Interaction pathways promote spliceosome module integration and network-level robustness to cascading effects

Paulo R. Guimarães Jr[1], Mathias M. Pires[2], Maurício Cantor[3,4], Patricia P. Coltri[5]

[1] Departamento de Ecologia, Instituto de Biociências, Universidade de São Paulo, Rua do Matão, Travessa 14, 05508-900, São Paulo, SP, Brazil;

[2]Departamento de Biologia Animal, Instituto de Biologia, Universidade Estadual de Campinas, Rua Monteiro Lobato 255, 13083-862, Campinas, SP, Brazil;

[3] Departamento de Ecologia e Zoologia, Centro de Ciências Biológicas, Universidade Federal de Santa Catarina, Trindade, Caixa Postal 5102, CEP 88040-970, Florianópolis, SC, Brazil;

[4] Centro de Estudos do Mar, Universidade Federal do Paraná, Av. Beira-mar, s/n, Caixa Postal 61, CEP 83255-976, Pontal do Paraná, PR, Brazil;

[5] Departamento de Biologia Celular e do Desenvolvimento, Instituto de Ciências Biomédicas, Universidade de São Paulo, Av Prof Lineu Prestes, 1524, ICB-I, 05508-000, São Paulo, SP, Brazil.

Abstract: Biological systems are organized as networks. A central problem in the study of biological networks is to understand if and how the network structure affects the fragility of biological systems to multiple types of perturbations. For example, the functionality and fragility of protein networks may depend on their network structure, and mutations and other errors may generate cascading effects that, in turn, lead to system malfunctioning. Spectral graph theory studies the structural and dynamical properties of a system based on the mathematical properties of matrices associated with the networks, providing tools, which can reveal the fragility of biological networks to cascading effects. We combined two of such tools to explore the fragility to cascading effects of the network describing protein interactions within a key macromolecular complex, the *S. cerevisiae* spliceosome. The spliceosome network shows a higher number of indirect pathways connecting proteins than expected for random networks. The multiplicity of pathways may promote routes to cascading effects to propagate across the network. However, analytical results derived from the spectral graph theory and numerical simulations of a minimal mathematical model suggest that the modular structure of the spliceosome network constrains the propagation of cascading effects due to the concentration of pathways within modules. We hypothesize that the concentration of pathways within modules favors robustness of the spliceosome against failure but may lead to a higher vulnerability of functional subunits, which may affect the temporal assembly of the spliceosome. Our results illustrate the usefulness of spectral graph theory in identifying fragile domains in biological systems and predicting their implications, which can become a useful as a roadmap for the development of new therapies within the emerging field of network medicine.

1. Introduction

Multiple biological systems can be characterized by networks[1,2]. Genes form regulatory networks[3,4], proteins are connected through a network of pathways[5,6], individuals are embedded in social networks[7], and species are linked to each other in food webs[8]. In the past two decades, we have learned about the main structural aspects of multiple biological networks [8-11]. Simultaneously, a wide range of empirical and theoretical studies explored the dynamical implications of network structure [5,12-14].

Within the cell, analyzing the structure of protein networks may provide information about the underlying processes shaping the organization and function of macromolecular complexes. Also structural aspects of protein networks may influence the vulnerability of these macromolecular processes to distinct types of perturbations[15-18]. Mutations may lead to non-functional proteins that in some cases may imperil primal functions, causing the death of the cells or affecting tissue functioning and leading to multiple types of diseases[19,20]. Alternatively, some mutations may not impair macromolecular processes. For instance, despite the importance of PRP8, a protein involved in the formation of the spliceosome catalytic core, some mutations on yeast PRP8 affect splicing efficiency and fidelity but did not impair spliceosome formation[21,22]. Such examples suggest that the underlying protein network is resilient. Hence, a key challenge to the study of protein networks is to understand if and how network structure may impact the fragility of protein interactions to perturbations.

One of the main venues to explore how network structure influence biological dynamics is by means of mathematical modelling[23]. However, mathematical modelling of complex networks requires enourmous amount of empirical data to parameterize

dynamical models, which makes the use of mathematical modelling challenging. Fortunately, because multiple dynamical processes in networks are shaped by the architecture of the underlying network[12], we can circumvent this challenge by focusing on the possible role of network structure on dynamics. This way is possible to assess the consequences of particular interaction patterns in spreading perturbations through the system[24]. To this end, spectral graph theory is a powerful tool[25-29]. Spectral graph theory is the study of the eigenvalues and eigenvectors of matrices that describe or are associated with the networks. Specifically, the distribution of eigenvalues (spectra) of two matrices associated with the underlying graph describing the network, namely the adjacency matrix and the Laplacian matrix, contain information of the potential role of network structure in favoring or constraining cascading effects in distinct systems[25-27,30,31].

Here, we explored the fragility of the network describing protein interactions within a key macromolecular complex, the *Saccharomyces cerevisiae* spliceosome (Fig 1). Spliceosomes catalyze splicing, an essential process for gene expression regulation in eukaryotic cells[32]. We used tools from the spectral analysis to investigate if the spliceosome network is vulnerable to cascading effects. Since pathways linking proteins indirectly provide propagation routes for cascading effects, we first used results related to the spectra of the adjacency matrix to estimate how these indirect pathways are distributed within the spliceosome network. Then, we used the spectra of the Laplacian matrix associated with the spliceosome network to estimate the vulnerability of the network to cascading effects, in which changes in the state of a protein (mutations, unfolding, misprocessing, malfunctioning) may propagate across the network. Finally, we simulated cascading effects in these networks using a simple mathematical model

and compared the simulated dynamics with the analytical predictions derived using spectral graph theory.

2. The spliceosome network

The spliceosome is composed of 5 snRNAs (small nuclear RNAs U1, U2, U4, U5 and U6) and more than 100 proteins. The structure of this macromolecular complex suggests that proteins and snRNAs are organized in different sub-complexes or modules, important for the formation of the complex catalytic core[33]. We analyzed the protein network of the *S. cerevisiae* spliceosome, comprised by N=103 proteins and their pairwise interactions [2,34]. Proteins and putative protein-protein interactions were recovered from the STRING database[35]. Pairwise interactions can be inferred using different approaches and these approaches may provide different levels of support to a putative pairwise interaction between two proteins[2,34]. A reliability score was assigned (varying from zero to one) to each putative protein-protein interaction according to the level of evidence suggesting if the interaction occur and provided by different experimental approaches[36]. The proteins are depicted as nodes and we assume there is a link connecting two proteins if the level of support to that putative interaction is higher than 0.5 (additional details at [2,34]). This level of support represents a heuristic cutoff since lower cutoffs imply, by definition, the record of weakly supported protein-protein interactions and higher cutoffs do not imply in structural changes to spliceosome network structure[34]. The dataset is available at [2] and all the analyses were run using MATLAB scripts that are available upon request.

In this network, there are 881 interactions (connectance C = 0.168) and each protein has, on average, 17.10 ± 13.04 interactions[34]. In this sense, the spliceosome network is similar to other protein networks in which most proteins interact with a few other proteins and there is a small set of highly-connected proteins[15,37,38]. Previous studies show that the spliceosome network is nonrandomly structured, combining patterns of nestedness (proteins with fewer interactions and highly connected proteins interact with the same proteins) and modularity (cohesive groups of proteins that interact more with each other than with the rest)[2,34]. The modular structure of the spliceosome network analyzed here was previously characterized in four modules identified using the maximization of Q index of modularity in a simulated annealing framework[34] (Fig. 1A). These modules are associated with functional subunits of the complex (additional details in [34]). Modularity was also recorded in a distinct dataset for *S. cerevisiae* spliceosome and in human spliceosome network, supporting the notion that the modular structure of spliceosome is a general pattern[34].

