Algebraic connectivity may explain the evolution of gene regulatory networks

Zoran Nikoloski, Patrick May, Joachim Selbig

www.elsevier.com/locate/yjtbi

Algebraic connectivity may explain the evolution of gene regulatory networks

Zoran Nikoloski, Patrick May, Joachim Selbig

PII: S0022-5193(10)00385-1
DOI: doi:10.1016/j.jtbi.2010.07.028
Reference: YJTBI6090

To appear in: *Journal of Theoretical Biology*

Received date: 18 November 2009
Revised date: 21 July 2010
Accepted date: 21 July 2010

Cite this article as: Zoran Nikoloski, Patrick May and Joachim Selbig, Algebraic connectivity may explain the evolution of gene regulatory networks, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2010.07.028

Algebraic connectivity may explain the evolution of gene regulatory networks

Zoran Nikoloski[1,2*], Patrick May[2,3,4], Joachim Selbig[1,2]

1 Institute of Biochemistry and Biology, University of Potsdam, Potsdam, Brandenburg, Germany

2 Max-Planck Institute for Molecular Plant Physiology, Potsdam, Brandenburg, Germany

3 Luxembourg Centre for Systems Biomedicine, Université du Luxembourg, Luxembourg

4 Institute for Systems Biology, Dudley and Shmulevich Labs, Seattle, WA, USA ∗ **E-mail: nikoloski@mpimp-golm.mpg.de**

Abstract

Gene expression is a result of the interplay between the structure, type, kinetics, and specificity of gene regulatory interactions, whose diversity gives rise to the variety of life forms. As the dynamic behavior of gene regulatory networks depends on their structure, here we attempt to determine structural reasons which, despite the similarities in global network properties, may explain the large differences in organismal complexity. We demonstrate that the algebraic connectivity, the smallest non-trivial eigenvalue of the Laplacian, of the directed gene regulatory networks decreases with the increase of organismal complexity, and may therefore explain the difference between the variety of analyzed regulatory networks. In addition, our results point out that, for the species considered in this study, evolution favours decreasing concentration of strategically positioned feed forward loops, so that the network as a whole can increase the specificity towards changing environments.

Moreover, contrary to the existing results, we show that the average degree, the length of the longest cascade, and the average cascade length of gene regulatory networks cannot recover the evolutionary relationships between organisms. Whereas the dynamical properties of special subnetworks are relatively well understood, there is still limited knowledge about the evolutionary reasons for the already identified design principles pertaining to these special subnetworks, underlying the global quantitative features of gene regulatory networks of different organisms. The behavior of the algebraic connectivity, which we show valid on gene regulatory networks extracted from curated databases, can serve as an additional evolutionary principle of organism-specific regulatory networks.

Key words: algebraic connectivity, evolution, gene regulatory networks, dynamics

Introduction

Recent evidence from fully-sequenced genomes suggests that organismal complexity arises much more from the elaborate regulation of gene expression than by the genome size itself (Levine and Tjian, 2003). The ever-increasing throughput in experimental manipulation of gene activity coupled with the methods for quantitative assessment of perturbed transcriptome, proteome, and metabolome have begun to identify the effects of transcription factors, binding ligands, and post–translational modifications on regulated genes (Papp and Oliver, 2005). In addition to structural information regarding the regulatory interactions, comprehensive understanding of the behavior of these interactions also requires specification of: (1) the type of regulation (*i.e.*, activation or inhibition) (Albert and Othmer, 2003), (2) kinetics of interactions (Ronen *et al.*, 2002), and (3) the specificity of the interactions with respect to investigated tissue and/or stress condition (Luscombe *et al.*, 2004). The elucidation of complete network of regulatory interactions parametrized with

kinetic information leading to a particular type of gene expression is, at present, still a challenging task even for the well–studied model organisms whose networks have been partially assembled for few selected processes, conditions or on the level of the entire genome (Davidson *et al.*, 2002; Lee *et al.*, 2002; Shen-Orr *et al.*, 2002; Zhang *et al.*, 2006).

Nevertheless, recent theoretical investigations have established that the qualitative behavior of dynamic processes on complex networks is closely related to the structural network properties (Conradi *et al.*, 2007; Craciun *et al.*, 2006; Elowitz and Leibler, 2000; Feinberg, 1987; Madan Babu *et al.*, 2006). Consequently, the existing studies of gene regulatory networks have attempted to determine unifying design principles in order to understand and make biologically relevant conclusions solely from the network structure. The reported results mainly fall in the following two categories: (1) identifying local structures, such as the basic building blocks of networks, which, if present at statistically significant concentrations, are termed motifs (Milo *et al.*, 2002; Shen-Orr *et al.*, 2002), or particular subnetworks, called cascades (Rosenfeld and Alon, 2003), and (2) developing models which can explain global structural properties related to the salient network properties (*e.g.*, degree distribution, average path length, clustering coefficient) (Albert and Barabasi, 2002).

Network motifs are defined as patterns that occur more often in the real network than in randomized networks (Milo *et al.*, 2002). Three types of recurring network motifs were found to describe most of the *Escherichia coli* and *Saccharomyces cerevisiae* transcriptional regulatory networks (Milo *et al.*, 2002; Shen-Orr *et al.*, 2002). Each of these network motifs was shown to exhibit quantitative differences in the levels and timing of gene expression (Alon, 2007). Here, we are interested only in the concentration of a particular subnetwork, rather than its statistical significance. Another important feature of regulatory network architectures is the distribution of transcription cascade lengths

(Rosenfeld and Alon, 2003; Shen-Orr *et al.*, 2002). Transcription cascades are defined by a set of transcription factors that regulate each other sequentially. In a study of the architecture of the transcription network of *Escherichia coli* and *Saccharomyces cerevisiae*, it was observed that these networks have a strikingly shallow architecture, with most genes regulated by overlapping cascades of length one or two, which is different for developmental regulatory interactions of *Drosophila melanogaster* (Rosenfeld and Alon, 2003). Thus, it was consequently suggested that the response delay, governed by the length of cascades, may act as a design principle of gene regulatory networks.

With respect to the salient network properties, gene regulatory networks, like the majority of social and technological networks, have been shown to belong to the class of scale-free networks (also called *complex* networks) (Albert, 2005). In addition to the scale-free degree distribution, in which the number of highly connected nodes (called hubs) is small, these networks are marked with a large clustering coefficient and small average path length. Recently, it was demonstrated that precisely such combination of network properties allows ultra fast emergence of consensus on the network (where consensus of opinions is achieved only through communication of a network node with its direct neighbours) and synchronization of coupled oscillators robust to edge removal (Olfati-Saber, 2005; Wang and Chen, 2002).

