Comparative topological analysis of the Escherichia coli and Homo sapiens interactome

David Altamirano

(Dated: August 16, 2023)

The study of complex networks have become to a powerful tool for understanding any type of interaction in any field of science. In biology, networks are used, among other things, to study protein-protein interactions (PPI), where proteins are the nodes and interaction are the links. In the present work we use Escherichia coli and Homo sapiens interactomes to analyze the network topology. We explore the three most robust measures: average shortest path length, clustering coefficient and degree distribution. These measures allow to determine that networks do not tend to form dense clusters so diseases and information spread slowly. Also, the measures allow to determine that EC network is entirely scale-free while HS network has heterogeneous degree distribution but it is only scale-free when $k_{min} > 97$ and not for the entire network. Finally, the most important hubs of the networks was identified.

INTRODUCTION

In the current century the complex networks have been studied deeply in many different fields due to the importance for understanding the world from another point of view. The fields with more impact using networks are social, technological and biological systems, such as protein-protein interactions that will be studied in the current work. In this context, the network is made by nodes which correspond to the protein itself and by edges that correspond to the interaction between proteins. The most common networks models are the Erdős–Rényi (ER) model [1] which states that there are N nodes in the network connected between them with a probability p, in such a way that the every connection is independent. The second important model is described by Barabási–Albert (BA) [2] which explain the evolution of free-scale networks. It means that the probability $P(k)$ that a node interacts with k other nodes of the network decays as a power law, $P(k) \sim k^{-\gamma}$. The model is based on two main features, growth and preferential attachment mechanism. It states the idea that as the network grows, the probability that new nodes will connect with existing nodes is proportional to the degree of the existing node. These two ingredients play an important role in the formation of many real complex systems.

A well-documented example of a biological network is the protein-protein interaction (PPI) inside cells. PPIs is essential for understanding cell physiology both in normal and disease states [3], as well as drug development. The totality of interactions that happen in a cell is known as the “interactome” [4]. Large-scale PPI screening techniques as the yeast two-hybrid assay [5, 6] or affinity purification combined with mass-spectroscopy [7] have allowed the construction of ever more complex and complete interactomes. Despite this, our knowledge about the interactome is still incomplete. However, there are enough data sets to study the protein-protein interaction networks (PPIN). Networks can show the small world effect meaning that there is great connectivity between proteins or a small network's diameter. Besides that, networks also can show scale-free behaviour where the majority of the proteins have only a few connections, whereas some hub proteins are connected to many other proteins. Hubs might contribute to robustness of the network and can be 'party' and 'date' hubs [8], where party hubs interact with most of their neighbours simultaneously, while date hubs bind their different neighbours at different times or locations. Wang et. al. have reviewed the importance to analyze the PPIN for biological mechanisms [9].

The goal of this work is analyze the protein network of Escherichia coli and Homo sapiens organisms in order to identify hub proteins, and compare their network topology using a python script. The comparison will finish with a brief analysis to determine if each network is scale-free, it means that present a power law behaviour. For this, in Section II is described the methods employed for the analysis and comparison of the network topology. Section III shows the results obtained in a detailed way. Section IV summarizes the work with a clear conclusion.

METHODS

In this work, we have used an interactome-wide set of protein interaction obtained from Interactome INSIDER [10] genomic information center. Specifically, we have downloaded the data sets of Escherichia coli, Fig. 1 (above), and Homo sapiens, Fig. 1 (below), organisms.

In order to perform the study of the structure and functions of complex networks we employ the *networkx* package of Python. We have to take into account that the networks are unweighted and undirected. The first network feature that we compute is the density which is the ratio between the actual interactions number with respect to the maximum number of possible interactions and is given by,

