UNIVERSIDAD
POLITÉCNICA
DE YUCATÁN

BIS
Universities

Analysis of Protein-Protein Interaction Network in Yeast

Jair Armando Martinez Castillo
Project 1 - Social Network Analysis
2109165

Abstract

This document presents an analysis of the protein-protein interaction (PPI) network in yeast, *Saccharomyces cerevisiae*. The study focuses on understanding the network's structure, its functional implications, and potential applications in various biological and medical fields. We aim to map the complex interactions that govern cellular functions by utilizing high-throughput techniques and computational tools. The results highlight the importance of PPI networks in functional genomics, disease mechanisms, and drug discovery.

I. Introduction

The protein-protein interaction (PPI) network in yeast, *Saccharomyces cerevisiae*, serves as a model system for understanding the functional organization of the cell. These interactions are crucial for various biological processes such as signal transduction, cellular organization, and metabolic pathways. Mapping these interactions allows researchers to unravel the complex web of relationships that dictate cellular functions.

A. Importance of the PPI Network

PPIs play some critical roles in almost every aspect of cellular function. They facilitate the formation of protein complexes that are essential for numerous cellular processes. The yeast PPI network functions as a model system due to its relatively simple and well-characterized proteome. Insights obtained from studying yeast can often be extrapolated to more complex organisms, making it a valuable resource in biomedical research.

With all this in mind, it can be mentioned the distinct applications of PPI network analysis.

Some of the applications are fundamental when analyzing PPI networks. For example, predicting the function of uncharacterized proteins based on their interaction patterns. Also, disease mechanisms, which can be by understanding how disruptions in protein interactions can lead to diseases. Furthermore, drug target discovery, which can function as identifying proteins that can be targeted for therapeutic intervention. Finally, systems biology, can function by integrating PPI data with other omics data to build comprehensive models of cellular function.

B. Datasets and Literature

Regarding the network dataset and the references the main page has given, the yeast PPI network has been analyzed using a vast amount of high-throughput techniques. Which refers to tools that are used to analyze millions of samples for biological activity at the model organism, cellular, pathway, or molecular level.

One notable study by Ito et al. Employed the yeast two-hybrid system to explore approximately 6,000 proteins, identifying 4,549 interactions among 3,278 proteins [1].

Another comprehensive analysis by Uetz et al. used a different high-throughput approach, revealing a large network of 2,358 interactions among 1,548 proteins [2]. These studies have significantly expanded our knowledge of the yeast interactome, highlighting both the robustness and the limitations of current interaction mapping techniques.

C. Classification of the Network

The yeast PPI network is an undirected network because the interactions between proteins are mutual and do not have a specific direction. It can also be classified as a non-planar network, due to the complexity and density of the interactions that cannot be embedded in a plane without crossing edges.

D. Visualization

The visualization of the yeast PPI network reveals a densely connected core surrounded by sparser peripheral regions. This structure reflects functional modules within the cell, such as complexes involved in metabolic pathways or signal transduction cascades.

The network has many nodes located on the periphery, connected to the core by fewer interactions, these nodes represent proteins that are likely involved in more specialized or less central functions within the cell. They might participate in specific pathways or localized processes. As previously explained, the central part, the dense core, indicates the presence of hub proteins that have many interactions. These hub proteins are crucial for maintaining the network's connectivity and are involved in essential cellular functions, this may be more easy to understand when analyzing the Network metrics and the explanation of each one.

The overall layout suggests a hierarchical structure where a few highly connected hubs are central, and many less connected nodes form the periphery. The network's modular organization can be inferred from the clusters of nodes, indicating functional modules or protein complexes that carry out specific cellular functions.

For easy understanding of the network, the nodes here represent a protein yeast proteome. This means, "the collection of 6,607 protein sequences predicted based on the genome of yeast" [5].

Each edge represents a physical interaction between two proteins. The presence of an edge indicates that the connected proteins physically interact within the cell.

