are easily comparable for the graphs of different sizes. It's worth noting that eigenvalues are not always smoothly distributed over [0, 2]. One can obtain a high structural difference between two graphs which are structurally the same but have very different sizes. We convolve the spectrum with a kernel before computing D to solve this issue. Smoothing the distribution of the spectrum does not correspond to a different structure (for the reasons see Ipsen and Mikhailov, 2002; Chen et al., 2004) and thus the size difference is not a problem for our metric D to measure structural distance between two networks.

2.4. Cluster of the metabolic networks by constructing an unrooted tree

We are interested in extracting the clusters among all metabolic networks from their structural distances which can be implemented by using an unrooted tree and invoking a neighbor-joining method. We calculate $D(\Gamma_i, \Gamma_j)$ for each pair of those networks (Γ_i, Γ_j) and build a distance matrix. We use the software package PHYLIP (Felsenstein, 1996) and SplitsTree (Huson, 1998) for the tree construction. Since our interest is to find the evolutionary relationships but not to reconstruct the phylogenetic tree, the branch length is not important for our purpose. Hence we ignore the branch length when we plot the tree.

We also use PHYLIP to compute the *symmetric difference* (Robinson and Foulds, 1981) between two trees.

2.5. Compute the normalized Z score of a motif

The normalized Z score of a motif for a given network is the normalized relative frequency of that motif compared to its expression in the randomized version of the same network. The statistical significance of a motif σ is presented by its Z score,

$$Z_\sigma = \frac{N_\sigma^{real} - \langle N_\sigma^{rand} \rangle}{SD(N_\sigma^{rand})}, \tag{8}$$

where N_σ^{real} is the number of times the motif σ appears in the network, and $\langle N_\sigma^{rand} \rangle$ and $SD(N_\sigma^{rand})$ are the mean and standard deviation of its appearance in the ensemble of randomized networks. Hence, the normalized Z score of a motif σ is $Z_\sigma / (\sum_\sigma Z_\sigma^2)^{1/2}$. Here, with the help of the software mfinder1.2 (freely available on http://www.weizmann.ac.il/mcb/UriAlon/), the Z score of each motif of size 3 and 4 are computed and normalized.

2.6. Data source

We enquired the freely accessible database http://www.barabasilab.com/rs-netdb.php to obtain metabolic data of the 43 species used in Jeong et al. (2000).

At the time of database construction, genomes of 25 species (18 bacteria, 2 eukaryotes and 5 archaea) had been completely sequenced while it was partially sequenced for the remaining 18 species. The analysis of the errors (Jeong et al., 2000) suggests that there would not be a drastic change in the final result. We utilized network data for protein–protein interaction of *Helicobacter pylori* from http://www.cosinproject.org/ and neuronal connectivity (used in Watts and Strogatz, 1998; White et al., 1986) of *C. elegans* from http://cdg.columbia.edu/cdg/datasets.

3. Results and discussion

First, we applied our measure on networks from different classes and observed the efficiency our measure in capturing the similarities and dissimilarities between the intra-class networks. We chose:

- Metabolic networks of *Pyrococcus horikoshii, Escherichia coli, Saccharomyces cerevisiae* where nodes are metabolites and metabolic reactions (network sizes are 945, 2859 and 1812 respectively)
- Protein–protein interaction network of *H. pylori* where nodes are proteins (network size 710)
- Neuronal connectivity (network) of *Caenorhabditis elegans* where nodes are represented by the neuronal cells (network size 297)
- Western States power-grid of the United States where generator, transformers, substations are considered as nodes (of size 4941)

For further reference, we denote these networks by Γ_{Ph}, Γ_{Ec}, Γ_{Sc}, Γ_{Hp}, Γ_{Ce} and Γ_{PG} respectively. Due to similar mechanisms (many metabolites or proteins have the same neighbors) of the network formation, it is expected that the metabolic networks will have a similar architecture with the protein–protein interaction networks rather than neuronal or power-grid networks. This phenomenon is clearly visible in the spectral plots (Fig. 1) of the above networks. Then we measured the structural distances between those networks using our metric D. The differences and similarities between those networks (inter- and intra-class) are clearly captured by this measurement (see the Table 1). Note that each network has a different size, but nevertheless, we can measure the structural distance from the difference of their spectral distributions. All the distances between these three metabolic networks are closer to each other than the protein–protein interaction network but far from the neuronal and power-grid network. The results are similar for the protein–protein interaction network. The relative distance between neuronal and power-grid networks compared to the other networks, is less but not as close as the one between the protein–protein interaction network and metabolic networks. Although the neuronal network is a biological network but it is structurally more similar to the power-grid network than the metabolic and protein–protein interaction networks. Unlike other biological networks, neuronal network exhibits small-world property like the power-grid network (Watts and Strogatz, 1998). These structural similarities are well captured by our defined measure. Our results show that we can consider our suggested metric as a suitable measure for structural differences.