3. Computing indirect pathways among proteins

One crucial implication of the network structure of biological networks is the emergence of pathways connecting distinct elements of the network directly or indirectly[30,39]. In the spliceosome network, these pathways connect otherwise spatially and temporally isolated pairs of proteins. Given that cascading effects may propagate through such pathways[39], quantifying how distinct proteins are connected in the network may inform the fragility of the spliceosome to cascading effects. In Graph Theory, a pathway can be depicted as a walk. A walk is defined as a set of nodes (here,

proteins) and links (here, protein interactions), starting and finishing with nodes (Fig. 2A). The same node or link may be represented multiple times in a pathway and, therefore, longer pathways may be formed by a combination of smaller pathways (Fig. 2A). As a consequence, the number of pathways increases exponentially with pathway length, ζ, which is the number of links (distinct or not) in a pathway, a phenomenon called pathway proliferation[39].

The denser the pathway in a network with a given length ζ, the higher the number of possible routes that allow perturbations to cascade through proteins in the network. Therefore, the rate of increase in the number of pathways of length ζ (hereafter rate of pathway proliferation) is a useful statistic to describe the multiplicity of routes allowing cascading effects in a given network[39]. The rate of pathway proliferation depends on the eigenvalues of the adjacency matrix **A**[39,40]. In the adjacency matrix **A,** a given element a_{ij} describes if the interaction between two proteins, i and j, occurs ($a_{ij} = 1$) or not ($a_{ij} = 0$). Although all eigenvalues may affect the rate of pathway proliferation for low values values of ζ, for large values of ζ the rate of pathway proliferation is governed by the leading eigenvalue of **A**, λ_1, and the number of pathways with a given length ζ, $\psi^{(\zeta)}$, is

$$\psi^{(\zeta)} \simeq \psi\lambda_1^{\zeta-1}$$

(1),

in which ψ is the number of protein interactions recorded in the spliceosome network. The higher the λ_1, the higher the number of pathways of a given length ζ connecting proteins in the network.

We first computed the leading eigenvalue of the spliceosome network, which is $\lambda_1 = 25.84$. We used the leading eigenvalue to predict the accumulation of pathways

with the increase of ζ and we compared to actual accumulation of pathways in the spliceosome network. The total number of direct interactions between pairs of proteins (pathways of length, $\zeta = 1$) in a network with N proteins is equal to $\psi = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}$. Likewise, the number of pathways of length $\zeta = 2$, $\psi^{(2)}$, can be estimated by computing $\mathbf{A}^2$. The element $a_{ij}^{(2)} = \frac{1}{2}\sum_{k=1}^{N} a_{ik}a_{kj}$, which is nonzero if there is at least one protein k interacting with both proteins i and j, and $\psi^{(2)} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}^{(2)}$. Hence, the number of pathways of length $\zeta = h$ is the sum of all elements of matrix $\mathbf{A}^h$, $\psi^{(h)} = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}^{(h)}$. In the spliceosome network, our results show the rate of pathway proliferation follows the analytical prediction derived from spectral graph theory even for short pathways, $\zeta < 4$(Fig 2B), even though equation (1) represents a prediction for the number of pathways assuming large ζ.

We now turn our attention to the role of the nonrandom structure of the spliceosome network in shaping the proliferation of pathways. We computed the leading eigenvalues of 1,000 random networks with the same number of proteins and the same number of interactions but in which the probability that a protein i interacts with protein j is constant and equal to the connectance C. Spectral graph theory predicts that for a random graph in which $NC >> \log N$ the expected value for the leading eigenvalue for a random network (an Erdos-Renyi graph) is $\lambda_R = [1 + o(1)]NC$, in which $o(1)$ is a function that converges to a value close to zero[41]. Thus, the expected leading eigenvalue for a random network with the same number of proteins and interactions is $[1 + o(1)]NC \simeq NC = 17.24$. This prediction is close to the mean leading eigenvalue for our simulated random networks (17.94 ± 0.12, 1000 random networks) and both the analytical prediction for random networks and the estimated value for simulated

random values are much smaller than the leading eigenvalue of the spliceosome network ($\lambda_1 = 25.84$; Fig. 2C, $P < 0.001$). This difference in the leading eigenvalues of spliceosome network and random networks implies that, for a given pathway length, the empirical spliceosome network is much more connected than expected by chance. For example, for $\zeta = 3$, there are 602,250 pathways connecting the spliceosomal proteins in the network. In contrast, for a random network with the same number of proteins and protein interactions, the expected number of pathways is just one third of the number of pathways observed in the empirical network ($282{,}780 \pm 3{,}614.6$; $P < 0.001$).

4. Pathways and modular structure

The spliceosome network has a modular structure, in which interacting proteins are concentrated into cohesive subgroups[34]. Such modular structure counterbalances the role of pathways in facilitating cascading effects. Since there are no disconnected components in the spliceosome network, there are pathways connecting all pairs of proteins. However, modularity may prevent the system-level effects of cascading effects since most of the long pathways start and end in proteins of the same module[30,39]. If this is true, then cascading effects are expected to propagate among modules less efficiently and the consequences of indirect pathways are expected to be greater locally, i.e., within modules, than between modules[39].

The spliceosome network analyzed here has four modules[34]. We investigated if pathways are more concentrated within modules than between modules. To do so, we first ascribe all pathways in two groups: (1) pathways in which the starting and ending protein are within the same module (including pathways starting and ending with the

same protein); and (2) pathways starting and ending with proteins from different modules. To compare the role of pathways in connecting proteins from the same or from distinct modules, we computed the mean number of pathways of a given length ζ starting and ending with proteins within the same module, $\psi_m^{(\zeta)}$, and starting and ending with proteins from distinct modules, $\psi_d^{(\zeta)}$. We then computed the ratio $\delta^{(\zeta)} = \psi_m^{(\zeta)}/\psi_d^{(\zeta)}$. The direct interactions, $\zeta = 1$, are concentrated within modules: 36.03% of all possible protein interactions within modules are observed, whereas just 8.94% of the possible protein interactions between modules were observed, leading to $\delta^{(1)} = 4.03$. Because the modules are not disconnected, indirect pathways connect proteins from different modules, leading to a reduction of $\delta^{(\zeta)}$ as ζ increases (Fig. 3). That said, the effects of the modular structure still affect the pathway distribution. For example, when $\zeta = 2$, there were twice as much pathways per pairs of proteins within modules than between modules ($\delta^{(2)} = 1.99$). The values of $\delta^{(\zeta)}$ converge to values close to one as ζ increases, but even for very large values of ζ, there is still a small concentration of these long pathways within modules when compared with pathways between modules ($\delta^{(\zeta\to\infty)} = 1.02$). Hence, although the spliceosome network structure favors a higher density of pathways than expected in random networks, the modular structure implies that most of these pathways are concentrated within rather than between modules.

5. How likely are cascading effects to spread across the network?

We showed that most of the pathways are concentrated within the modules of the spliceosome network. Modularity may limit the cascading effects of perturbations[42-44]. Therefore, a fundamental feature of a given network is to describe how isolated the

network modules are. There are multiple approaches to measure the isolation of modules. The spectral analyses of Laplacian matrix is an approach directly rooted in the implications of connectivity to the dynamics within networks, e.g., network flow, homogenization of states, and synchronization[45]. The Laplacian matrix is defined as $L = (K - A)$, in which K is a diagonal matrix in which all elements are zero but the main diagonal $k_{ii} = \sum_{j=1}^{N} a_{ij}$. The spectral properties of the Laplacian matrix inform on the connectivity of groups within networks[26]. For example, the number of zero eigenvalues inform the number of isolated groups of proteins within a network (i.e., the network components)[26]. In the spliceosome network, there are pathways connecting any pairs of proteins and therefore there is a single zero eigenvalue. In this case, the smallest non-zero eigenvalue, also called algebraic connectivity or Fiedler number, describes how well connected are modules within networks[46].