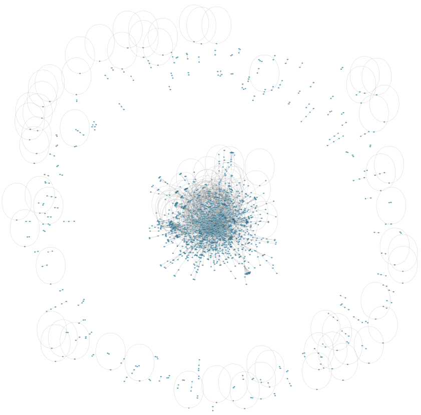

Fig. 1. Yeast PPI Network from Dataset

E. Observations Based on Graph Theory

The yeast PPI exhibits several interesting properties from a graph theory perspective.

The network typically follows a power-law distribution, with a few highly connected hub proteins and many proteins with fewer connections. Furthermore, key proteins can be identified using centrality measures such as degree centrality and betweenness centrality, indicating their importance in maintaining the network’s connectivity. Moreover, the network contains distinct clusters or communities, often corresponding to functional modules within the cell. For example, proteins involved in the same biological process tend to cluster together, reflecting their cooperative roles in cellular functions. By understanding these properties, which will be mentioned further, it can be gain deeper insights into the organizational principles of cellular networks and the biological significance of protein interactions.

II. Network Characteristics

This section presents the key characteristics and distance metrics of the largest connected component of the protein-protein interaction (PPI) network in yeast.

A. Size of Network

The size of the network was 1647 nodes. This can be due to the large number of proteins that compose a PPI network, which reflects the complexity and diversity of protein interactions in yeast. The fact that 1647 nodes form the largest connected components suggests that these proteins are part of a core network that interacts extensively.

The number of links obtained was 2682 edges. The high number of interactions, which are represented by the edges, indicates that the proteins in this network have multiple interaction partners. This is typically for biological networks where proteins often participate in various complexes and pathways.

The value obtained from the Average Path length was 5.612. This means that it is relatively short, which suggests that the network is well-connected. In biological terms, this means that most proteins can interact or influence each other through a small number of intermediaries, facilitating efficient communication and functional integration within the cell.

The clustering coefficient value obtained was 0.057. This is relatively low, indicating that while there are many interactions, they do not form many tightly-knit clusters. This could be due to the diverse functional roles of the proteins, where many interactions are spread out across different functional modules rather than being concentrated within specific groups.

B. Distance Metrics

The average distance value obtained was 5.612. This mirrors the average path length, reinforcing the idea of a well-connected network where proteins are relatively close to each other in terms of interaction steps.

The diameter value was 14, which is the longest and shortest path between any two nodes, indicating the maximum interaction distance in the network. A diameter of 14 suggests that even the most distant proteins can be connected through a reasonable number of interactions, highlighting the network’s robustness.
The radius obtained was 8; The radius being the minimum eccentricity, shows the shortest maximum distance from a central node to all other nodes. A radius of 8 suggests that some central proteins can reach all others in the network within 8 steps, underscoring their central role in maintaining network connectivity.
Now some important and noticeable measures in the graph, the periphery result nodes were: [’45’, ’1244’, ’77’, ’159’, ’227’, ’334’, ’677’, ’795’, ’1643’, ’1533’]. The periphery includes nodes with the highest eccentricity, meaning they are the furthest from the center of the network. These proteins might be involved in more specialized or less central functions, leading to fewer direct interactions with the core proteins.

The center measurement output nodes were: [’1701’, ’330’, ’1400’, ’1186’, ’267’, ’515’, ’1883’, ’752’, ’1356’, ’88’, ’90’, ’1668’, ’1888’, ’528’, ’919’, ’1339’, ’156’, ’124’, ’716’, ’223’, ’178’, ’434’, ’862’, ’876’, ’918’, ’1084’, ’1088’, ’1439’, ’1592’, ’1776’, ’1833’, ’1890’, ’1726’, ’1460’, ’733’, ’939’, ’1386’, ’1714’, ’993’, ’986’, ’863’, ’1841’, ’1866’, ’873’, ’387’, ’648’, ’637’, ’768’, ’818’, ’1094’, ’1384’, ’1967’].
As noticed in the visualization, the center includes nodes with the lowest eccentricity, indicating they are the most central and can reach all other nodes within the shortest maximum distance. These central proteins are likely crucial for the network’s integrity and play key roles in major cellular processes.

