Hierarchical Community Detection on Co-expression Networks for Functional Classification of Cell-Cycle Regulated Genes of *Saccharomyces cerevisiae*

Jaimielle Kyle Calderon[1] and Princess Angel Ventures[2]

[1] Algorithms and Complexity Lab,
University of the Philippines Diliman
jccalderon@up.edu.ph
[2] Algorithms and Complexity Lab,
University of the Philippines Diliman
pventures@up.edu.ph

Abstract. This study explored hierarchical community detection in gene co-expression networks to enhance the functional classification of cell-cycle regulated genes in *Saccharomyces cerevisiae*. Biological networks, which illustrate complex interactions among genes and proteins, required sophisticated methods to uncover their intricate structures. We evaluated three hierarchical community detection algorithms—Girvan-Newman (GN), Paris, and Local Optimization Function Model (LFM)—on a gene co-expression network of *Saccharomyces cerevisiae*. Our findings showed that the GN algorithm effectively identified distinct community structures at various hierarchical levels, demonstrating high modularity and closeness centrality. The Paris algorithm provided a comprehensive view of the network's hierarchical structure, while the LFM algorithm revealed detailed sub-communities sensitive to the alpha parameter. These algorithms offered valuable insights into the hierarchical structures within biological networks, contributing uniquely to our understanding of gene co-expression patterns. The study's results were validated using modularity, closeness centrality, and the Adjusted Rand Index (ARI), highlighting their potential applications in genomics and cellular biology. This research advanced the field by presenting new perspectives and methods for biological network analysis, paving the way for more targeted and effective approaches to understanding gene functions.

Keywords: Networks · Community Detection · Gene Co-expression · Hierarchical Structures · *Saccharomyces cerevisiae* · Biological Networks · Modularity · Closeness Centrality · Adjusted Rand Index (ARI)

1 Introduction

Prior research highlighted the urgent need for more sophisticated methods to unravel the complexities of biological networks. These networks, illustrating interactions among genes, proteins, and other biological entities, were incredibly

intricate. This complexity made it difficult to accurately interpret and analyze these networks, as noted by Besas and Redeakar. Their work underscored the importance of community detection in uncovering hidden structures and functional relationships within these networks.

Our study focuses on one specific challenge: exploring hierarchical community structures. This area holds great promise for uncovering the layered intricacies of biological networks, potentially revealing new insights into how different elements within these networks are interconnected and function. Building on foundational research by Rossetti and others, we aim to expand upon their methodologies and findings.

The goal of our study is to assess the capability of predicting hierarchical sub-clusters within the *Saccharomyces cerevisiae* gene co-expression dataset. By addressing this challenge, we hope to open new pathways for research and application in understanding biological networks.

We evaluated the performance of three hierarchical community detection algorithms—Girvan-Newman (GN), Paris, and Local Optimization Function Model (LFM)—on the gene co-expression network of *Saccharomyces cerevisiae*. The GN algorithm effectively identified distinct community structures at multiple hierarchical levels, demonstrating high modularity and closeness centrality. The Paris algorithm provided a comprehensive overview of the hierarchical structure, while the LFM algorithm revealed detailed sub-communities sensitive to the alpha parameter. Our findings indicate that these algorithms offer valuable insights into the intricate hierarchical structures within biological networks, each contributing uniquely to our understanding of gene co-expression patterns. Detailed evaluations using modularity, closeness centrality, and Adjusted Rand Index (ARI) further validated the efficacy of these methods, highlighting their potential for practical applications in genomics and cellular biology.

2 Background of the Study

2.1 Network Analysis

Network analysis served as our analytical beacon, guiding us through the maze of biological complexities. It's a versatile framework applied in diverse realms such as social networks, transportation pathways, and biological systems (Hanneman & Riddle, 2005; Hevey, 2018; Ma & Gao, 2012; X. Zhang et al., 2015).

In the context of biological networks, network analysis revealed patterns and relationships among biological entities. Our study focuses on the nuances of community detection in these networks, a critical step in understanding and interpreting the complex web of biological interactions.

