Uncovering the organizing principles behind signed biological networks

Gonçalo Archer Franco Frazão

Instituto Superior Técnico - Universidade de Lisboa

November, 2018

Abstract

Networks are useful tools to describe biological systems, encoding elements of the system and their interactions in nodes and edges, respectively. Signed networks encode further information about the interaction in the form of a signal in each edge.

Despite extensive studies of local patterns in the network structure, especially in social networks, to the best of our knowledge, these have not yet been studied in undirected signed biological networks. Local patterns have proved useful to predict new interactions in networks built from noisy and incomplete datasets, supporting the delimitation of functional modules and even the prediction of single-element function.

This project's goal is to identify the typical patterns in genetic interaction networks, more specifically triangles and squares. Two networks describing the interactions between essential and non-essential genes in yeast were constructed, from data publicly available.

For the essential network, we found that triangles with a negative edge product are favored in the network structure, while triangles with a positive edge product are deprecated. Thus, to correctly define the structural rules in genetic networks, for triangles, one must invert the signals of Structural Balance, a very well-known and long-standing theory for balance in social networks. For squares, the opposite is verified, patterns with positive edge product appear to be favored, while negative product squares appear deprecated, in accordance with what expected for social networks.

These results are a first step to tackle a very important issue in Network Science, link prediction, in the context of signed biological networks.

1 Introduction

A network consists on a set of vertices, V, and a set of edges, E, where each edge connects two vertices. When the edges have no direction, but connect the nodes without distinguishing a source and a target node, the network is called undirected. Furthermore, a network may have both positive and negative links, corresponding to positive and negative interactions between the elements of the system. In this case it is called a signed network, and each edge is represented as $e = (u, v, s)$ where $e \in E$; $u, v \in V$ and $s \in \{+, -\}$. The degree of a node is the number of edges connect to it.

Network theory is a powerful tool to describe and study complex biological systems. In fact, the study of signed networks' structure allowed the delimitation of modules in gene co-expression networks in embryonic stem cells [1], the prediction of gene function in yeast genetic interaction networks [2] and even predicting previously unknown drug synergies [3].

The structural organization of signed networks has been extensively studied in social sciences and and the distribution of 3-node motifs can be described by the Structural Balance (SB) theory. The strong formulation of SB states that only the configurations of signed triangles with a positive product of the edges are stable, whereas the weak formulation states that the configuration with one negative signal is unbalanced.

Local patterns, such as 3-node signed motifs and 4-node unsigned, have proved useful for the prediction of new interactions in both social [4] and biological networks [5, 6], respectively. The biologists from the Boone Lab, University of Toronto, working with signed genetic interaction networks, are very interested in a link prediction algorithm. However, it seems there is a gap in the literature regarding the structure of undirected signed biological networks, and more specifically on which are the preponderant local patterns - motifs.

The goal of this project is to perform a thorough statistical analysis of the motifs present in yeast genetic interaction networks, with the aim of gaining some insight on the principles behind its structural organization and to guide future link-prediction algorithms in signed biological networks.

2 Datasets

We will study the genetic interaction data constructed in [2], obtained using synthetic genetic array analysis, a high-throughput technique to quantitatively screen genetic interactions. A negative interaction is identified when a double mutant displays a fitness defect that is more extreme than expected. Conversely, a positive interaction describes a double mutant exhibiting a fitness greater than expected. Negative interactions connect functionally related genes while positive interactions map general regulatory connections. We construct a essential genetic interaction network (ExE) from genes indispensable to the yeast viability, and a non-essential genetic interaction network (NxN).

There are different protocols to generate a network from the raw data. We constructed two slightly different protocols, _31 and _41, and also used protocol _7 from the Boone Lab, all the protocols are described in detail in the dissertation. The filtering of the data to generate a network can be performed at a stringent confidence, yielding a reduced number of false positives at the cost of augmented false negatives, at an intermediate confidence or at lenient confidence. In [2] and other papers from the Boone Lab, networks are thresholded at intermediate confidence, found to be the biologically more meaningful cutoff.

