Comparison of proteins interactions graphs in different species
Final report

D'Alberton Enrico, Girotto Pietro, Qiu Yi Jian

November 26, 2024

1 Abstract

Protein-protein interactions (PPI) play a crucial role in cellular functions and influence the evolutionary trajectories of organisms. This report employs graph networks to explore and compare PPI graphs across six different species. Using key metrics such as closeness centrality, betweenness centrality, and clustering coefficients, the study identifies interesting proteins and examines their significance in diverse PPI networks. The analysis, conducted on public datasets using NetworkX reveals interesting patterns in averaged metrics, z-scores and top proteins for each species. Notably, the closeness and betweenness centrality, in relation to the z-score, suggest the presence of densely connected subregions. Finally, we found out that some of the top betweenness proteins play a crucial role on the biological functions of the species.

2 Introduction

Protein to protein interactions (PPI) are pivotal for cellular functions, influencing the evolutionary trajectory of organisms. This study employs graph networks to explore the evolutionary relationships and functional roles of proteins across six different species. Focusing on key metrics, we aim to identify recurrent proteins and discern their significance in diverse PPI graphs. While avoiding intricate biological pathways due to limited expertise, we concentrate on recurrent proteins and their functions. Our analysis prioritizes the top 5 proteins for each graph and metric used considering closeness centrality, betweenness centrality, clustering coefficients, as well as the z-scores with regard to generated random graphs for the averaged values of those metrics across the whole network.
From a technical perspective, we gathered different graphs from the Graph Web public datasets [2], run the before mentioned analysis on a commodity PC and made some suggestions on why proteins might be recurrent or not based on biological literature.
The code repository used for this analysis can be found in the references [4].

3 Datasets

We decided to investigate betweenness, closeness and clustering coefficient on 6 different living species: Homo Sapiens, Saccharomyces cerevisiae (Yeast), Rattus norvegicus (Brown Rat), Bos taurus (Cattle), Caenorhabditis Elegans (Worm), Mus Musculus (Mouse). The datasets in detail are available here:

- **Homo Sapiens**: 8077 nodes and 26590 edges [URL];
- **Saccharomyces cerevisiae**: 5718 nodes and 48259 edges [URL];
- **Rattus norvegicus**: 1375 nodes and 1671 edges [URL];
- **Bos taurus**: 268 nodes and 306 edges [URL];
- **Caenorhabditis Elegans**: 3274 nodes and 5736 edges [URL];
- **Mus Musculus**: 2929 nodes and 4328 edges [URL];

4 Methods

The methods used to get the analytical results were grouped into the following macro-areas:

- **Graph Data Representation**: the analysis employed the NetworkX library [3] to represent protein-protein interaction (PPI) networks as graphs. Each interaction was modeled as an edge between protein nodes.
- **Metric Calculation**: various metrics were calculated to assess the importance of proteins within PPI graphs. Closeness centrality, betweenness centrality, and clustering coefficients were computed using NetworkX's built-in functions.
- **Averaged Metrics and Top Proteins Extraction**: average closeness, average betweenness, and global clustering coefficients were computed to provide an overview of the entire PPI network. Top 5 proteins for closeness and betweenness metrics were extracted based on their centrality values.
- **Z-score Calculation**: z-scores were computed to evaluate the significance of the averaged metrics for the graphs compared to randomly generated graphs. The mean and standard deviation of metric values from 10 random graphs were used to standardize the previously averages obtained.
- **Parallel Processing**: parallel processing, facilitated by the `concurrent.futures` module [5], expedited the analysis of multiple PPI graphs concurrently. This approach enhanced the efficiency of the study, especially when dealing with large datasets.

