CONSTRUCTOR
UNIVERSITY

Analysis of BioInteractions from STRING, BioGrid, and IntAct using Multilayer Networks

by

Cenhan Du

Bachelor Thesis in Computer Science

Submission: August 31, 2023 Supervisor: Prof. Dr. Marc-Thorsten Hütt

Constructor University | School of Computer Science and Engineering

CONSTRUCTOR
UNIVERSITY

Statutory Declaration

Family Name, Given/First Name	Du, Cenhan
Matriculation number	30002444
Kind of thesis submitted	Bachelor Thesis

English: Declaration of Authorship

I hereby declare that the thesis submitted was created and written solely by myself without any external support. Any sources, direct or indirect, are marked as such. I am aware of the fact that the contents of the thesis in digital form may be revised with regard to usage of unauthorized aid as well as whether the whole or parts of it may be identified as plagiarism. I do agree my work to be entered into a database for it to be compared with existing sources, where it will remain in order to enable further comparisons with future theses. This does not grant any rights of reproduction and usage, however.

This document was neither presented to any other examination board nor has it been published.

German: Erklärung der Autorenschaft (Urheberschaft)

Ich erkläre hiermit, dass die vorliegende Arbeit ohne fremde Hilfe ausschließlich von mir erstellt und geschrieben worden ist. Jedwede verwendeten Quellen, direkter oder indirekter Art, sind als solche kenntlich gemacht worden. Mir ist die Tatsache bewusst, dass der Inhalt der Thesis in digitaler Form geprüft werden kann im Hinblick darauf, ob es sich ganz oder in Teilen um ein Plagiat handelt. Ich bin damit einverstanden, dass meine Arbeit in einer Datenbank eingegeben werden kann, um mit bereits bestehenden Quellen verglichen zu werden und dort auch verbleibt, um mit zukünftigen Arbeiten verglichen werden zu können. Dies berechtigt jedoch nicht zur Verwendung oder Vervielfältigung.

Diese Arbeit wurde noch keiner anderen Prüfungsbehörde vorgelegt noch wurde sie bisher veröffentlicht.

31, August, 2023

....

Date, Signature

Abstract

In this thesis, we explore the complexities of biological networks and their potential insights through the lens of multilayer networks. Leveraging integrated data from STRING, IntAct, and BioGrid, we construct a three-layer multilayer network that encapsulates diverse protein-gene interactions. Employing centrality measures and clustering coefficient analysis, we reveal intricate connectivity patterns and local neighborhoods. Correlation analysis across layers uncovers co-evolution tendencies among different interaction types. Additionally, layer-specific sub-networks showcase key nodes and network co-evolution dynamics. Through varying score thresholds, we illuminate structural changes and connectivity dynamics within the network. Our findings underscore the significance of multilayer networks in uncovering hidden interactions and capturing the interplay between different types of interactions. Overall, this work helps the understanding of the complexities of biological systems and provides valuable insights into the intricate web of protein-gene interactions on both local and global scales.

Contents

1 Introduction

Protein or genetic interactions are crucial in the medical field for understanding the mechanistic basis of cellular functions and the molecular basis of biological processes. For example, interpreting cancer and other diseases, such as chronic infections, rely highly on the observation of biological interactions in living organisms[18][12]. These interactions are scattered across many databases, which collect information using different techniques and for a number of objectives. In particular, databases like STRING, BioGrid, and IntAct cover all types of interactions, detection methods, and scoring systems. They are categorized in gene-phenotype relationships, protein interactions, genetic interactions, and chemical interactions[18], and they are collected from various methods and sources like text mining, experiments, and computational predictions[22]. Therefore, in order to obtain useful and analytical results, efforts in developing advanced analysis tools and corresponding measures are required for deeper and more detailed research.

By bringing these three databases to a unified data structure, a larger scale of molecular interactions is created for further investigation. However, the new biological interaction network merged by such databases faces numerous challenges in its complexity. One of the key complexities is the scale and heterogeneity, which demand sophisticated computational methods and multidisciplinary insights to unravel their intricacies. Moreover, the dynamic nature of biological systems further contributes to the complexity because interactions can vary under different conditions and over time.

One remarkable development in network science is the emergence of multilayer networks that address these aspects of complexity, which is a more general framework that accommodates the evolution or interaction of distinct networks[2]. That is, a multilayer network extends beyond the conventional single-layer representation, which makes the framework more realistic and able to capture the essence of interconnectedness by incorporating multiple types of interactions, each layer reflecting a distinct aspect of the network's dynamics. In essence, a multilayer network serves as a multidimensional canvas, where nodes are not confined to mere dots but become part of an intricate web of relationships across layers[14]. Multilayer network possesses the capacity to fight against scale complexity by analyzing each layer's structure and dynamics independently while also considering their interdependence[11].

The main investigation of this thesis delves into the interaction dynamics within biological systems. The key questions are: Have different types of interaction 'co-evolved' to compensate each other or to enhance each other? Do different types of interaction operate on different scales (local vs. global)?

2 Statement and Motivation of Research

2.1 Introduction to Protein-Gene Interaction

Over the years, the number of identified protein-protein interactions has significantly increased, leading to the creation of numerous databases serving different annotations and purposes. Thus, more and more complicated biological interaction networks are aggregated, which serves as a framework for understanding the mechanism of protein interactions as well as genetic interactions[18]. Among these interactions, proteins, biomolecules, or macromolecules perform a wide range of functions in organisms. Not only do they carry out the roles of interacting with other molecules like DNA, RNA, membranes, carbohydrates, and small molecule metabolites[18], but they also have intentional physical contact with one or more proteins as a result of biochemical events or electrostatic forces[1]. Therefore, protein-protein interactions and protein-gene interactions, in particular, will be the main focus of this thesis.

With the development of high-throughput screening technologies, computational predictions, and data-mining processes, more protein-protein interactions are identified. A full understanding of the processes a protein is involved in and the mechanisms one is regulated is crucial for its association with the development of diseases and determining potential drug targets. Furthermore, this knowledge aids in the design of drugs that can disrupt harmful protein interactions or enhance beneficial ones[20]. Protein-gene interactions, on the other hand, have less attention from researchers, but they play a very important role in cellular functionality. Proteins, the molecular workhorses that facilitate most biological processes in a cell, including gene expression, cell growth, and many others, execute a myriad of tasks ranging from catalyzing chemical reactions to transmitting signals. Genes, the building blocks of life, encode the instruction of the information for making specific molecules and proteins that allow cells to function and operate. It is through the dynamic interrelationship between both proteins and genes that life has its mysterious charm. Protein-gene interaction constitutes the foundation of living organisms. At the molecular level, proteins act as the messengers and executors of genetic information, catalyzing reactions, regulating gene expression, and responding to signals from the environment[1]. Genes, on the other hand, encode the information required for protein construction and assembly. However, protein-gene interaction is not a linear process; instead, it is characterized by networks of relationships where proteins and genes collaboratively shape the dynamics of cellular systems. While their combined effort creates biological functions, each component plays a unique role as we go deeper into these interactions, where we will meet the challenges of understanding the mechanisms and the implications for health and disease.

2.2 Challenges in Studying Protein-Gene Interactions

One of the most well-known challenges in studying protein-gene interactions is data integration. That includes extracting biologically relevant information from massive amounts of microarray data, gene acronym redundancies, protein synonyms, incomplete information, and potential false positives, as well as noise data, bias towards disease nodes, and the need for robust methods to capture the interplay among disease-associated genes[**bi10**][7][8]. Also, integrating different sources, such as experimental studies, omics data(genomics, transcriptomics, proteomics), and literature mining, can make it difficult to create a comprehensive view of the interactions[22]. The unstructured nature of biomedical literature makes it hard to extract computationally tractable data elements such as protein or genetic interactions and identify physically interacting protein pairs, as the text does not necessarily imply physical interaction[18]. In addition, the need to reconcile and unify various genetic interaction terminologies used within different model organism research communities is a challenge in accurately representing and integrating genetic interaction data across multiple species

Another aspect of protein-gene interaction challenges points towards data complexity, network heterogeneity, and the dynamic nature of biological systems. Since high-throughput technology generates a massive amount of data, including gene expression profiles, protein-protein interaction networks, and epigenetic modifications, the study of protein-gene interactions confronts an extremely large complexity of biological data. However, these data are often noisy and heterogeneous, which makes it difficult to extract meaningful insights from such complex datasets[3]. Additionally, multiple omics data sources, each capturing different aspects of cellular behavior, demand sophisticated algorithms that can account for variations and correlations among diverse data types. Protein-gene interactions form a heterogeneous network, where nodes represent proteins or genes and edges represent their interactions. In order to construct accurate and comprehensive network models, obstacles like the diversity caused by heterogeneous spans across various scales must be overcome. Network heterogeneity underscores the importance of considering multiple layers of interactions and adopting advanced algorithms that can capture the complex nature of biological networks. Inherently, biological systems are dynamic. Many factors, like cellular response to stimuli, development stages, and environmental conditions, can lead to altered protein functions[6]. This dynamic nature challenges the static representation of interactions in traditional network models. Moreover, capturing the dynamics requires time-series data and modeling approaches that can simulate the temporal evolution of interactions. The challenges listed above are the existing literature discovered and should be addressed properly in this paper.