3 Randomization model

We wish to find which structural patterns are typical, or statistically more frequent, in the networks. However, to make a judgment on the statistical relevance of a motif, we need a null model to which we can compare the network to assess. The null model is an ensemble of networks preserving some property of the original one (e.g. keeping the degree distribution), and then compare the characteristic in study (e.g. frequency of a specific motif) in the original network against the ensemble. If the characteristic of the original network is not replicated in the ensemble, then it must be consequence of some aspect other than the preserved structural property.

Following [7], we consider the signed-degrees, i.e. the number of positive and negative edges connecting a node, of vital importance for the functional characterization of each node. Thus, we chose the signed-degrees as the characteristic to preserve. After a thorough literature review, we found no model preserving the signed-degrees on undirected networks. So we built ourselves a signed-degree preserving method, however, the synthetic networks generated often contained parallel edges, which have no biological meaning in the

context of genetic interaction networks. We ended up using a still unpublished method, kindly provided by István Kovács, from Northeastern University, which keeps the network framework unchanged and shuffles the signs, preserving, on average, the signed-degree distributions on the ensemble.

4 Results

All the results presented in this section used the randomization model mentioned in the previous section to generate the random populations with 100 synthetic networks. The measure used to evaluate each signed motif is the Z score, the number of standard deviations a value is from the mean. If the frequency of a motif in the input network is more than 2 standard deviations above the population average, we consider that motif to be over-represented in the network, and highlight its Z score in green. Similarly, if the frequency of a motif in the input network is more than 2 standard deviations below the population average, we consider that motif to be under-represented in the network, and highlight its Z score in red. Z scores between -2 and 2 are considered as no signal, and highlighted in gray. These motifs, neither over- nor under-represented, are said to be balanced.

The different signed motifs studied are presented in Fig. 1. The 3-node motifs in the top row and 4-node motifs in the bottom row.

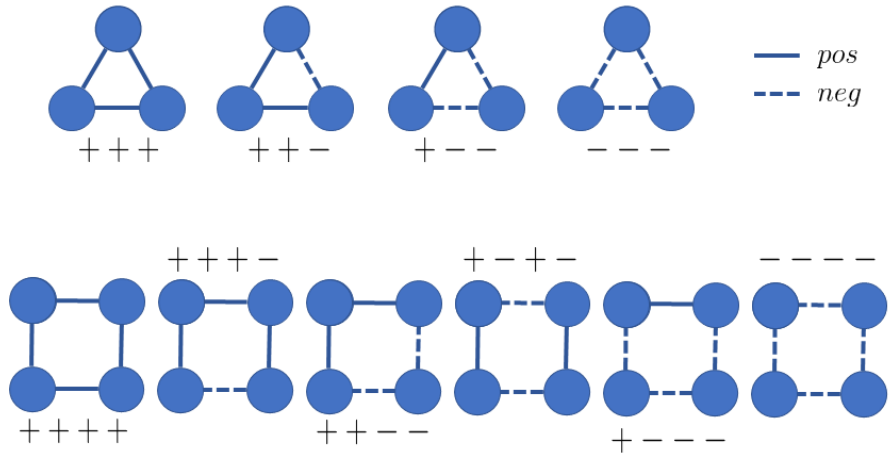

Figure 1: 3 and 4-node patterns

A summary of the Z scores computed for 3-node signed motifs in the essential and non-essential networks can be found in Table 1, the networks were constructed using each of the filtering protocols with intermediate threshold to test the consistency of the motifs distribution among protocols For method _7, we also tested stringent (ExE_7 s) and lenient (ExE_7 l) thresholds, to evaluate the variation of the results with the different false positive and false negative rates. We also present the expected outcomes for the strong and weak formulations of social Structural Balance.

For a biological interpretation of the individual Z scores please refer to the dissertation.

With respect to NxN, it appears that there are some faint principles underlying the organization of the genetic interaction network. However, this structural organization vanishes with fluctuations of the false positive and false negative rates. Therefore, we consider that there is no evidence to support an empirical formulation for the 3-node Balance in the non-essential genetic interaction network.