$$D = \frac{M}{\frac{N(N-1)}{2}} = \frac{2M}{N(N-1)}, \quad (1)$$

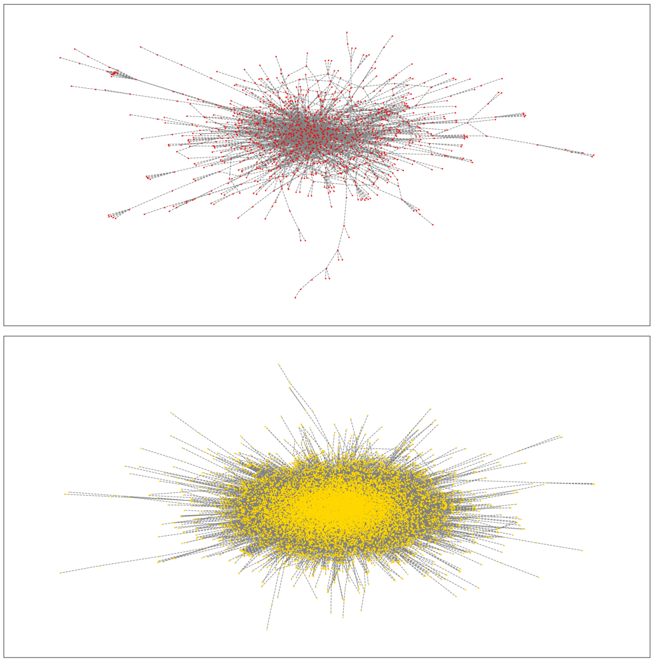

FIG. 1. Above: Escherichia coli network where red dots are the proteins and gray dashed lines are the interactions. Below: Homo sapiens network where yellow dots are the proteins and gray dashed lines are the interactions.

where M is the actual number of interactions and N is the total number of proteins. Also, we compute the average shortest path length which measure the efficiency of information transport on the network. The diameter which indicates the longest path of all the shortest paths in the network. The clustering coefficient measures the degree of a protein tend to cluster with other proteins in the network and is given by,

$$C_i = \frac{2e_i}{k_i(k_i - 1)}, \tag{2}$$

where e_i is the number of links between the neighbours of protein i and k_i is the degree (number of connections) of protein i. The clustering coefficient of the whole network is the average of clustering coefficients of all the nodes. A high clustering coefficient indicates a small world behaviour. The betweenness centrality shows the amount of influence a node has over the information flow in the network and it is calculated by,

$$b(s,t,v) = \sum_{s \neq t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{3}$$

where $\sigma_{st}(v)$ is the number of shortest paths from source s to target node t through the *between* node v and σ_{st} is the number of shortest paths from source s to target node t. Assortativity [11] is the preference of nodes to interact with other similar nodes, namely the hubs (nodes with high degree) tend to link each other and small-degree nodes tend to link to other small-degree nodes. The assortativity coefficient is the Pearson correlation coefficient of degree between pairs of linked nodes. Another way to show the assortativity is using the average degree connectivity or k *nearest neighbours*, which is expressed as follow,

$$k_{nn} = \sum_{k'} k' P(k'|k), \tag{4}$$

where $P(k'|k)$ is the conditional probability starting from a link of k degree reaching a node of k' degree. Applying the logarithm, the above equation can be written as,

$$k_{nn}(k) = Ck^{\mu}, \tag{5}$$

where μ is the correlation exponent. If $\mu > 0$ the network is *assortative*, if $\mu < 0$ the network is *disassortative*, while $\mu = 0$ is a *neutral* network.

Finally, we study, probably, the most important feature of the network topology, the degree distribution. A degree is the number of interactions that a protein has with other proteins, then the degree distribution is a probability distribution, $P(k)$, of these degrees, k. The distribution can be described by an uniform exponential topology, namely on average the proteins posses the same number of links, or by a scale-free topology with proteins having widely different connectivities following a power law,

$$p_k \sim k^{-\gamma}. \tag{6}$$

If we randomly choose a node and has $2 < \gamma < 3$, then is considered as *scale-free*. In other words, a network is scale-free if the vast majority of nodes have very few connections, while a few hubs have a huge number of connections. A similar study was done by Jeong et. al using *S. cerevisiae* organism [12]. Or a review by Typas et. al for bacterial networks [13].

RESULTS

In the following section we are going to present the results. The Escherichia coli network is made of 1207 nodes and 2133 edges where the density is 0.00293, the average shortest path length is 5.26 and the diameter is 16. While Homo sapiens network is made of 14709 nodes and 117268 edges where the density is 0.00108, the average shortest path length is 3.64 and the diameter is 10. In both cases we have low density networks with a relatively small average shortest path length which indicates a high connectivity and an efficient transport of the information. It could suggests the existence of hubs detailed the 10 most important in Table I. Also we can realize that in the case of Homo sapiens network which is larger, the connections between nodes are shorter and faster.