We computed the algebraic connectivity of the spliceosome network and we compared with the smallest nonzero eigenvalue of 1,000 random networks with the same number of proteins and the same number of interactions. The algebraic connectivity of the spliceosome network is $\lambda_2^{(L)} = 0.69$, a value much smaller than expected for a similar random network (7.24 ± 0.93, $\boldsymbol{P} < 0.001$, Figure 4a). Therefore, the connectivity among modules in the empirical spliceosome network is much weaker than expected for a random network with the same number of interactions and proteins.

The spectral properties of the Laplacian matrix also describe how fast cascading effects impact the network by means of approaches derived from the study of flow of networks and of the emergence of synchronization. We used a set of difference equations to explore the consequences of the spectral properties of Laplacian matrix for the dynamics of the system. This minimal model does not aim to reproduce the

dynamics of a given biological process in detail, but to explore the potential role of network structure in shaping cascading effects in very simple dynamics. Our minimal model assumes that there is a state associated to a given protein i, ϕ_i, and the dynamics describe how fast direct and indirect effects lead to the homogenization of state values. The state ϕ_i is binary, $\phi_i = 1$ if the protein performs its function in the spliceosome and $\phi_i = 0$ if the protein is mutated in a nonfunctional form or if the malfunctioning of interacting proteins lead to a cascading effect that inhibits the performance of an otherwise functional protein. We assume that the probability of a given protein to be affected by a nonfunctional interacting protein is:

$$P(\phi_i^{t+1} = 0 | \phi_i^t = 1) = c^{k_{ii} - \sum_{j=1}^{N} a_{ij}\phi_j} \quad (2),$$

in which k_{ii} is the number or proteins interacting with protein i and c is a constant between zero and one that controls the propagation of cascading effects. In this model, given enough time, all the proteins lose their functionality so the spliceosome becomes unviable. In reality, not all errors and mutations may lead to the collapse of the spliceosome[20,47].

However, by estimating how fast the functionality of proteins collapse due to errors, this minimal model allows one to estimate the effect of network structure in favoring cascading effects, without the complexity of real, empirical dynamics of biological processes. The faster the convergence in state values, the higher is the contribution of network structure to fuel cascading effects. The dynamics of the model can be approximated by a mean-field model in which protein states, ϕ_i^t, are described as continuous variables:

$$\phi_i^{t+1} = \phi_i^t - \frac{c}{k_i}\left(k_i \phi_i^t - \sum_{j=1}^{N} a_{ij}\, \phi_j^t\right) \quad (3)$$

The equation (3) can be generalized to all proteins of the network. In matrix form, the resulting set of equations represent all the difference equations of the network is:

$$\mathbf{\Phi^{t+1}} = \left(\mathbf{I} - c\mathbf{L^{(n)}}\right)\mathbf{\Phi^{t}} \tag{4}$$

in which $\mathbf{\Phi^{t}}$ is a vector with the states of all proteins at time t and $\mathbf{L^{(n)}}$ is the Laplacian matrix normalized for random walks. Note that c could assume distinct values for each protein but for the sake of simplicity we used the same value for all proteins without any loss of generality. We can then define the matrix $\mathbf{M} = \mathbf{I} - c\mathbf{L^{(n)}}$ and the dynamics of states as:

$$\mathbf{\Phi^{t}} = \mathbf{M^{t}}\mathbf{\Phi^{0}} \tag{5}$$

All absolute values of the eigenvalues of **M** are bounded between zero and one and, as a consequence, they converge to zero with large t. The eigenvalues of **M** controls how fast the network would favor the homogenization of states due to its own structure. In an empirical system, such time-to-homogenization is an estimate of the effect of network structure on cascading effects. The eigenvalues of **M** is defined by the eigenvalues of $\mathbf{L^{(n)}}$:

$$\lambda_i^{(M)} = 1 - c\lambda_i^{(L^n)} \tag{6}$$

in which $\lambda_i^{(M)}$ is the i-th eigenvalue of **M** and $\lambda_i^{(L^n)}$ is the i-th eigenvalue of $\mathbf{L^{(n)}}$. The smallest eigenvalue of **M** is always $\lambda_i^{(L^n)} = 0$, defining the leading eigenvalue of M as $\lambda_1^{(M)} = 1$. For large t, the states of the system converge to eigenvector associated to $\lambda_1^{(M)}$. In a connected network, the effects of all N-1 nonzero eigenvalues of **M** will decays to zero with large t, because their absolute value is bounded between zero and one, controlling the time to collapse in the model. The higher the time to collapse, the longer it would take to a cascade propagate across the network. Thus, the eigenvalues of

Laplacian matrices link the structural connectivity of a network with its dynamical consequences.

We tested if the eigenvalues of the Laplacian matrix normalized for random walks predicts the transient dynamics described by equation (4) in simulations parameterized with information of the structure of empirical spliceosome network. At t=0, when all proteins are functional, $\phi_i = 1$ for any protein i. At the equilibrium, all states converge to zero, leading to the complete collapse of the spliceosome.

We ran 1,000 simulations of the model and computed the mean number of functional proteins in a given time step and assuming $c = 0.1$ (Fig. 4b). We fit a log-linear model to estimate the rate of the decay of functional proteins. To reduce statistical fluctuations we truncated the analysis for time steps in which the mean number of functional proteins is higher than one. Despite the small system size (N=103 proteins) and the binary nature of protein states in our numerical simulation (equation 2), the analytical predictions assuming continuous states and based on the spectra of $\mathbf{L^{(n)}}$, $\sum_{i=1}^{N} \phi_i{}^t = \sum_{i=2}^{N} \left(\lambda_i^M\right)^t$ predicted the time to collapse in the model. The simulated rate of decay in the number of functional proteins is 0.917^t, close to the analytical prediction of $\sum_{i=2}^{N} \left(\lambda_i^M\right)^t = 0.928^t$ (Figure 4b). Taken together, these results suggest that the modular structure of the interactions among spliceosomal proteins lead to a network in which modules of proteins are loosely connected by interaction pathways, making the system much less prone to cascading effects across the whole networks than expected for a random network with the same number of interactions and proteins.

6. Discussion

One of the main consequences of network organization is the formation of pathways connecting otherwise isolated elements of the system[48]. Pathways create the routes for cascading effects to propagate through the system, coupling the dynamics of non-interacting elements across multiple levels of biological organization[30,49-51]. In protein networks, these pathways of interactions are fundamental to a series of intracellular processes with functional consequences for the cell[52]. Here, we used spectral analysis of graphs to explore the consequences of the structural organization of the spliceosome, a key macromolecular complex, and improved the understanding of the vulnerability of this protein-protein network against perturbations in three key ways.

First, we reveal that the network organization of the spliceosome favors the proliferation of pathways. This feature leads to more pathways of a given length than expected by similar-sized random networks. We show that the number of pathways of a given length connecting proteins is predicted by the leading eigenvalue of the adjacency matrix describing the spliceosome. The high proliferation rate in the spliceosome network is a consequence of the large variation in the number of interactions per protein[34]. One key finding of spectral graph theory is that the upper bound value of the largest leading eigenvalue is the largest number of interactions recorded for a protein in a network[53]. Biologically, the large variation in the number of interactions per protein implies that some highly connected proteins will be central for the organization of the network, creating pathways among poorly connected proteins. Highly-connected proteins are central to a number of processes in the cells, and deletions of such hubs in the protein networks may be lethal due to cascading effects[5]. The relevance of highly connected proteins—the centrality-lethality rule—may stem from multiple factors, such

as participation in essential, direct pairwise interactions[54] or in the organization and reorganization of the large protein network[55]. In this sense, we suggest that pathway proliferation is another structural consequence of highly-connected proteins that may favor cascading effects. Thus, changes in the network structure via experimental protein deletions can be predicted by the leading eigenvalue, which can inform the relative contribution of individual proteins to indirect pathways in the protein networks.