2.3 Complexity of Biological Networks

Biological networks exhibit an intricate structure and dynamic behavior. While the complexity of biological networks presents in multiple components and their interactions, such as proteins, genes, and signal molecules, it is worth noting that the differential expression of genes in different cell types, which is regulated by extracellular signals and influences cellular behavior and specialization, has also raised the complexity[15]. This dynamic behavior of biological networks is one of the key complexities as the networks are not static entities; they respond to various internal and external cues, adapting and changing over time. This dynamic nature adds another layer of complexity, as it requires the integration of temporal and spatial information to capture the full spectrum of network dynamics. Additionally, biological networks exhibit scale-free properties, meaning that a small number of highly connected nodes (hubs) coexist with numerous nodes that have fewer connections. This feature contributes to the resilience and robustness of the network against perturbations.

Within the intricate landscape, biological networks often are like spider webs, weaving together proteins, genes, and other biomolecules in a harmonious symphony that orchestrates life's processes. Each thread of interaction, represented by nodes and edges, contributes to connections that dictate the behavior of cells and organisms. Nodes within these networks can represent a diverse array of entities, from individual genes and proteins to entire cellular pathways. Each node carries a biological significance, and their connections form the basis of functional relationships. Edges, which link nodes, capture an assortment of interactions, from physical binding events to regulatory influences[11]. This intricate web of interactions leads to the emergence of complex network structures.

Studying biological networks presents both challenges and opportunities. As we delve more into the field of complexity of biological networks, more insights into fundamental biological processes and disease mechanisms are gained. By understanding how interactions within these networks give rise to emergent properties, we can unlock new ways for drug discovery, disease diagnosis, and personalized medicine. The study of biological networks transcends disciplinary boundaries, uniting researchers from biology, mathematics, computer science, and other fields in a collaborative endeavor to unravel the mysteries of life's intricate connectivity.

2.4 Emergence of Multilayer Networks

A multilayer network is a framework for studying complex systems that have many interdependencies not properly captured by single-layer networks. These networks consist of a set of entities that interact with each other in various patterns, encompassing multiple types of relationships and changes over time. The study of multilayer networks aims to understand the interconnectedness and dynamics of these systems, taking into account their multilayer features[11].

The concept of multilayer networks has emerged from multiple disciplines and has become an important direction in network science. In a multilayer network, nodes can be connected to other nodes within the same layer (intralayer links) or across different layers (interlayer or coupling links)[2]. This framework allows for a more accurate representation of real-world systems, such as online social systems, where different layers may represent different platforms or types of interactions. For example, in the context of protein-gene interactions, these layers could correspond to different databases or experimental techniques. The layers are interconnected by cross-layer links, enabling researchers to explore connections between nodes across different layers. This arrangement creates a rich interactional representation that spans multiple dimensions, unveiling relationships that might remain hidden in single-layer networks.

As for biological systems, multilayer networks stand by with several advantages. By generalizing traditional network theory and extending to a more comprehensive framework, multilayer networks shed light on how genes and proteins collaborate across multiple contexts. They facilitate the identification of nodes that play critical roles across different layers, pinpointing key regulators and drivers of network dynamics. Moreover, multilayer networks help in understanding the evolution of biological systems by considering the interplay between different layers and their impact on evolutionary processes.

The utilization of multilayer networks offers a compelling framework for extracting deeper insights from complex biological data. By integrating multiple layers of interaction data, we can uncover hidden patterns and emergent properties that are often overlooked when analyzing each layer in isolation. Through this approach, we are not only able to comprehend the individual components of the system but also find out how they interact and influence one another across different dimensions. The power of sophisticated computational techniques present in multilayer network frameworks that are drawn from fields such as network science, graph theory, and machine learning empowers researchers to identify functional modules and characterize the robustness of network structures. The emergence of multilayer networks marks a transformative step towards understanding the multifaceted nature of cellular networks and the mysteries they hold.

2.5 Existing Approaches and Limitations

Traditionally, the study of protein-gene interactions has relied on conventional methods that often overlook the nuanced complexities of these intricate networks. Conventional approaches often focus on analyzing single-layer networks that capture specific types of interactions, such as protein-protein or protein-DNA interactions. For protein-protein interactions, computational methods, such as graph representation learning and deep learning architectures, have been developed to predict protein-protein interactions based on sequence and surface representations[23]. These methods have shown promising results in predicting interaction sites and interactions of proteins, which can be valuable for screening protein pairs in drug development. On the other hand, experimental methods, including affinity purification, cross-linking, and gel electrophoresis, have been used to purify protein complexes and investigate protein-protein interactions in vivo[17]. For DNA-protein interactions, various methods and techniques are used in different fields. For example, DNA-footprinting to find DNA-sites, filtering binding assay for studying binding kinetics and affinities, and combined techniques to identify potent aptamers that bind to specific proteins[9]. While these methods have provided valuable insights into individual interaction types, they often fall short of capturing the holistic nature of interactions within the context of a dynamic cellular environment.

Protein-gene interaction approaches often face the limitation of lack of precision and accuracy as the techniques get more complicated and sensitive to certain interactors. The intricate interplay between different types of interactions and the inherent cross-talk that occurs across various sources can also be challenging when it comes to collecting these interactions. Such interdependencies require more advanced computational analytical tools and functions to gain insights from the complex biological networks. Multilayer networks provide a promising avenue to overcome these limitations, as they allow us to simultaneously consider diverse types of interactions and their temporal dynamics. By moving beyond the constraints of single-layer analyses, we can unlock new insights into the co-evolution, interdependence, and emergent properties of protein-gene interactions within the intricate context of cellular systems.

2.6 Research Goals and Objectives

The goal of this research is to harness the potential multilayer networks to unravel the intricate landscape of protein-gene interactions and understand more about the dynamic interplay that shapes cellular behavior. The central focus lies in addressing key questions related to co-evolution, interaction scales, and the integrative potential of multilayer networks.

The very first but most crucial objective of this thesis is to integrate data from multiple sources, which are STRING, IntAct, and BioGrid, into a unified data structure based on their interactor identifiers. This process allows us to consolidate information from different sources and enhance the accuracy and scope of our investigation.

Then, aligned with our research goals, the project is rooted in harnessing the potential of multilayer networks to provide a holistic understanding of protein-gene interactions. By integrating diverse interaction types into a unified framework with proper consideration, we aim to capture the complex interdependencies and cross-layer correlations that shape the functional landscape of biological systems.

With the constructed multilayer networks, it is possible to go deep into the concept of co-evolution of protein-gene interactions. This thesis is interested in whether different types of interactions among genes and proteins have evolved in a way that they work together or complement each other. In biological systems, as mentioned previously, interactions are diverse, ranging from physical interactions between proteins to regulatory interactions between genes. The question is whether these different types of interactions have developed in a coordinated manner to achieve specific functions or outcomes. For example, consider a scenario where a protein-protein interaction is crucial for a specific cellular process, but at the same time, a gene regulatory interaction (transcriptional regulation) is also necessary to fine-tune the expression of the involved proteins. The research would seek to understand whether these types of interactions tend to occur together more often than expected by chance, suggesting a potential co-evolution to achieve functional synergy. To answer this question, we would need to analyze the accumulated interaction network and look for patterns of co-occurrence between different types of interactions. Statistical methods can be applied to assess whether certain types of interactions tend to appear together more frequently than expected. If we find a significant correlation between certain types of interactions, it could suggest a form of co-evolution or functional interdependence.