3.1. Evolutionary relationship from the distance measure

We explored ways of finding networks which are evolutionary close to each other based on their structural differences. It is very likely that the networks constructed from the same evolutionary process are structurally close to each other and the architectures of the networks that share the same evolutionary path are expected to be more similar than others. Hence to a large extent, one can elucidate the evolutionary relationships between the networks within the same system from their structural distances. To verify this conviction, we develop a graph along a tree (see Fig. 2) and predict the evolutionary relations among the graphs of a generation. Here we choose the initial graph $A0$, a scale-free network constructed by Barabási–Albert's model (Barabási and Albert, 1999) ($m_0 = 5$ and $m = 3$) and apply two different evolutionary processes

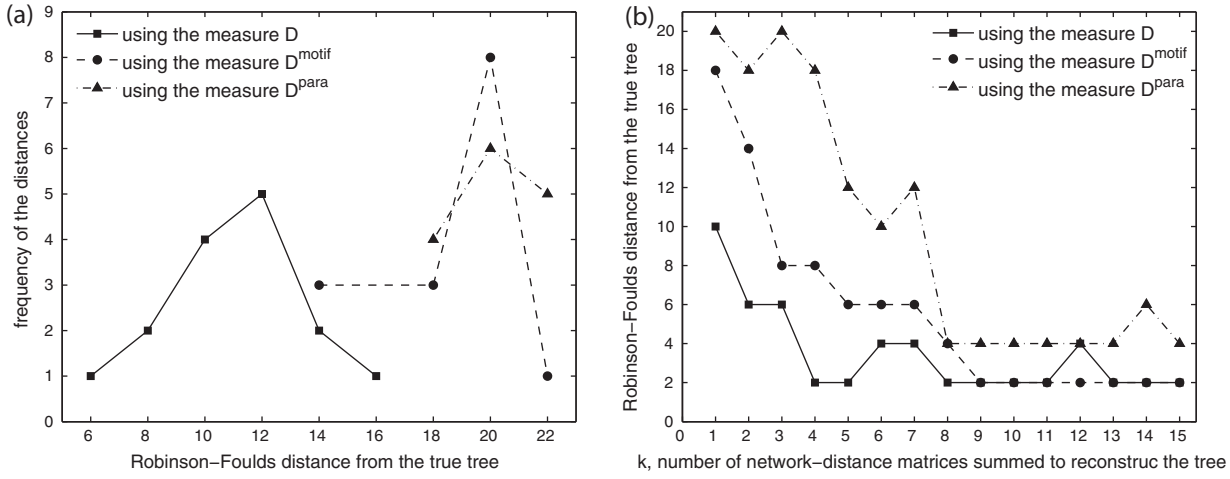

Fig. 4. The measure D is more accurate than D^{motif} and D^{para}. Here all the measures have been applied on the graphs constructed by the 'Edge-rewiring' evolutionary mechanism. (a) Frequency distributions of the Robinson–Foulds (R–F) distances of the trees that are constructed from graph structural-distances using D, D^{motif}, and D^{para} from the true tree (in Fig. 2). (b) The (R–F) distances are plotted along the vertical axis. We generate the graph distance matrices using D, D^{motif}, D^{para} for every k realization of graph evolution. Then we sum all the k distance matrices for each measure and compute the R–F distances of the trees reconstructed from these summed matrices from the true tree.

separately to produce the graphs A1111,. . .,A2222 in the 5th generation (Fig. 2).

Edge-rewiring: This is a very general evolutionary mechanism where the number of vertices, number of edges, and the degree sequence remain the same for all graphs produced by this evolution. We rewired a certain number of edges while keeping the degree of each node the same and produce a graph for the next generation.

Duplication–divergence: The duplication and divergence evolutionary mechanism is taken from the model (Vázquez et al., 2003) and it performs better in predicting the structure of protein–protein interaction networks compared to many other models (Middendor et al., 2005). We duplicated a randomly chosen vertex i and added an edge between i and its duplicate i' with the probability p (we choose $p = .1$). For each neighbor j of i, one of the randomly chosen edges (i, j) or (i', j) is removed with the probability q ($q = .5$).

We applied our measure D (in (7)) on the graphs A1111,. . .,A2222 (produced by both of the evolutionary mechanisms) separately. With these distances, we generated a splits network (Huson, 1998) for each evolutionary mechanism which can extract phylogenetic signals that are missed in other tree-representations. Fig. 3 shows the splits networks constructed from the distances (measured by D) between the graphs A1111,. . .,A2222 which are produced by the 'Edge-rewiring' and 'Duplication–divergence' evolutionary mechanisms respectively. This tree-like network (Fig. 3) shows that the distances contain a prominent phylogenetic signal and clearly demonstrates the evolutionary relationships between those graphs.

Next, we explored the performance of the measure D for many realization (of the evolutionary processes) and the ability to retrace