Whereas the dynamical properties of special subnetworks (*e.g.*, cascades and network motifs) are relatively well understood, there is still limited knowledge about the evolutionary reasons for the already identified design principles pertaining to these special subnetworks, underlying the global quantitative features of gene regulatory networks of different organisms. Moreover, despite tremendous progress in the development of models (via simple principles/rules) that mimic the global properties of complex regulatory networks (Albert and Barabasi, 2002), the identified rules cannot explain the differences

between regulatory architectures from different organisms.

Here, we use the adjacency matrix representation of **directed** regulatory networks which allow network comparison via the spectrum, *i.e.*, the set of eigenvalues, of their corresponding matrices. Graph spectra have already been proposed as a systematic tool in computational biology and other fields, and have been applied in the analysis of undirected networks (Banerjee and Jost, 2009). Moreover, it has been observed that a spectrum of a graph in fact encodes information about many structural properties (*e.g.*, average path length, diameter, treewidth) as well as local substructures which may lead to positing hypotheses regarding the evolution of the networks (Chung and Lu, 2006). We demonstrate that the algebraic connectivity (smallest non-trivial eigenvalue) of the **directed graph** representation of gene regulatory networks is a network characteristic which can explain the difference between the regulatory networks of different organisms. Our findings about the algebraic connectivity suggests that information flow in gene networks of simple organisms is generally less spatially restricted that in the networks of more complex organisms. The behavior of the algebraic connectivity, which we show valid on gene regulatory networks from different organisms, is an additional hypothesis for the evolutionary principles of organism-specific regulatory networks giving rise to the variety of life forms.

Methods

A gene regulatory network comprises the transcription factors and two types of genes: (1) genes coding for transcription factors, called *regulatory* genes, and (2) genes regulated by transcription factors, known as *target* genes. In the case of auto–regulation and cascades, a regulatory gene also plays the role of a target gene. We model the structure of regulatory

interactions as a directed bipartite graph G, in which the two partitions, $V_{tf}(G)$ and $V_g(G)$, comprise the transcription factors and the genes, respectively. Such a graph will be called *TF–gene regulatory* network. From a TF–gene regulatory network G, one can obtain a *gene–gene* regulatory network G'. The graph G' contains only those nodes from G which are in $V_g(G)$. The transformation from G to G' is accomplished according to the following rule: A directed edge is drawn from node u to node v in G' if and only if there is a directed path of length two between the corresponding nodes in G. The effects of this transformation on the properties of the resulting network are illustrated and disccused in Supplementary File 1.

Understanding gene expression, as already described, requires analysis of kinetic and structural information pertaining to gene regulatory interactions. In the absence of the former, we model gene expression as a signal propagation process (Anchang *et al.*, 2009; Stetter *et al.*, 2003), by which a regulatory gene attempts to modify the activity of particular target genes. Therefore, we model the signal propagation, itself, as a random walk on a directed graph. Diffusion processes (such as propagation) have already been considered and modeled via diffusion kernels, mostly with focus on undirected graphs (Kondor and Lafferty, 2002; Simonsen *et al.*, 2004).

For a vector x, let $diag(x)$ denote the diagonal matrix, which has the entries of vector x on its diagonal, and all other entries equal zero. Given a directed graph $G = (V, E)$, a random walk on G with adjacency matrix A is a Markov process with transition matrix $P = \mathcal{D}^{-1}A$, where $\mathcal{D} = diag(Ae)$ is a diagonal matrix with entries on the diagonal corresponding to the out-degrees of the nodes in $V(G)$ (e is the vector in which all entries are 1). Assume, for now, that G is strongly connected *i.e.*, for every $u, v \in V(G)$, there exist directed paths, from u and v and from v to u. The Perron-Frobenius theorem for matrix P then states that there exists a unique left eigenvector which is strictly positive

with eigenvalue 1 (since $Pe = e$) (Ninio, 1976; Seneta, 1981). Note that P has a unique normalized left eigenvector with eigenvalue 1 if G is aperiodic (Langville and Meyer, 2004). Let π be the unique normalized left eigenvector such that

$$\pi P = \pi \tag{1}$$

and $\sum_{u \in V(G)} \pi(u) = 1$. Furthermore, the row-vector π corresponds to the stationary distribution of the random walk defined by P. From the Eq. (1), we have that

$$\pi(u) = \sum_{v, v \to u} \pi(v) P(v, u), \tag{2}$$

i.e., the probability of finding the random walk at u is the sum of all incoming probabilities to nodes v which are neighbours of u. We can now define a circulation of the directed graph G.

Definition 0.1. (Chung, 2005) A function $F : E(G) \to R_0^+$ that assigns each directed edge to a non-negative value is called a *circulation* if

$$\sum_{u, u \to v} F(u, v) = \sum_{w, v \to w} F(v, w),$$

for each node v.

The circulation can be easily interpreted as a flow in the graph. The flow at each node must be conserved, hence, the flow in is equal to the flow out. One such circulation is defined in terms of the stationary distribution of the random walk on G (see Eq. (2)) and is given by $F_\pi(u, v) = \pi(u) P(u, v)$. If the random walk is reversible, then $F_\pi(u, v) = F_\pi(v, u)$. We can now examine the matrix:

$$\tilde{A} = \frac{\Pi P + P^T \Pi}{2}, \tag{3}$$

where $\Pi = diag(\pi)$. In contrast to A, the adjacency matrix of the directed graph G, the matrix $\tilde{A}$, is symmetric and, thus, has an undirected graph $\tilde{G}$ corresponding to G. In contrast to G, which is unweighted, the obtained undirected graph $\tilde{G}$ is edge-weighted. The weights of the edges in $\tilde{G}$ correspond to the defined circulation in terms of the random walk on G.

The Laplacian of the graph $\tilde{G}$ is $L(\tilde{G}) = \Pi - \tilde{A}$. To eliminate the influence of the number of nodes and edges, we investigate the normalized directed graph Laplacian, which is defined in terms of $L(\tilde{G})$ as follows:

$$\mathcal{L}(G) = \Pi^{-1/2} \left(\Pi - \tilde{A}\right) \Pi^{1/2} = I - \frac{1}{2}\left(\Pi^{1/2} P \Pi^{-1/2} + \Pi^{-1/2} P^T \Pi^{1/2}\right). \tag{4}$$

Initially, we assumed that the graph G is strongly connected and aperiodic, and therefore that G has a unique stationary distribution. If G is not strongly connected, one can define the Page rank transformation matrix, which is a transformation of the transition matrix P to ensure aperiodicity and strong connectivity (*i.e.*, there is a path between any two nodes) (Bianchini *et al.*, 2005; Langville and Meyer, 2004). The transformation can be formalized as follows:

$$P_{pr} = \alpha P + \frac{(1-\alpha)}{n} e e^T, \tag{5}$$

where $\alpha = 0.85$, to ensure convergence of the random walk. An example of the transformation imposed by Eq. (5) appears in Supplementary File 2.

