COURSE:
MATH 225

COMPREHENSIVE REPORT

Network Analysis of Human Apoptosis Protein Interaction

By:
Fatema Al Ali 2023006479
Amir Besher 2023006017

1. Introduction to Apoptosis and Protein Interaction Networks:

Apoptosis, a conserved and tightly controlled form of programmed cell death, is essential for the proper development and maintenance of multicellular organisms. It plays a pivotal role in shaping tissues during embryonic development, eliminating damaged or infected cells, and preserving cellular balance. Disruptions in apoptotic mechanisms are closely linked to numerous diseases, such as cancer, where resistance to cell death fosters uncontrolled growth, and neurodegenerative disorders, where excessive apoptosis contributes to the progressive loss of neurons. Gaining a thorough understanding of the molecular pathways governing apoptosis is thus vital for advancing therapeutic strategies.

Central to apoptosis is a complex web of interacting proteins. Protein interaction networks (PINs) offer a systems-level framework for examining how proteins collaborate to control cell fate. In these networks, proteins are modeled as nodes, and their physical or functional associations are represented as edges. Leveraging graph theory on these networks helps uncover organizational patterns in the apoptotic pathway. Through network analysis, researchers can identify key proteins acting as hubs or bottlenecks, detect collaborative protein clusters, and assess how the pathway might respond to various disruptions. This report details a comprehensive analysis of the human apoptosis protein interaction network, aiming to characterize its structure, spotlight influential proteins, and outline potential regulatory mechanisms that shape the decision-making process behind cell death. The insights derived from this analysis may significantly improve our understanding of apoptosis and suggest new therapeutic targets for conditions involving abnormal cell death regulation.

2. Methods Used for Analysis

The apoptosis protein interaction network was built and analyzed using Python's NetworkX library — a versatile platform for managing and investigating complex networks. The dataset used for this project was gathered from several Excel files within a designated directory. These files, including "Edge List.xlsx", "Interactions.xlsx", "Sub Networks Edge List.xlsx", and "Centrality & Degrees & List of Edge List.xlsx", contained verified and/or predicted interactions relevant to apoptosis.

The first step involved importing and preprocessing the data. Custom Python functions were created to extract interaction data from various sheets in the Excel files. These functions, like load_edge_list and load_all_edge_lists, were designed to accommodate formatting differences and merge the information into a single edge list. To maintain consistency, a standardized protein name list was imported from the "Protein degree" sheet in one of the core files. Protein names in the compiled edge list were then aligned to this reference (using standardize_protein_names) to address issues like alternate naming conventions or isoforms. With this cleaned edge list, an undirected graph was generated in NetworkX, where proteins were represented as nodes and documented interactions as edges. Self-loops — connections where a protein interacts with itself — were removed to avoid skewing the analysis.

- Basic network statistics were then computed using NetworkX's built-in functions (via calculate_basic_stats). These included metrics such as the number of proteins (nodes), interactions (edges), network density, average protein interactions (degree), clustering coefficient, and network diameter (or the largest connected component's diameter if disconnected). Additionally, the degree assortativity coefficient was calculated to see whether proteins with similar numbers of interactions tend to associate.

- To better visualize the network's structure, layouts like the spring layout (nx.spring_layout) and Kamada-Kawai layout (nx.kamada_kawai_layout) were used to graphically display the network. These layouts can help reveal patterns such as tightly knit clusters or key central nodes. The distribution of protein interactions (degree distribution) was also plotted using plot_degree_distribution to check for characteristics of a scale-free network, which typically features a few highly connected hubs amid many proteins with fewer connections. The most influential proteins were then identified through centrality analysis (calculate_centrality_measures). Several measures were considered: degree centrality (number of connections), betweenness centrality (proteins that bridge different network regions), closeness centrality (how accessible a protein is from others), eigenvector centrality (influence based on connections to other influential proteins), and PageRank (importance based on the significance of connected proteins). Top-ranking proteins for each metric were highlighted using get_top_proteins, and the network was visualized again with node sizes reflecting their centrality values (visualize_network_with_centrality). Correlations between these centrality metrics were also explored through scatter plots and correlation analysis (compare_centrality_measures).

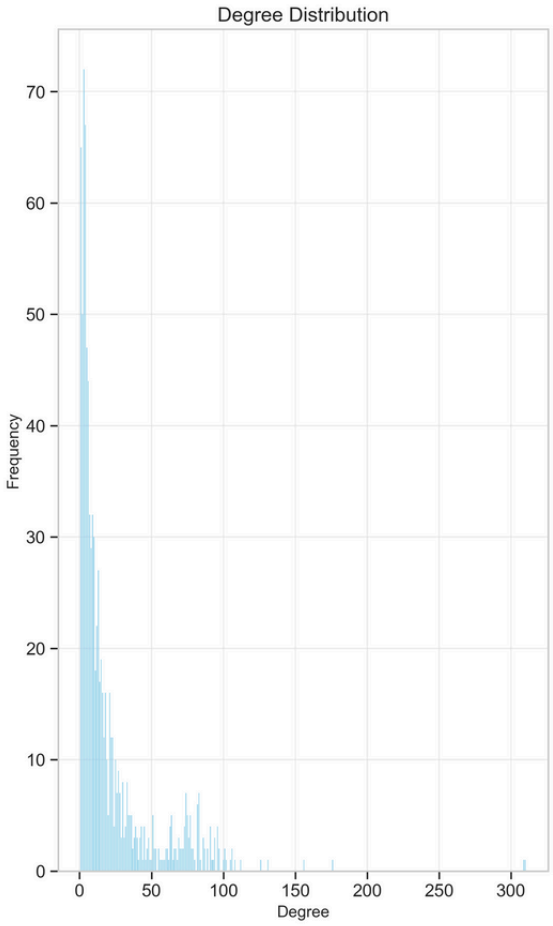
- To uncover functional groupings within the network, community detection was performed using algorithms like Louvain (preferred) or label propagation (detect_communities_louvain, detect_communities_label_propagation). These methods divide the network into groups of proteins that interact more with each other than with the rest of the network. The resulting communities were visualized by coloring nodes according to their group (visualize_communities), and their structure was evaluated with modularity scores and other related statistics (analyze_communities).