Index	EC Protein	EC degree	HS Protein	HS degree
1	P75679	65	P19320	627
2	P0ACL5	56	P08238	523
3	Q46864	52	P62993	498
4	P68646	48	Q86SX1	457
5	Q79E92	43	P02751	454
6	P45577	42	P78362	442
7	P30750	41	P63279	431
8	P0ACJ8	32	P00533	380
9	P28638	32	P68400	346
10	P0A805	31	Q08379	344

TABLE I. The first 10 proteins with the highest number of interactions of Escherichia coli and Homo sapiens networks.

Regarding to the global clustering coefficient, for EC network the coefficient is 0.0712 and for HS network is 0.0914. A low global clustering coefficient suggests that nodes of the networks do not tend to form dense clusters, namely that the topology of the network is not as small-world but is disperse. In relation with the betweenness centrality, a low global clustering coefficient means that the shortest paths may not necessarily traverse dense connections. In Table II is described the proteins with the highest betweenness centrality coefficient of both networks. If we compare with the Table I, we can realize which proteins are really important. In the case of EC network, the proteins P0ACL5 and P75679 are the most important. While in the HS network, the most important proteins are P19320 and P62993. In the next section we are going to discuss its biological importance.

Index	EC Protein	EC coeff.	HS Protein	HS coeff.
1	P0ACL5	0.1396	P19320	0.0293
2	P75679	0.1303	P62993	0.0269
3	Q46864	0.1161	P78362	0.0225
4	P30750	0.0765	P08238	0.0222
5	P28630	0.0762	P63279	0.0218
6	P45577	0.0656	P02751	0.0200
7	P39409	0.0613	P00533	0.0185
8	P68646	0.0612	Q86SX1	0.0151
9	P0A805	0.0556	P54253	0.0140
10	Q79E92	0.0518	P38398	0.0138

TABLE II. The first 10 proteins with the highest betweenness centrality coefficient of Escherichia coli and Homo sapiens networks.

To further quantify and compare correlation patterns, we calculated the assortativity or the average connectivity of nearest neighbors of a node, as a function of its own connectivity (Fig. 2). For the two studied networks, the average connectivity shows a gradual decline, which can be fitted with a power law, $k_{nn} \propto k^{-0.218}$ for EC network and $k_{nn} \propto k^{-0.170}$ for HS network. In both cases the network is disassortative which means that the hub proteins tend to link with small-degree proteins. This is another evidence that these networks do not tend to form dense clusters while the information can travel through all the nodes but on a slow way, such that if a disease appears the propagation takes time.