Although the modifications by Eqs. (3) and (5) alter the directed graph G, when

applied in the same fashion across different directed graphs, they allow for systematic network comparison. One way to compare different directed graphs is through the spectra of their Laplacians. The spectrum of $\mathcal{L}(G)$ is given by n eigenvalues $\lambda_0 = 0 \leq \lambda_1 \leq \ldots \leq \lambda_{n-1}$. If the graph G is strongly connected, then $\lambda_1 > \lambda_0 = 0$ and λ_1 is known as the *algebraic connectivity* of the directed graph G.

The algebraic connectivity λ_1 of a graph G is connected to another graph invariant called Cheeger constant (Chung, 1997). The Cheeger constant is a numerical measure of whether or not a graph has a "bottleneck". To define the Cheeger constant for a directed graph, we need to introduce the concept of out-boundary:

Definition 0.2. Given a graph G and a subset of nodes $S \subset V(G)$, the *out-boundary* of S, denoted by ∂S, is comprised of all edges going from a node in S to a node in $\bar{S}$, $\bar{S} = V(G) - S$, *i.e.*,

$$\partial S = \{(u, v) \in E(G) | u \in S, v \in \bar{S}\}.$$

Definition 0.3. The volume of a node u is defined as $vol(u) = \sum_v P(u, v)$, the volume of set S is defined as $vol(S) = \sum_{u \in S} vol(u)$, and the volume of an out-boundary is defined as $vol(\partial S) = \sum_{u \in S, v \in \bar{S}} P(u, v)$.

The Cheeger constant for a graph G, denoted by $h(G)$, can then be defined in terms of the Cheeger constant of a cut $(S, \bar{S})$ implied by a given node-subset $S \subset V(G)$:

Definition 0.4. The Cheeger constant of a cut $(S, \bar{S})$ in a graph G is

$$h_G(S) = \frac{vol(\partial S)}{\min\{vol(S), vol(\bar{S})\}}.$$

The Cheeger constant of the graph G is

$$h(G) = \min_{S \subset V(G)} h_G(S).$$

The Cheeger constant is strictly positive if and only if G is a connected graph. Intuitively, if the Cheeger constant is small but positive, then there exists a "bottleneck", in the sense that there are two large sets of nodes with few edges between them. The Cheeger constant is large if any possible bi-partition of the node set has many edges between the two partitions. Since computing the Cheeger constant is NP-hard, we rely on its approximation in terms of the algebraic connectivity expressed by the following theorem:

Theorem 0.5. *(Chung, 1997) The algebraic connectivity, λ_1, of the normalized directed Laplacian of the graph G is related to the Cheeger constant $h(G)$ by*

$$\frac{h(G)^2}{2} \leq \lambda_1 \leq 2h(G).$$

With these notions, we can now analyze the spectrum of the directed TF–gene and gene–gene regulatory networks via the generalization of graph Laplacian for directed graphs. As explained, this generalization symmetrizes the graph (*i.e.*, turns it into an undirected graph) without loss of information. From the directed graph Laplacian, we determine the *algebraic connectivity*, and establish connections between this graph invariant for both, TF–gene and gene–gene, regulatory networks and their corresponding average path length (from the distribution of cascades), diameter, and average degree. In addition, we determine the local structures shown to have effect on the dynamics of gene expression, namely the concentrations of all 2– and 3–node directed subgraphs, and then

calculate their Pearson correlation with the algebraic connectivity across the analyzed organisms. The concentration of a subgraph i on n nodes occurring M_i^n was calculated by $c_i^n = \frac{M_i^n}{\sum_i M_i^n}$. Based on the correlation we can rank the importance of a particular property and local structure in the evolution of regulatory architectures. Note that we do not assess the significance of a subnetwork, since we would like to determine how the abundance of each of these subnetworks correlates with the algebraic connectivity.

In addition, we calculate the evolutionary tree from the considered network properties by first determining the distance matrix and then using it to perform hierarchical clustering. We define the distance between two organisms with respect to a network property as the absolute difference between the values of the property for the networks of the two organisms. This is the simplest distance measure which already shows results in good agreement with the established phylogenetic relationships of organisms analyzed here. For distance-based methods for building phylogenetic trees from metabolic networks the reader is directed to (Mazurie *et al.*, 2008) and reference therein.

Data set

We analyze the data set comprising the interactions between regulatory genes, transcription factors, and target genes, extracted from TRANSFAC 7.0 database (Matys *et al.*, 2003), for 97 organisms from all kingdoms of life. TRANSFAC contains data on transcription factors, their experimentally proven binding sites, and regulated genes. In the present study, an edge between a gene and a transcription factor is established if the gene encodes the transcription factor; moreover, an edge is established between a transcription factor and a gene if the transcription factor is experimentally proven to bind to the gene.

In addition to the networks from TRANSFAC we analyzed curated gene regulatory networks for *Escherichia coli*, extracted from Regulon DB and the developmental gene

regulatory networks of Sea urchin's ectoderm and endomesoderm, obtained from (Davidson *et al.*, 2002). In total, our preliminary analysis is conducted on 100 gene regulatory networks. The analysis was performed by using the *igraph* R package.

The assembled TF–gene and gene–gene regulatory networks were first subjected to preprocessing in order to identify the weakly connected components (WCCs) and their corresponding distributions across all organism-specific networks. From the 100 analyzed TF–gene regulatory networks, 17 have a largest weakly connected component which contains more than 40% of the nodes from the originally assembled network. The other connected components do not contain more than 4 nodes, and were excluded from further consideration to reduce the bias due to different number of small WCCs. From the 17 identified networks, only 9 contain more than 80 nodes (see Supplementary File 3). The remaining 8 networks contain small number of nodes, between 4 and 20, and were also excluded from further analysis. Finally, our analysis was performed on the largest WCCs of 8 networks from the following 7 organisms: *Escherichia coli*, *Saccharomyces cerevisiae*, *Sea urchin* (two networks—ectoderm and endomesoderm), *Drosophila melanogaster*, *Mus musculus*, *Rattus norvegicus*, and *Homo sapiens*.