- Network resilience was tested by simulating protein removal scenarios (analyze_network_resilience). Two strategies were applied: targeted attacks, where high-centrality proteins were systematically removed, and random failures. The effects of these removals on the size of the largest connected component were tracked and visualized (visualize_network_resilience), offering insights into the network's robustness under different stresses.

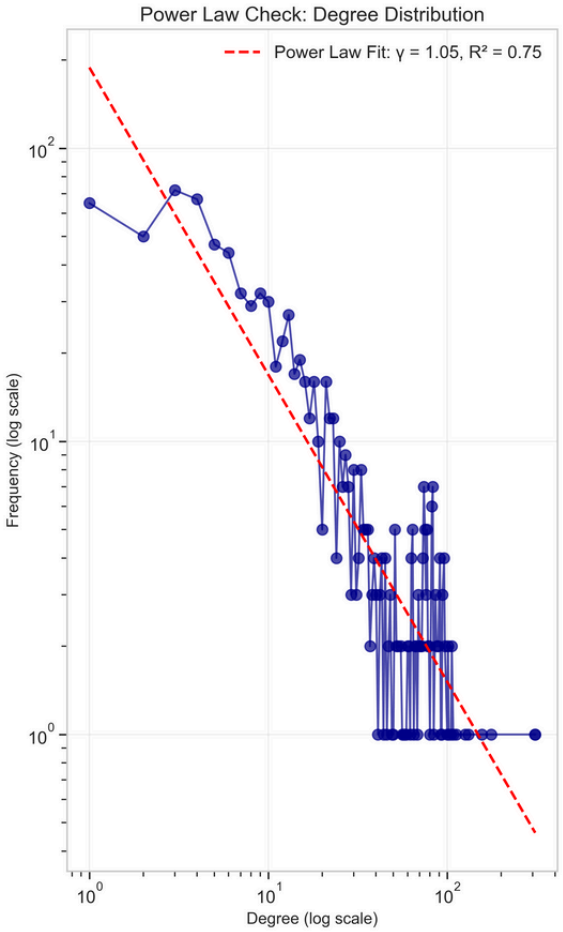
- Path analysis was then conducted to examine potential signaling routes between key apoptosis proteins (like BCL2, CASP9, FAS, CASP3). Using NetworkX's shortest path algorithms, the shortest paths and intermediary proteins were identified to infer possible regulatory interactions.

- bottleneck proteins — those essential for maintaining overall network connectivity — were identified, typically showing high betweenness centrality and a significant effect on network fragmentation during resilience testing.

3. KEY FINDINGS AND THEIR BIOLOGICAL SIGNIFICANCE



The histogram shows the frequency distribution of node degrees (i.e. the number of interactions per protein). The x-axis represents the degree, and the y-axis represents the number of proteins with that degree. The distribution is right-skewed – most proteins have a low number of interactions, while a few have very high degrees. This long-tailed pattern is a typical signature observed in biological networks.



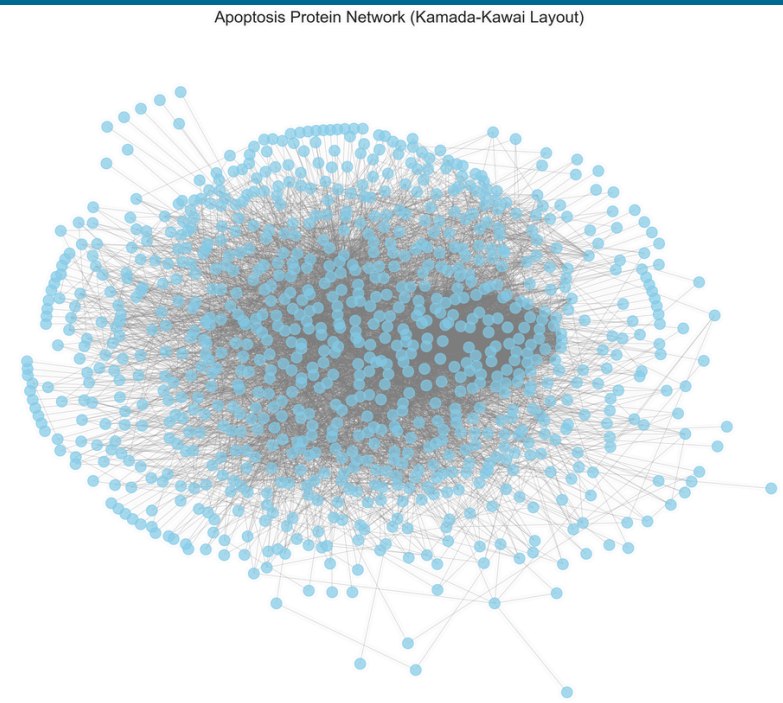
This log-log plot evaluates whether the degree distribution follows a power-law distribution, often characteristic of scale-free networks. The x-axis represents the degree (on a log scale), and the y-axis represents the frequency (also on a log scale). The data points are plotted in blue, and a fitted power-law line is shown in red. The power-law model has a scaling exponent (γ) of 1.05 and an R^2 value of 0.75.