Another aspect of research in this thesis is to understand whether different types of interactions predominantly influence local cellular processes or have a broader impact on the entire cellular network. Local interactions might involve direct physical interactions between neighboring proteins, while global interactions could involve regulatory relationships that affect multiple components across the network. For instance, consider a gene regulatory network where certain transcription factors control the expression of genes involved in diverse pathways. The research question would explore whether these regulatory interactions have a more global influence on the network, affecting multiple cellular processes, or if they primarily regulate specific local functions. To address this question, we would need to analyze the interaction network and assess the scale of influence for different types of interactions. This could involve measuring the number of direct neighbors for each node in the network (degree distribution) or calculating network centrality measures.

In essence, these research questions aim to uncover the dynamics of how different types of interactions have evolved and how their effects are distributed across various scales within biological networks. By addressing these questions, the goal is to gain insights into how different types of interactions contribute to the overall functionality and behavior of the biological system.

3 Description of the Investigation

3.1 Construction of Three-Layer Multilayer Network

The investigation of protein-gene interactions starts by constructing the multilayer network that has the data across various sources, in this case, three prominent biointeraction databases: STRING, IntAct, and BioGrid. Every database has different sources of data that align with the cellular dynamic of biological networks and, therefore, be able to provide valuable insights into distinct types of interactions. The STRING Database The STRING database integrated information on known and predicted associations between proteins. It collects and scores evidence from various sources, including automated text mining of scientific literature, databases of interaction experiments and annotated complexes/pathways, computational interaction predictions, and systematic transfers of interaction evidence from one organism to another[22]. The database content is pre-computed, stored in a relational database, and available for separate download. Each interaction contains seven distinct evidence channels: neighborhood, fusion, co-occurrence, experiments, co-expression, knowledge, and text-mining, with interactor identifiers using the Ensembl genome database.

3.1.1 The IntAct Database

The IntAct database curates mostly protein-protein interaction data. These interactions are either curated from the literature or directly deposited by researchers. Each interaction in IntAct undergoes quality control, including peer review by a senior curator and additional rule-based checks at the database level. The original authors of the publications are also contacted to ensure the accuracy of the representation of their data[13]. With their strictness, IntAct provides the most detailed information in many formats; for example, IntAct PSI-MI TAB format 2.7 contains 42 columns for each interaction, with interactor identifiers using the UniprotKB database.

3.1.2 The Biogrid Database

The BioGrid database focuses on protein, genetic, and chemical interactions for various model organism species and humans. It contains records for biological interactions manually annotated from publications, including protein-protein, protein-gene, and chemical-protein interactions. Additionally, BioGRID annotates genome-wide CRISPR/Cas9-based screens that report gene-phenotype and gene-gene relationships[18]. The BioGrid database also provides many formats to store the data in; for simplicity, we choose PSI$_M ITAB, justlikeIntAct, which$

3.1.3 Protein Identifiers Mapping

After introducing three databases, it is thorough to integrate them into a unified data structure by the alignment of interactor identifiers. This approach not only allows us to create a unified framework for analysis but also facilitates the comparison of interactions across databases. In practice, it is possible to map the databases together with any interactor identifiers possible, such as gene symbols other than the ones from the given databases. However, in this thesis, we are using UniProtKB for alignment out of simplicity.

3.1.4 Network of Networks vs Multiplex Networks

Choosing which type of multilayer networks should also be considered for their different properties. In a network of networks, the individual networks themselves are treated as nodes in a higher-level network. Each of these individual networks is referred to as a layer. The connections between layers represent interactions or relationships between the networks. This framework allows for a more comprehensive understanding of the dynamics and behavior of complex systems that involve multiple interconnected networks. On the other hand, A multiplex network is a multilayer network where multiple types of relationships exist between the same set of nodes (entities). Each layer in a multiplex network corresponds to a different type of interaction or relationship between the nodes. Nodes can have connections in one or more layers, and the connections in each layer can represent different types of interactions[14]. In this paper, the aim is to investigate the relationship between interactions for the same genes or proteins performed across databases. Therefore, it is wise to go with the framework of multiplex networks.

The edges that cross layers(inter-layer links) act as bridges, connecting nodes across layers and enabling us to trace their roles and interactions across different contexts. By examining the connections between layers, we can uncover how nodes engage in diverse relationships and how these relationships evolve over time. These inter-layer connections are particularly valuable when exploring questions related to co-evolution and the interplay between local and global interactions.

3.2 Computation of Analytical Measures

The quantification of nodes' importance within biological networks is a cornerstone of network analysis. Centrality measures in multilayer networks are quantitative metrics used to identify the importance or influence of nodes within a multilayer network. These measures play a crucial role in assessing the significance of individual nodes, uncovering key players, and identifying potential hubs in the network. In our investigation, we look into three fundamental centrality measures: degree centrality, betweenness centrality, and closeness centrality, as well as the clustering coefficient. These measures enable us to analyze the roles of nodes within and across layers of the multilayer network.

3.2.1 Degree Centrality

Degree centrality measures the number of connections a node has across all layers, indicating its overall importance in the network. Calculating degree centrality involves tallying the connections a node has, irrespective of their weight or direction. The computation is defined as

$$C_D\ (i) = \frac{(degree(i))}{(n-1)}$$

where degree is the number of edges a node has, and n is the total number of nodes.

In our multilayer network, we compute the degree centrality for each node in each layer. This enables us to compare the participation of nodes in distinct interaction types, shedding light on their overall connectivity within the biological system.

Degree centrality directly addresses our research question regarding the co-evolution of different interaction types. Nodes with a high degree centrality in one layer but a low degree centrality in another may indicate differences in the co-evolution of interactions. Such variations could imply compensatory mechanisms or enhanced functional roles in one type of interaction compared to another.

3.2.2 Betweenness Centrality

Betweenness centrality quantifies the extent to which a node lies on the shortest paths between other nodes in the network, considering all layers. Betweenness centrality focuses on the role of nodes in mediating the flow of information across the network. Nodes with high betweenness centrality act as crucial intermediaries in communication paths between other nodes. Computation of betweenness centrality considers the fraction of shortest paths passing through a given node, which is defined as

$$C_B\ (v)\ =\ \sum_{s,\ t \in V} \frac{(\sigma(s,\ t|v))}{(\sigma(s,\ t))}$$

where V is the set of nodes, $\sigma(s,t)$ is the number of shortest (s, t)-paths, and $\sigma(s,t|v)$ is the number of those paths passing through some node v other than s, t. If s = t, $\sigma(s,t) = 1$, and if $v \in s,t$, $\sigma(s,t|v) = 0$ [4][5].

The calculation of betweenness centrality directly relates to our exploration of different interaction scales, specifically local versus global interactions. However, due to the limitation of computation on personal computers, we are not able to compute the complete betweenness centrality for every node in the network; instead, only 1 percent of the nodes in each layer are calculated. Nodes with high betweenness centrality may exhibit global influence by connecting disparate parts of the network. Conversely, nodes with low betweenness centrality could signify nodes predominantly engaged in local interactions. Through betweenness centrality, we aim to identify nodes that play significant roles in maintaining communication across the network, potentially revealing insights into the interplay of local and global interactions.

3.2.3 Closeness Centrality

Closeness centrality measures how close a node is to all other nodes in the network, taking into account the shortest paths across layers. Nodes with high closeness centrality are able to quickly transmit information to their neighbors, making them pivotal for local interactions and efficient communication. Closeness centrality is calculated as the reciprocal of the sum of the shortest path lengths from a node to all other nodes. The computation is defined as

$$C(u) \ = \frac{n-1}{\sum_{v=1}^{n-1} d(v,\ u)}$$

where d(v, u) is the shortest-path distance between v and u, and n-1 is the number of nodes reachable from u[10]. In our investigation, assessing closeness centrality for each node in each layer unveils nodes that are strategically positioned to influence nearby nodes, fostering efficient communication within specific interaction contexts. In practice, the computation for closeness centrality for each node is time-consuming; therefore, we used another approach, which leverages the adjacency matrix and computes the shortest paths using the Floyd-Warshall Method. Then, use the shortest path Matrix to calculate the closeness metric for each node.

```
A = nx.adjacency_matrix(G).tolil()
D = scipy.sparse.csgraph.floyd_warshall( \
        A, directed=False, unweighted=False)
n = D.shape[0]
closeness_centrality = {}
for r in range(0, n):

    cc = 0.0

    possible_paths = list(enumerate(D[r, :]))
    shortest_paths = dict(filter( \
        lambda x: not x[1] == np.inf, possible_paths))

    total = sum(shortest_paths.values())
    n_shortest_paths = len(shortest_paths) - 1.0
    if total > 0.0 and n > 1:
        s = n_shortest_paths / (n - 1)
        cc = (n_shortest_paths / total) * s
        closeness_centrality[r] = cc
```

Closeness centrality assists in addressing our research question concerning different interaction scales. Nodes with high closeness centrality can indicate their potential to engage in local interactions, contributing to the formation of clusters or sub-networks. Conversely, nodes with lower closeness centrality may participate in more global interactions, connecting distant regions of the network. Through closeness centrality analysis, we aim to distinguish nodes that are hubs of local interactions from those that facilitate global communication.