In the following, we present the results for the properties of the largest weakly connected component of the 8 organism-specific TF–gene and gene–gene regulatory networks. The biological interpretations of the obtained results ultimately depend on the quality and scope of the considered data. Therefore, the application of the proposed method on these and other data sets warrants revisiting the analysis in a future study when the quality of known gene-regulatory networks may be considerably improved.

Results and Discussion

Network properties of TF–gene and gene–gene regulatory networks

The average degree of the TF–gene regulatory networks was found to be in the range from 1.235 for *Saccharomyces cerevisiae* to 4.821 for *Sea urchin* ectoderm (see Table 1). Since for any directed graph, the average in- and out-degrees are the same, we report only one value. There is an apparent difference in the average degree between TF–gene networks over the entire genome and regulatory networks involved in developmental processes. If the average degree in both types of networks played a role in the evolution of these regulatory networks, one would not have expected the similar behavior of this network property on networks from organisms of different kingdoms, such as: *Saccharomyces cerevisiae* (1.235) and *Mus musculus* (1.370) or *Escherichia coli* (2.122) and *Sea urchin* endomesoderm (3.588) with closest values to each other. The difference in the average degree across organisms becomes more pronounced for the gene–gene regulatory networks due to the densification implied by the transformation from bipartite TF–gene to gene–gene networks. For the gene–gene regulatory networks, the average degree ranges from 1.322 for *Saccharomyces cerevisiae* to 6.836 for *Sea urchin* ectoderm, as shown in Table 2. However, like for the TF–gene networks, the evolutionary relationships cannot be recovered from this network property, as shown in Fig. 1 (A).

Previous analyses have demonstrated differences in the distribution of cascades between *Escherichia coli* and *Drosophila melanogaster*, and speculated the biological implications of this finding (Rosenfeld and Alon, 2003). Namely, Rosenfeld and Alon (2003) found that the longest cascade in the network of *Escherichia coli* is of length 4, and that the majority of cascades are of unit length; in the network of *Drosophila melanogaster*,

the largest cascade is of length 9 and the majority of cascades are of length between 6 and 9. Our results for the distribution of cascade lengths in TF–gene regulatory networks, analyzed here, are shown in Fig. 2. The shape of the distributions for *Saccharomyces cerevisiae* and *Rattus norvegicus* point out at their shallow regulatory architectures, with predominantly small length cascades. The longest cascades are found in *Homo sapiens* (17), *Sea urchin* endomesoderm (14), and *Escherichia coli* (12). Moreover, the shape of the distribution of cascade lengths in *Escherichia coli* (red) and *Drosophila melanogaster* (blue) are very similar, as shown in Fig. 2, which is contrary to the previously reported results. Furthermore, the average path length is almost identical for *Escherichia coli* (3.804) and *Drosophila melanogaster* (3.884). The largest average path lengths are found in *Homo sapiens* (5.249) and *Sea urchin* endomesoderm (3.973), followed by *Drosophila melanogaster* and *Escherichia coli*. Finally, the shape of the cascade length distribution for *Escherichia coli* closely follows those of *Sea urchin*'s networks. These findings point out that the distribution of cascade lengths may not play a significant role in shaping the differences between the investigated organisms.

The results for gene–gene regulatory networks differ by a factor of 2, due to the transformation from TF–gene to gene–gene networks, which turns every directed path of length 2 into a directed edge, as shown in Table 2. From Fig. 1 (B) and (C), we can conclude that neither the length of the longest cascade nor the average cascade length can be used to recover the evolutionary relationship between the investigated organisms. A typical result of discordance is the clustering of *Escherichia coli* and *Homo sapiens*, in the case of the longest cascade, and *Homo sapiens* and *Sea urchin*, in the case of the average path length.

Connected subnetworks in gene–gene regulatory networks

The distribution of connected subnetworks on n nodes can be regarded as a feature of a gene–gene regulatory network. As the number of nonisomorphic directed connected graphs on n nodes is a fast growing function (Harary and Palmer, 1973), we determined the concentrations for each of 2- and 3-connected subnetworks. As in (Milo *et al.*, 2002), a subnetwork was considered to be *significant* if it appears at least 4 times in an analyzed network. Note that we are not interested in establishing the significance of the concentration of a particular subnetwork, but only on the possible correlation between the subgraph concentration and other structural properties. The results for the distribution of 3-connected subnetworks appear in Fig. 3.

The two 2-connected subnetworks, *i.e.*, a directed edge and a pair of directed edges in the investigated networks, cannot discriminate the complexity of the organisms and behave similarly to the average degree. The significant subnetworks across all organisms include the following: M1, M2, M4, and M5 (see Fig. 3). The subnetworks M3, M4, M6, and M11 are significant for all organisms except *Saccharomyces cerevisiae* and *Rattus norvegicus*, while subnetworks M8, M9, M10, M12, and M13 are not significant in all networks except the two networks of *Sea urchin*. While the last five enumerated subnetworks do separate the developmental regulatory networks from those on the entire genome, we still face the same problem as in the case of the distribution of cascade lengths: Namely, the distributions of subnetwork concentrations of *Saccharomyces cerevisiae* and *Homo sapiens* are closest to each other and the distribution of *Drosophila melanogaster* is closest to those of *Sea urchin*'s networks.

Algebraic connectivity of gene–gene regulatory networks

To account for the various endogenous and exogenous conditions which may affect the expression of genes in a given organism, we model gene expression as signal propagation, *i.e.*, a random walk by which a regulatory gene can affect a target gene. The convergence of this random walk to its stationary distribution is characterized by the algebraic connectivity of the directed Laplacian of the graph on which the random walk takes place. The algebraic connectivity is related to the Cheeger constant (see Methods) which characterizes the existence of bottlenecks in the underlying network structure. The bottlenecks can be seen as the means for throttling or slowing down signal propagation on the graph.

In graphs with large Cheeger constant, signals can easily propagate between any two nodes, since any two node partitions are connected with a large number of edges. Such a property is desirable for gene regulatory networks which must settle in a stationary state in a shorter time, rendering, ultimately, shorter response times on a system's level. On the other hand, if a subset of regulatory interactions is condition-specific, its connection to the rest of the network should be limited, and only a portion of the network is affected (*i.e.*, the random walk, on average, remains localized in one subnetwork).

The seemingly absent link between algebraic connectivity of TF–gene networks and evolutionary relatedness, which follows from Table 1, is due to the fact that TF–gene networks are bipartite . As stated in the seminal paper of Lovasz (1996), the Markov chain P does not converge to a stationary distribution. Moreover, the PageRank algorithm used in the derivation of the algebraic connectivity has the same degeneracy on bipartite graphs, yielding different ranking for nodes of equal indegree (see Theorem 4 (Meghabghab and Kandel, 2008)). This drawback of using TF-gene networks is overcome precisely by the transformation of the bipartite TF-gene graphs into non-bipartite gene-gene graphs.

