OPEN ACCESS Freely available online

PLoS one

Failure Tolerance of Motif Structure in Biological Networks

Baharan Mirzasoleiman, Mahdi Jalili*

Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

Abstract

Complex networks serve as generic models for many biological systems that have been shown to share a number of common structural properties such as power-law degree distribution and small-worldness. Real-world networks are composed of building blocks called motifs that are indeed specific subgraphs of (usually) small number of nodes. Network motifs are important in the functionality of complex networks, and the role of some motifs such as feed-forward loop in many biological networks has been heavily studied. On the other hand, many biological networks have shown some degrees of robustness in terms of their efficiency and connectedness against failures in their components. In this paper we investigated how random and systematic failures in the edges of biological networks influenced their motif structure. We considered two biological networks, namely, protein structure network and human brain functional network. Furthermore, we considered random failures as well as systematic failures based on different strategies for choosing candidate edges for removal. Failure in the edges tipping to high degree nodes had the most destructive role in the motif structure of the networks by decreasing their significance level, while removing edges that were connected to nodes with high values of betweenness centrality had the least effect on the significance profiles. In some cases, the latter caused increase in the significance levels of the motifs.

Citation: Mirzasoleiman B, Jalili M (2011) Failure Tolerance of Motif Structure in Biological Networks. PLoS ONE 6(5): e20512. doi:10.1371/journal.pone.0020512

Editor: Matjaz Perc, University of Maribor, Slovenia

Received April 9, 2011; **Accepted** April 28, 2011; **Published** May 26, 2011

Funding: This work was funded by Sharif University of Technology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mjalili@sharif.edu

Introduction

Many real-world complex systems can be described as networks. Examples include the Internet, World Wide Web, the brain functional/anatomical networks, genetic regulatory networks, metabolism of biological species, ecological systems, and networks of author collaborations [1,2,3]. Scholars have found that many real-world networks from physics to biology, engineering and sociology have some common structural properties such as power-law degree distribution [4] and small-worldness [5]. Studying the properties of such networks could shed light on understanding the underlying phenomena or developing new insights into the system. For example, studying biological networks helps us to better understand the organization and evolution of their units [6]. Recent developments in computing facilities let researchers mine the data of real-world networks to discover their topological properties.

In its simplest form, a network consists of a set of discrete elements called nodes (or vertices), and a set of connections linking these elements called edges (or links). One of the tricky parts of research in this field is to extract the graph of system under study that is to identify the individual nodes and reconstruct the links connecting them. As network structure is identified, its structural and dynamical properties are investigated. Network motifs are among such attributes that are usually tested for natural networks. It has been shown that networks in various fields exhibit interesting features in terms of repeated occurrences of certain subgraphs, i.e. motifs [7,8]. Network motifs are patterns (particular subgraphs) that statistically overrepresented or underrepresented within the network. The significance of a particular subgraph in a network is usually measured by comparing its occurrences in the original network against some properly randomized networks. Network motifs have been identified in networks from different branches of science and are suggested to be the basic building blocks of most complex networks [9]. Analysis of this over/under abundant substructures can help us in determining different network properties and functions such as its hierarchal structure. The motif structure of a network might be important in determining its dynamical properties. For example, the evolution of cooperativity [10,11], has been linked to the motif structure in real networks [12].

One of the important features of many engineering and biological networks is robustness against component failure [13,14]. Real-world networks may undergo random or systematic failures and consequently lose some of their components, i.e. nodes and/or edges. Therefore, it is essential to investigate the tolerance of critical network properties to errors– failures of randomly chosen nodes and/or edges of the networks and attacks– systematic failures of components that play a critical role in the network [15,16]. It has been shown that many biological networks exhibit high degrees of robustness against random errors that might happen in their structure [13,14,15,17,18]. In general, it has been shown that scale-free networks, i.e. networks whose node-degree distribution follows a power-law, are robust against errors, but, at the same time, they are fragile in response to systematic attacks [15,19,20,21]. Several measures have been proposed for

measuring robustness of networks against attacks and errors. One of the frequently used ones is the largest connected component whose size scales linearly with the number of nodes in the network [15,20,22]. Efficiency is another important measure that is studied in the context of robustness of complex networks against attacks/errors [19]. The errors/attacks influence the evolution of dynamical processes happening on the networks. Network cooperativity, for instance, has been shown to be extremely robust against random failures, while it is fragile when nodes with maximum degree are removed from the network [23].

In this paper we investigated the influence of link failures in the profile of network motifs. We considered protein structure network [8] and functional network of human brain extracted through functional magnetic resonance imaging technique [24]. A number of strategies for choosing candidates edge for removal were taken into account that included random removal, removing edges based on the degrees of the end nodes, based on the betweenness centrality of the nodes, and based on the closeness centrality of the nodes. We then compared the profile of the network motifs as a function of the percentage of removed edges. Interestingly, different failure strategies resulted in different pattern of changes in the motif structure where the strategy based on the betweenness centrality was the most different with the other three.