3.2.4 Clustering Coefficient

Clustering coefficient in multilayer networks is a measure of the extent to which nodes in a network tend to form clusters or groups. The clustering coefficient offers insights into the nature of nodes to form tightly connected groups or clusters. This measure quantifies the extent to which a node's neighbors are interconnected. High clustering coefficients indicate nodes participating in tightly-knit neighborhoods, suggesting the presence of functional modules or pathways. By calculating clustering coefficients across layers, we can identify nodes that foster local interactions and modular structures within the network. For weighted networks, the computation is defined as

$$C_u = \frac{1}{deg(u)(deg(u)-1)} \sum_{vw} (\hat{w}_{uv}\hat{w}_{uw}\hat{w}_{vw})^{1/3}$$

where deg(u) is the degree of u and edge weight $\hat{w}_{uv}$ are normalized by the maximum weight in the network[19][16].

The clustering coefficient contributes to our exploration of local interactions and network structure. Nodes with high clustering coefficients may contribute to the formation of local clusters, potentially indicating functional cooperation or pathway cohesiveness. On the other hand, nodes with lower clustering coefficients might play roles in facilitating global connections.

3.3 Correlation Analysis Across Layers

The investigation into the correlations between centrality measures across different layers serves as a beacon guiding us through the complexities of multilayer networks. The methodology and significance of correlation analysis decipher the connections that exist across diverse interaction types. It is necessary to choose correlation methods depending on the characteristics of the nature of our analysis.

Pearson correlation is used to measure the linear relationship between two continuous variables. It assumes that the data is normally distributed and that the relationship between the variables is linear. It quantifies the strength and direction of a linear relationship, where a positive correlation indicates that as one variable increases, the other tends to increase, and vice versa for a negative correlation.

Spearman correlation, also known as rank correlation, measures the strength and direction of a monotonic relationship between two variables. It does not assume a linear relationship and is appropriate for variables that are ordinal or not normally distributed.

Spearman correlation is calculated based on the ranks of the data rather than the actual values.

For this thesis, consider the quantified linear relationships of the data, the simple interpretability, and the standardization in the research field. In the context of studying multilayer networks, where nodes are linked across different layers, Pearson's correlation coefficient can be extended to quantify the relationships between centrality measures across layers. This helps researchers understand how nodes' importance or centrality changes across different types of interactions. The computation is defined as

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

where m_x is the mean of the vector x and m_y is the mean of the vector y[21].

Like other correlation coefficients, this one varies between -1 and +1, with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases. In the context of this thesis, a high positive correlation between centrality measures across layers indicates that the importance or prominence of nodes in one layer is consistently similar to their importance in another layer. A low or near-zero correlation between centrality measures across layers implies that the importance of nodes in one layer does not consistently correspond to their importance in another layer. Similar interpretations can be applied to clustering coefficients.

By analyzing high and low correlation values and clustering coefficients across layers, we can draw insights into how nodes' importance and local connectivity vary across different interaction types. These patterns can contribute to answering your research questions about co-evolution, local vs. global interactions, and the potential differences between interaction types in shaping network structure and node roles.

3.4 Sub-networks on Different Thresholds

It is also a good approach to perform the whole analysis on interactions for a high score threshold (accepting only highly significant interactions) and a low score threshold. Varying score thresholds influence the inclusivity of interactions in the network, allowing us to explore interactions that range from well-established and confidently supported to those that are more speculative or tentative. By analyzing networks constructed at different score thresholds, we can capture a spectrum of interactions that represent varying degrees of data confidence.

This approach provides insights into our research questions. For instance, when examining the co-evolution of different interaction types, exploring how interactions evolve and strengthen across different score thresholds can reveal patterns of compensatory or synergistic co-evolution. Additionally, by analyzing interactions on both local and global scales using different thresholds, we can recognize whether certain types of interactions predominantly operate in specific contexts.

There are a few reasons to choose different thresholds, namely high and low score thresholds. By considering only highly significant interactions, we emphasize strong relationships in the network. This can help us identify the most crucial nodes and interactions in the system. With such nodes, Analyzing centrality measures and clustering coefficients with these interactions can reveal the core of the network, providing insights into its robustness and key communication pathways. Lower score thresholds, on the other hand, serve as a representation of those weaker and less understood relationships. There might be emergent properties or unexpected relationships invisible to high score interactions. This can lead to the discovery of novel pathways or regulatory mechanisms. By comparing the analytical measures between high and low score threshold networks, insights into how networks change over time or under specific conditions can be drawn.

In summary, performing network analysis with different score thresholds allows us to explore the network's behavior, structure, and functional modules under varying levels of interaction significance. By incorporating various centrality measures and clustering coefficients, we can uncover different aspects of node importance, information flow, and network organization for both strong and weak interactions. This comprehensive approach can provide a more nuanced understanding of the network's dynamics and functionality.

4 Evaluation of the Investigation

4.1 Multilayer Network

The multilayer network constructed in this thesis is a multiplex network, which means every node from each of the three databases is present in every layer, where nodes represent the gene or protein identifiers and layers are the databases. Because of this property, it is common that some nodes in one layer have many intra-layer edges but remain unlinked in other layers, which would be really interesting for us to investigate. Every node in one layer has an inter-layer edge connecting to its counterparts in other layers, forming a diagonal pattern. This means that the coupling edges and their weights are independent of the nodes, and the adjacency matrix is weighted accordingly[14].

Figure 1 shows a better visualization of how the multiplex network should look like. The figure vividly captures the layer-specific clustering and interplay, offering a visual narrative that resonates with the underlying intricacies of biological interactions. Amid this complexity, certain nodes emerge unlinked, underscoring the multifaceted nature of interactions and highlighting that not all nodes are universally connected.

4.2 Centrality Measures and Clustering Coefficient

Comparing different centrality measures(degree centrality, betweenness centrality, closeness centrality) and the clustering coefficient for the same node across different layers(database) in the multilayer network can provide valuable insights and address the key questions in this thesis. By choosing the appropriate interactor that has a relatively high centrality measure, it can be convenient to understand its position in the networks by comparing the same centrality in other layers for the same interactor. As Figure 2 shows, a relatively high degree centrality in the STRING layer does not guarantee degree centrality in the other two layers. It is also clear that a centrality measure can be correlated across layers for the same node, such as closeness centrality. Therefore, it is in our interest to research the true relationships between the centrality measures for the same node across layers.

4.2.1 Degree Centrality

Comparing degree centrality for the same node across layers can reveal how connected and influential a node is within each layer. As shown in figure 3, it is clear that high degree centrality nodes in the STRING layer do not have a significant impact on the degree centrality in the IntAct and BioGrid layers.

It is not enough to see only the highest nodes in the STRING database. Thus figure 4 and 5 show the highest nodes with Biogrid and IntAct databases, respectively. Since it is not clear to see and thus not able to make precise judgment, a rough one, like the top nodes with the highest degree centrality in each database, has their distinct importance in the local layer. Although it is hard to draw the conclusion of co-evolution based on degree centrality, it is fair to suggest that some nodes are globally important while others are locally influential within specific interaction types.