2.1.1 Biological Networks

Important biological activities resulted from the coordinated effects of multiple molecules, cells, and genes interacting with each other (Ma & Gao, 2012). Initially, it was believed that biological processes operated in isolation and followed

linear sequences of events. However, as research progressed, it became clear that these processes form a complex, interconnected web of interactions (Bhalla & Iyengar, 1999). This revelation provided valuable insights into the integrated nature of biological functions, making network analysis a powerful tool in studying biology.

Biological networks can represent a wide range of factors such as protein-protein interactions, metabolic pathways, and gene co-expression networks (Yu et al., 2013). Several topological properties are usually analyzed in these networks, including degree and centrality for identifying hub genes or disease biomarkers, and motifs and modules/sub-networks for identifying functional units and recurring patterns (Ma & Gao, 2012; Rahiminejad et al., 2019).

Schwikowski et al. analyzed a global yeast protein network of 1,548 proteins with 2,358 interactions to predict functions. They successfully predicted functions for 72% of characterized proteins and assigned functions to 364 previously uncharacterized proteins. This network analysis revealed that proteins with established functions and shared locations tended to form clusters (Schwikowski et al., 2000).

Tang et al. used weighted gene co-expression analysis to identify potential prognostic biomarkers for breast cancer. Using gene expression data from the Gene Expression Omnibus (GEO) database, they found five distinct modules, with one module heavily correlated with the severity or grade of breast cancer. Within that module, they identified several hub genes, including CASC5 and RAD51, whose higher expression levels are linked to poorer survival outcomes for individuals with breast cancer (Tang, 2018).

2.1.2 Construction of Gene Co-expression Networks

The heart of our study lies in gene co-expression networks. These networks are intricate maps representing the functional relationships among genes, revealing how their expression patterns correlate under various conditions (Montenegro, 2022). This approach is crucial for understanding gene functions and plays a vital role in identifying novel metabolic pathways and regulatory networks.

Central to our understanding of these networks is the concept of communities non-overlapping groups of nodes within a network where nodes are more densely connected internally than with the rest of the network (Fortunato & Newman, 2022). Identifying these communities is instrumental in understanding the functional clustering of genes and can reveal previously unknown aspects of gene function and interaction.

However, community detection is fraught with challenges. It involves the automated discovery of groups of nodes that are either strongly connected or share similar features. The complexity of this task varies with the sparsity of the graph and the presence of generative noise. In some cases, polynomial-time algorithms suffice, while in others, the problem may require exponential time, making the hardness of community detection highly case-specific (Moore, 2017).

There are two methods for graph construction: rank-based and value-based. In graph representation, each vertex represents a gene, and an edge connects

two genes if they are co-expressed. With value-based construction, we compute the pairwise similarity of genes using Pearson's correlation coefficient to form the gene co-expression network.

Gene expression data, reflecting gene activity levels in a sample, offers insights into the activation levels of genes and their co-expression patterns. Gene co-expression networks, showing relationships between different genes based on collective up-regulation or down-regulation, provide valuable information on functional relationships. Clustering gene expression data and analyzing it can offer possible annotations for genes with unknown functions based on other known genes in the same cluster (Oyelade et al., 2016; Yeung, 2001).

2.2 Community Detection in Networks

Exploring biological systems is challenging due to their inherent complexity and dynamic nature. Network science has revolutionized this field, especially with the introduction of community detection in gene co-expression networks. This methodology uncovers patterns and relationships within biological data, paving the way for meaningful discoveries.

Community detection in gene co-expression networks has evolved into a fundamental approach, with notable contributions from researchers such as Besas et al., who leveraged community detection algorithms to predict the biological classification of cell-cycle regulated genes in *Saccharomyces cerevisiae*.

Community detection involves identifying and classifying clusters where nodes (genes, in this context) exhibit higher connectivity within their group than with nodes outside it (Fortunato & Newman, 2022; Girvan & Newman, 2002). In biological networks, such as gene co-expression networks, community detection has been instrumental in unraveling hidden structures and functional associations.

A diverse array of community detection algorithms exists, each tailored to decipher specific types of networks. Fortunato classifies these algorithms into several categories:

- **Optimization approach:** This approach assigns scores to potential divisions of a network into communities, prioritizing divisions that yield high scores. Notable methods include Modularity, Spectral Clustering, and Hierarchical Clustering.
- **Statistical inference approach:** Here, communities are seen as primary drivers of network structure, with nodes connecting due to shared group affiliations. Methods such as Stochastic Block Model (SBM), Bayesian Inference, and Maximum Likelihood Estimation are popular in this category.
- **Dynamical process approach:** This method links community structure to dynamical processes occurring on networks. InfoMap and Random Walk Methods are exemplary techniques in this category.