The algebraic connectivity of the gene–gene regulatory networks demonstrated an in-

teresting property—the more complex the organism, the smaller the algebraic connectivity (see Table 2). Therefore, organisms whose gene regulatory interactions should react with higher specificity to different stimuli depending on the tissue/conditions, exhibit smaller algebraic connectivity, effectively rendering localization of the random walk (*e.g.*, the networks of *Drosophila melanogaster*, *Rattus norvegicus*, *Mus musculus*, and *Homo sapiens* have algebraic connectivity of 0.009, 0.007, 0.009, and 0.007, respectively). Moreover, the developmental networks of *Sea urchin* have high algebraic connectivity (0.109 for ectoderm and 0.010 for endomesoderm), comparable to that of *Escherichia coli* (0.08) and *Saccharomyces cerevisiae* (0.02). This implies that the gene regulatory interactions are structured to facilitate faster convergence of gene expression to a stationary state for organisms and processes that should adapt faster to exogenous conditions. Furthermore, as shown in Fig. 1 (D), the value of the algebraic connectivity does reflect the evolutionary relationship between the investigated organisms.

The Pearson correlation between the average degree and the algebraic connectivity over the investigated organisms is 0.640, which implies that the average degree does have influence on the algebraic connectivity. Recent mathematical result points at the rules for building a graph model for which the algebraic connectivity can be made arbitrarily close regardless of the given (prescribed) degree distribution (Atay *et al.*, 2006). Therefore, our finding suggests that, although the degree distribution of the regulatory networks from different organisms may be scale-free (even with equal exponents, Fig. 4 (A) and (B)), the edges in the graph are distributed in a way that the increased complexity is matched with smaller algebraic connectivity.

On the other hand, the correlation between the longest cascade and the algebraic connectivity over the investigated organisms is -0.19, which shows that the longest cascade has only a negligible reciprocal effect on the algebraic connectivity. The same can be

observed for the correlation between the average path length and algebraic connectivity. Recent theoretical results demonstrated that synchronizability of the complex networks worsens with the increase of the average path length (Zhao *et al.*, 2006). Since the Laplacian matrix arises naturally in the study of models of synchronization (Watts, 2003), our empirical results point out that gene regulatory networks form a class of complex networks whose synchronizability may not be affected by the average path length, but is tightly coupled with the algebraic connectivity.

Next, we determine to what extent the local structures determine the existence of the bottlenecks which may have effects on gene expression. To this end, we calculated the correlation between the 3-connected subnetworks (see Fig. 3) and algebraic connectivity across the investigated organisms. The algebraic connectivity exhibits the largest correlation of 0.934 with the concentration of subnetworks M6, followed by M5 (0.886), M8 (0.769), M10 (0.766), M9 and M12 (0.766). From these ranking of subnetworks and the previous results about their significance, we can conclude that subnetworks M5 and M6 have the highest influence on the value of the algebraic connectivity.

Although, in gene–gene networks, the concentrations of subgraphs M5 and M6, see Table 2, show hight correlation with the algebraic connectivity, only the latter reflects the evolutionary relatedness between the considered organisms (see Supplementary File 4). The reason for the discrepancy in the evolutionary relatedness according to the two properties is due to the differences in the type of structural information employed for calculating them. For instance, the concentration of a particular 3-subnetwork is calculated based on a local property of how 3 nodes are connected to each other. Therefore, it can only capture a limited portion of the network topology. Moreover, the concentration of a subgraph states neither how the copies of the subnetwork are distributed across the network, nor how they are positioned in conjunction to the remaining types of sunet-

works. Hence, deriving principles of network evolution on a local property may represent an ill-founded approach, as already illsutrated.

On the other hand, algebraic connectivity is a global property, whose calculation depends on global information captured in the matrices involved in its calculation. It is likely that the algebraic connectivity, as a global property, reflects the information about not only the concentration of a particular subnetwork, but also of how the copies of the subnetwork are distributed and positioned in the network. Making a formal statement and empirical investigation about the relation of the algebraic connectivity and the concentration of subnetworks in complex networks is far beyond the scope of this paper, and remains as an interesting open problem.

A closer look at the subnetworks M5 and M6, identified to have the highest rank with respect to their correlation with the algebraic connectivity, we conclude that M5 is precisely the feed forward loop (FFL) and M6 contains the FFL as an embedded subnetwork. The feed forward loop is determined by three genes, say u, v, and w, such that u regulates v, while both u and v regulate the gene w. The FFL has been demonstrated theoretically and experimentally to perform a basic information–processing function, *i.e.*, it shows a delay following ON steps of an input inducer, but not after OFF steps. Moreover, it was recently demonstrated that the FFL is selected over simpler subnetworks in environments where the distribution of the input pulse duration is sufficiently broad and contains both long and short pulses (Dekel *et al.*, 2005; Mangan and Alon, 2003). In addition, recent simulation studies have demonstrated that in directed graphs, synchronization (directly expressed via the algebraic connectivity) can be correlated with the abundance of FFL as a specific 3-node subnetworks (Brede, 2008; Lodato *et al.*, 2007).

With regard to the ranking of subnetworks, our result on the positive correlation between the concentration of subnetworks M5 and the algebraic connectivity implies that

the smaller algebraic connectivity requires smaller concentration of strategically embedded subnetworks isomorphic with the FFL. Therefore, in addition to the identified function of single FFL to detect non-temporary change of the environment, our results suggest that evolution favours decreasing concentration of FFLs combined in such a way that the network, as a whole, can throttle a random walk on the network and, thus, increase the specificity towards changing environments.

Our results point out that the algebraic connectivity is a salient property of gene-gene regulatory networks and encodes other network properties (*e.g.*, average degree and subnetwork concentration) governing the network evolution. Although the FFL subnetwork and scale-free degree distributions have already been identified as salient properties of gene regulatory networks, it was not known what is the evolutionary principle which select these over other network properties. Moreover, thus far, it was not known why the FFLs appear as significant in virtually all gene regulatory networks as well as their relations with regulatory networks coding for different organismal complexity. Here, we identified the decrease in the algebraic connectivity of the directed regulatory network as one possible evolutionary principle.