Materials and Methods

Motif Structure

Many real-world complex networks have been shown to be composed of well-defined building blocks called *motifs*. Network motifs are patterns of interconnection or subgraphs that occur in natural networks much more frequent than those in randomized networks [7,8]. They can be thought of as simple building blocks of complex networks [8], which can provide valuable information about structural design principles of networks. First discovered in the gene regulation (transcription) network of the bacteria *Escherichia coli* by Alon and his team [8,25], they have been found in many networks ranging from biochemistry to neurobiology networks, ecology, and engineering [9,26,27]. Study of network motifs is therefore propitious for revealing the basic building blocks of most complex networks.

Some studies have related the function of networks to the structure of their motifs. Transcription networks are among those heavily studied both theoretically and experimentally. For example, negative-autoregulation which is one of the simplest and most abundant motifs in *Escherichia coli* has been shown to be response-acceleration and repair system [28]. Positive-autoregulation motif is important in biomodal distribution of protein levels in cell population [29]. Feed-forward loop that is commonly found in many gene systems and organisms is important in speeding up the response time of the target gene expression following stimulus steps, pulse generation and cooperativity [30]. Dense Overlapping Regulons that occur when several regulators combinatorially control a set of genes with diverse regulatory combinations, has also been shown to be important in the function of *Escherichia coli* [31].

Although subgraphs of different sizes can be studied in natural networks, among them, biological networks contain three and four-node substructures far more often compared to randomized networks with similar structural properties. Many beneficial outcomes have been ensued from these observations. Often the network motifs are detected by comparing the network against a null hypothesis, that is, the number of appearance of a specific subgraph is counted in the networks and is subsequently compared with the number of appearances in properly randomized networks. The randomized networks can be constructed in various ways. However, they should at least share some common properties with the original network. For example, the randomized networks should have the same number of nodes and edges with the original network. One possible method is to build the corresponding Erdos-Renyi version for the networks [32]. A better way of constructing the randomized networks is to preserve not only their size and average degree but also their degree distribution or at least degree sequence. This can be simply done by shuffling the adjacency matrix [33]. Many of the motif detection strategies use this algorithm for constructing the randomized version of the original network under study. The motif detection algorithm can be summarized as follows [7,8]:

1) Consider a specific subgraph i
2) Count the number of appearances of the subgraph i in the network N_i
3) Generate sufficiently large number of randomized networks with the same number of nodes and degree distribution as the original network
4) Count the number of appearances of the subgraph i in each of the randomized networks
5) Compute the average number of appearances of the subgraph i in the randomized networks $\langle N\mathrm{rand}_i\rangle$ and its standard deviation $\mathrm{std}(N\mathrm{rand}_i)$
6) Compute the significance of appearances of the subgraph i as

$$Z_i = \frac{N_i - \langle N\mathrm{rand}_i\rangle}{\mathrm{std}(N\mathrm{rand}_i)}. \quad (1)$$

7) The networks motifs are subgraphs for which the probability P of appearing in the randomized networks an equal or greater number of times than in the original network is lower than a cutoff value (e.g. $P<0.01$). Thus, higher absolute values of Z-scores correspond to more significant network motifs.

Note that the Z-score of a motif can be positive or negative; positive when it is highly overrepresented in the original network as compared to randomized ones and negative when it is highly underrepresented.

It has also been proposed to normalize the Z-scores [7]. The Z-score of an specific motif may depend on the network size and it tends to be higher in larger networks [7]. Since complex networks may vary widely in size, one can take an approach that enables to compare different network's local structure. To this end, the normalized Z-scores can be calculated as

$$Z_i = \frac{Z_i}{\sqrt{\sum_i Z_i^2}}. \quad (2)$$

The normalization emphasizes the relative significance of subgraphs rather than the absolute significance, which is important for comparison of subgraph of different sizes [7].

A motif of size k is called a k-motif. The runtime of counting process grows very fast with k. This is one of the reasons why only small k-motifs (usually three- or four-nodes) have been studied in most of the works. Different tools have been developed for the detection and analysis of network motifs such as Mfinder [34],

MAVisto [35], and FANMOD [36]. In this work we used Mfinder, which uses a semi-dynamic programming algorithm in order to reduce the running time [34]. It also uses an efficient sampling algorithm that significantly reduces the running time compared to the cases where all edges are visited.

Two Biological Networks

Techniques from complex networks have been widely applied to many biological systems (e.g. see reviews [6,13,37]). Recent developments in designing efficient techniques in molecular biology have led to extraordinary amount of data on key cellular networks in a variety of simple organisms [8,38,39,40,41]. This allowed scholars to study networks such as protein interaction, transcriptional regulatory, and metabolic in different organisms. Networks have also been widely studied in neurosciences [42,43,44]. The brain networks can be studied on a micro-scale containing a number of neurons with some excitatory/inhibitory connections in-between [45,46,47]. However, this approach cannot be used for studying the whole-brain connectivity network. For such cases, one should use functional magnetic resonance imaging, diffusion imaging, magnetocephalography, or electroencephalography techniques to extract the large-scale functional/anatomical brain connectivity networks [48,49,50,51].