Second, we demonstrate that these pathways are not randomly distributed across network, but rather concentrated within modules. Modularity is one of the main features of protein networks and evolutionary processes may favor the emergence of modularity by a combination of gene duplication, horizontal gene transfer, and natural selection[15,37]. Selection may favor modularity allowing both specificity and autonomy of functionally distinct subsets of proteins[15,38]. In this sense, the concentration of pathways within modules provide a way to increase module integration, favoring distinct functional roles developed by proteins in distinct modules. In the spliceosome, modules are associated with subcomplexes that act in distinct steps of the spliceosome assembly and function[34]. Thus, pathway proliferation may favor the emergence of highly integrated subunits, in which effects of pairwise interactions may also activate indirect effects on non-interacting proteins associated with the same function or step of splicing process. More than promoting within-module integration, pathway proliferation also integrates distinct modules. Modularity does reduce the potential for cascading effects across the system[43]. However, modularity does not imply module isolation. System functioning also depends on the indirect effects between functional modules[56], allowing complex tasks to be completed by distinct subunits of the system[57]. Pathway proliferation indirectly connects distinct modules in the network through multiple

pathways, allowing indirect effects. Hence, selection is not expected to favor complete module isolation and these pathways that allow system functioning may also lead to routes for cascading effects triggered by mutations and deletions. Based on these results, we would expect that proteins connecting modules are more conserved across evolution or, at least, less prone to failures that alter their function.

Third, we suggest that the local integration of modules promoted by pathway proliferation and their semi-independence to other modules provide robustness to mutations on specific proteins, as seen in human cells with SR and hnRNP families of proteins. The hnRNP proteins were previously associated with intronic miRNAs, probably facilitating splicing reactions on these pre-RNA substrates[58,59]. Some SR proteins (hnRNP-A2/B1 and hnRNP-U) are specific modulators of splicing in SMN1 and SMN2 genes[60]. As a consequence, splicing defects on SMN1 and SMN2—and the emergence of the neurodegenerative disorder spinal muscular atrophy—might be associated with a subset of hnRNP proteins. By constraining pathways to modules, the network structure may represent an additional layer of robustness to the functioning of spliceosome.

It is important to notice that the semi-independence of modules does not imply network-level robustness of the spliceosome to all types of failures and errors. Nevertheless, these results provide a theoretical benchmark that help predict which kinds of failures are likely to cause network-level collapse in the spliceosome network. For instance, we can expect that errors in proteins that are simultaneously highly connected and link distinct modules will lead more often to network-level collapse of the spliceosome. For example, PRP8 is a large, highly-connected protein acting as a core component of U5 snRNP and essential for efficiency and fidelity of splicing reactions[22].

Interestingly, human PRP8 expression is reduced with an increase in cell proliferation, possibly affecting splicing globally[19]. Moreover, because the spliceosome network has a temporal structure and subunits are assembled and disassembled sequentially during the splicing reaction, we could expect that local collapse of the early subunits that join the complex is more likely to cause the largest disruptions with the spliceosome functioning. In fact, mutations in a group of proteins that associate to assemble early spliceosome complexes, among which are U2AF35, SF3A1 and SF3B1, are frequently associated with development of myeloid neoplasms[61]. Thus, our study provides a network-based explanation for alterations that might lead to splicing collapse: it will be a consequence of the local collapse of modules due to cascading effects propagating across pathway proliferation.

The interplay between modularity and pathway proliferation provides an hypothesis on how protein networks preserve the interconnectivity among functional modules and constrains deleterious cascading effects to propagate across the system. Future studies should investigate the role of modular structure in robustness by combining experiments *in vivo* in which key proteins are deleted to the network analysis of protein roles in the organization of spliceosome network. By now, our results support that the existence of multiple, indirect pathways connecting proteins is a potentially relevant consequence of the network structure for protein-protein interactions. Mapping the fragile and robust points of such networks could aid the development of new therapies, given that the misregulation of the spliceosome resulting from mutations in their proteins and from single-point mutations changing the splicing of a given gene are linked to many human diseases, such as spinal muscular atrophy, retinitis pigmentosa and several types of cancer, such as lymphocytic leukaemia and myelodysplasia[19,62]. We

suggest that the approach introduced here to uncover the distribution of pathways and their potential dynamical implications to spliceosome may help to characterize other types of molecular networks. In this sense, the characterization of indirect pathways and their possible consequences in multiple molecular networks may provide insights on the role of the network structure in shaping the emergence of complex diseases, contributing to emerging field of network medicine[63,64].

7. Acknowledgments

We thank Ana Paula Assis and Pâmela C. Santana for helpful comments. PRG was supported by CNPq and FAPESP (2017/08406-7). PPC was supported by FAPESP (2017/06994-9). MC was supported by a PMP/BS postdoctoral fellowship (UFPR/UNIVALI 46/2016).

8. Literature cited

1. Alon, U. Biological networks: the tinkerer as an engineer. *Science* **301**, 1866-1867 (2003).
2. Cantor, M. *et al.* Nestedness across biological scales. *PloS One* **12**, e0171691 (2017).
3. Huang, S., Eichler, G., Bar-Yam, Y. & Ingber, D.E. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical Review Letters* **94**, 128701 (2005).
4. Gerstein, M.B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91 (2012).
5. Jeong, H., Mason, S.P., Barabási, A.-L. & Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **411**, 41 (2001).
6. Vinayagam, A. *et al.* Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. *Nature Methods* **11**, 94 (2014).
7. Lazer, D. *et al.* Life in the network: the coming age of computational social science. *Science* **323**, 721 (2009).
8. Pascual, M. & Dunne, J.A. *Ecological networks: linking structure to dynamics in food webs*, (Oxford University Press, 2006).
9. Girvan, M. & Newman, M.E. Community structure in social and biological networks. *Proceedings of theNnational Academy of Sciences* **99**, 7821-7826 (2002).

10. Guimera, R. & Amaral, L.A.N. Functional cartography of complex metabolic networks. *Nature* **433**, 895 (2005).
11. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298**, 824-827 (2002).
12. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**, 47 (2002).
13. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U. Complex networks: Structure and dynamics. *Physics Reports* **424**, 175-308 (2006).
14. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378 (2000).
15. Typas, A. & Sourjik, V. Bacterial protein networks: properties and functions. *Nature Reviews Microbiology* **13**, 559 (2015).
16. Fanning, A.S. & Anderson, J.M. Protein–protein interactions: PDZ domain networks. *Current Biology* **6**, 1385-1388 (1996).
17. Taylor, I.W. *et al.* Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology* **27**, 199 (2009).
18. Mosca, R., Céol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nature Methods* **10**, 47 (2013).
19. Scotti, M.M. & Swanson, M.S. RNA mis-splicing in disease. *Nature Review Genetics* **17**, 19-32 (2016).
20. Shukla, G.C. & Singh, J. Mutations of RNA splicing factors in hematological malignancies. *Cancer Letters* **409**, 1-8 (2017).
21. Liu, L., Query, C.C. & Konarska, M.M. Opposing classes of prp8 alleles modulate the transition between the catalytic steps of pre-mRNA splicing. *Nature Structural and Molecular Biology* **14**, 519-26 (2007).
22. Kurtovic-Kozaric, A. *et al.* PRPF8 defects cause missplicing in myeloid malignancies. *Leukemia* **29**, 126-36 (2015).
23. Macnamara, C.K. & Chaplain, M.A. Diffusion driven oscillations in gene regulatory networks. *Journal of Theoretical Biology* **407**, 51-70 (2016).
24. Newman, M.E. The structure and function of complex networks. *SIAM review* **45**, 167-256 (2003).
25. Spielman, D.A. Spectral graph theory and its applications. in *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on* 29-38 (IEEE, 2007).
26. Chung, F.R. *Spectral graph theory*, (American Mathematical Soc., 1997).
27. Yeakel, J., Moore, J., Guimarães, P. & Aguiar, M. Synchronisation and stability in river metapopulation networks. *Ecology Letters* **17**, 273-283 (2014).
28. de Aguiar, M.A.M. & Bar-Yam, Y. Spectral analysis and the dynamic response of complex networks. *Physical Review E* **71**, 016106 (2005).
29. Staniczenko, P.P., Kopp, J.C. & Allesina, S. The ghost of nestedness in ecological networks. *Nature Communications* **4**, 1391 (2013).
30. Guimarães Jr, P.R., Pires, M.M., Jordano, P., Bascompte, J. & Thompson, J.N. Indirect effects drive coevolution in mutualistic networks. *Nature* **550**, 511 (2017).
31. Gibert, J.P., Pires, M.M., Thompson, J.N. & Guimarães Jr, P.R. The spatial structure of antagonistic species affects coevolution in predictable ways. *The American Naturalist* **182**, 578-591 (2013).
32. Staley, J.P. & Guthrie, C. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92**, 315-26 (1998).