Although at present we cannot exhibit a model to mimic these findings, our results point out at two necessary rules for developing gene regulatory networks which encode growing complexity: (1) the degree distribution has a heavy-tail with average degree positively correlated with the decreasing algebraic connectivity, and (2) the distribution of edges is such that the concentration of the FFL subnetwork is positively correlated with the decreasing algebraic connectivity. Initial steps for developing the first rule on undirected graphs have already been made in (Atay *et al.*, 2006). The second rule can be addressed by developing a model of directed networks which considers the effect of gene duplication events and subsequent sub- and neo-functionalization of genes on the

algebraic connectivity.

Acknowledgments

Z.N., P.M., and J.S. would like to acknowledge the support from the GoFORSYS project funded by the German Federal Ministry of Education and Research, Grant Nr. 0313924.

References

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, **118**(Pt 21), 4947–4957.

Albert, R. and Barabasi, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, **74**.

Albert, R. and Othmer, H. G. (2003). The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster. *Journal of Theoretical Biology*, **223**, 1–18.

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, **8**(6), 450–461.

Anchang, B., Sadeh, M. J., Jacob, J., Tresch, A., Vlad, M. O., Oefner, P. J., and Spang, R. (2009). Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models. *Proceedings of the National Academy of Sciences of the USA*, **106**(16), 6447–6452.

Atay, F. M., Biyikoglu, T., and Jost, J. (2006). Synchronization of networks with pre-

scribed degree distributions. *IEEE Transactions on Circuits and Systems I: Regular Papers*, **53**(1), 92–98.

Banerjee, A. and Jost, J. (2009). Graph spectra as a systematic tool in computational biology. *Discrete Applied Mathematics*, **157**(10), 2425–2431.

Bianchini, M., Gori, M., and Scarselli, F. (2005). Inside PageRank. *ACM Transactions on Internet Technology*, **5**(1), 92–128.

Brede, M. (2008). Synchronization on directed small wolrds: Feed forward loops and cycles. *EPL*, **84**, 40004.

Chung, F. (2005). Laplacians and the cheeger inequality for directed graphs. *Annals of Combinatorics*, **9**(1), 1–19.

Chung, F. and Lu, L. (2006). *Complex Graphs and Networks*. Number 107 in CBMS Regional Conference Series in Mathematics. American Mathematical Society.

Chung, F. R. K. (1997). *Spectral Graph Theory*. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Society.

Conradi, C., Flockerzi, D., Raisch, J., and Stelling, J. (2007). Subnetwork analysis reveals dynamic features of complex (bio)chemical networks. *Proceedings of the National Academy of Sciences of the USA*, pages 19175–19180.

Craciun, G., Tang, Y., and Feinberg, M. (2006). Understanding bistability in complex enzyme-driven reaction networks. *Proceedings of the National Academy of Sciences of the USA*, **103**(23), 8697–8702.

Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C. H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C. T., Livi, C. B.,

Lee, P. Y., Revilla, R., Rust, A. G., Pan, Z., Schilstra, M. J., Clarke, P. J., Arnone, M. I., Rowen, L., Cameron, R. A., McClay, D. R., Hood, L., and Bolouri, H. (2002). A genomic regulatory network for development. *Science*, **295**(5560), 1669–1678.

Dekel, E., Mangan, S., and Alon, U. (2005). Environmental selection of the feed-forward loop circuit in gene-regulation networks. *Physical biology*, **2**(2), 81.

Elowitz, M. B. and Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, **403**(6767), 335–338.

Feinberg, M. (1987). Chemical reaction network structure and the stability of complex isothermal reactors—i. the deficiency zero and deficiency one theorems. *Chemical Engineering Science*, **42**(10), 2229–2268.

Harary, F. and Palmer, E. M. (1973). *Graphical Enumeration*. Academic Press.

Kondor, R. I. and Lafferty, J. (2002). Diffusion kernels on graphs and other discreet structures. *Proceedings of the International Conference on Machine Learning*, pages 315–322.

Langville, A. N. and Meyer, C. D. (2004). Deeper inside PageRank. *Internet Mathematics*, **1**(3), 335–400.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, **298**(5594), 799–804.

Levine, M. and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, **424**(6945), 147–151.

Lodato, I., Boccaletti, S., and Latora, V. (2007). Synchronization properties of network motifs. *EPL*, **78**, 28001.

Lovasz, L. (1996). Random walks on graphs. *Combinatorics*, **2**(1), 1–46.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**(7006), 308–312.

Madan Babu, M., Teichmann, S. A., and Aravind, L. (2006). Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *Journal of Molecular Biology*, **358**(2), 614–633.

Mangan, S. and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences of the USA*, **100**(21), 11980–11985.

Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D. U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003). Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, **31**(1), 374–378.

Mazurie, A., Bonchev, D., Schwikowski, B., and Buck, G. A. (2008). Phylogenetic distances are encoded in networks of interacting pathways. *Bioinformatics*, **24**(22), 2579–2585.

Meghabghab, G. and Kandel, A. (2008). *Search Engines, Link Analysis, and Users Web Behavior: A Unifying Web Mining Approach*. Springer, Berlin.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science*, **298**(5594), 824–827.

Ninio, F. (1976). A simple proof of the Perron-Frobenius theorem for positive symmetric matrices. *Journal of Physics A: Mathematical and General*, **9**(8), 1281–1282.

Olfati-Saber, R. (2005). Ultrafast consensus in small-world networks. *American Control Conference, 2005*, pages 2371–2378.

Papp, B. and Oliver, S. (2005). Genome-wide analysis of context-dependence of regulatory networks. *Genome Biology*, **6**(2), 206.

Ronen, M., Rosenberg, R., Shraiman, B. I., and Alon, U. (2002). Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *PNAS*, **99**(16), 10555–10560.

Rosenfeld, N. and Alon, U. (2003). Response delays and the structure of transcription networks. *Journal of Molecular Biology*, **329**(4), 645–654.

Seneta, E. (1981). *Non-negative matrices and Markov chains*. Springer Series in Statistics. Springer, New York.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, **31**(1), 64–68.

Simonsen, I., Eriksen, K. A., Maslov, S., and Sneppen, K. (2004). Diffusion on complex networks : A way to probe their large scale topological structures. *Physica A: Statistical and Theoretical Physics*, **336**(1–2), 163–173.

Stetter, M., Deco, G., and Dejori, M. (2003). Large-scale computational modeling of genetic regulatory networks. *Artificial Intelligence Review*, **20**(1–2), 75–93.

Wang, X. F. and Chen, G. (2002). Synchronization in scale-free dynamical networks: robustness and fragility. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, **49**(1), 54–62.

Watts, D. J. (2003). *Six degrees: The science of a connected age*. WW Norton & Company.