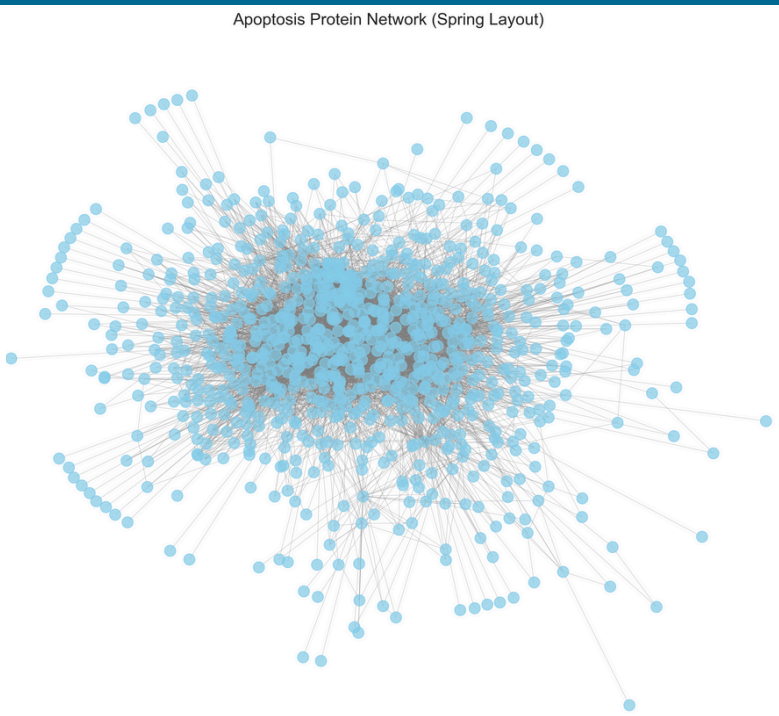
To assess whether the apoptosis protein interaction network follows a power-law distribution, characteristic of many biological networks, the degree distribution was analyzed through both a histogram and a log-log plot. The histogram revealed a long-tailed distribution, where most proteins exhibited a low number of interactions while a small number of proteins demonstrated significantly higher degrees. This pattern is indicative of a heterogeneous network structure often observed in biological systems. Further evaluation using a log-log plot compared the empirical degree distribution to a fitted power-law model. The resulting fit produced a scaling exponent (γ) of 1.05 and an R^2 value of 0.75, suggesting a reasonably strong alignment with a power-law distribution. Although not a perfect fit, this level of correlation is acceptable given the inherent noise and incompleteness in biological data. Collectively, these results indicate that the apoptosis protein interaction network exhibits scale-free properties, where a few highly connected hub proteins play central roles in maintaining network structure and communication. This organizational principle suggests resilience to random failures but potential vulnerability to targeted disruptions of key hub proteins, consistent with the behavior of many biological networks.

THE STRUCTURE OF THE HUMAN APOPTOSIS PROTEIN INTERACTION NETWORK

Kamada-Kawai layout; arranges the nodes by treating the connections like springs, aiming to position highly connected nodes closer together while evenly spacing others to reduce visual clutter. As a result, this layout reveals a dense core of interactions in the center with peripheral nodes branching outward, making it easier to identify highly interconnected proteins at the heart of the network.

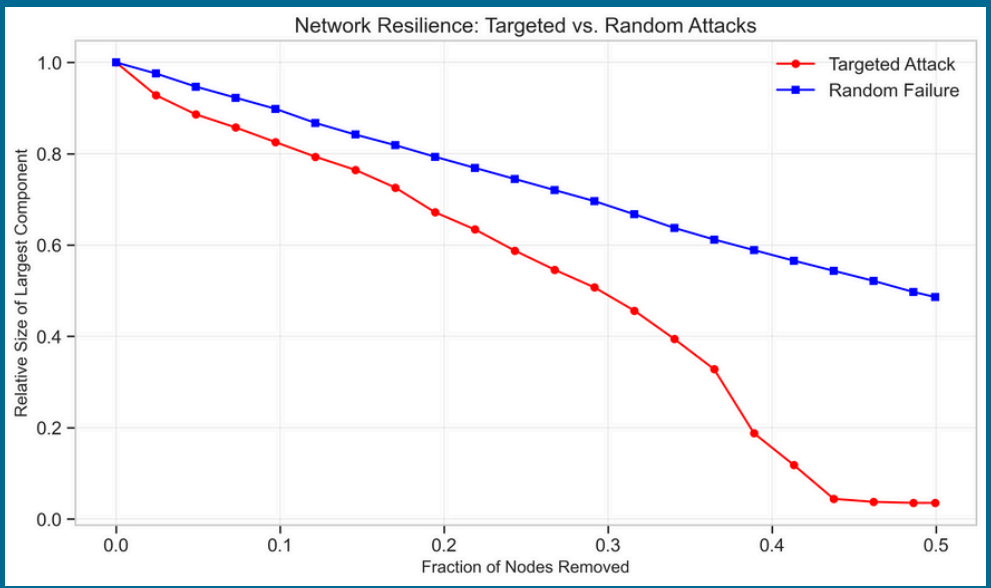
Spring layout: is also based on a force-directed algorithm but with a slightly different optimization approach. This layout similarly pulls connected nodes together while pushing unconnected nodes apart, though the overall structure appears slightly more compressed horizontally, giving a more oval shape to the network. The Spring layout highlights the dense interactivity within the central region but allows for slightly clearer separation of smaller clusters and less-connected proteins around the edges.





When comparing the two, the Kamada-Kawai layout provides a cleaner and more interpretable representation of the network's hierarchical core-periphery structure. It better emphasizes the central hub of interactions and the peripheral nodes radiating outward. The Spring layout, while effective in illustrating dense clustering, tends to compress the overall structure, potentially making it harder to distinguish between closely spaced groups of nodes. For this reason, the Kamada-Kawai layout might be preferred for visually interpreting the organizational pattern of the apoptosis protein network.