2.2.1 Application of Community Detection in Biological Networks

Our challenge lies in choosing the right algorithm for community detection in biological networks. Selecting an appropriate community detection algorithm is

crucial as it significantly impacts the quality and validity of the insights derived from the analysis of biological networks. These networks, which can represent relationships between genes, proteins, or other biological entities, often exhibit complex structures that necessitate specialized approaches for accurate interpretation.

In recent years, the application of community detection to biological networks has been extensively explored. Redekar and Varma (2022) categorizes the commonly used algorithms for biological networks as shown in Table 1

Table 1. Comparison of Community Detection Methods (Redekar & Varma, 2022)

Category	**Type**	**Methods**
Type of community	Disjoint Community	Traditional Method (Graph Partitioning, Hierarchical Clustering, Spectral Clustering), Divisive Method (Girvan-Newman Method), Modularity-Based (Greedy Techniques, Simulated Annealing, External Optimization, Spectral Optimization)
	Overlapping Community	Clique Percolation Method, Line graph and Link Partitioning, Local Expansion and Optimization, Fuzzy Detection
Nature of network	Static Community	Graph Partitioning, Spectral Bisection, Hierarchical Clustering
	Dynamic Community	Random Walk, Spin Models, Synchronization

The Table 1 aids researchers in selecting the most appropriate method for specific network analysis needs, especially in the context of analyzing biological data like gene co-expression networks in *Saccharomyces cerevisiae*. Each method has its unique strengths and is suitable for different types of network structures and community detection requirements.

We identified three crucial open problems in existing literature: the identification of dynamic communities, an exploration of hierarchical community structures, and an analysis of overlapping communities. In response, we have curated a comprehensive list of algorithms tailored to address specific open problems in biological network analysis. This list, presented in Table 2, was compiled with insights from Vieira et al. (2020), Jaguzovic et al. (2022), Wagenseller III and Wang (2017), Rossetti et al. (2017), and Rossetti (2019). It provides a comparative overview of various community detection methods and approaches, facilitating the selection of the most suitable algorithm for each open problem.

The Table 2 includes a variety of community detection algorithms, each backed by significant research (Bonald et al., 2018; Coscia et al., 2014; Girvan & Newman, 2002; Gregory, 2010; Lancichinetti et al., 2009; Palla et al., 2005; Pons & Latapy, n.d.; Reichardt & Bornholdt, 2006; Rossetti et al., 2017; Rosvall & Bergstrom, 2008; Xie et al., 2011; Yang & Leskovec, n.d.). Each of these

Table 2. Comparison of Community Detection Methods for the Identified Open Problems

Open Problem	Algorithm	Approach	CDLib	Network	Communities
Identification of Dynamic Communities	Spinglass	Dynamical	spinglass	Undirected, Unweighted, Temporal	Crisp
	InfoMap	Dynamical	infomap	Directed, Weighted	Crisp
	Random Walk	Dynamical	walktrap	Undirected, Unweighted	Crisp
	TILES	Dynamical	tiles	Undirected, Unweighted, Temporal	Temporal
Examination of Hierarchical Community Structures	GN-Method	Optimization	girvan_newman	Undirected, Unweighted	Crisp, Hierarchical
	Local Optimization	Optimization	lfm	Undirected, Unweighted	Overlaps, Hierarchical
	Hierarchal Clustering	Optimization	paris	Undirected, Weighted	Crisp, Hierarchical
Analysis of Overlapping Communities	Clique Percolation Method	Optimization	kclique	Undirected, Unweighted	Overlaps
	Demon	Optimization	demon	Undirected, Unweighted	Overlaps
	SLPA	Optimization	slpa	Undirected, Unweighted	Overlaps
	Bigclam	Optimization	big_clam	Undirected, Unweighted	Crisp, Overlaps, Nested
	CFinder	Optimization	-	Undirected, Unweighted	Overlaps
	Copra	Optimization	-	Undirected, Unweighted	Overlaps

algorithms, with their unique approaches and applications, contributes significantly to the field of community detection in biological networks, offering diverse methodologies to address the identified open problems in the study.