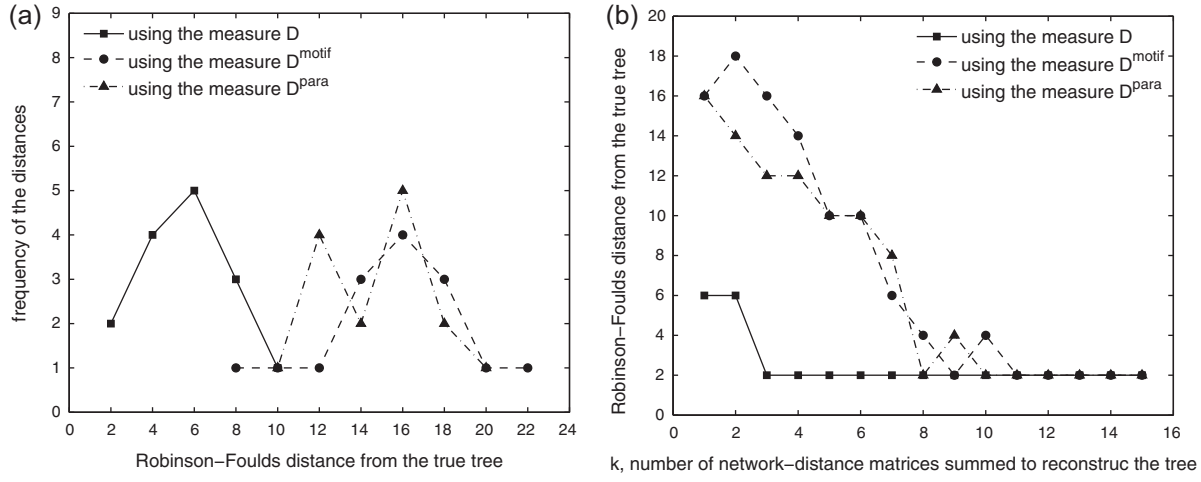

Fig. 5. The measure D is more accurate than D^{motif} and D^{para}. Here all the measures have been applied on the graphs constructed by the 'Duplication–divergence' evolutionary mechanism. (a) Frequency distributions of the Robinson–Foulds (R–F) distances of the trees that are constructed from graph structural-distances using D, D^{motif}, and D^{para} from the true tree (in Fig. 2). (b) The (R–F) distances are plotted along the vertical axis. We generate the graph distance matrices using D, D^{motif}, D^{para} for every k realization of graph evolution. Then we sum all the k distance matrices for each measure and compute the R–F distances of the trees reconstructed from these summed matrices from the true tree.

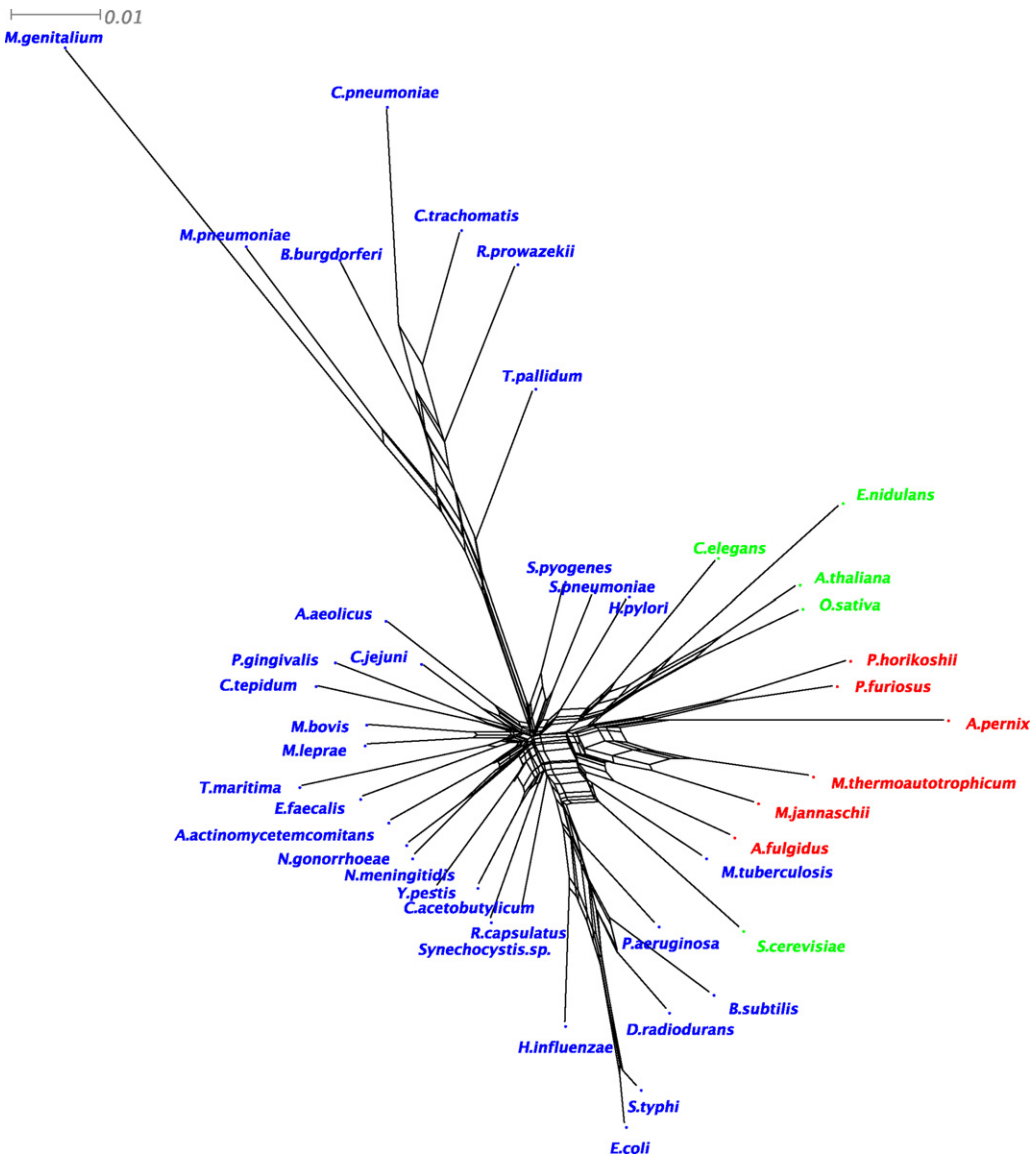

Fig. 6. The splits network for the structural distances (calculated by the metric D) between the metabolic networks of 43 species. This network shows that the distance-data is tree-like and has some phylogenetic signal. The colors, blue, green and red indicate Bacteria, Eukaryotes and Archaea respectively. We used Neighbor-Net (Bryant and Moulton, 2004) to generate this figure.

the evolutionary relationships for a higher amount of input data. We also compared D with the other structural difference measures which consider different network parameters.