5 Experimental Results

Species	Avg Closeness	Avg Betweenness	Global Clustering
Cattle	0.081822	0.001520	0.133686
Rat	0.208850	0.001403	0.065791
Mouse	0.172893	0.001027	0.100123
C.elegans	0.183504	0.000990	0.026928
Yeast	0.308946	0.000396	0.123421
Human	0.227639	0.000376	0.083431

Table 1: Averaged metrics per graph

Species	Z-scores - Closeness	Z-scores - Betweenness	Z-scores - Clustering
Cattle	-0.729450	-0.672724	29.947760
Rat	2.875699	-0.473574	95.244507
Mouse	1.694762	-0.396655	226.686354
C.elegans	1.366033	-0.351993	67.849667
Yeast	1.349070	-0.093678	1157.226873
Human	2.229288	-0.310289	644.498301

Table 2: Z-scores

Species	**Top 5 Closeness**	**Top 5 Betweenness**
C. elegans	GEI-4 (0.305) GEI-16 (0.294) MIG-5 (0.292) ATN-1 (0.291) ALP-1 (0.289)	GEI-4 (0.127) ATN-1 (0.075) GEI-16 (0.063) PAL-1 (0.057) K09B11.9 (0.051)
Cattle	AATM_BOVIN (0.270) ACADV_BOVIN (0.238) NDUA2_BOVIN (0.210) ATPB_BOVIN (0.203) ATPA1_BOVIN (0.201)	AATM_BOVIN (0.131) ACADV_BOVIN (0.059) ACPM_BOVIN (0.050) FREQ (0.030) ATPB_BOVIN (0.023)
Human	GRB2 (0.350) IKBKE (0.339) TRAF6 (0.335) EBI-397435 (0.334) ARRB2 (0.333)	GRB2 (0.097) TRAF6 (0.045) IKBKE (0.042) ARRB2 (0.040) EBI-397435 (0.037)
Mouse	YWHAB (0.338) GRIN2B (0.292) GRIN1 (0.291) YWHAZ (0.283) KCNMA1 (0.283)	YWHAB (0.445) KCNMA1 (0.099) GRIN2B (0.076) YWHAZ (0.064) PRKCE (0.062)
Rat	SLC2A4 (0.433) ACTB (0.334) DYL1_RAT (0.326) TUBA1A (0.323) ATP5O (0.320)	SLC2A4 (0.606) EBI-2257702 (0.107) TNF (0.103) EBI-1638146 (0.045) HNRPK (0.044)
Yeast	SSA2 (0.477) SSB1 (0.457) EBI-6314 (0.451) VMA2 (0.435) UBC7 (0.426)	SSA2 (0.113) UBC7 (0.106) SSB1 (0.055) EBI-6314 (0.052) JSN1 (0.044)

Table 3: Top 5 Closeness and Betweenness Proteins with Values for Each Species

6 Analysis

Based on the provided graph analysis results, several key observations can be made across the different species. Therefore, they have been divided by topic.

6.1 Averaged Metrics

The average betweenness values are relatively comparable across the graphs, indicating a similar level of node centrality. Closeness follows a similar patter, with the exception of the Cattle graph which has a notably lower averaged value, perhaps suggesting a different network structure. Surprisingly, the global clustering coefficients show little variance across the different species, despite their diverse nature.

6.2 Z-Scores

The z-scores for closeness and betweenness reveal interesting patterns. While closeness does not exhibit significant deviations, the z-scores for betweenness are notably low, indicating a resemblance to random graphs.
The Cattle graph has a very low z-score for closeness, this might implicate that the graph has no particular structure and the value is lower than the others due to the high nodes to edges ratio. In contrast, the z-scores for global clustering are remarkably high, especially for Mouse, Yeast, and Human. This implies a high propensity for proteins that share a common protein to interact with each other.
We can suggest that proteins which have a shared interaction will interact more likely. This is counter intuitive given the low z-scores for the other two metrics.
This could be due to the presence of highly connected sub regions which do not share a great number of connections between each other.

6.3 Top 5 proteins

Starting from the obvious, no protein is repeated from one specie to the other.
The top 5 closeness values demonstrate a consistent similarity among the species, with relatively high values compared to the average. On the other hand, the distribution of the top 5 betweenness values is intriguing, with the first node often exhibiting a betweenness approximately 1000 times higher than the average, while the subsequent nodes in the top 5 display much lower values.
This is particularly interesting as it supports the "dense regions" theory previously illustrated. The proteins with high closeness centrality might resemble the centers of those dense regions, whereas the top protein for betweenness is probably a graph-hub to which most dense regions are connected.