2.2.2 Evaluation Methods of Community Detection Algorithms

Evaluating the success of community detection algorithms is typically done using two types of metrics: internal criteria and external criteria.

Internal criteria, such as Modularity and Closeness Centrality, assess the quality of detected communities without reference to any external structure, providing a blind evaluation of the community structure.

External criteria, on the other hand, involve comparing the detected community structure against a known reference, using metrics like Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Purity. These metrics offer a benchmarked evaluation, essential for validating the detected communities against a ground truth (Liu et al., 2019).

3 Related work

Having explored the complexities of biological networks and the nuances of community detection, we now delve deeper into the realm of gene expression data, hierarchical community detection algorithms, and statistical analysis of these algorithms. This section uncovers the gaps and opportunities in the current body of knowledge. Specifically, we are interested in the sophisticated analysis of hierarchical community detection algorithms, particularly in their application to the *Saccharomyces cerevisiae* gene expression data.

3.1 Gene Expression Data

In the realm of publicly accessible resources, we identified a particularly pertinent dataset comprising a subset of 384 genes, observed across 17 distinct time points.

This dataset, initially used in the study by Yeung et al. (2001) and originating from Cho et al. (1998), is a rich repository of information, accessible via the supplementary website for Yeung's paper. The utilization of gene co-expression networks in our study was motivated by the accessibility and richness of such online datasets.

Cho et al. (1998) conducted a meticulous analysis of the *Saccharomyces cerevisiae* gene expression data derived from laboratory experiments. Their study categorized approximately 380 genes into five primary functional groups—early G1 phase, late G1 phase, synthesis phase, second growth phase, and mitosis—adding further depth by identifying subgroups within each of these categories. This classification provided granular insights into the functions associated with each gene ID.

Building upon this foundational work, Clemente (2022) conducted an analysis of the *Saccharomyces cerevisiae* gene co-expression network. Using community detection techniques on the same gene set examined by Cho et al. (1998), Clemente's study identified five primary clusters. These clusters broadly align with the five main functional groups identified by Cho et al. (1998), offering a macroscopic view of gene activity during various phases of the gene cycle. However, this approach, while informative, leaves much to be desired in terms of specificity. The broad nature of these clusters obscures the nuanced functions of individual genes, presenting a significant limitation in our understanding of gene functionality within these groups.

This leads us to a conspicuous gap in the existing literature: a comprehensive analysis and comparison of results obtained from hierarchical community detection algorithms, particularly applied to the *Saccharomyces cerevisiae* gene expression data, remains unexplored. Such an analysis, especially one that focuses on identifying sub-clusters within the main functional groups, could provide a more detailed understanding of gene functions.

Recognizing the immense potential of sub-cluster analysis to yield finer, more detailed insights into gene functions—as suggested by H. Zhang et al. (2018)—our study aims to bridge this gap. By employing and comparing various hierarchical community detection algorithms, we aspire to uncover a more nuanced understanding of the *Saccharomyces cerevisiae* gene co-expression network, thus contributing a significant piece to the puzzle of gene functionality in biological research.

3.2 Hierarchical Community Detection Algorithmns

In the context of community detection within complex networks, hierarchical algorithms not only detect communities but also reveal the multi-level structure inherent within networks. In Table 2, we have highlighted three such hierarchical algorithms available in CDLib: Girvan-Newman, Paris, and LFM. Each of these algorithms offers a unique approach to uncovering the layered structure of networks.

In the context of our study, these three algorithms – Girvan-Newman, Paris, and LFM – offer a diverse range of analytical lenses through which we can ex-

amine the gene co-expression network of *Saccharomyces cerevisiae*. Their unique characteristics and methodologies provide us with a comprehensive toolkit for dissecting and understanding the intricate web of relationships and structures within our dataset.

3.2.1 Girvan-Newman Algorithm

The Girvan–Newman algorithm stands out for its methodical approach in community detection. It progressively dissects a network by systematically removing its edges. At each iteration, the algorithm identifies the 'most valuable' edge, typically the one with the highest betweenness centrality, and removes it. This process of edge removal continues until a clear community structure emerges within the network. The communities are revealed as the graph fragments into more tightly connected groups. The algorithm's effectiveness lies in its ability to highlight the natural divisions within a network, which can be visualized through a dendrogram. The algorithm is particularly suited for undirected and weighted networks, though it does not accommodate directed networks. Its process, characterized by the systematic deconstruction and subsequent revelation of tightly-knit communities, provides a detailed and insightful view of the network's architecture (Bonald et al., 2018; Clemente, 2022).