Table 1: Balance for signed triangles

Z score	+++	++ −	+ − −	− − −
ExE_31	-5.1	6.7	-19	8.6
ExE_41	-5.7	7.8	-32	16
ExE_7	-5.6	7.4	-37	15
ExE_7 s	-6.3	0.1	-19	13
ExE_7 l	-4.8	12	-20	9.7
SB strong				
NxN_31	1.6	2.4	-18	3.9
NxN_41	1.5	2.0	-19	5.2
NxN_7	1.6	2.2	-18	3.7
NxN_7 s	-0.8	-4.2	2.8	-0.1
NxN_7 l	-2.7	4.7	-4.5	1.3
SB weak				

When we compare the Z scores obtained for the 3-node motifs in ExE, with what is expected by the strong formulation of SB, we see that these are inverted. In fact, the motifs predicted to be over-represented by SB ('+ + +' and '+ − −') are under-represented in the network, whereas the unstable motifs according to SB ('+ + −' and '− − −') are the more prevalent! This suggests that the characteristics we attribute to the '+' and '−' signs in social networks might be reversely applied to the genetic '+' and '−' interactions.

In social networks, "positive ties are more likely to be clumped together, while negative ties tend to act more like bridges between islands of positive ties" [8]. In the genetic networks, however, genes related to the same pathway form clusters with a very high density of negative interactions, whereas the majority of positive interactions connects genes from different clusters. In short, both friendship and negative genetic interaction are mainly an intra-cluster bond, while enmity and positive genetic interaction tend to be a connection inter-clusters.

Hence, to correctly define the rules of balance in genetic networks, for triangles, one must invert the strong formulation of Structural Balance. This is, to the best of our knowledge, a completely new result for both the Biology and Network Science communities!

Similarly, we computed the Z scores for each of the 4-node signed motifs. Protocol _41 is not shown, due to problems in the randomization, and "NxN_7 l" was not computed due to time restrictions (for more details, refer to the dissertation).

Despite the long known differences between the essential and non-essential networks, such as network density and many other functional divergences [2], we would like to emphasize the congruence of their signed square distributions between ExE and NxN, i.e. there was not a single motif simultaneously over-represented for some network and over-represented in another.

In conclusion, we sum up the motif distributions observed to propose a Genetic Structural Balance. Regarding the signed triangles, we verified that the essential genetic networks followed the reverse of the strong formulation of social Structural Balance (SB social), whereas the distributions in the non-essential networks, thresholded with intermediate confidence, followed the reverse of the weak formulation. However, since the latter results were not consistent for the different confidence thresholds, we will not propose a Structural Balance for the non-essential network (SB N) on the 3-node motifs, but only for the essential network (SB E). The signed triangle results are summarized in table 3.

Table 2: Balance for signed squares

Z score	++++	+++−	++−−	+−+−	+−−−	−−−−
ExE_31	-1.2	-3.0	8.2	11	-2.2	1.8
ExE_7	0.04	-4.1	11	10	-6.0	4.3
ExE_7 s	-0.4	-5.0	0.1	4.6	-12	5.4
ExE_7 l	-0.1	-2.5	106	13	-3.7	1.3
NxN_31	4.9	-0.3	13	13	-12	-0.5
NxN_7	4.4	-0.4	12	13	-10	-0.4
NxN_7 s	11	-6.9	2.9	10	-5.4	0.8

Table 3: Genetic Structural Balance: triangles

	+ + +	+ + −	+ − −	− − −
SB social				
SB E				
SB N	xxx	xxx	xxx	xxx

We emphasize one more time that the need to reverse the signs of the classic SB to match the genetic Balance is, to the best of our knowledge, a completely new result.

As for the squares, the empirically verified distributions for the genetic networks can be found in table 4. We also included the SB social, considering positive edge product squares as over-represented and negative ones as under-represented. Now there is consistency between the genetic and social SB's, without having to reverse signs, due to the even number of edges in the squares.