3.2. Efficiency of the measure D and the comparison with other structural difference measures

To measure the efficiency of our distance measure, we studied the reproducibility of the true tree (Fig. 2). We used the *symmetric difference* measure, defined by Robinson–Foulds (R–F) (Robinson and Foulds, 1981), between the tree constructed from a distance matrix using neighbor-joining (N-J) method and the true tree shown in Fig. 2. The R–F distance between two trees is the number of bipartitions that can be found in one tree but not in the other one. Since our true tree contains two internal nodes (A12 and A221) of degree 4, the N-J tree with all the internal nodes have degree 3 always has two bipartitions which are never present in the true tree. A N-J tree that resembles the true tree most will have a R–F distance of 2 to the true tree (even the R–F distance of the true tree itself will be 2). Hence the more similar the trees are, the closer is the R–F distance to 2.

Other methods can also be used to quantify the structural similarities of the networks. A common way to compare two graph structures is to collate the independent heuristic parameters defined on them. For this purpose, we chose the following parameters: transitivity, diameter, radius, average path length, average edge-betweenness centrality, and average node-betweenness centrality for this purpose (see Newman, 2003 for the details on these parameters). We constructed a vector V_{Γ}^{para}, with the values of the parameters mentioned above from a graph Γ as the components and computed the structural difference D^{para} between two graphs Γ_1 and Γ_2 as

$$D^{para}(\Gamma_1, \Gamma_2) = \|V_{\Gamma_1}^{para} - V_{\Gamma_2}^{para}\| \quad (9)$$

The other measure (D^{motif}) we considered is based on the normalized Z score (Milo et al., 2002) of the motif of size 3 and 4. It has been shown that the networks can be categorized in different superfamilies (Milo et al., 2004) based on the characteristic

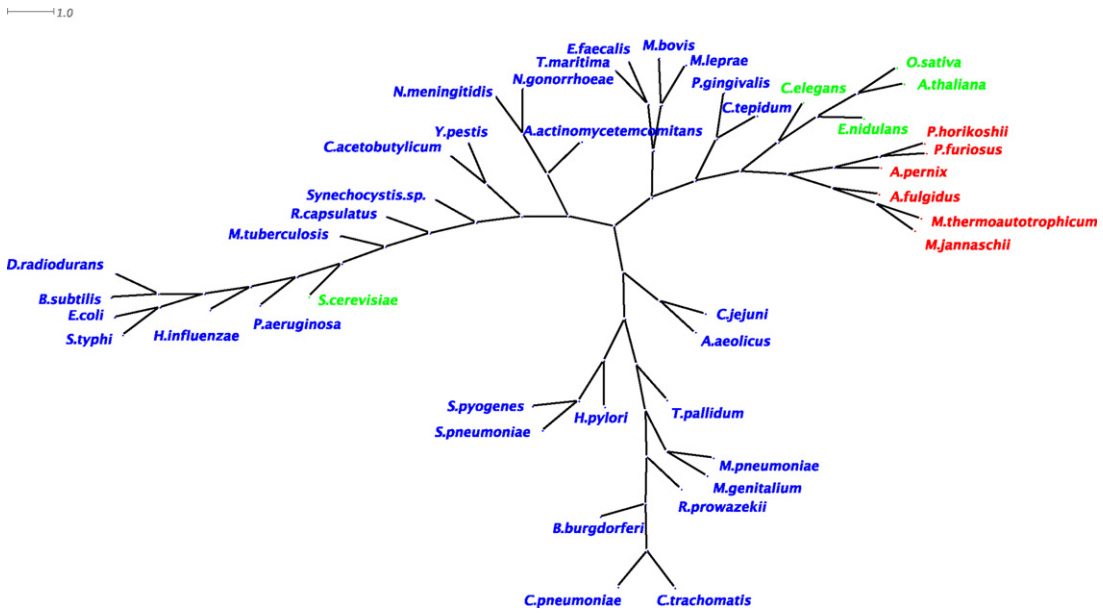

Fig. 7. The un-rooted tree of metabolic networks of 43 species constructed with their structural distances (calculated by our proposed metric) using the neighbor-joining method. Bacteria, Eukaryotes and Archaea are shown by the color, blue, green and red respectively and all of them form separate clusters within the tree. Only *S. cerevisiae* belongs to a different group, Bacterium.

distribution of the relative frequency of their motifs. In a similar way, we constructed a vector V_{Γ}^{motif} from a graph Γ with the values of the normalized Z score for the motif of size 3 and 4 as the components and computed the structural difference between two graphs Γ_1 and Γ_2 as

$$D^{motif}(\Gamma_1, \Gamma_2) = \|V_{\Gamma_1}^{motif} - V_{\Gamma_2}^{motif}\| \quad (10)$$

Next, we compared the efficiency of the measure D with D^{motif} and D^{para} to predict the evolutionary relationships among the graphs. As before, we computed the matrix with the distances estimated by a particular measure mentioned above between the graphs that are produced in the 5th generation of the graph evolution along the tree (Fig. 2) by both of the evolutionary mechanism (Edge-rewiring and Duplication–divergence). We repeated the process 15 times and computed the R–F distance for each time for all the distance measures. Figs. 4(a) and 5(a) shows the three frequency distributions of such R–F distances for every measure (for the evolutionary mechanism, Edge-rewiring and Duplication–divergence respectively). This clearly demonstrates that the measure D is more accurate than the other two. The limited accuracy can be explained by the stochasticity in the process of graph evolution.