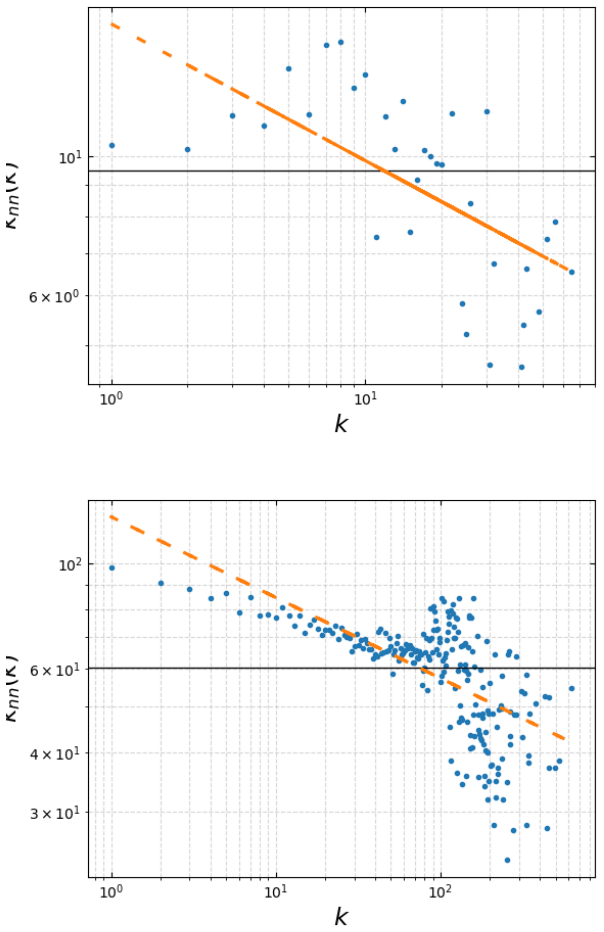

FIG. 2. Correlations in connectivities of neighbors. The blue dots are the average connectivity of nearest neighbors of a protein with the connectivity in the physical interaction. The solid black line correspond to a neutral behaviour. The dashed orange line is a power law fit, $\propto k^{-0.218}$ for Escherichia coli network (above) and $\propto k^{-0.170}$ for Homo sapiens network (below).

Until this point, the above studied features indicate that the networks look like scale-free networks because they do not tend to form dense clusters. In order to answer concretely we calculate the degree distribution which follows a power law in both networks (Fig. 3). For EC, the γ parameter is 2.827 ± 0.154 which is less than 3, then the entire network is scale-free as the above results suggested. This is supported when we calculate k_{min} which indicates that since $k_{min} > 7$ the networks behaves as power law, namely the most of the network. On the other side, the HS network shows $\gamma = 3.583 \pm 0.141$ which is greater than 3, then the entire network is not scale-free despite above results. When we compute the minimum k, we realize that only for $k_{min} > 97$ the network is scale-free. However the network is still heterogeneous.

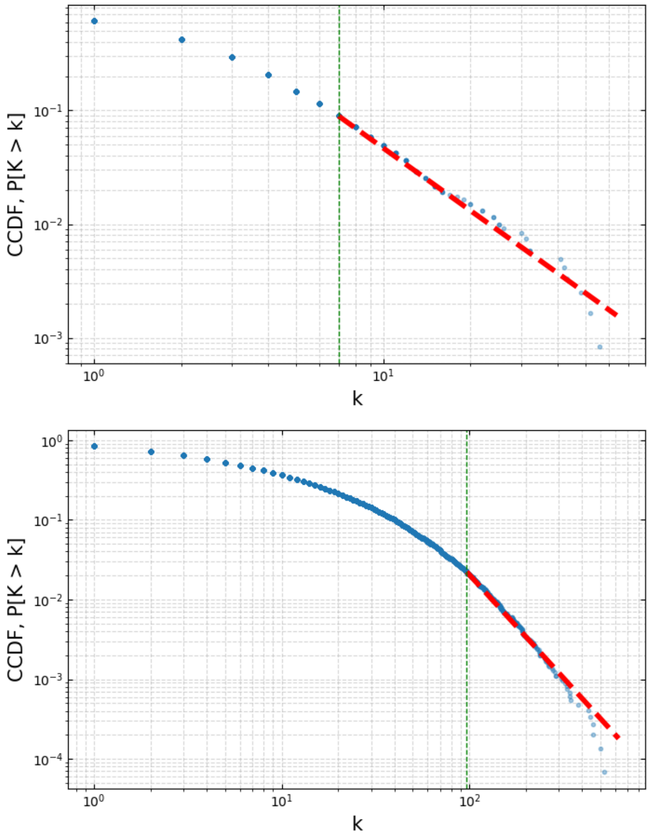

FIG. 3. Degree distribution. The y-axis correspond to the cumulative distribution function of the degree k, for $k > 0$. The blue dots are the observed data. The green vertical line represents the minimum k from where power law behaviour starts. The red line is the power law fit. For EC network (above) $k_{min} > 7$ and $k^{\gamma} = k^{2.827 \pm 0.154}$. And for HS network (below) $k_{min} > 97$ and $k^{\gamma} = k^{3.583 \pm 0.141}$.