Zhang, Z., Liu, C., Skogerb, G., Zhu, X., Lu, H., Chen, L., Shi, B., Zhang, Y., Wang, J., Wu, T., and Chen, R. (2006). Dynamic changes in subgraph preference profiles of crucial transcription factors. *PLoS Computational Biology*, **2**(5), e47+.

Zhao, M., Zhou, T., Wang, B.-H., Yan, G., Yang, H.-J., and Bai, W.-J. (2006). Relations between average distance, heterogeneity and network synchronizability. *Physica A: Statistical and Theoretical Physics*, **371**(2), 773–780.

Figure Legends

Figure 1. Evolutionary relationship from network properties of gene–gene networks. Heatmaps together with the corresponding hierarchical clustering based on: (A) average degree, (B) longest cascade, (C) average cascade length, and (D) algebraic connectivity for *Escherichia coli* (Ec), *Saccharomyces cerevisiae* (Sc), *Sea urchin* ectoderm (SUect), *Sea urchin* endomesoderm (SUend), *Drosophila melanogaster* (Dm), *Mus musculus* (Mm), *Rattus norvegicus* (Rn), and *Homo sapiens* (Hs).

Figure 2. Distribution of cascade lengths in TF–gene networks across organisms. The distribution for *Saccharomyces cerevisiae* (green line) confirms its shallow architecture, with *Rattus norvegicus* being its closest neighbour. Moreover, the distribution of cascades between *Escherichia coli* and *Drosophila melanogaster* differ only slightly, in contradiction to already reported results.

Figure 3. Distribution of 3–connected subnetworks in gene–gene networks of studied organisms. The thirteen 3–connected directed subnetworks are shown on the right. Subnetwork M5 is the feed forward loop. Two groups of organisms can be observed based on the correlation between the distributions of 3–connected subnetworks: the first is given by *Drosophila melanogaster* and Sea urchin, the second by *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, and *Saccharomyces cerevisiae*. The network of *Escherichia coli* stands on its own, due to the large concentration of the subnetwork M1.

Figure 4. Degree distributions. Distribution of the (A) in-degree and (B) out-degree of the analyzed networks. All networks, except for those of *Sea urchin*, follow the same trend in the degree distribution for both in– and out-degrees.

Tables

Table 1. Network properties for TF–gene networks

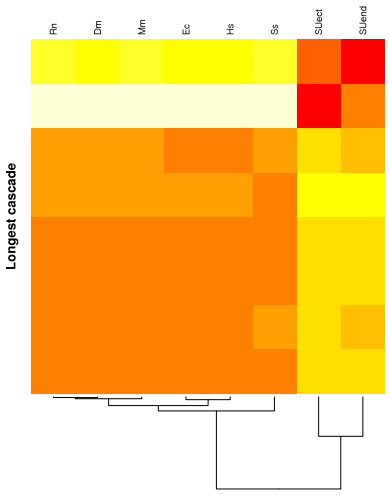
	network properties					
organism	n	m	λ_1	D	D_{avg}	d_{avg}
Escherichia coli	1578	3349	0.023	12	3.804	2.122
Saccharomyces cerevisiae	531	656	0.007	6	1.784	1.235
Sea urchin ectoderm	84	405	0.011	9	3.409	4.821
Sea urchin endomesoderm	119	427	0.032	14	3.973	3.588
Drosophila melanogaster	343	485	0.009	10	3.884	1.413
Mus musculus	1652	2264	0.015	10	3.251	1.370
Rattus norvegicus	700	1019	0.023	8	2.343	1.455
Homo sapiens	2979	4686	0.009	17	5.249	1.573

Network properties of TF–gene regulatory networks. Six network properties: number of nodes (n), number of directed edges (m), algebraic connectivity (λ_1), longest cascade (D), average cascade (D_{avg}), and average degree (d_{avg}) for TF–gene regulatory network of *Escherichia coli*, *Saccharomyces cerevisiae*, *Sea urchin*'s ectoderm and endomesoderm, *Drosophila melanogaster*, *Mus musculus*, *Rattus norvegicus*, and *Homo sapiens*.

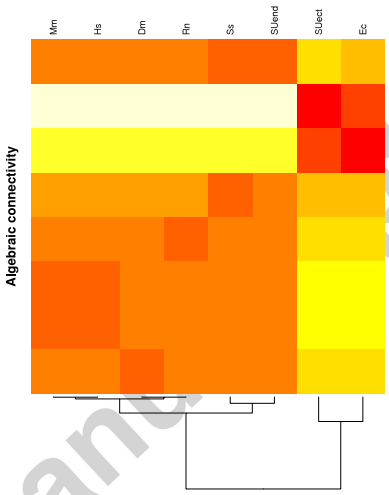
Table 2. Network properties of gene–gene networks

	network properties							
organism	n	m	λ_1	D	D_{avg}	d_{avg}	c_{M_5}	c_{M_6}
Escherichia coli	1434	3206	0.085	6	2.142	2.235	0.00401	0.00084
Saccharomyces cerevisiae	387	512	0.024	3	1.154	1.322	0.00367	0.00000
Sea urchin ectoderm	55	376	0.109	4	1.902	6.836	0.10238	0.03171
Sea urchin endomesoderm	79	387	0.018	7	2.170	4.898	0.08696	0.01125
Drosophila melanogaster	204	376	0.008	5	2.163	1.843	0.02984	0.00421
Mus musculus	831	1455	0.007	5	1.834	1.750	0.00627	0.00067
Rattus norvegicus	392	711	0.009	4	1.468	1.813	0.01107	0.00000
Homo sapiens	1596	3354	0.007	8	2.816	2.101	0.00590	0.00122

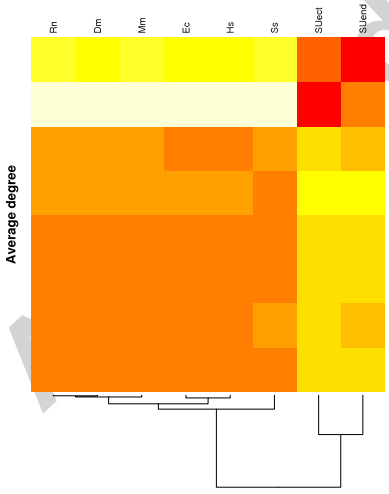
Network properties of gene–gene regulatory networks. Six network properties: number of nodes (n), number of directed edges (m), algebraic connectivity (λ_1), longest cascade (D), average cascade (D_{avg}), average degree (d_{avg}), concentration of M5 subnetwork (c_{M_5}), and concentration of M6 subnetwork (c_{M_6}) for gene–gene regulatory network of *Escherichia coli*, *Saccharomyces cerevisiae*, *Sea urchin*'s ectoderm and endomesoderm, *Drosophila melanogaster*, *Mus musculus*, *Rattus norvegicus*, and *Homo sapiens*.