In this work, we have considered two biological networks: protein structure network [7], and human brain functional network extracted through functional magnetic resonance imaging [24]. Figure 1 shows their structure by representing the nodes and edges connecting them. Their properties including, size, average degree, standard deviation of the degrees, average path length and clustering coefficient is represented in Table 1. We used Mfinder to determine the significance of all three- and four-nodes subgraphs of these networks. In order to obtain a high level of accuracy, we set the parameters of random network generation algorithm and counting motifs in the tool as follows [34]

- Number of random networks = 10000
- Uniqueness threshold is ignored
- No threshold on mfactor to use when counting motifs
- No threshold on Z-score to use when counting motifs
- Default values were considered for other parameters, including switching method for generating random networks.

Table 2 summarizes the set of three- and four-node motifs with their corresponding normalized and non-normalized Z-scores in the networks. As we can see motif #7 — a four-node motif with five edges — has the highest positive Z-score, and thus, is the most significance motif structure in both of the networks and can be considered as the dominant motif. On the other hand, motif#1 has the highest negative Z-score in both of the networks, and thus, is the most significant anti-motif in the set of three- and four-node subgraphs. There is a significant direct correlation between the Z-scores of the motifs in these two networks ($r = 0.9328$, $P<0.001$; Pearson linear correlation and $r = 0.9286$, $P<0.0025$; Spearman rank correlation). This indicates the similarity of these two networks in the structure of their building blocks, i.e. #2, #5, #7, and #8, have always positive Z-score, i.e. they are significantly more abundant in these networks as compared to random networks. As the clustering coefficient of the real networks is relatively large (see Table 1), it seems natural that the subgraphs that include a triangle structure have a positive Z-score. In some sense, the Z-score of motifs #5, #7 and #8 seems strongly dependent on the Z-score of motif #2. The negative Z-score of motif #1 seems also correlated to the positive Z-score of motif #2. Subgraph #1 and #4 (motif #6 that has small Z-score and is not a significant motif) has always negative Z-score meaning that they are anti-motifs appearing much less in the original networks as compared to random ones.

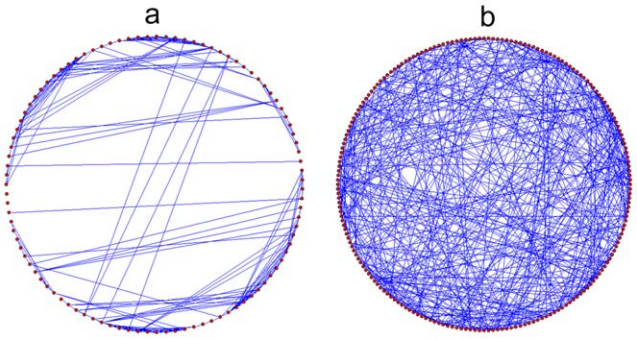

Figure 1. Topology of sample biological networks. (a) Protein structure network [7] and (b) human brain functional network extracted through functional magnetic resonance imaging [24].
doi:10.1371/journal.pone.0020512.g001

Table 1. Characteristics of considered biological networks.

Network Type	N	<k>	std(k)	P	C
Protein structure	99	4.2828	0.4748	5.2607	0.3600
Functional human brain	200	4.5400	0.5690	5.2200	0.2858

First columns: the name of the networks. Second to sixth columns: network size (N), average node-degree (<k>), standard deviation of node-degree (std(k)), average characteristic path length (P), and clustering coefficient (C).
doi:10.1371/journal.pone.0020512.t001

Random and Systematic Failures in the Edges

Random or systematic failures can occur in some of the networks' components, i.e. nodes and edges. For example in protein-protein interaction network, while attacking nodes may correspond to breakdown of polypeptides by appropriate enzymes, attacking edges of the network can be interpreted as preventing physical interaction between two polypeptides in order to prevent carrying out their biological functions. In this work we considered failures in the edges and investigated its influence on the profile of the motif structure of the networks. Failures in the networks are of two types, in general: random failures that are called errors or systematic failures that are called attacks.

Let define some preliminary metrics of graph theory. Consider an undirected and unweighted network with adjacency matrix $A = (a_{ij})$, $i, j = 1, \ldots, N$, where N is the size of the network. Let denote the edges between the node i and the node j by e_{ij}. The degree of the node i can be obtained as

$$k_i = \sum_{j=1}^{N} a_{ij}. \quad (3)$$

Edge betweenness centrality (load) is a centrality measure of an edge in a graph, which counts the number of shortest paths passing through the edge. The betweenness centrality L_{ij} of the edge e_{ij} between nodes i and j that is defined by [52]