33. Yan, C. *et al.* Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* **349**, 1182-91 (2015).
34. Pires, M.M. *et al.* The network organization of protein interactions in the spliceosome is reproduced by the simple rules of food-web models. *Scientific Reports* **5**, 14865 (2015).
35. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research* **39**, D561-D568 (2010).
36. Von Mering, C. *et al.* STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Research* **33**, D433-D437 (2005).
37. Wagner, G.P., Pavlicev, M. & Cheverud, J.M. The road to modularity. *Nature Reviews Genetics* **8**, 921 (2007).
38. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47 (1999).
39. Borrett, S.R., Fath, B.D. & Patten, B.C. Functional integration of ecological networks through pathway proliferation. *Journal of Theoretical Biology* **245**, 98-111 (2007).
40. Borrett, S.R. & Patten, B.C. Structure of pathways in ecological networks: Relationships between length and number. *Ecological Modelling* **170**, 173-184 (2003).
41. Chung, F., Lu, L. & Vu, V. Spectra of random graphs with given expected degrees. *Proceedings of the National Academy of Sciences* **100**, 6313-6318 (2003).
42. Wu, J.-j., Gao, Z.-y. & Sun, H.-j. Cascade and breakdown in scale-free networks with community structure. *Physical Review E* **74**, 066111 (2006).
43. Gilarranz, L.J., Rayfield, B., Liñán-Cembrano, G., Bascompte, J. & Gonzalez, A. Effects of network modularity on the spread of perturbation impact in experimental metapopulations. *Science* **357**, 199-201 (2017).
44. Stouffer, D.B. & Bascompte, J. Compartmentalization increases food-web persistence. *Proceedings of the National Academy of Sciences* **108**, 3648-3652 (2011).
45. Newman, M. *Networks: an introduction*, (Oxford university press, 2010).
46. Gross, J.L., Yellen, J. & Zhang, P. *Handbook of graph theory*, (Chapman and Hall/CRC, 2013).
47. Hirabayashi, S. *et al.* Spliceosomal gene aberrations are rare, coexist with oncogenic mutations, and are unlikely to exert a driver effect in childhood MDS and JMML. *Blood* **15**, e96-9 (2012).
48. Watts, D.J. & Strogatz, S.H. Collective dynamics of 'small-world'networks. *Nature* **393**, 440 (1998).
49. Fowler, J.H. & Christakis, N.A. Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences* **107**, 5334-5338 (2010).
50. Pedraza, J.M. & van Oudenaarden, A. Noise propagation in gene networks. *Science* **307**, 1965-1969 (2005).
51. Matozaki, T., Nakanishi, H. & Takai, Y. Small G-protein networks:: Their crosstalk and signal cascades. *Cellular Signalling* **12**, 515-524 (2000).
52. Petrakis, S. & Andrade-Navarro, M.A. Protein Interaction Networks in Health and Disease. *Frontiers in Genetics* **7**, 111 (2016).
53. Arumugam, S., Brandstädt, A., Nishizeki, T. & Thulasiraman, K. *Handbook of graph theory, combinatorial optimization, and algorithms*, (Chapman and Hall/CRC, 2016).

54. He, X. & Zhang, J. Why do hubs tend to be essential in protein networks? *PLoS Genetics* **2**, e88 (2006).
55. Asensio, N.C., Giner, E.M., De Groot, N.S. & Burgas, M.T. Centrality in the host–pathogen interactome is associated with pathogen fitness during infection. *Nature Communications* **8**, 14092 (2017).
56. Han, J.-D.J. *et al.* Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430**, 88 (2004).
57. Sauro, H.M. Modularity defined. *Molecular Systems Biology* **4**, 166 (2008).
58. Guil, S. & Caceres, J.F. The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. *Nature Structural and Molecular Biology* **14**, 591-6 (2007).
59. Paiva, M.M., Kimura, E.T. & Coltri, P.P. miR18a and miR19a Recruit Specific Proteins for Splicing in Thyroid Cancer Cells. *Cancer Genomics Proteomics* **14**, 373-381 (2017).
60. Wee, C.D., Havens, M.A., Jodelka, F.M. & Hastings, M.L. Targeting SR Proteins Improves SMN Expression in Spinal Muscular Atrophy Cells. *PLoS One* **9**, e115205 (2014).
61. Yoshida, K. *et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64-9 (2011).
62. Chen, W. & Moore, M.J. The spliceosome: disorder and dynamics defined. *Current Opinion in Structural Biology* **24**, 141-149 (2014).
63. Goh, K.-I. *et al.* The human disease network. *Proceedings of the National Academy of Sciences* **104**, 8685-8690 (2007).
64. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12**, 56 (2011).