C. Probable Reasons for the Observed Metrics

As some observations to understand more thoroughly the network, in the case of short average path length and average distance, the data indicates a highly interconnected network, which is typical for biological systems where efficient interaction and signal transduction are necessary. The low clustering coefficient suggests that while the network is well-connected, it does not form many tightly-knit clusters. This could be due to the diverse roles of proteins and their involvement in multiple pathways.
And mentioning periphery and center nodes, highlights the hierarchical structure of the network, with central proteins playing key roles and peripheral proteins participating in more specialized functions.

III. Centrality Measures

By doing an exploratory analysis applying the Centrality Measures given by the *NetworkX* Python library. We analyzed the dataset using the following metrics:

- Degree Centrality.
- Katz Centrality.

- PageRank.
- Betweenness Centrality.
- Closeness Centrality.

A. Degree Centrality

What was obtained by applying the Degree Centrality measure was the following:

Top 5 nodes by Degree Centrality: [(’1356’, 0.0451165096678235), (’1400’, 0.0406544372830937),(’1637’, 0.04015865146256817), (’1017’, 0.025780862667327712), (’528’, 0.022806147744174516)]
What can be identified from here is that nodes with a high degree of centrality have the most direct connections. In this case, proteins ’1356’, ’1400’, and ’1637’ are highly connected, making them potential hubs within the network. Given their extensive interaction with other proteins, these hubs are likely to be essential for cellular processes.

B. Katz Centrality

The output obtained applying the Katz Centrality was: Top 5 nodes by Katz Centrality: [(’1400’, 0.5319081422199414), (’1017’, 0.22555470947900189), (’514’, 0.15354368751373154), (’806’, 0.12385093908069741), (’1892’, 0.12039695198517276)]

Katz Centrality accounts for the overall reach of a node within the network. In this case Node ’1400’ has a significant reach, influencing many proteins either directly or indirectly. This can habilitate the identification of proteins involved in widespread regulatory functions.

C. PageRank

The output obtained was the following: Top 5 nodes by PageRank: [(’1400’, 0.010201593267471647), (’1356’, 0.010066154733762652), (’1637’, 0.009427933800099448), (’528’, 0.006521527301045293), (’1017’, 0.006139321429989983)]
In this case, similar to Kartz centrality, PageRank identifies important nodes based on their incoming links’ quality. As seen before, the node ’1400’ stands out, suggesting its crucial role in cellular processes. This measure helps pinpoint proteins that might be crucial for maintaining the network’s robustness.

D. Betweeness Centrality

The data obtained from the measurement application was: Top 5 nodes by Betweenness Centrality: [(’1356’, 0.18151976215555363), (’1400’, 0.15323158610041307), (’1637’, 0.1102454307700271), (’528’, 0.06596389322291873), (’161’, 0.04306488085334644)]
The Beetweness centrality highlights nodes that act as bridges within the network. High scores for nodes ’1356’ and ’1400’ indicate these proteins facilitate communication between different parts of the network. These nodes are crucial for information flow and might be potential targets for therapeutic interventions, as disrupting them can significantly impact the network.

E. Closeness Centrality

The output obtained was: Top 5 nodes by Closeness Centrality: [(’1356’, 0.24498275690734178), (’1400’, 0.23446333672943884), (’1637’, 0.22338939898934893), (’330’, 0.21966319805771953), (’1888’, 0.2156776583370191)]
Closeness centrality measures how quickly a node can interact with all other nodes. High closeness centrality for nodes ’1356’ and ’1400’ suggests these proteins can rapidly disseminate signals throughout the network, making them vital for efficient cellular communication.

F. Observations

Nodes previously mentioned, ’1356’, ’1400’, and ’1637’, consistently appear in the top ranks of our analysis across the multiple centrality measures. This consistency may indicate that these proteins are central to the network’s structure and function, most likely playing critical roles in essential cellular processes. Also, nodes with high betweenness centrality, such as ’1356’ and ’1400’, could be potential targets for drug development. Disrupting these nodes might disrupt key pathways, providing therapeutic benefits. Finally, Nodes with high degrees and betweenness centrality serve as hubs and bridges respectively. These nodes are crucial for maintaining the network’s connectivity and facilitating interactions between different protein clusters.