DISCUSSION AND CONCLUSION

For the present work, the topology of the networks have been studied. We explore the betweenness centrality which is a way to determine the hub proteins. Identify hubs are important to understand pathways, regulatory mechanisms, and diseases. In the above section we identified the most important proteins of the networks, now we are going to discuss its importance. For this we use UniProt [14]. The protein P0ACL5 is a transcriptional activator and also negatively regulates the transcription of its own gene. While P75679 is involved in the transposition. And the common characteristic is that both protein are related with the DNA-binding. In the case of HS the most important protein is P19320 which is Vascular cell adhesion protein 1 and, how its name indicates, has a big importance for the cell adhesion predominantly expressed on the surface of endothelial cells that plays an important role in inflammation [15]. The other hub is P62993 which corresponds to growth factor receptor-bound protein 2. This is an adapter protein that provides a critical link between cell surface growth factor receptors and the Ras signaling pathway [16].

For conclude the work, a brief summary. We explore the three most robust measures for determining the network topology: average shortest path length, clustering coefficient and degree distribution using Escherichia coli and Homo sapiens networks obtained from INSIDER Interactome. Both cases present a relatively small average shortest path length which suggest a high connectivity. As well as low global clustering coefficient and negative assortativity (dissasortative network) which suggest that the network do not tend to form dense clusters. The information can flux to all proteins but on a slow way. In general, both networks present features of scale-free that is completely cleared with the degree distribution analysis. The EC network presents $2 < \gamma < 3$ for consequence is an entire scale-free network. While for HS $\gamma > 3$, then the network is not scale-free but its degree distribution is still heterogeneous i.e. many nodes with few links and few hub nodes connected with the most of the other nodes. However the network presents a power law behaviour for $k_{min} > 97$.

For further works we can study the exact proteins in the HS network which follow a power law.

[1] Erdos, P.L., and Rényi, A. On the evolution of random graphs. Transactions of the American Mathematical Society. 1984; 286, 257-257.

[2] Barabasi, A.L. and Albert, R. Emergence of scaling in random networks. Science. 1999;286(5439): p. 509-12.

[3] Ideker, T. and Roded Sharan, R. Protein networks in disease. Genome Res. 2008. 18: 644-652.

[4] Koh, G., Porras, P., Aranda, B., Hermjakob, H. and Orchard, S. Analyzing Protein–Protein Interaction Networks. J. Proteome Res. 2012, 11, 4, 2014–2031.

[5] Rual JF, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein-protein interaction network. Nature. 2005 Oct;437(7062):1173-1178.

[6] Uetz, P., Giot, L., Cagney, G. et al. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Nature. 2000;403, 623–627.

[7] Qin G, Ma J, Chen X, Chu Z, She YM. Methylated-antibody affinity purification to improve proteomic identification of plant RNA polymerase Pol V complex and the interacting proteins. Sci Rep. 2017 Feb 22;7:42943.

[8] Han, JD., Bertin, N., Hao, T. et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. Nature. 2004; 430, 88–93.

[9] Wang, S., Wu, R., Lu, J., Jiang, Y., Huang, T., and Cai, Y.-D. Protein-protein interaction networks as miners of biological discovery. Proteomics. 2022; 22, e2100190.

[10] Interactome INSIDER: a structural interactome browser for genomic studies. MJ Meyer, JF Beltrán, S Liang, R Fragoza, A Rumack, J Liang, X Wei, H Yu - Nature Methods, 2018.

[11] Newman, M. E. J. Mixing patterns in networks. Physical Review E. 2003;67 (2): 026126.

[12] Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001 May;411(6833):41-42.

[13] Typas, A., and Sourjik, V. Bacterial protein networks: properties and functions. Nature Reviews Microbiology. 2015; 13(9), 559–572.

[14] The UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023, Nucleic Acids Research, Volume 51, Issue D1, 6 January 2023, Pages D523–D531.

[15] Li Y, Huang X, Guo F, Lei T, Li S, Monaghan-Nichols P, Jiang Z, Xin HB, Fu M. TRIM65 E3 ligase targets VCAM-1 degradation to limit LPS-induced lung inflammation. J Mol Cell Biol. 2020 Apr 24;12(3):190-201.

[16] Pao-Chun L, Chan PM, Chan W, Manser E. Cytoplasmic ACK1 interaction with multiple receptor tyrosine kinases is mediated by Grb2: an analysis of ACK1 effects on Axl signaling. J Biol Chem. 2009 Dec 11;284(50):34954-63.