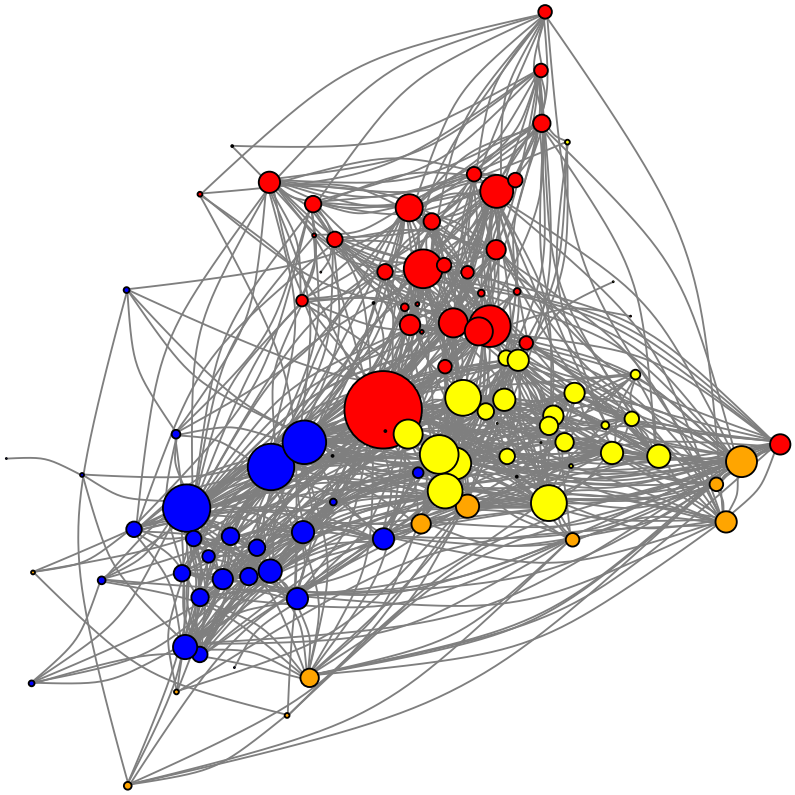
9. Figure legends

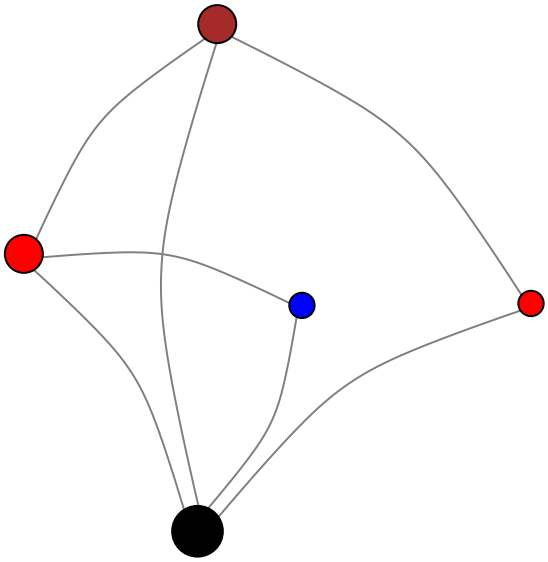
Fig 1. A protein network describing protein-protein interactions (links) distributed among 103 proteins (nodes) from *Saccharomyces cerevisiae* spliceosome. Links are defined based on the reliability of the evidence favoring protein-protein interaction (see text for further details). Node size is proportional to the number of interactions per protein and proteins with the same color are components of the same module in the network[34]. Modules represent groups of proteins that are more densely connected to each other than to other proteins in the network.

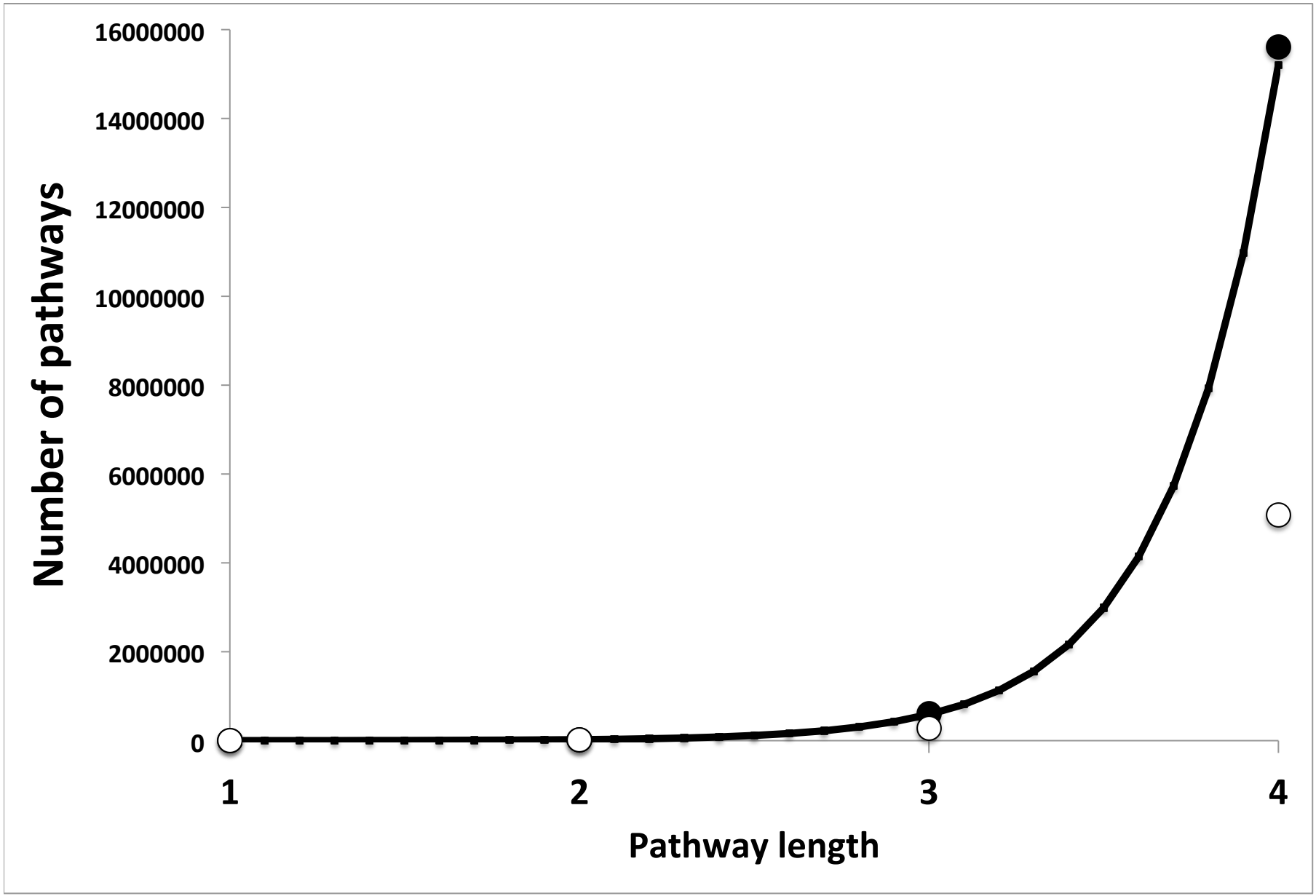
Fig 2. Pathways in a network. A) A hypothetical network with five proteins in which links represent protein interactions. The two (left and right) red nodes are connected by multiple pathways of different lengths, including left-black-right, right-blue-black-left, left-brown-right-black-left, and left-brown-left-brown-right. B) The number of pathways increases faster with pathway length form the spliceosome network (closed nodes) than for random networks with the same number of proteins and protein-protein interactions (open nodes). The trendline describes the analytical prediction for the spliceosome network based on the leading eigenvalue of the adjacency matrix (see text for further details). C) The higher pathway proliferation of the spliceosome network is a consequence of the larger leading eigenvalue of its adjacency matrix (red arrow) when compared with the expected leading eigenvalue of random networks (black bars, 1000 random networks).