NETWORK RESILIENCE ANALYSIS



network resilience: compares the impact of targeted attacks and random failures on the size of the largest connected component of the protein interaction network. The x-axis represents the fraction of nodes (proteins) removed from the network, ranging from 0 to 0.5 (50%). The y-axis shows the relative size of the largest connected component, normalized by the initial size of the network, ranging from 0 to 1.0. The red line with circular markers illustrates the effect of a targeted attack, where nodes are removed based on their centrality (presumably the highest centrality nodes are removed first). The blue line with square markers shows the effect of random failure, where nodes are removed randomly from the network.

Network Resilence Key Findings:
Targeted Attacks:

- Removing even a small fraction of central proteins causes a rapid decline in network connectivity. At around 30% removal, the largest connected component drops to ~45% of its original size, collapsing almost entirely after 40%.

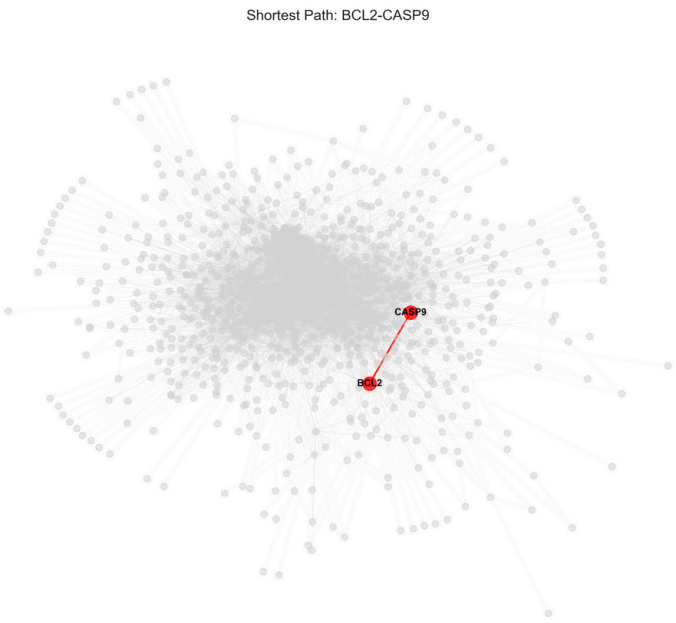
Random Failures:

- Random removal leads to a gradual, less severe decline. Even after 50% of proteins are removed, a substantial connected component remains.

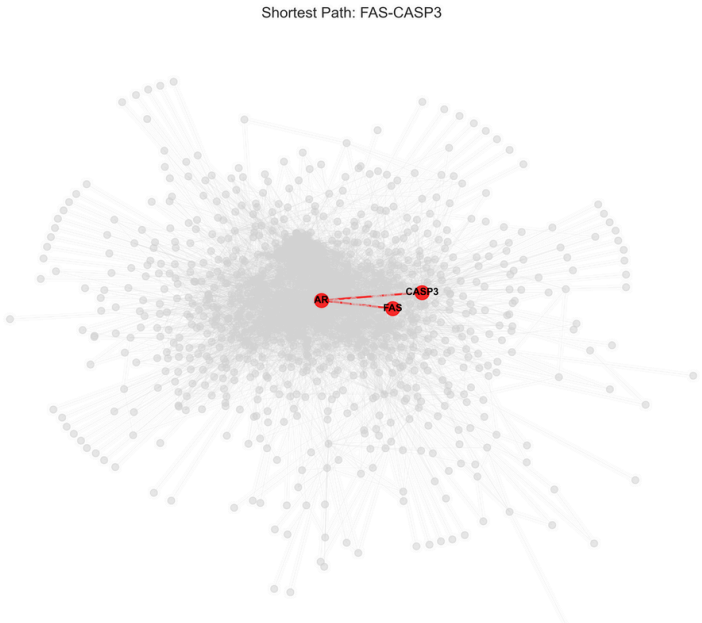
Implications:

- The network is highly vulnerable to targeted attacks, as central hub proteins are critical for maintaining connectivity.
- It shows resilience to random failures, suggesting robustness against incidental protein loss.
- Hub proteins play a crucial role in structural integrity and could be valuable therapeutic targets, especially for conditions like cancer where manipulating apoptosis is beneficial.
- Understanding resilience to random failures is also important for anticipating side effects of broadly acting drugs.

PATH ANALYSIS:



- The shortest path is direct: BCL2 -- CASP9. This suggests a direct interaction or regulatory relationship between these two proteins within the context of the network.



- The shortest path is indirect: FAS -- FADD -- CASP3. This indicates that the signal or influence from FAS to CASP3 likely involves FADD as an intermediary.