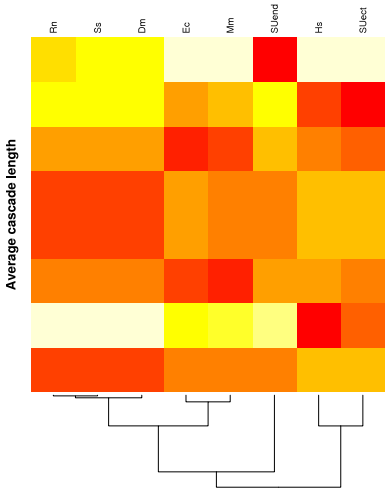
(B)



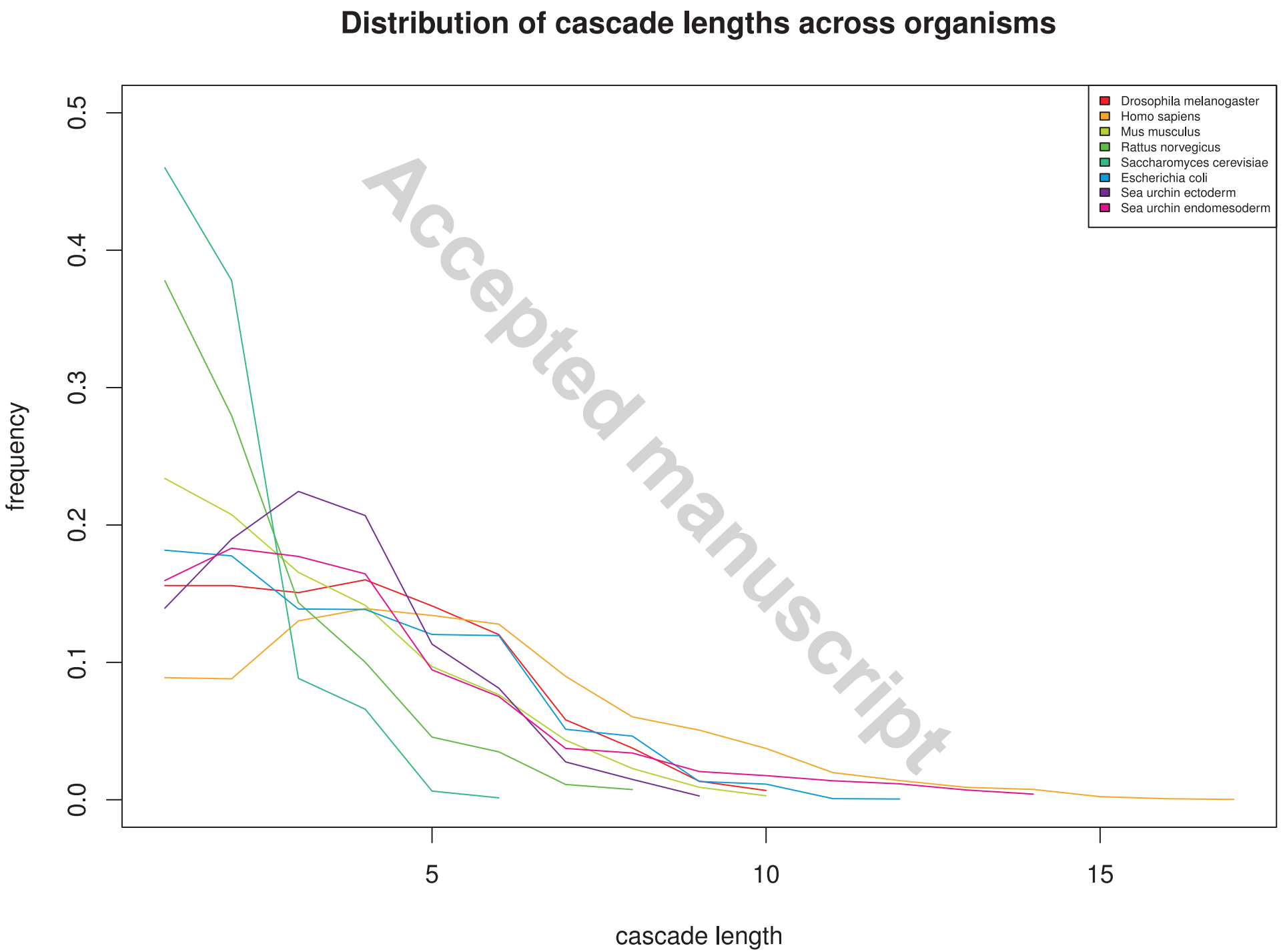
(D)

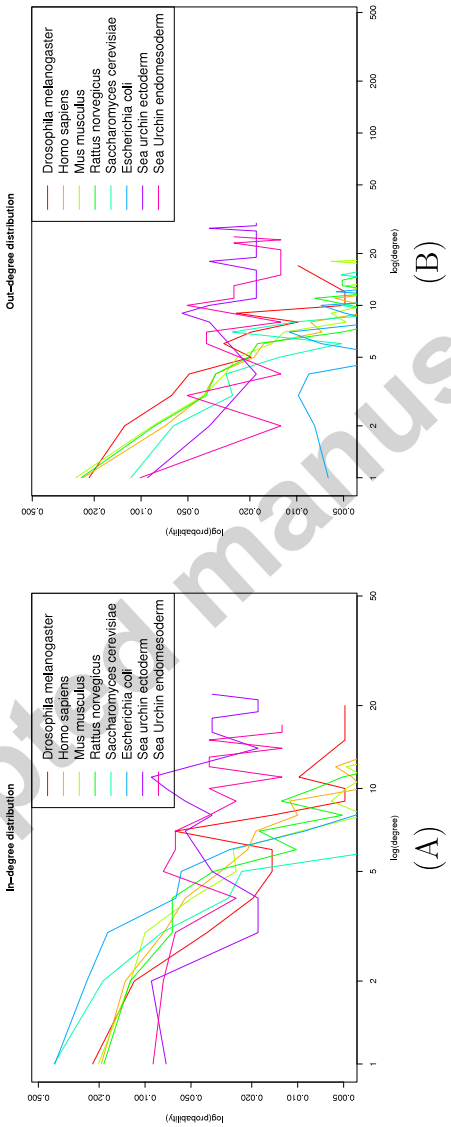


(A)



(C)





(A) (B)