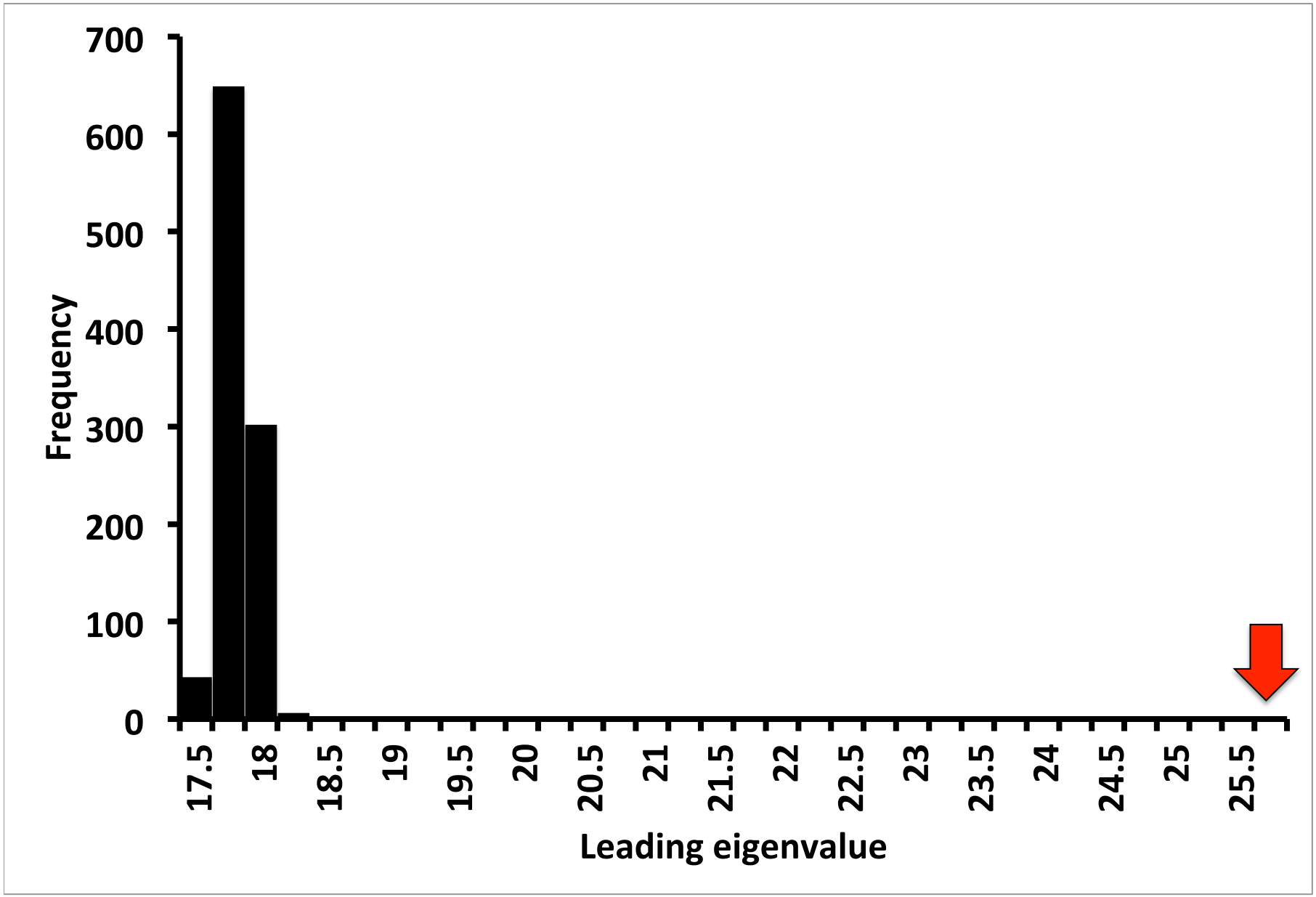
Fig 3. Pathways within and across modules. Pathway ratio is the ratio between the number of pathways starting and ending with proteins from the same module to the number of pathways starting and ending with proteins from different modules. The concentration of pathways within modules decays with pathway length. Nevertheless, even for long pathways most of pathways connect two proteins from the same module (pathway ratio >1).

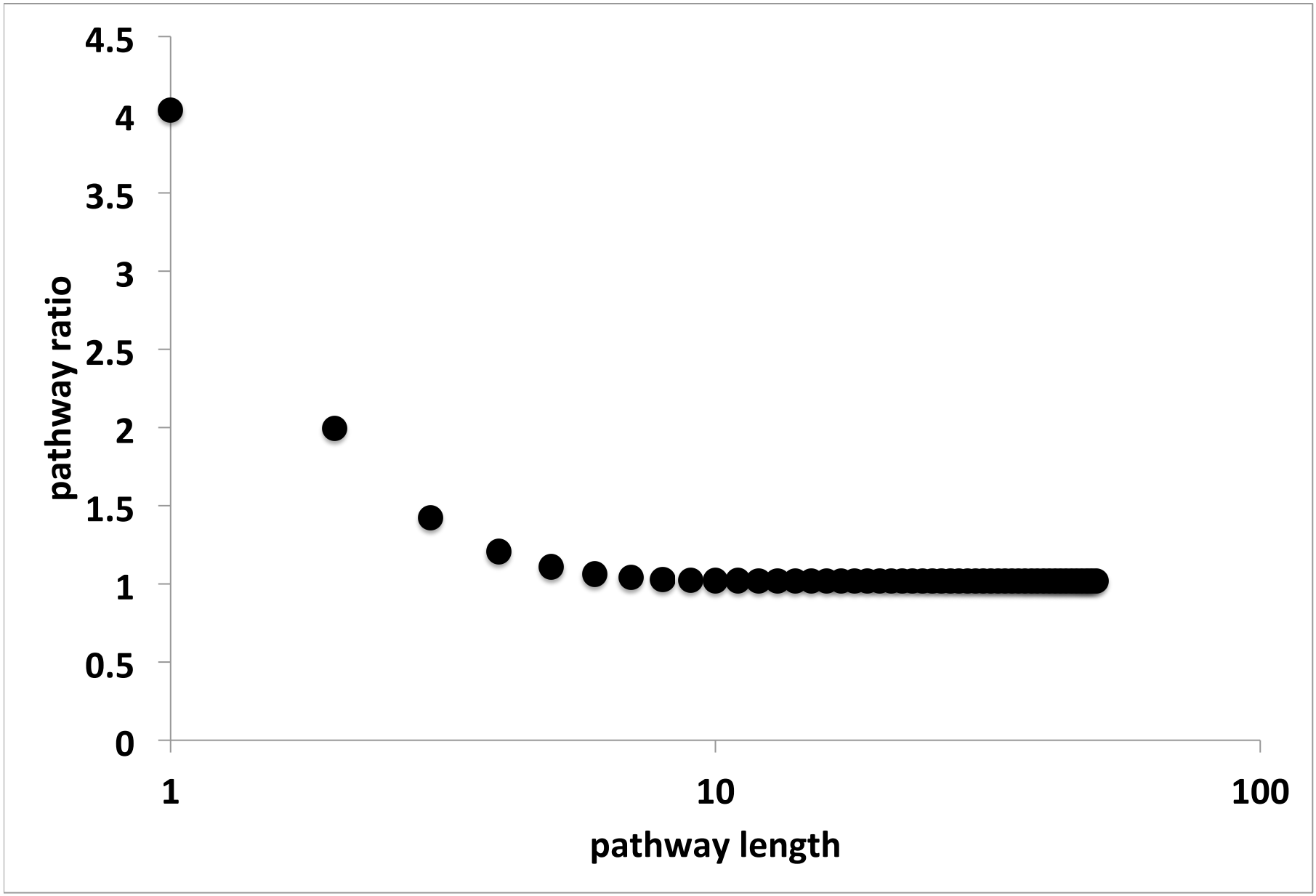
Fig. 4. Cascading effects in the spliceosome network. A) Spectral graph theory predicts that the rate of spreading of cascading effects will be governed by the algebraic connectivity (see text for further details). The spliceosome network show lower algebraic connectivity (red arrow) when compared with the expected algebraic connectivity of random networks (black bars, 1000 random networks). B) The time-to-collapse of the spliceosome network in a minimal mathematical model of failure spreading is predicted by the algebraic connectivity of the spliceosome network, each open square is the mean number of functional proteins (1000 simulations) and the solid line is the analytical prediction derived from the spectra of the Laplacian matrix normalized for random walks.