4.2.2 Betweenness Centrality

A similar analysis is done to the multilayer network but in the frame of betweenness centrality, which measures how often a node acts as a bridge along the shortest paths between other nodes. Figures 6, 7, and 8 plot the nodes with the highest betweenness centrality by each database. Similar conclusions like degree centrality are drawn, which supports the idea that different interactions operate on different scales. Especially having observed that high betweenness centrality for STRING databases leads to low betweenness centrality in BioGrid and, unsurprisingly, the same pattern for BioGrid to STRING.

4.2.3 Closeness Centrality

The clustering coefficient measures how well a node's neighbors are connected to each other. Unlike degree centrality and betweenness centrality, nodes from both the BioGrid and IntAct layers show a pattern that high closeness centrality in this layer leads to high value in another layer. See figure 9, 10, 11. This means that these nodes can quickly spread information across the network, which indicates that nodes are in the global scope.

4.2.4 Clustering Coefficient

The clustering coefficient measures how well a node's neighbors are connected to each other. Nodes with consistently high clustering coefficients across layers form tightly connected local neighborhoods, indicating local interactions. As shown in figures 12, 13, and 14, top nodes for one layer do not lead to similar high values in clustering coefficient for other layers, thus implying globally scoped.

4.3 Results of Correlation Analysis and Implications

With the almost contradictory conclusions we observed from different centrality measures and the clustering coefficients, it is obvious that node-to-node comparison is not a good approach for analysis. Therefore, correlation methods are required for a comprehensive and holistic view of the network. Consider the correlation coefficients in Figure 15 and link to our key research questions:

4.3.1 Key question 1: Have different types of interaction 'co-evolved' to compensate each other or to enhance each other?

Looking at the correlation results for degree centrality:

BioGrid and STRING: Moderate positive correlation (0.20)
BioGrid and IntAct: Moderate positive correlation (0.26)
STRING and IntAct: Strong positive correlation (0.44)

These correlation coefficients suggest that there is a degree of co-evolution or interplay between different types of interactions. Nodes that are highly connected in one database tend to be relatively highly connected in another database as well. The varying strengths of correlations indicate that there might be some compensation or enhancement between different types of interactions, with the strongest co-evolution between STRING and IntAct.

4.3.2 Key question 2: Do different types of interaction operate on different scales (local vs. global)?

Examining the correlation results for closeness centrality:

BioGrid and STRING: Moderate positive correlation (0.31)
BioGrid and IntAct: Strong positive correlation (0.50)
STRING and IntAct: Moderate positive correlation (0.42)

These correlation coefficients suggest that nodes with high closeness centrality in one layer tend to have relatively high closeness centrality in another layer. This could indicate that nodes with significant global influence in one layer tend to maintain similar levels of global influence in other layers. This consistency across different types of interactions implies some degree of commonality in their global roles.

Examining the correlation results for the clustering coefficient:

BioGrid and STRING: Weak positive correlation (0.21)
BioGrid and IntAct: Moderate positive correlation (0.41)
STRING and IntAct: Weak positive correlation (0.17)

These correlation coefficients suggest that there might be some similarity in the local clustering patterns of nodes across different types of interactions. Nodes with similar tendencies to form local clusters in one layer also exhibit similar behavior in other layers, albeit with varying strengths of correlation. This implies that certain local clustering characteristics might be shared across different interaction databases.

4.4 Threshold Analysis

Compared with correlation computation across different layers, it might make more sense to do the correlation computation for different thresholds for one database, as shown in figure 16. It is obvious that the correlations between different thresholds in one layer are much stronger than those in different layers. The strength of the correlation increases as the thresholds become more restrictive, implying that nodes with high connectivity in one threshold are likely to maintain their high connectivity in other thresholds.

The positive correlations observed in centrality measures and clustering coefficient across thresholds suggest that certain nodes maintain their relative importance and local interactions consistently, regardless of the data confidence level represented by the threshold. This has implications for understanding how different types of interactions co-evolve and how interaction types operate on different scales within biological networks.

4.5 Limitations and Outlooks

In our research, there are a few improvements that can be made with better consideration and design of analysis. First, not all data from the STRING database are used in this network; due to the limited storage and RAM space, only data from homo sapiens are being used. Second, while the current multilayer network only uses the two interactors and their confidence score, the integrated data structure actually contains many other useful information, namely the interaction type, detection methods, and interaction source if researchers show special interest in a certain pair of interactions. Also, a new multilayer network can be constructed by defining layers as different interaction types, which provides a more complicated network matching the dynamic and complex nature of biological systems.

Other limitations relate to the limit of the computational speed and RAM space of the used computer. By using a high-performance computer like a supercomputer, it is then possible to compute the full closeness centrality and betweenness centrality. In the context of this thesis, the closeness centrality is limited to the first 1 percent of the nodes, and betweenness centrality is limited to a few hundred.

However, regardless of the speed and space limitations, the integration of data structure and construction of multilayer networks are easy to perform, and with simple changes, it can help researchers perform further analyses without difficulties.

5 Conclusions

The correlation analysis between different types of interactions, as indicated by the degree centrality values, suggests that there is a certain level of co-evolution or interplay between these interactions. This phenomenon implies that different types of interactions might be influencing each other to some extent, potentially compensating or enhancing their roles in the network. Therefore, our findings suggest that different interaction types in biological networks are not isolated but rather interrelated, contributing to the overall network structure.

The consistent positive correlations across thresholds for analytic measures also suggest that different types of interactions in the network tend to co-evolve. This indicates that nodes with high centrality and strong local interactions in one layer are likely to exhibit similar characteristics in other layers as well. This co-evolution implies a functional interplay between interaction types, where nodes that are central in one layer also maintain their importance in others, reinforcing the idea that biological processes are driven by coordinated interactions.

The examination of closeness centrality values across different interaction databases provides insights into the global influence of nodes in the network. The positive correlations observed imply that nodes with significant global influence in one layer tend to maintain similar levels of influence in other layers. This suggests that the scale of interaction, whether local or global, might have some degree of commonality across different types of interactions. While local clustering and community formation might vary, our findings suggest that certain nodes play important roles on a global scale across all interaction types. Thus, our investigation indicates that the scale of interaction is not strictly segregated by interaction databases but shows overlapping patterns of influence.

The positive correlations across thresholds also shed light on the scale of interactions. Nodes that exhibit high centrality and clustering coefficient values are likely to play major roles in both local and global contexts. These nodes serve as key connectors between different parts of the network while also forming dense local clusters. As data confidence increases, some nodes may transition from being intermediaries between different network regions to becoming more locally focused, as evidenced by the variations in correlation strength for betweenness centrality. This suggests a dynamic interplay between local and global interactions.

In conclusion, the positive correlations observed across layers and thresholds provide strong support for the idea that interactions within biological networks co-evolve and that nodes with high analytic measures maintain their importance across interaction types and scales. These findings enhance our understanding of the intricate relationships within biological systems and highlight the dynamic nature of node roles in shaping network structure and function.

References

[1] Saliha Ece Acuner Ozbabacan et al. “Transient protein–protein interactions”. In: *Protein Engineering, Design & Selection* 24.9 (2011), pp. 635–648.

[2] Alberto Aleta and Yamir Moreno. “Multilayer networks in a nutshell”. In: *Annual Review of Condensed Matter Physics* 10 (2019), pp. 45–62.

[3] Mounia Bouabdellah et al. “Hybrid very high throughput satellites: Potential, challenges, and research directions”. In: *2020 IEEE Eighth International Conference on Communications and Networking (ComNet)*. IEEE. 2020, pp. 1–6.

[4] Ulrik Brandes. “A faster algorithm for betweenness centrality”. In: *Journal of mathematical sociology* 25.2 (2001), pp. 163–177.

[5] Ulrik Brandes. “On variants of shortest-path betweenness centrality and their generic computation”. In: *Social networks* 30.2 (2008), pp. 136–145.

[6] Paolo Carloni, Ursula Rothlisberger, and Michele Parrinello. “The role and perspective of ab initio molecular dynamics in the study of biological systems”. In: *Accounts of Chemical Research* 35.6 (2002), pp. 455–464.

[7] Juan Casado-Vela et al. “Protein-Protein Interactions: Gene Acronym Redundancies and Current Limitations Precluding Automated Data Integration”. In: *Proteomes* 1.1 (2013), pp. 3–24.

[8] Juan Casado-Vela et al. “Protein-Protein Interactions: Gene Acronym Redundancies and Current Limitations Precluding Automated Data Integration”. In: *Proteomes* 1.1 (2013), pp. 3–24.