In order to address whether the accuracy is also influenced by systematic effects, we investigated the trend in the R–F distances of the trees that are constructed by the sum of k distance matrices produced by a particular measure over k realizations of graph evolution from the true tree. When k increases (Figs. 4(b) and 5(b)), the R–F distance decreases and assumes its minimum value 2. For these two particular graph evolutions, the evolutionary relationships can be perfectly recovered from information obtained from the D-measure if the input size is large enough.

Evidently, the spectral distribution, which contains more qualitative properties of a network than the heuristic parametric values and the expression of the small motifs, captures the traces of an evolutionary relationship better when compared to a set of structural parameter values. Obviously the metric D which we considered, in essence, is of higher dimensionality (N, size of the network) compared to the other two measures and that is added advantage for better prediction.

3.3. Evolutionary relationships between metabolic networks of 43 species

We applied our structural difference measure D to estimate the distances between the metabolic networks of 43 species. Here, we consider metabolites and metabolic reactions as vertices. A metabolic reaction is connected to a metabolite with an edge if the metabolite is an educt or product of that reaction. We constructed a distance matrix between all 43 metabolic networks using D. Fig. 6, which is a splits network for these distances, supports that the data contained in that matrix has a substantial amount of phylogenetic signal and some parts of the data are tree-like. Due to the non-uniform evolutionary rate of topological change, we constructed an unrooted tree from the mentioned distance matrix by using the neighbor-joining method to analyze the structural similarities among the networks of all those species.

This tree, which highly resembles the phylogenetic tree of those 43 species, shows different clusters according to the structural similarities of the metabolic networks (see Fig. 7). Except Yeast which belongs to the group of bacteria, we see the prominent separation of three groups: bacteria, archaea and eukaryotes. This is well captured in our findings and support the other cladistic results based on gene content (Snel et al., 1999) and ribosomal RNA sequences (Woese et al., 1990).

We also used the metabolic-centric networks (where metabolites are the vertices and two metabolites are connected by an edge if they participate in a same metabolic reaction) and reaction-centric networks (where metabolic reactions are considered as vertices and two reactions are connected if a product of one reaction becomes the educt of another) constructed from the same data set. Fig. 8 shows the splits networks for the distances between the metabolic-centric networks and reaction-centric networks respectively as measured by D. The figure supports the existence of substantial amount of phylogenetic signal in the data contained in the distance matrices and shows separate clusters for bacteria, archaea and eukaryotes. The separation is prominent for the metabolic-centric networks. Only *A. pernix* belongs to the group of eukaryotes. For the reaction-centric networks, the clusters of archaea and eukaryotes are not clearly separated from each other