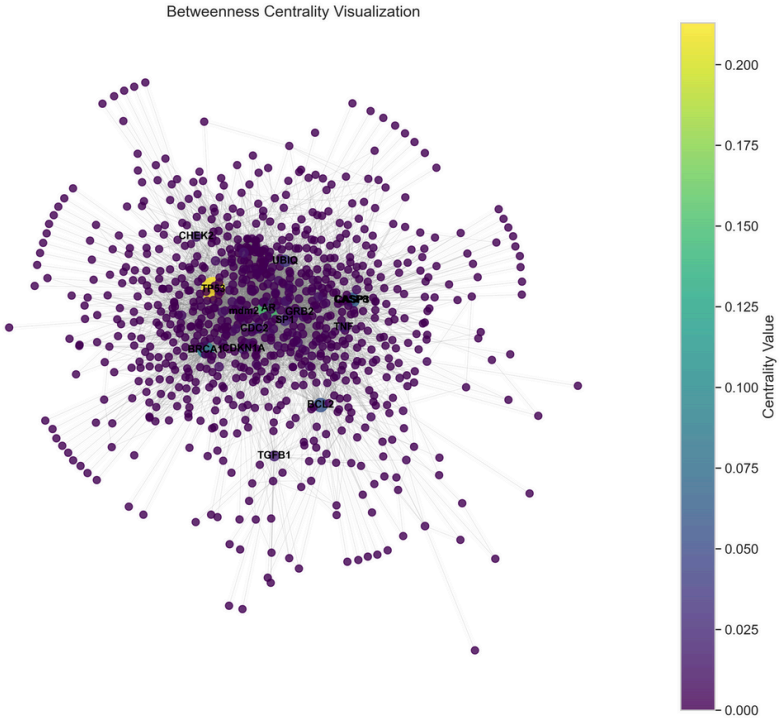
Role as Signal Transducers:

- BCL2: Can transduce survival signals by inhibiting pro-apoptotic signals involving CASP9.
- CASP9: Transduces pro-apoptotic signals in the intrinsic pathway, activating downstream caspases.
- FAS: Acts as a receptor, initiating and transducing death signals in the extrinsic pathway.
- FADD: Crucially transduces the death signal from FAS to downstream caspases like CASP3.
- CASP3: Transduces the apoptotic signal by executing cell death through substrate cleavage.

Proteins with High Betweenness as Critical Links:

- The central, densely connected proteins in the network likely have high betweenness and serve as critical links for communication.
- FADD is a potential high-betweenness protein, acting as a key intermediary in the FAS-CASP3 pathway and potentially others.
- Proteins that connect the central hub to the outer parts of the network are also likely to be critical links with high betweenness.

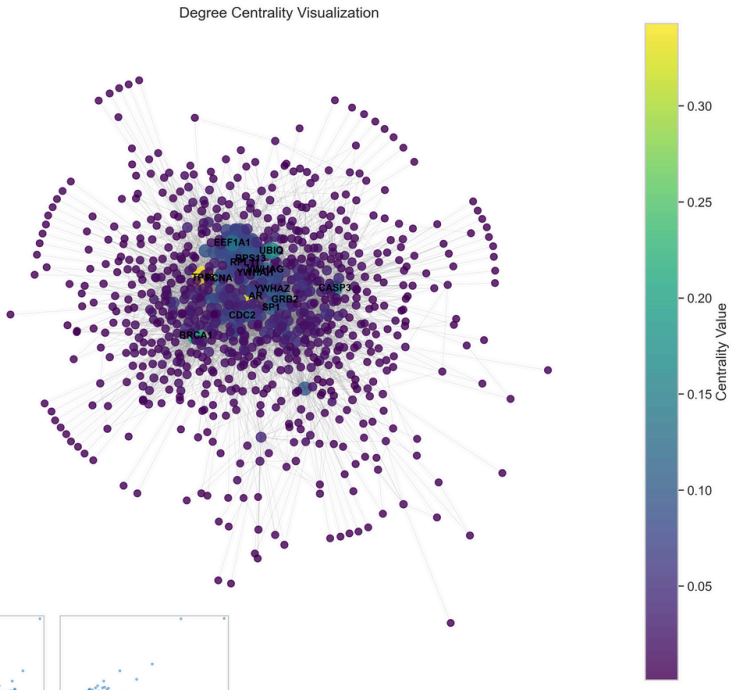
CENTRALITY ANALYSIS

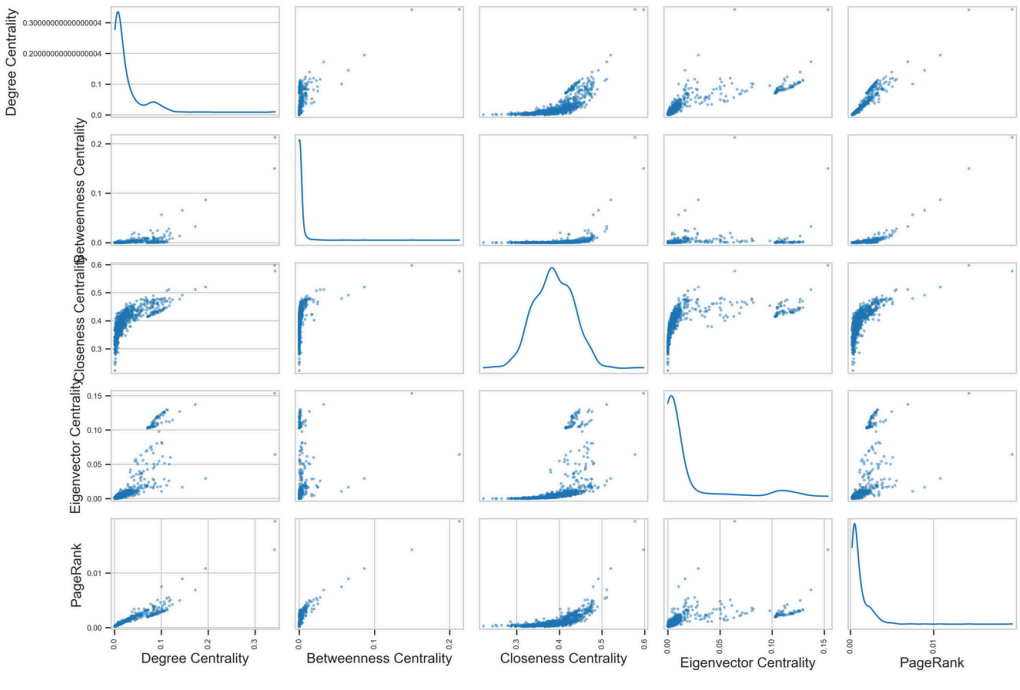


This visualization depicts a protein-protein interaction network where each node (dot) represents a protein. The color of each node represents its betweenness centrality, as indicated by the color scale on the right. Proteins with higher betweenness centrality are shown in warmer colors (yellow), while those with lower betweenness are in cooler colors (purple/dark blue).