Table 4: Genetic Structural Balance: squares

	+ + + +	+ + + −	+ + − −	+ − + −	+ − − −	− − − −
SB social						
SB E						
SB N						

Despite the differences between the essential and non-essential empirical rules, it is important to notice that there is never direct contradiction. There is no motif simultaneously over-represented in one formulation and under-represented in the other. Thus, in the future, it may be possible to merge both definitions of the genetic SB, with broader studies including genetic networks of organisms other than the yeast.

5 Conclusions

The main result of the present work is the empirical derivation of rules of Balance for the genetic interaction networks of the budding yeast. We found that, for most networks, triangles with a positive edge product are stable and thus over-represented in the network structure, when compared to a null model, while triangles with a negative edge product are deprecated. The only exception was the non-essential network, when imposed a strict false positive rate on the interactions.

This means that, to correctly define the rules of balance in genetic networks, for triangles, one must invert the signals in the strong formulation of Structural Balance, a very well known and long-standing theory for balance in social networks. This is, to the best of our knowledge, a novel result for both the Biology and Network Science communities.

As for the squares, 4-node motifs with a negative edge product were never over-represented, when compared to the null model, and thus are probably unstable and deprecated by the network structure. Squares with a negative edge product were never found to be under-represented, and thus appear to be stable or balanced. These results match the ones expected in social networks.

One important motivation of this study was to find out which type of motif, triangles or squares, would reveal more informative for future sign prediction algorithms.

On one hand we have triangles, the 3-node signed motifs in the essential network presented not only the more stronger Z scores but also the greater consistency between different thresholding confidences and different filtering protocols. The latter are important arguments since biological data is frequently noisy, and biological networks have a big variety of false negative and false positive rates. However, in the non-essential network, the Z scores of the triangles yielded contradictory results, depending on the thresholding confidence. Thus, we find 4-node motifs a more useful tool for sign prediction.

Squares have the initial advantage of being the chosen pattern in the state-of-the-art link-prediction algorithms in biology [5, 6], since sign prediction will most likely be associated with link-prediction in signed networks. Although there is some variation on the intensity of the Z scores signals with the filtering, which may difficult the choice of each motifs' weight in a putative sign prediction algorithm, there is a great consistency in the qualitative results (over-represented, balanced or under-represented motif). Even though, as also for the triangles, some structural differences between the essential and non-essential genetic networks have been detected, these were never contradictory. Finally, a third argument is that, due to its even parity, there is no need to worry about how the edge signals were defined. A useful characteristic when applying the algorithm to a network of which we have little *a priori* information.

As motivated in the beginning of the project, biologists are interested in link-prediction algorithms to either guide their experimental assays and reduce costs, or to make accurate predictions in the search space currently inaccessible due to fundamental technical limitations. Our result is a first step in the way of designing a new successful link-prediction algorithm for genetic interaction networks.

We will continue to develop this project, in partnership with István Kovács, from the Northeastern University, in Boston, and in dialogue with the Boone Lab, from the University of Toronto, with the aim of publishing an article on the structural organization of biological networks and its correlation to function.

References

[1] Mason, M. J., et al. *Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells*, BMC Genomics **327** (2009), 327-51.

[2] Costanzo, M., et al. *A global genetic interaction network maps a wiring diagram of cellular function*, Science **353(6306)** (2016), 31381.

[3] Li F. et al. *Network-based computational drug combination prediction*, bioRxiv (2016).

[4] Liben-Nowelland, D. and Kleinberg, J. *The link-prediction problem for social networks*, J. Am. Soc. Inf. Sci. Tec. **58(7)** (2007), 1019-31.

[5] Kovács, I. A. et al. *Network-based prediction of protein interactions*, bioRxiv (2018).

[6] Muscoloni, A., Abdelhamid, I. and Cannistraci, C. *Local-community network automata modelling based on length-three-paths for prediction of complex network structures in protein interactomes, food webs and more*, bioRxiv (2018).

[7] Iorio, F. et al. *Efficient randomization of biological networks while preserving functional characterization of individual nodes*, BMC Bioinformatics **74**, (2016) 542-55.

[8] Leskovec, J., Huttenlocher, D., and Kleinberg, J. *Signed networks in social media*, in ACM SIGCHI Conf. Hum. Fact. Comp. Syst. (2010), 1361-70.