In terms of hardness, it has been known that Girvan-Newman is an NP-Hard problem (Brandes et al., 2008). Specifically, it is $(1 + \varepsilon)$-inapproximability for dense graphs and logarithmic approximation for sparse graphs (DasGupta & Desai, 2013).

3.2.2 Paris Algorithm

Paris is an algorithm that combines the principles of modularity-based clustering with a hierarchical approach. It operates agglomeratively, basing its process on a unique distance metric between clusters. This metric is induced by the probability of sampling node pairs within the network. Paris is particularly effective in its simplicity and its ability to reveal community structures at various levels of granularity. This agglomerative algorithm is adept at handling undirected networks, although it does not support directed or weighted graphs. Its simplicity and the intuitive nature of its distance metric allow for a nuanced and layered understanding of a network's community structure, making it an excellent tool for exploring gene co-expression networks where intricate relationships are often hidden within the data (Girvan & Newman, 2002).

3.2.3 LFM Algorithm

The LFM algorithm is a robust tool for detecting overlapping communities and understanding hierarchical structures within networks. It operates on the principle of local optimization of a fitness function. One of its key features is its adaptability, as it can uncover communities of varying sizes based on the alpha parameter. This parameter effectively controls the granularity of the community detection process, with larger values leading to smaller communities and vice

versa. This flexibility makes LFM suitable for undirected networks that require a nuanced approach to community detection, allowing researchers to tailor the algorithm to the specific needs of their data analysis (Lancichinetti et al., 2009).

3.3 Statistical Analysis of Hierarchical Community Detection

Having explored the capabilities of the Girvan-Newman, Paris, and LFM algorithms in unveiling the hierarchical structures within the *Saccharomyces cerevisiae* gene co-expression network, we now turn our focus to the critical aspect of evaluating their effectiveness. The success of any community detection algorithm lies not only in its ability to identify patterns but also in how accurately these patterns reflect the true underlying structure of the data. To this end, we employ both internal and external criteria in our statistical analysis, ensuring a comprehensive and rigorous assessment of the algorithmic outputs.

3.3.1 Internal Criteria

Internal criteria evaluate the quality of the communities detected by the algorithms based solely on the structure of the network itself, without reference to external data. This self-contained analysis is important in determining how well the algorithms have managed to capture the intrinsic clustering within the gene co-expression network.

One of the primary metrics that we utilize in this study is Modularity. Modularity measures the strength of division of a network into communities by assessing the density of links inside communities compared to links between communities. A high modularity score indicates that the network has a strong community structure, with more connections within communities than between them. This metric is particularly useful in evaluating algorithms like Girvan-Newman, which are designed to identify distinct and tightly-knit communities (Hajibabaei et al., 2023).

Additionally, we consider the Closeness Centrality, which measures the degree to which an individual node is near all other nodes in the network. In the context of community detection, this metric helps in understanding how central or peripheral each node is within its assigned community, thus providing insights into the community structure's coherence and the algorithm's ability to allocate nodes appropriately (Jarukasemratana et al., 2014).

By employing these internal criteria, we can critically assess the intrinsic effectiveness of our chosen algorithms. This assessment not only validates the detected communities in terms of the network's own topology but also ensures that the algorithms are effectively deciphering the complex, layered relationships inherent in the *Saccharomyces cerevisiae* gene co-expression network. Such an analysis forms the bedrock of our evaluation process, ensuring that the communities we identify are not only statistically significant but also meaningful in the context of the biological data.

3.3.2 External Criteria

Central to our external validation process is the the Adjusted Rand Index (ARI). ARI provides a quantitative measure of the similarity between two data clusterings – one generated by our hierarchical community detection algorithms and the other representing the true clustering, as established by prior biological knowledge.