[9] Ricardo André Campos Ferraz et al. “DNA–protein interaction studies: a historical and comparative analysis”. In: *Plant Methods* 17.1 (2021), pp. 1–21.

[10] Linton C Freeman et al. “Centrality in social networks: Conceptual clarification”. In: *Social network: critical concepts in sociology. Londres: Routledge* 1 (2002), pp. 238–263.

[11] Zaynab Hammoud and Frank Kramer. “Multilayer networks: aspects, implementations, and application in biomedicine”. In: *Big Data Analytics* 5.1 (2020), p. 2.

[12] Lun Hu et al. “A novel network-based algorithm for predicting protein-protein interactions using gene ontology”. In: *Frontiers in Microbiology* 12 (2021), p. 735329.

[13] Samuel Kerrien et al. “The IntAct molecular interaction database in 2012”. In: *Nucleic acids research* 40.D1 (2012), pp. D841–D846.

[14] Mikko Kivelä et al. “Multilayer networks”. In: *Journal of complex networks* 2.3 (2014), pp. 203–271.

[15] Avi Ma’ayan. “Complex systems biology”. In: *Journal of the Royal Society Interface* 14.134 (2017), p. 20170391.

[16] Jukka-Pekka Onnela et al. “Intensity and coherence of motifs in weighted complex networks”. In: *Physical Review E* 71.6 (2005), p. 065103.

[17] Oliver Orasch et al. “Protein–Protein Interaction Prediction for Targeted Protein Degradation”. In: *International Journal of Molecular Sciences* 23.13 (2022), p. 7033.

[18] Rose Oughtred et al. “The BioGRID interaction database: 2019 update”. In: *Nucleic acids research* 47.D1 (2019), pp. D529–D541.

[19] Jari Saramäki et al. “Generalizations of the clustering coefficient to weighted complex networks”. In: *Physical Review E* 75.2 (2007), p. 027105.

[20] Malgorzata Skwarczynska and Christian Ottmann. “Protein–protein interactions as drug targets”. In: *Future medicinal chemistry* 7.16 (2015), pp. 2195–2219.

[21] Student. “Probable error of a correlation coefficient”. In: *Biometrika* 6.2-3 (1908), pp. 302–310.

[22] Damian Szklarczyk et al. “The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets”. In: *Nucleic acids research* 49.D1 (2021), pp. D605–D612.

[23] Yuko Tsuchiya, Yu Yamamori, and Kentaro Tomii. “Protein–protein interaction prediction methods: from docking-based to AI-based approaches”. In: *Biophysical Reviews* 14.6 (2022), pp. 1341–1348.

A Appendix

A.1 Figures

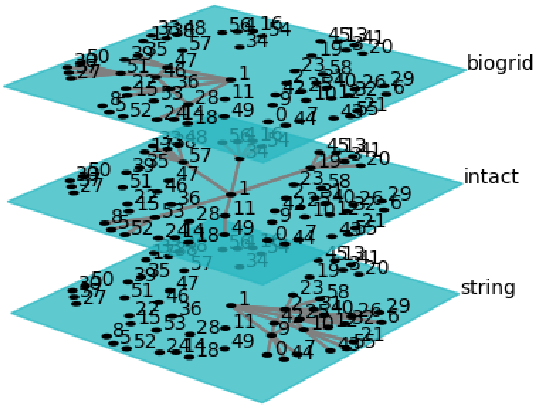

Figure 1: The network starts from a common node and links to different nodes in each layer. As those nodes connected by node ”1” also connect each other and further nodes, a dynamic hierarchy of interactions becomes apparent.

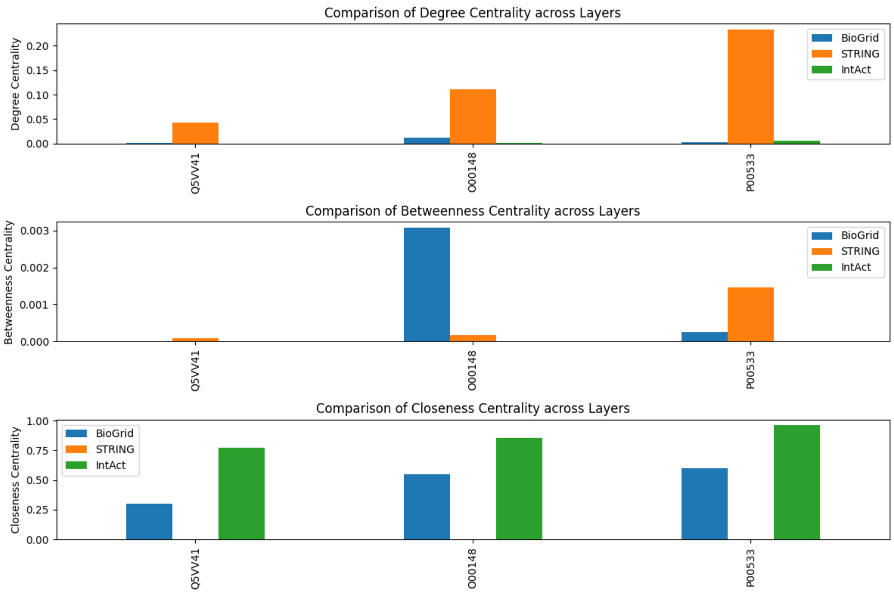

Figure 2: three distinct interactors are presented with their centrality measures comparison across three layers.

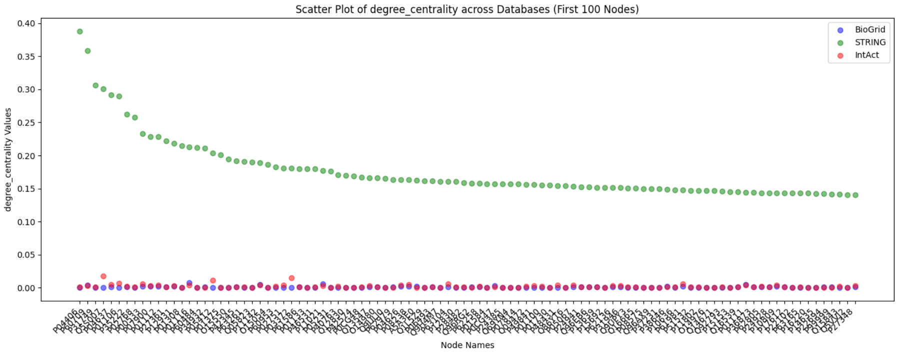

Figure 3: the first 100 nodes with the highest degree centrality in the STRING database are presented in ascending order, with their counterparts in other layers.

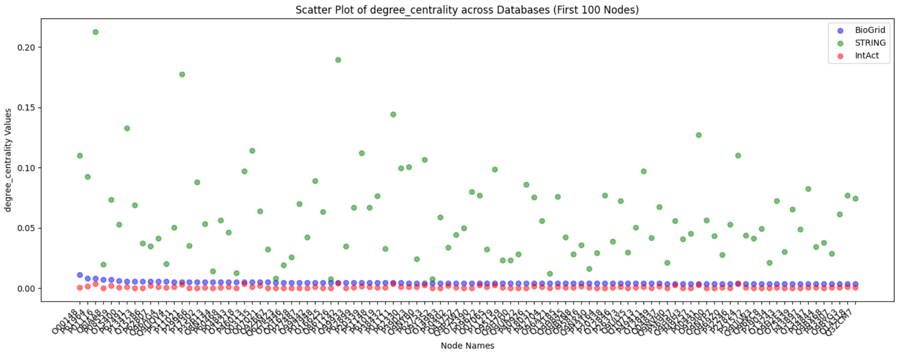

Figure 4: the first 100 nodes with the highest degree centrality in the Biogrid database are presented in ascending order, with their counterparts in other layers.