Proteins with high betweenness centrality act as bridges, lying on many of the shortest paths between other protein pairs. The image suggests that proteins like CHEK2 and UBM have relatively high betweenness centrality within this network. The protein TGFB1 is labeled, but appears to have lower betweenness centrality. The central region of the network generally shows higher betweenness values than the periphery.

Degree centrality visualization of a protein-protein interaction network: Each dot represents a protein, and its color indicates its degree centrality – how many other proteins it interacts with. Yellow nodes have the highest degree centrality, acting as hubs with many connections, while purple/dark blue nodes have the lowest, indicating fewer interactions. The central area shows several high-degree hub proteins like EEF1A1, UBB, and CDC2, suggesting their potential importance in network organization. The outer regions consist mainly of lower-degree proteins.





CENTRALITY ANALYSIS

Biological Significance of These Top Proteins:
TP53 (Tumor Protein p53): Often called the "guardian of the genome," TP53 is a transcription factor involved in regulating cell cycle, DNA repair, apoptosis, and senescence. Its high degree and betweenness centrality suggest it's a crucial hub and bridge in many cellular processes, making it central to maintaining cellular homeostasis and preventing cancer.

AR (Androgen Receptor): A nuclear receptor that binds to androgens, mediating their effects on gene expression. AR plays a critical role in the development and maintenance of male sexual characteristics and is implicated in prostate cancer. Its high degree and closeness centrality indicate it interacts with many proteins and can efficiently spread signals across the network related to androgen signaling.

BRCA1 (Breast Cancer 1, early onset): A tumor suppressor protein involved in DNA repair, particularly homologous recombination. BRCA1 interacts with many other proteins involved in genome maintenance. Its high degree and betweenness highlight its role as a significant player in DNA damage response and its potential to influence communication between different repair pathways.

UBIQ (Ubiquitin): A small regulatory protein that is attached to other proteins to mark them for degradation (by the proteasome) or to modify their function, localization, or interactions. Its high degree and closeness centrality reflect its involvement in a vast number of cellular processes through protein tagging and degradation.

CASP3 (Caspase-3): An effector caspase that plays a central role in the execution phase of apoptosis (programmed cell death). Its presence in the top 5 for both degree and betweenness indicates it's a key player in the apoptotic network, both interacting with many proteins involved in the process and potentially bridging different apoptotic signals.

BCL2 (B-cell lymphoma 2): An anti-apoptotic protein that inhibits programmed cell death. Its appearance in the top 5 for betweenness centrality suggests it might act as a crucial regulator of apoptotic signals, positioned to influence communication within the cell death network.

CDC2 (Cell Division Cycle 2): A key regulatory kinase that controls entry into mitosis (M phase) of the cell cycle. Its high closeness centrality suggests it can rapidly influence many other proteins involved in cell cycle progression.

Top 5 Proteins by Degree Centrality:

	Protein	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality	PageRank
147	TP53	0.342920	0.212927	0.577267	0.064228	0.019471
17	AR	0.341814	0.149784	0.597488	0.153470	0.014262
107	BRCA1	0.194690	0.086557	0.520438	0.029267	0.010812
614	UBIQ	0.172566	0.032580	0.511023	0.137200	0.006910
414	CASP3	0.144912	0.065313	0.491572	0.016536	0.008933

Top 5 Proteins by Betweenness Centrality:

	Protein	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality	PageRank
147	TP53	0.342920	0.212927	0.577267	0.064228	0.019471
17	AR	0.341814	0.149784	0.597488	0.153470	0.014262
107	BRCA1	0.194690	0.086557	0.520438	0.029267	0.010812
414	CASP3	0.144912	0.065313	0.491572	0.016536	0.008933
172	BCL2	0.100664	0.056355	0.479321	0.010402	0.007486

Top 5 Proteins by Closeness Centrality:

	Protein	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality	PageRank
17	AR	0.341814	0.149784	0.597488	0.153470	0.014262
147	TP53	0.342920	0.212927	0.577267	0.064228	0.019471
107	BRCA1	0.194690	0.086557	0.520438	0.029267	0.010812
614	UBIQ	0.172566	0.032580	0.511023	0.137200	0.006910
199	CDC2	0.116150	0.027919	0.511023	0.025790	0.005535

Comparison of Results from Different Centrality Measures:

Overlap: TP53, AR, BRCA1, and CASP3 appear in the top 5 for both Degree and Betweenness Centrality, highlighting their importance as both highly connected hubs and crucial bridges in the network.

Hubs vs. Bridges: UBIQ is a top hub (high degree) and has high closeness, reflecting its widespread involvement in many local interactions and efficient signal propagation. BCL2 appears in the top 5 for Betweenness but not Degree, suggesting it might be more critical for controlling information flow between different apoptotic components than being one of the most directly connected proteins overall. CDC2 is in the top for Closeness but not Degree or Betweenness in the top 5, indicating its potential to quickly reach other nodes in the network, likely within the cell cycle regulation module.

Closeness Centrality: The top proteins by closeness (AR, TP53, BRCA1, UBIQ, CDC2) tend to be central figures that can quickly influence or be influenced by other proteins in the network, potentially reflecting their involvement in core regulatory or signaling pathways that need rapid communication.