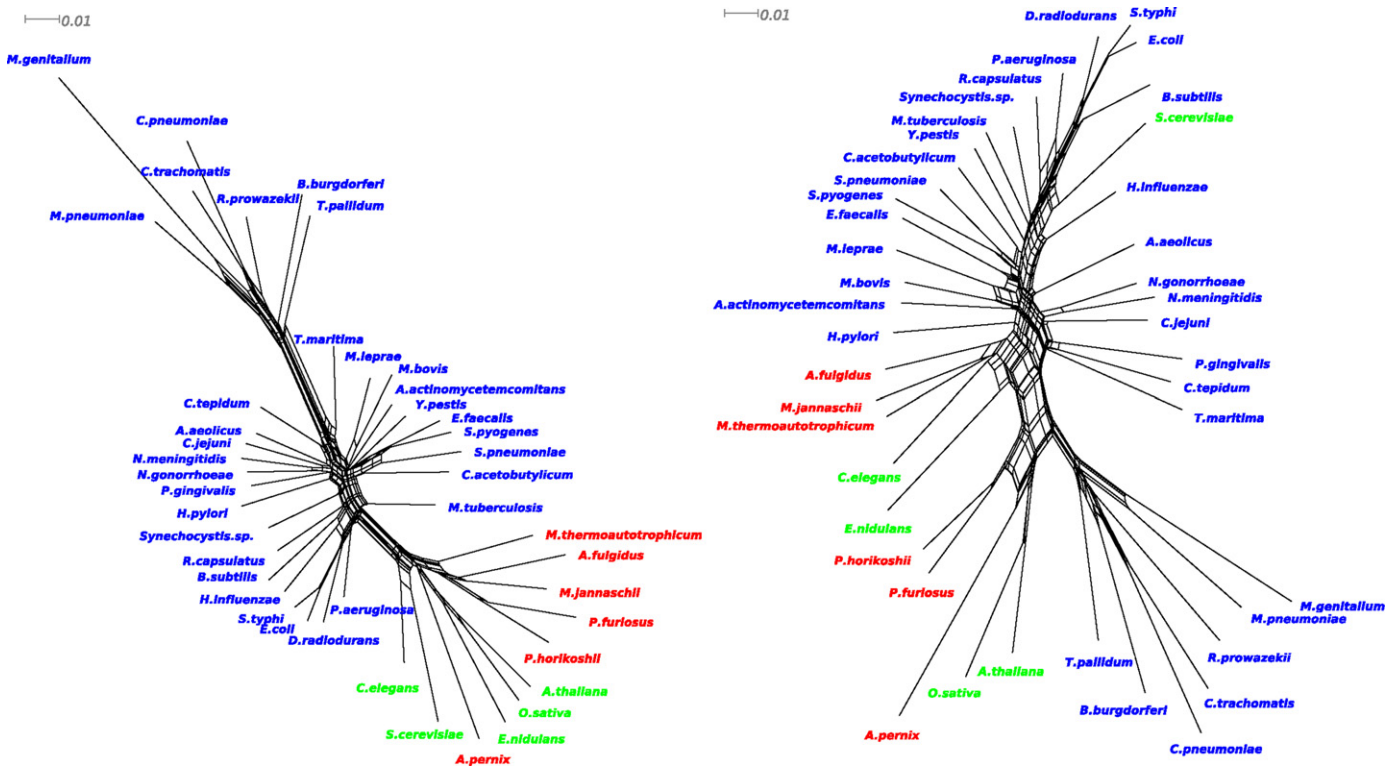

Fig. 8. The splits networks constructed from the structural distances (calculated by the metric *D*) between the metabolic-centric networks (on the left) and the reaction-centric networks (on the right) of 43 species. This network shows that the distance-data is tree-like and has some phylogenetic signal. The colors, blue, green and red indicate Bacteria, Eukaryotes and Archaea respectively and all of them roughly form separate clusters within the tree. We used Neighbor-Net (Bryant and Moulton, 2004) to generate this figure.

while Yeast belongs to the group of bacteria. All the figures, Figs. 6–8, show that eukaryotes and archaea are relatively close to each other than bacteria at the level of metabolic networks. This is a strong evidence how evolutionary relationship is reflected from the structural similarities which are clearly captured by our metric *D*.

To a large extent, the above result also holds for metabolite-centric networks where the currency metabolites (like ATP, water,

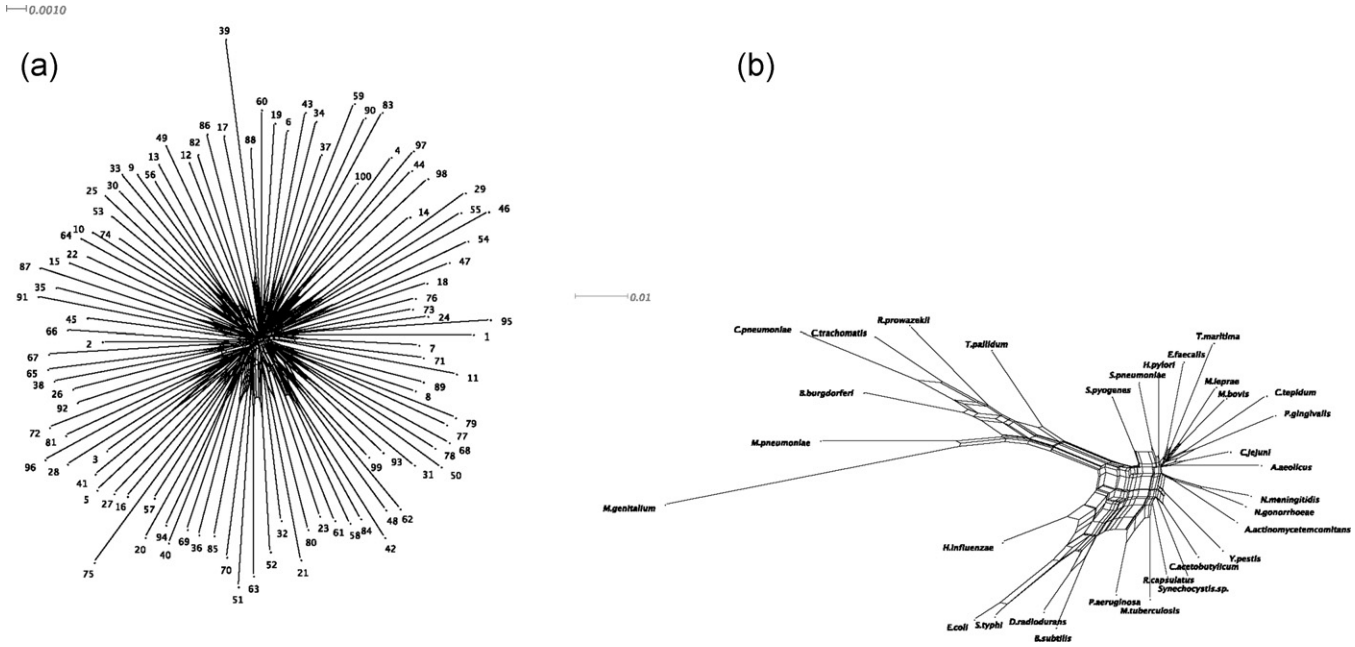

Fig. 9. The splits network of the structural distances between (a) 100 networks constructed by randomly deleting 5 percent of the reactions from the metabolic network of *E. coli* and (b) metabolic networks of 32 bacteria. The star-like structure of the splits network in (a), which is very different from the splits network of bacteria in (b), shows that the data of the distance matrix has a vague phylogenetic signal and the metabolic networks of bacteria are not constructed only by mapping from the *E. coli*. We have used Neighbor-Net (Bryant and Moulton, 2004) to construct both of the splits networks.

etc.) are excluded from the networks. In this case the species having similar characteristics are clustered together, e.g. thermophilic and mesophilic are clustered separately, and the species are in the symbiotic relation clustered together.