6.4 Specific Proteins Analysis

As suggested earlier, the top proteins based on betweenness centrality might illustrate a very important protein for such graph, given the dense sub regions supposition. We have therefore decided to briefly research each of them to give more context on their functions. Still, this is just a brief introduction to what we have found as software engineers. A proper interpretation of these data would require definitely more domain specific expertise.

- **GEI-4 (C. Elegans)**: the GEI-4 protein is a novel protein that has been implicated in the Wnt signaling pathway, which is crucial for asymmetric cell division in multicellular organisms [6];
- **AATM_BOVIN (Cattle)**: it is a protein that is found in the mitochondria of cells. It helps to maintain the balance of molecules inside the cell. It also helps to move molecules between the mitochondria and the rest of the cell, and is important for breaking down amino acids. Additionally, it helps to transport long-chain free fatty acids into the cell [9];

- **GRB2 (Human)**: it is widely expressed and is essential for multiple cellular functions. Inhibition of Grb2 function impairs developmental processes in various organisms and blocks transformation and proliferation of various cell types. [7] [8];
- **YWHAB (Mouse)**: it is a protein that helps cells communicate with each other. It does this by binding to other proteins that have a specific chemical tag on them. YWHAB also helps control the growth and development of cells by interacting with other proteins that are involved in the cell cycle. It can also help prevent cells from dying when they shouldn't. [10];
- **SLC2A4 (Rat)**: the primary function of SLC2A4 is to transport glucose across the cell membrane. This process is crucial for providing cells with the energy they need to function. [11]
- **SSA2 (Yeast)**: it helps in the folding of proteins, which is crucial for their proper function. [12];

7 Conclusions

In conclusion, the objective of our study on protein-protein interactions in six species was to underscore the role these networks play in cellular functions and evolutionary processes analyzing the top proteins of each network based on the key metrics we used to examine.
The research revealed key patterns and variations, providing insights into the structural and functional dynamics of these networks, in particular we can claim that there are few proteins, in each network, which have crucial role in the protein-protein interaction. Moreover the high values of clustering in Mouse, Yeast and Human networks, suggests that proteins which have a shared interaction will interact more likely.
Anyway, this investigation is not enough to fully understand the differences of interaction between proteins of the six species but could act as a foundation for collaborative efforts and future investigations.

8 Members' Contributions

Most of the work has been done together. As such it is quite difficult to clearly distinguish one's responsibilities from the others. Nevertheless, each of us was indirectly in charge of the following:

- **Yi Jian Qiu**: Reports' writer, revisionist
- **Enrico D'Alberton**: Code prototyping, parallelization.
- **Pietro Girotto**: Code polishing, ClusterDEI tester

The overall workload was evenly distributed (1/3 each). Most of the times we met together physically or virtually to accomplish tasks.

References

[1] *IntAct - Team* `https://www.ebi.ac.uk/intact/home`

[2] *Dataset - Source* `https://biit.cs.ut.ee/GraphWeb/welcome.cgi?t=examples`

[3] *NetworkX* https://networkx.org/

[4] *GitHub Repository* https://github.com/enricopro/learning_from_networks_unipd.git

[5] *concurrent.futures — Launching parallel tasks* https://docs.python.org/3/library/concurrent.futures.html

[6] *Investigating GEI-4, a DSH-2 interacting protein, in Wnt pathway regulation* https://theses.lib.sfu.ca/file/thesis/6939

[7] *Regulation of microRNA expression by the adaptor protein GRB2* https://www.nature.com/articles/s41598-023-36996-3.pdf

[8] *Giubellino, A. (2012). GRB2 Signaling as a Molecular Target for Cancer* https://doi.org/10.1007/978-1-4614-0730-0_1

[9] *Nucleotide sequence of a cDNA coding for bovine mitochondrial aspartate aminotransferase.* https://www.uniprot.org/citations/7641080

[10] *Nucleotide sequence of a cDNA coding for bovine mitochondrial aspartate aminotransferase.* https://www.proteinatlas.org/ENSG00000166913-YWHAB

[11] *Effects of short-term endurance and strength exercise in the molecular regulation of skeletal muscle in hyperinsulinemic and hyperglycemic Slc2a4 mice* https://link.springer.com/article/10.1007/s00018-023-04771-2

[12] *SSA2 / YLL024C Overview* https://www.yeastgenome.org/locus/S000003947