"Bottleneck" proteins are those whose removal would significantly impact network connectivity. High betweenness centrality is a strong indicator of potential bottleneck proteins.
Based on the Betweenness Centrality results:
TP53 and AR are strong candidates for bottleneck proteins due to their high betweenness scores. Removing them could disrupt communication pathways between different modules in the network.
BRCA1 and CASP3, with the next highest betweenness, are also likely to be important bottlenecks, though perhaps to a lesser extent than TP53 and AR.
BCL2, while lower than the others, still has a relatively high betweenness and could represent a bottleneck specifically within apoptosis-related signaling.
Potential Biological Significance of These Bottlenecks:
Disruption of Core Processes: Removing a protein like TP53 or AR, which bridge many interactions, could have cascading effects on multiple cellular processes they regulate or are involved in. This could lead to widespread dysfunction, potentially contributing to diseases like cancer.

Top 5 Proteins by Eigenvector Centrality:

	Protein	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality	PageRank
17	AR	0.341814	0.149784	0.597488	0.153470	0.014262
614	UBIQ	0.172566	0.032580	0.511023	0.137200	0.006910
37	RPS13	0.111726	0.000934	0.447082	0.129653	0.003222
29	RPL11	0.112832	0.001797	0.469610	0.129245	0.003293
63	EEF1A1	0.139381	0.013502	0.476793	0.126922	0.004972

Top 5 Proteins by PageRank:

	Protein	Degree Centrality	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality	PageRank
147	TP53	0.342920	0.212927	0.577267	0.064228	0.019471
17	AR	0.341814	0.149784	0.597488	0.153470	0.014262
107	BRCA1	0.194690	0.086557	0.520438	0.029267	0.010812
414	CASP3	0.144912	0.065313	0.491572	0.016536	0.008933
172	BCL2	0.100664	0.056355	0.479321	0.010402	0.007486

Biological Significance of These Top Proteins:
RPS13 (Ribosomal Protein S13) and RPL11 (Ribosomal Protein L11): These are components of the small (40S) and large (60S) subunits of the ribosome, respectively, and are essential for protein synthesis. Their high Eigenvector Centrality suggests they are connected to other influential proteins, potentially reflecting the central role of protein synthesis in cellular function and its regulation.

EEF1A1 (Eukaryotic Translation Elongation Factor 1 Alpha 1): This protein is crucial for the elongation step of protein synthesis, facilitating the binding of aminoacyl-tRNAs to the ribosome. Its high Eigenvector Centrality further emphasizes the importance of the protein synthesis machinery and its connections to other influential network members.

TP53 (Tumor Protein p53): Its top ranking in PageRank highlights its central and influential position in the network. PageRank, like Google's algorithm, assigns importance based on the number and quality (importance) of incoming links. TP53 receives connections from many important nodes, making it a highly ranked protein in terms of influence.

Comparison of Results from Different Centrality Measures:
AR as a Central Hub of Influence: AR consistently appears at the top or near the top in Degree, Betweenness, Closeness, and Eigenvector Centrality, indicating it is a highly connected, influential protein that bridges different parts of the network and is connected to other important nodes.

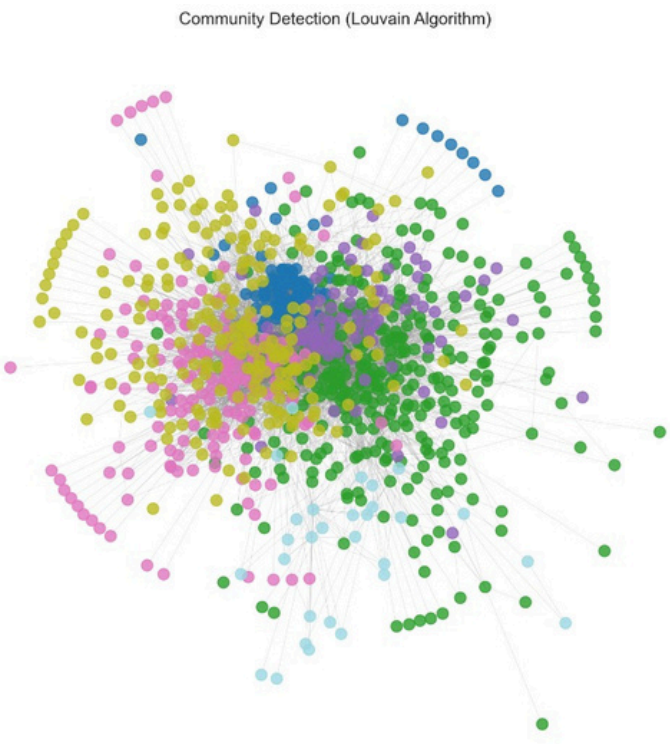
TP53 as a Broadly Important Regulator: TP53 is also consistently highly ranked across Degree, Betweenness, Closeness, and PageRank, solidifying its role as a central regulator with many interactions and a strong influence on the network.