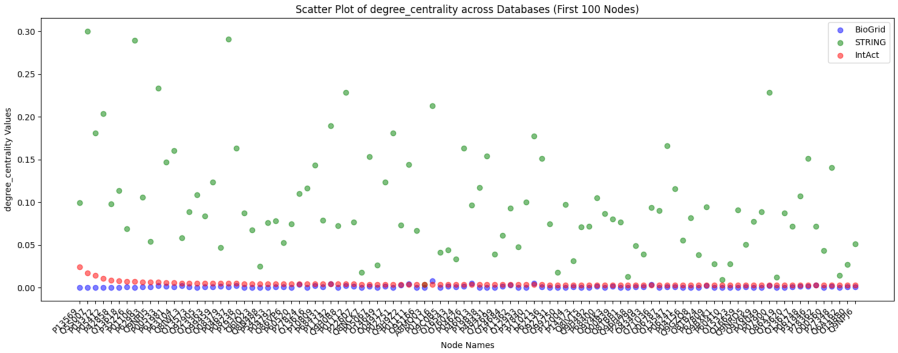

Figure 5: the first 100 nodes with the highest degree centrality in the IntAct database are presented in ascending order, with their counterparts in other layers.

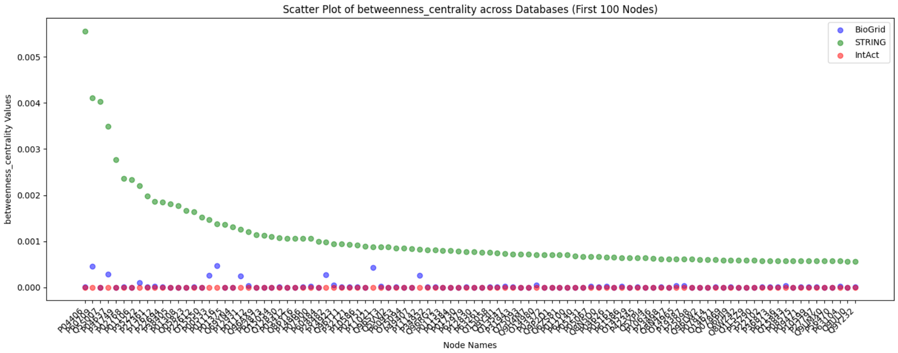

Figure 6: the first 100 nodes with the highest betweenness centrality in the STRING database are presented in ascending order, with their counterparts in other layers.

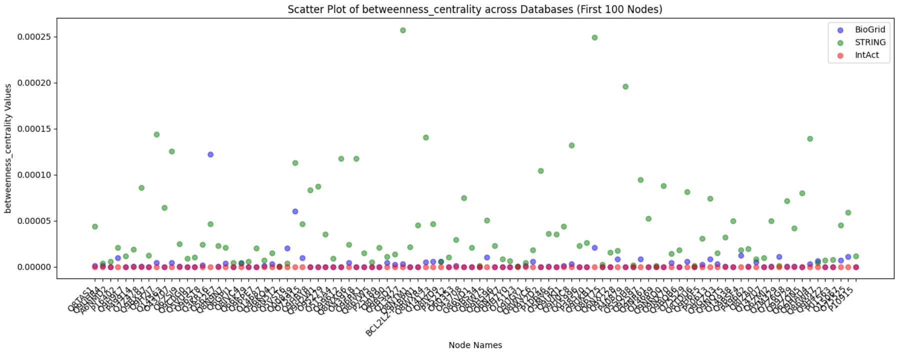

Figure 7: the first 100 nodes with the highest betweenness centrality in the IntAct database are presented in ascending order, with their counterparts in other layers.

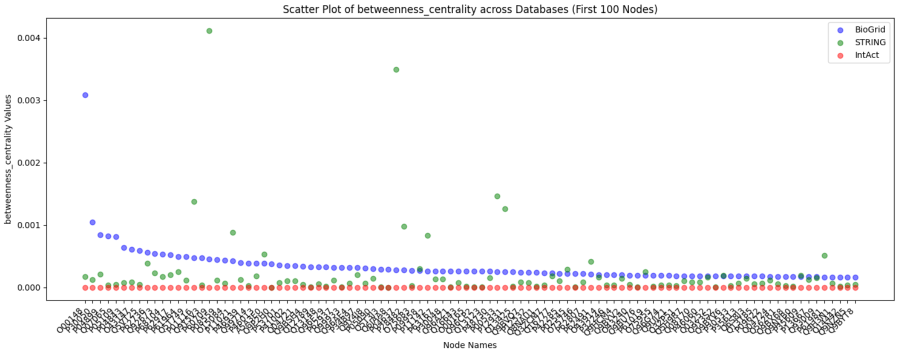

Figure 8: the first 100 nodes with the highest betweenness centrality in the BioGrid database are presented in ascending order, with their counterparts in other layers.

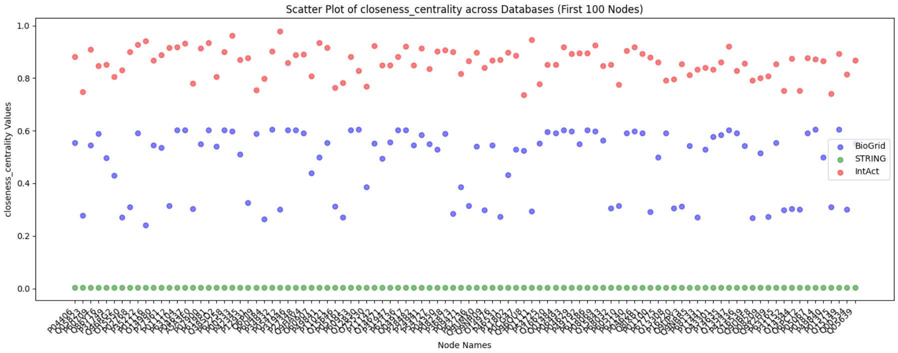

Figure 9: the first 100 nodes with the highest closeness centrality in the STRING database are presented in ascending order, with their counterparts in other layers.

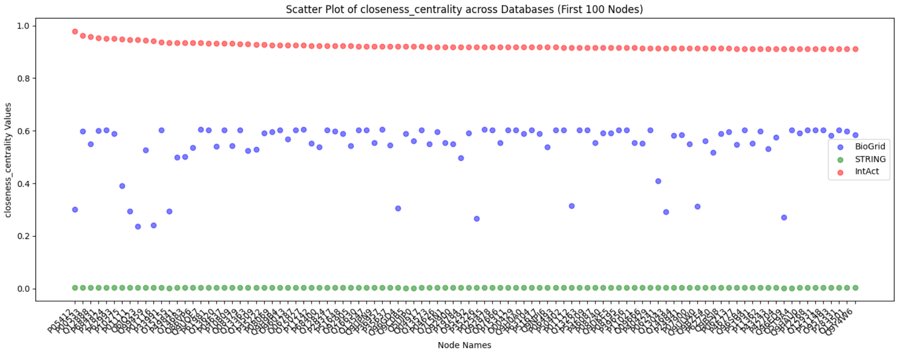

Figure 10: the first 100 nodes with the highest closeness centrality in the IntAct database are presented in ascending order, with their counterparts in other layers.

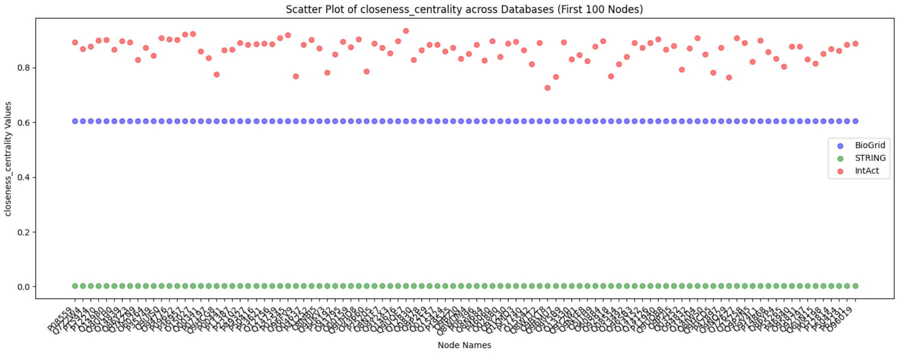

Figure 11: the first 100 nodes with the highest closeness centrality in the BioGrid database are presented in ascending order, with their counterparts in other layers.