The following visualization may be more helpful in identifying key nodes based on their degree of centrality. The nodes are colored according to their centrality values, with a color bar indicating the range of centrality scores. This specific visualization may aid in quickly spotting the most influential proteins within the network.

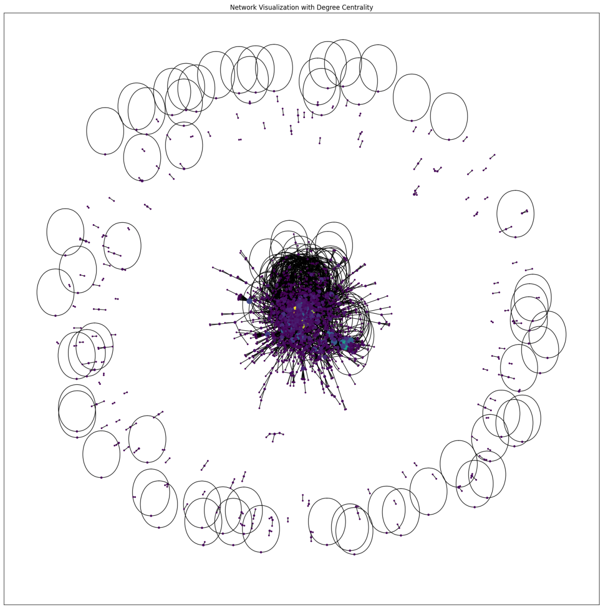

Fig. 2. Different Color Palette for easier Identification

To explain in more depth the color palette utilized, the nodes with higher degree centrality are assigned colors towards the yellow end of the spectrum, and the nodes with lower degree centrality are assigned colors towards the purple end of the spectrum. In this case, bodes that are yellow have a high degree centrality, meaning they have many direct connections to other nodes. These nodes are considered hubs within the network. In a PPI network, such hubs are often essential proteins involved in critical cellular functions, such as regulatory proteins or those involved in major metabolic pathways, and as we have seen already the nodes that appear more often, the yellow nodes inside the hub are most likely those mentioned. And on the other hand, nodes that are purple have a low degree of centrality, indicating they have fewer direct connections. These nodes may represent proteins involved in more specific, less central functions within the cell. They might be part of smaller, specialized pathways or processes.
Now the medium degree, which are the intermediate colors have a moderate degree of centrality. These nodes might be involved in connecting different parts of the network, playing a role in facilitating interactions between highly connected hubs and less connected peripheral nodes.

IV. Degree Distribution

The degree distribution plot represents the frequency of nodes (proteins) in the yeast protein-protein interaction (PPI) network based on their degree centrality.

The histogram shown is structured as follows:

- The X-axis represents the degree centrality of nodes in the network.
- The Y-axis represents the frequency (number) of nodes that have a particular degree centrality.

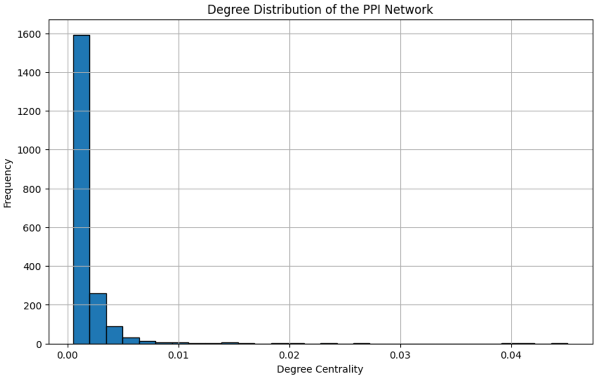

Fig. 3. Degree Distribution Plot

The histogram reveals that there are a few nodes with a very high degree centrality. These nodes are likely to be the hubs of the network, interacting with many other proteins. These hub nodes are essential for the structural integrity and functionality of the network. Most of the nodes have a low degree centrality, indicating they have few direct interactions with other proteins. These nodes might be involved in specific or specialized functions within the cell, as they do not require extensive interactions to perform their roles.
The presence of a few highly connected hubs and many low-degree nodes suggests that the yeast PPI network follows a scale-free distribution. This is a common characteristic of biological networks, where a small number of nodes (hubs) play critical roles in maintaining the network's connectivity [6].