3.4. Cross validation of the tree construction against the effect of the enzyme mapping from E. coli

All the metabolic pathways in *E. coli* have been constructed independently in wetlab. which is not always the case for the other bacteria. If an enzyme-specific gene that also exists in *E. coli* was detected, the same metabolic reactions catalyzed by that enzyme are incorporated into the database. If there are no other genes which have been reported from every other bacteria and that can make significant change in the network structure, all other metabolic networks will be very similar and the detection of the phylogenetic relationship can be an artifact. In order to verify this fact, we reconstructed 100 networks by randomly deleting 5 percent of the reactions from the metabolic network of *E. coli* and produce a splits network of the distances between those 100 networks. The star-like structure of this splits network, which is very different from the splits network constructed from the structural distances between the metabolic networks of 32 bacteria, shows that the distances of those 100 networks have a vague phylogenetic signal (Fig. 9). Hence the evolutionary relationships cannot be detected if all other metabolic networks are only mapped from the network of *E. coli*.

4. Conclusions

We suggested a method to compare the architecture of networks with different sizes. With a defined metric, we quantified their structural differences based on the spectral distribution. This captures the qualitative properties of the underlying graph topology which can emerge from the evolutionary process like motif duplication or joining, random rewiring, random edge deletion, etc. With our proposed measure, we showed that the architecture of the networks are more similar within the same class than between classes. Due to the interplay between the structure and dynamics in many self-organized systems, like biological systems, the networks constructed from the same evolutionary process have similar structures and vice versa. We applied our topological distance measure on 43 metabolic networks from different species and apply phylogenetic clustering. In spite of network reconstruction error (see source of the data), our method elucidate the evolutionary relationships between those metabolic networks constructed from 43 different species.

Due to incomplete sequencing of the genome of different species, many biological data are incomplete and they contain statistical errors. To capture a more appropriate (i.e. with less error) network architecture we focused on the big component. It is very probable that this part of the network is constructed from the most studied metabolic pathways, hence consists more complete data and capture most of the qualitative properties of the original complete network. Moreover, in our analysis we considered the underlying undirected graphs of the real networks which are directed in many cases. The reduced graph itself carries a lot of structural information and is quite informative about the network. One can easily extend this method to directed networks.

One can use the spectrum of non-normalized Laplacian matrix or adjacency matrix to solve these issues but in some cases the spectral density of these matrices are influenced by the degree distribution of the network (Zhan et al., 2010; Dorogovtsev et al., 2004).

Our approach can also be useful to explore evolutionary relationships in other domains like language and society structure and in other biological areas.

Authors' contributions

AB conducted the research and wrote the paper.

Acknowledgments

The author is thankful to Martin Vingron, Thomas Manke, Roman Brinzanik, Sitabhra Sinha, Monojit Choudhur for valuable discussions. A special thank to Hannes Luz for useful suggestions on phylogenetic tree construction. The author is also thankful to Antje Glück for the helpful comments on preparing the manuscript. This is the part of the project "Complex Self-Organizing Networks of Interacting Machines: Principles of Design, Control, and Functional Optimization" (I/82 697) and is funded by the VolkswagenStiftung.

References

Albert, R., Jeong, H., Barabási, A.L., 1999. Internet – diameter of the world-wide web. Nature 401, 130–131.
Atay, F.M., Bıyıkoğlu, T., Jost, J., 2006a. Synchronization of networks with prescribed degree distributions. IEEE Trans. Circuits Syst. I 53 (1), 92–98.
Atay, F.M., Bıyıkoğlu, T., Jost, J., 2006b. Network synchronization: spectral versus statistical properties. Physica D: Nonlinear Phenomena 224 (1–2), 35–41.
Atay, F.M., Jost, J., Wende, A., 2004. Delays connection topology and synchronization of coupled chaotic maps. Phys. Rev. Lett. 92 (14), 144101.
Banerjee, A., Jost, J., 2007a. Laplacian spectrum and protein–protein interaction networks. Preprint. E-print available: arXiv:0705.3373.
Banerjee, A., Jost, J., 2008a. On the spectrum of the normalized graph Laplacian. Linear Algebra Appl. 428, 3015–3022.
Banerjee, A., Jost, J., 2009a. Graph spectra as a systematic tool in computational biology. Discrete Appl. Math. 157 (10), 2425–2431.
Banerjee, A., Jost, J., 2007b. Spectral plots and the representation and interpretation of biological data. Theory Biosci. 126 (1), 15–21.
Banerjee, A., Jost, J., 2008b. Spectral plot properties: towards a qualitative classification of networks. NHM 3 (2), 395–411.
Banerjee, A., Jost, J., 2009b. Spectral characterization of network structures and dynamics. In: Ganguly, N., et al. (Eds.), Dynamics on and of Complex Networks; Modeling and Simulation in Science, Engineering and Technology, Springer Birkhäuser Boston, pp. 117–132.
Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286 (5439), 509–512.
Berg, J., Lässig, M., 2006. Cross-species analysis of biological networks by Bayesian alignment. PNAS 103 (29), 10967–10972.
Borenstein, E., 2008. Large-scale reconstruction and phylogenetic analysis of metabolic environments. PNAS 105 (38), 14482–14487.
Bryant, D., Moulton, V., 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21 (2), 255–265.
Chen, G., et al., 2004. An interlacing result on normalized Laplacian. SIAM J. Discrete Math. 18 (2), 353–361.
Chung, F., 1997. Spectral Graph Theory. AMS.
Dorogovtsev, S.N., et al., 2004. Random networks: eigenvalue spectra. Physica A 338, 76–83.
Erten, et al., 2009. Phylogenetic analysis of modularity in protein interaction networks. BMC Bioinformatics 10, 333.
Felsenstein, J., 1996. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 266, 418–427.
Forst, C.V., et al., 2006. Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. BMC Bioinformatics 7, 67.
Guimera, R., Mossa, S., Turtschi, A., Amaral, L.A.N., 2005. The worldwide air transportation network: anomalous centrality, community structure, and cities global roles. Proc. Natl. Acad. Sci. U.S.A. 102 (22), 7794–7799.
Heymans, M., Singh, A.K., 2003. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Bioinformatics 19 (1), i138–i146.
Huson, D.H., 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 14 (1), 68–73.
Ipsen, M., Mikhailov, A.S., 2002. Evolutionary reconstruction of networks. Phys. Rev. E 66, 046109.
Jeong, H., Mason, S.P., Baraási, A.L., Oltvai, Z.N., 2001. Lethality and centrality in protein networks. Nature 411, 41–42.
Jeong, H., Tombor, B., Albert, R., Oltval, Z.N., Barabási, A.L., 2000. The large-scale organization of metabolic networks. Nature 407, 651–654.
Jost, J., 2007. Dynamical networks. In: Feng, J.F., Jost, J., Qian, M.P. (Eds.), Networks: From Biology to Theory. Springer, pp. 35–62.
Jost, J., Joy, M.P., 2002a. Spectral properties and synchronization in coupled map lattices. Phys. Rev. E 65, 16201–16209.