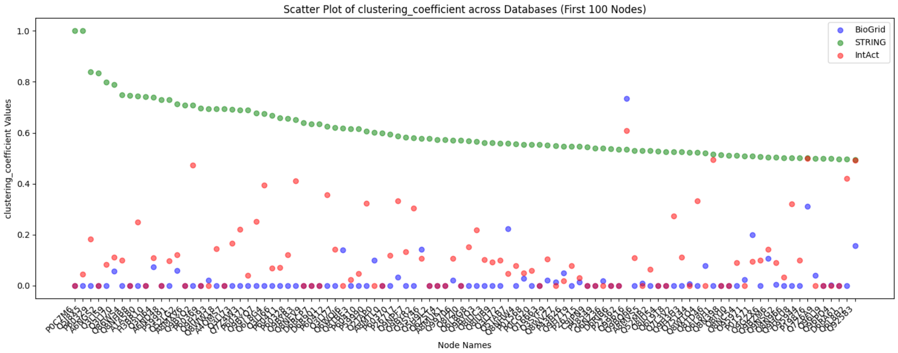

Figure 12: the first 100 nodes with the highest clustering coefficient in the STRING database are presented in ascending order, with their counterparts in other layers.

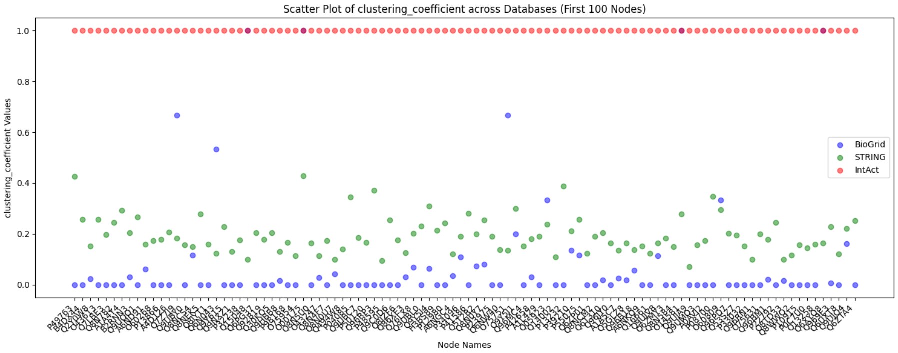

Figure 13: the first 100 nodes with the highest clustering coefficient in the IntAct database are presented in ascending order, with their counterparts in other layers.

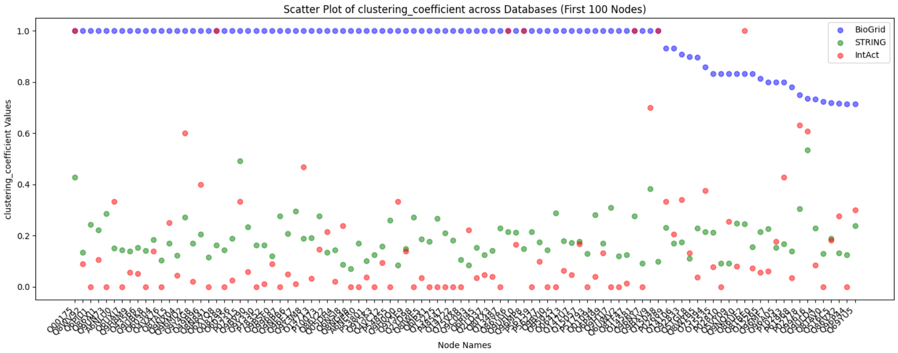

Figure 14: the first 100 nodes with the highest clustering coefficient in the BioGrid database are presented in ascending order, with their counterparts in other layers.

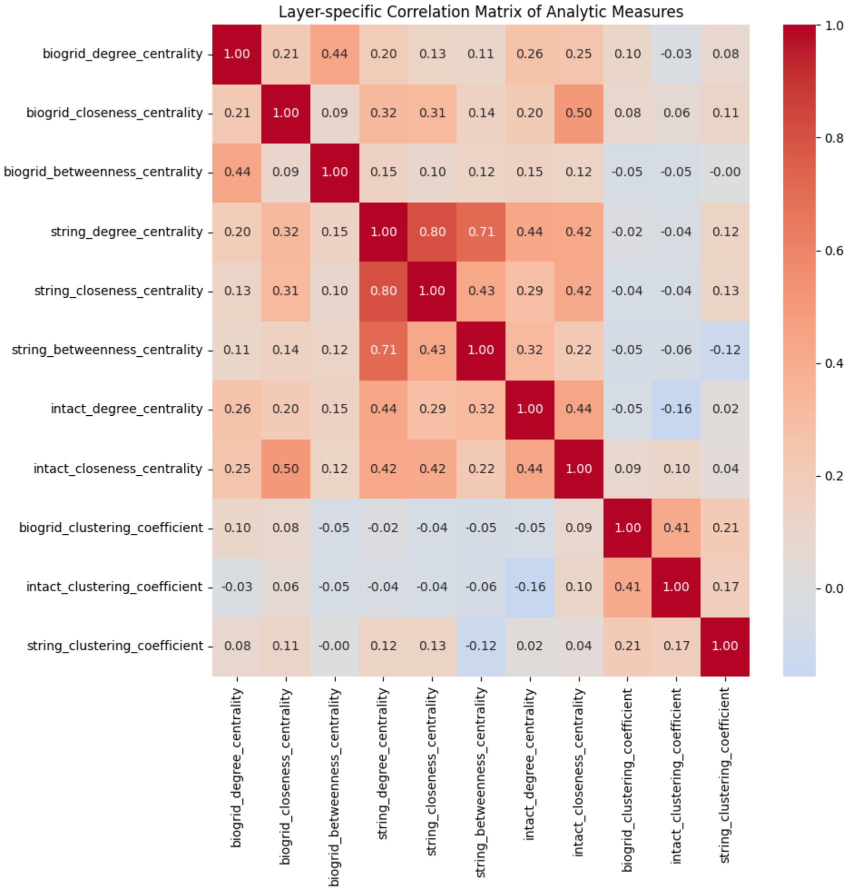

Figure 15: correlation on every analytic measure in every layer.

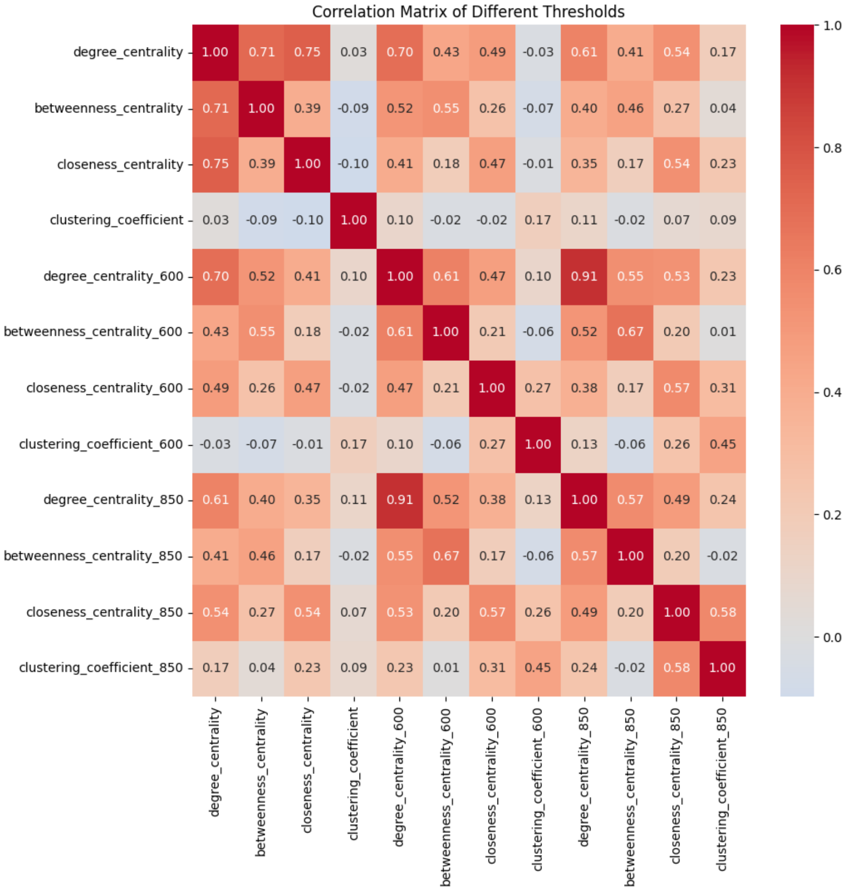

Figure 16: Correlation coefficients of different analytic measures on different score thresholds, namely no threshold, 600, and 850 in the STRING database.