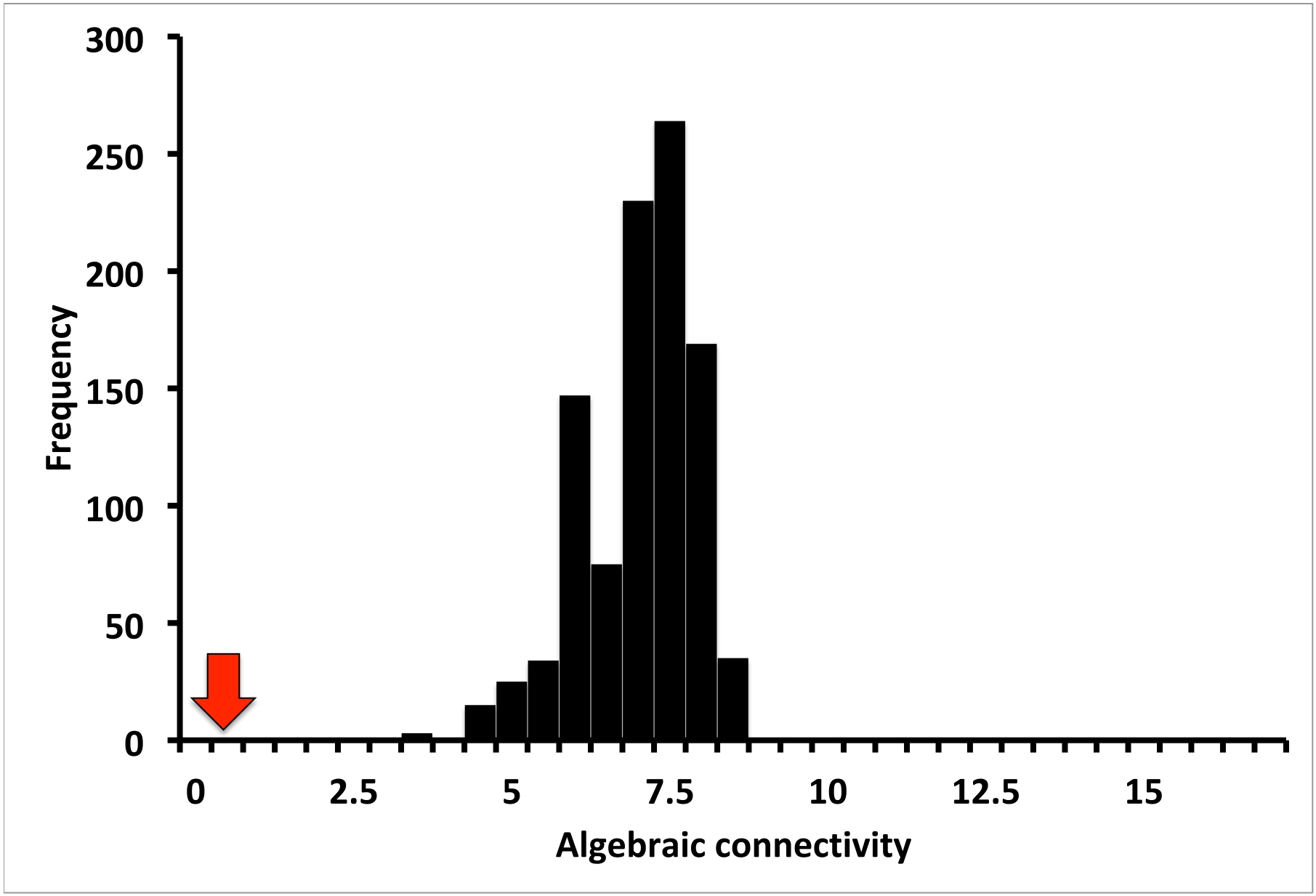


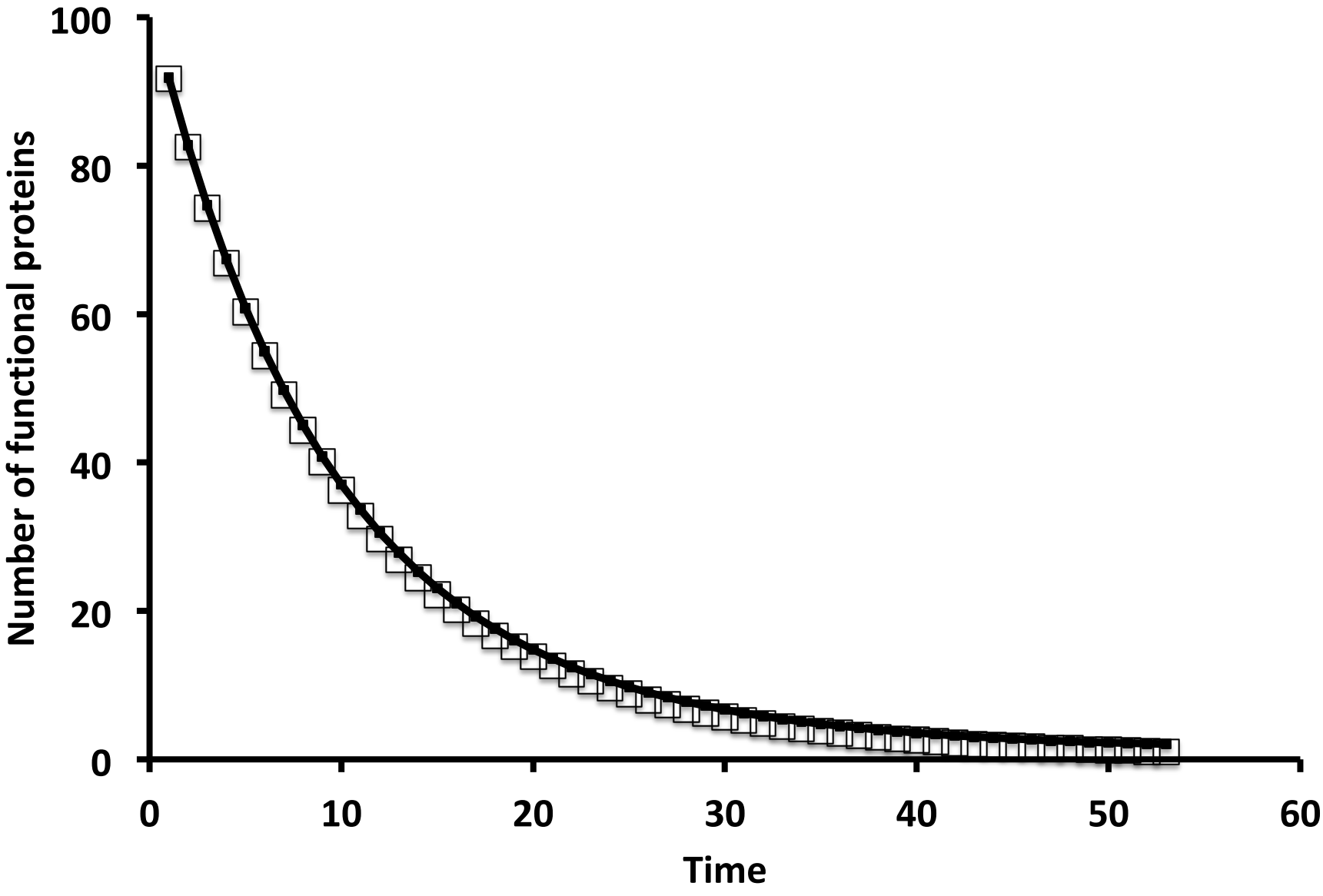



16000000
14000000
12000000
10000000
8000000
6000000
4000000
2000000
0
Number of pathways
1
2
3
4
Pathway length


700
600
500
400
300
200
100
0
Frequency
17.5
18
18.5
19
19.5
20
20.5
21
21.5
22
22.5
23
23.5
24
24.5
25
25.5
Leading eigenvalue


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
pathway ratio
1
10
100
pathway length


300
250
200
150
100
50
0
Frequency
0
2.5
5
7.5
10
12.5
15
Algebraic connectivity


100
80
60
40
20
0
Number of functional proteins
0
10
20
30
40
50
60
Time