Jost, J., Joy, 2002b. Evolving networks with distance preferences. Phys. Rev. E 66, 36126–36132.
Lin, J., 1991. Divergence measures based on the Shanon entropy. IEEE Trans. Inf. Theory 37 (January (1)), 145–151.
Mano, A., et al., 2010. Comparative classification of species and the study of pathway evolution based on the alignment of metabolic pathways. BMC Bioinformatics 11 (Suppl. 1), S38.
Mazurie, A., et al., 2008. Phylogenetic distances are encoded in networks of interacting pathways. Bioinformatics 24 (22), 2579–2585.
Ma, H.-W., Zeng, A.-P., 2004. Phylogenetic comparison of metabolic capacities of organisms at genome level. Mol. Phylogenet. Evol. 31, 204–213.
Middendor, M., Ziv, E., Wiggins, C.H., 2005. Inferring network mechanisms: the *Drosophila melanogaster* protein interaction network. PNAS 102 (9), 3192–3197.
Milo, R., et al., 2004. Superfamilies of evolved and designed networks. Science 303, 1538–1542.
Milo, R., et al., 2002. Network motifs: simple building blocks of complex networks. Science 298, 824–827.
Mohar, B., 1991. The Laplacian spectrum of graphs. In: Alavi, Y., Chartrand, G., Oellermann, O.R., Schwenk, A.J. (Eds.), Graph Theory, Combinatorics, and Applications, vol. 2, Wiley, pp. 871–898.
Newman, M.E.J., 2003. The structure and function of complex networks. SIAM Rev. 45 (2), 167–256.
Oh, S.J., et al., 2006. Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks. BMC Bioinformatics 7, 284.
Österreicher, F., Vajda, I., 2003. A new class of metric divergences on probability spaces and its statistical applications. Ann. Inst. Statist. Math. 55, 639–653.
Poldani, J., et al., 2001. Comparable system-level organization of Archaea and Eukaryotes. Nat. Genet. 29, 54–56.
Papin, J., et al., 2004. Comparison of network-based pathway analysis methods. Trends Biotechnol. 22 (8), 400–405.
Price, N.D., et al., 2002. Extreme pathways and Kirchhoff's second law. Biophys. J. 83, 2879–2882.
Rangarajan, G., Ding, M.Z., 2002. Stability of synchronized chaos in coupled dynamical systems. Phys. Lett. A 296, 204–212.
Redner, S., 1998. How popular is your paper? An empirical study of the citation distribution. Eur. Phys. J. B 4 (2), 131–134.
Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.
Schuster, S., et al., 2000. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat. Biotechnol. 18, 326–332.
Snel, B., Bork, P., Huynen, M.A., 1999. Genome phylogeny based on gene content. Nat. Genet. 21, 108–110.
Vázquez, A., Flammini, A., Maritan, A., Vespignani, A., 2003. Modeling of protein interaction networks. ComplexUs 1, 38–44.
Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of small-world networks. Nature 393, 440–442.
White, J.G., et al., 1986. The structure of the nervous-system of the nematode Caenorhabditis-Elegans. Philos. Trans. R. Soc. Lond. Ser. B-Biol. Sci. 314, 1–340.
Wilson, R.C., Zhu, P., 2008. A study of graph spectra for comparing graphs and trees. Pattern Recognition 41 (9), 2833–2841.
Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U.S.A. 87, 4576–4579.
Zhan, C., Chen, G., Yeung, L.F., 2010. On the distributions of Laplacian eigenvalues versus node degrees in complex networks. Physica A 389, 1779–1788.
Zhu, D., Qin, Z.S., 2005. Structural comparison of metabolic networks in selected single cell organisms. BMC Bioinformatics 6, 8.