V. Community analysis

Community detection was performed on the protein-protein interaction (PPI) network in yeast, to identify groups of proteins that interact more frequently with each other than those outside their group. This analysis helps in understanding the network's modular structure and identifying functional modules within the cell.

A. Method

The label propagation method was used for community detection, which detects communities by spreading labels based on the majority vote of neighbors. The resulting communities were visualized by assigning unique colors to each community.

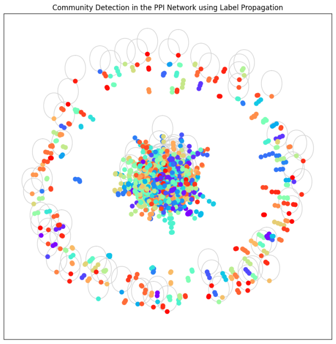
Community Detection in the PPI Network using Label Propagation

Fig. 4. Community Analysis: Label Propagation Algorithm

B. Observations

The community detection analysis revealed several distinct communities within the PPI network:

Modular Structure: The network is divided into multiple modules, each representing a group of proteins that interact more frequently with each other. These modules likely correspond to functional units within the cell, such as protein complexes or pathways.

Functional Implications: Identifying these communities can help understand the proteome's functional organization. Proteins within the same community may be involved in related biological processes or pathways.

Key Communities: Some communities are larger and more densely connected, indicating core functional modules essential for basic cellular functions. Smaller communities may represent specialized functions or localized processes within the cell.

VI. Conclusion

This networked analysis focused on the protein-protein Interaction (PPI) network in yeast, to understand its structure and functional implications. By utilizing and applying network analysis techniques it was managed to understand thoroughly the behavior and the structure of the network.

The network structure revealed a densely connected core with peripheral regions, indicating hierarchical organization. Centrality measures identified key hub proteins such as specific nodes which play crucial roles in maintaining network integrity and facilitating cellular processes.

The degree distribution followed a scale-free pattern, with a few highly connected hubs essential for network stability and functionality. Most proteins had fewer interactions, likely participating in specialized or localized processes within the cell.

Community detection was also applied, utilizing the label propagation algorithm, revealing distinct functional modules within the network. These modules likely correspond to protein complexes or pathways, providing valuable insights into the functional organization of the proteome.

Overall, this analysis enhances the understanding of the yeast PPI network, highlighting its significance in functional genomics, disease mechanisms, and drug discovery. The methodologies and findings can be applied to other organisms, contributing to broader biological and medical research efforts.

References

[1] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 8, pp. 4569-4574, 2001. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11283351/.

[2] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, M. Narayan, S. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg, "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae," *Nature*, vol. 403, no. 6770, pp. 623-627, 2000. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/10688190/.

[3] A. C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, C. Bastuck, B. Huhse, K. Leutwein, M. Heurtier, R. Copley, P. Edelmann, M. Querfurth, V. Rybin, G. Drewes, J. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga, "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141-147, 2002. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/11805826/.

[4] N. J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A. Punna, J. M. Peregrín-Alvarez, M. Shales, H. Zhang, R. Davey, J. Robinson, J. Raghibizadeh, M. P. Havugimana, J. Whetstone, P. W. Bellows, C. A. Chaudhry, H. S. Nesvizhskii, M. J. Dennis, B. Andrews, A. Emili, and J. F. Greenblatt, "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, pp. 637-643, 2006. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16554755/.

[5] P. Picotti et al., "A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis," Nature, vol. 494, no. 7436, pp. 266–270, Jan. 2013, doi: 10.1038/nature11835. Available: https://www.nature.com/articles/nature11835#:~:text=We%20defined%20the%20yeast%20proteome,www.yeastgenome.org).

[6] Broido, A.D., Clauset, A. Scale-free networks are rare. Nat Commun 10, 1017 (2019). https://doi.org/10.1038/s41467-019-08746-5