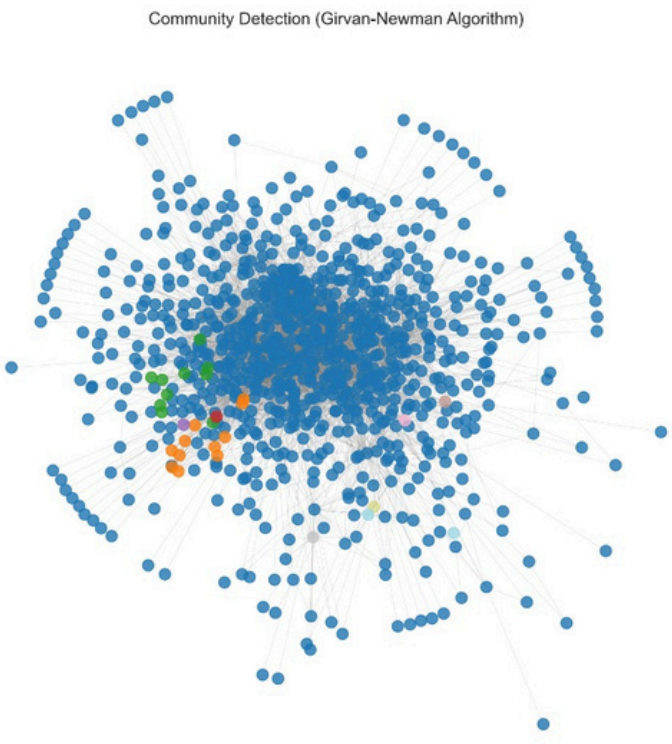
The Ribosomal Connection in Eigenvector Centrality: The prominence of ribosomal proteins (RPS13, RPL11) and a translation factor (EEF1A1) in the top Eigenvector Centrality list suggests that the protein synthesis machinery is highly interconnected with other influential proteins in the network. This highlights the fundamental importance and tight regulation of protein synthesis.

PageRank Emphasizes Key Regulators: PageRank highlights TP53, AR, BRCA1, CASP3, and BCL2, all of which are well-known regulators involved in critical cellular processes like DNA repair, apoptosis, and transcriptional control. This suggests that PageRank effectively identifies proteins that are "pointed to" by many other important proteins in the network.

UBIQ's Influence: UBIQ's high ranking in Degree, Closeness, and Eigenvector Centrality underscores its widespread involvement in cellular regulation through protein modification and degradation, connecting it to many key players.

COMMUNITY DETECTION





- Louvain Algorithm: This algorithm is a greedy optimization method that iteratively moves nodes between communities to maximize the modularity of the network. Modularity is a measure of the strength of division of a network into modules (also called communities, clusters or groups). High modularity indicates dense connections within communities and sparse connections between them. The visualization shows several distinct communities, each represented by a different color. The sizes of these communities appear varied.

- Girvan-Newman Algorithm: This algorithm takes a divisive approach. It iteratively removes the edges with the highest betweenness centrality in the network. As edges are removed, the network gradually breaks down into smaller, more cohesive communities. The visualization here shows a different community structure compared to the Louvain algorithm. It seems to have identified fewer, larger communities, with a few smaller ones highlighted in orange and green against a background of blue.

- Comparison of Detected Communities:

Visually comparing the two results, we can observe:

- Number of Communities: The Louvain algorithm appears to have detected a larger number of communities with varying sizes. The Girvan-Newman algorithm seems to have identified fewer, more encompassing communities, with some smaller, distinct sub-communities.

- Community Structure: The boundaries of the communities differ significantly between the two algorithms. Nodes grouped together by the Louvain algorithm are not necessarily in the same community according to the Girvan-Newman algorithm, and vice versa.

- Algorithm Sensitivity: This difference highlights the fact that community detection algorithms can be sensitive to their underlying methodologies and assumptions. Greedy optimization (Louvain) can lead to locally optimal solutions, while divisive hierarchical clustering (Girvan-Newman) can be computationally expensive for large networks and might miss finer-grained community structures.

4. Discussion of Limitations and Potential Improvements

- This analysis is limited by the inherent incompleteness and potential biases in protein interaction data, as experimental methods may not capture all interactions. Future studies should integrate additional data sources, such as gene expression profiles, post-translational modifications, and spatial proteomics, to provide a more comprehensive view of network dynamics.
- Advanced network analysis techniques, such as dynamic network modeling and network control theory, could be employed to explore the temporal evolution of signaling pathways and identify optimal intervention strategies, improving the analysis's predictive power.
- Experimental validation of key findings, including the functional roles of high-centrality proteins and the impact of targeted perturbations, is essential to translate network insights into therapeutic applications, bridging the gap between computational and experimental biology.

5. CONCLUSIONS AND INSIGHTS GAINED

- This systems-level analysis of the OsPPIN_CDPKs network has revealed critical architectural features and vulnerabilities. The network's scale-free topology renders it susceptible to targeted attacks on hub proteins, which may represent key regulatory nodes in apoptotic signaling. The identification of functional modules and potential signaling routes provides a framework for further experimental investigation.
- By integrating network analysis with experimental studies, we can gain a deeper understanding of the complex interplay of proteins that govern apoptosis and develop more effective therapeutic strategies, ultimately improving patient outcomes.

Refrences:

National Library of Medicine. (n.d.). PubMed. U.S. National Library of Medicine. Retrieved April 2, 2025, from https://pubmed.ncbi.nlm.nih.gov/

Research Collaboratory for Structural Bioinformatics. (n.d.). RCSB Protein Data Bank (PDB). Retrieved April 2, 2025, from https://www.rcsb.org/

UniProt Consortium. (n.d.). UniProt: The universal protein resource. Retrieved April 2, 2025, from https://www.uniprot.org/

NetworkX Developers. (n.d.). NetworkX: High-productivity software for complex networks. Retrieved April 2, 2025, from https://networkx.org/documentation/stable/index.html

European Bioinformatics Institute. (n.d.). IntAct molecular interaction database. Retrieved April 2, 2025, from https://www.ebi.ac.uk/intact/

The BioGRID Project. (n.d.). BioGRID: Biological General Repository for Interaction Datasets. Retrieved April 2, 2025, from https://thebiogrid.org/

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., ... & Jensen, L. J. (2019). STRING: a database of known and predicted protein-protein interactions. Nucleic acids research, 47(D1), D607-D613. Retrieved from https://string-db.org/

GeneCards. (n.d.). GeneCards: The human gene database. Weizmann Institute of Science. Retrieved April 2, 2025, from https://www.genecards.org/