The Rand Index (RI) measures the similarity between two data clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings. The Adjusted Rand Index (ARI), an enhanced version of RI, adjusts for the chance grouping of elements. It provides a more accurate measure of the similarity between the predicted and true clusterings. The ARI is calculated as $\text{ARI} = \frac{\text{RI} - E(\text{RI})}{\max(\text{RI}) - E(\text{RI})}$. A maximum ARI score of 1 indicates perfect alignment between the predicted and actual clusterings, while a negative ARI score implies an agreement level below random chance (Clemente, 2022; H. Zhang et al., 2018).

By incorporating ARI as our external criteria, we can gain a comprehensive understanding of how effectively the community detection algorithms categorize gene sets based on their biological functions. These metrics enable us to evaluate and compare the algorithms.

4 Research Aims

The comprehensive analysis we've undertaken thus far has not only shed light on the strides made in understanding gene networks but also brought into focus the critical gaps and uncharted territories in this field. These areas, brimming with potential for discovery and innovation, form the bedrock of our study's objectives. In the next sections, we articulate the specific challenges we addressed, alongside the objectives and scope of the study.

4.1 Problem Statement

In this study, we confront several critical gaps in the field of gene co-expression network analysis, particularly focusing on the *Saccharomyces cerevisiae* genome. Our objectives are threefold:

1. **Outdated Gene Function Classification:** To ensure the robustness and relevance of our findings, an updated and more extensive dataset was employed by us, building upon the foundations of past research while incorporating the latest advancements in genomic data. Since the classifications and sub-classifications from Cho et al. (1998) were outdated, we utilized existing databases for Yeast genome from sites like Kyoto Encyclopedia of Genes and Genomes (KEGG) and Saccharomyces Genome Database to update the gene function classifications.

2. **Limited Precision in Gene Function Identification:** Given that the five major classifications from clusters found in the study by Clemente (2022) were too general compared to the groups and subgroups seen in Cho et al. (1998), this research pinpointed the roles and contributions of individual genes with greater precision, addressing the limitations of existing classifications and paving the way for a better understanding of gene functions within the *Saccharomyces cerevisiae* gene co-expression network.
3. **Absence of Network Analysis on Hierarchical Sub-clusters within our Chosen Dataset:** A previous study by Salido (2016) has investigated hierarchical clusters within our chosen dataset, but it limited its analysis to traditional cluster methods, particularly Non-metric Multidimensional Scaling (nMDS). Our study aimed for a network-based analysis approach to identify the hierarchical clusters. The performance of each algorithm on the gene co-expression network of Yeast cell-cycle regulated genes was systematically compared, contributing to a nuanced analysis of different hierarchical community detection approaches in the context of gene co-expression networks.

4.2 Significance of the Study

The significance of this research lies in its potential to transform our understanding of hierarchical structures in biological networks, with far-reaching implications:

1. This research aims to contribute significantly to the theoretical underpinnings of biological network analysis. By focusing on hierarchical structures, it addresses a critical aspect of network science that is often overlooked but is crucial for understanding complex biological systems.
2. Understanding hierarchical structures in biological networks has direct implications for practical applications such as drug development, disease modeling, and personalized medicine. Insights gained from this research could lead to more targeted therapeutic strategies and a deeper understanding of disease mechanisms.
3. The findings from this study could have interdisciplinary impacts, bridging gaps between network science, biology, computational methods, and even extending into fields such as ecology and sociology, where hierarchical structures also play a crucial role.
4. This research can also enrich academic discourse, providing valuable material for educational purposes. It can inspire new curricula and teaching methods in the fields of biology and network science.

4.3 Objectives of the Study

The study seeks to provide a comprehensive analysis of hierarchical community structures within biological networks. The main objective is:

– To assess the capability of predicting hierarchical sub-clusters within the *Saccharomyces cerevisiae* gene co-expression dataset, which can be accessed through this link, using hierarchical community detection algorithms.

Alongside the main objectives, we have pinpointed specific objectives crucial to the overall methodology of the study:

1. To update the gene expression dataset to reflect the latest gene classification from Kyoto Encyclopedia of Genes and Genomes (KEGG) and Saccharomyces Genome Database.
2. To implement a Python script, integrating the CDLIB library and other necessary imports, to automate the detection of sub-clusters in the *Saccharomyces cerevisiae* gene co-expression network. This tool will also facilitate the comparative analysis of different algorithms using statistical methods such as the Adjusted Rand Index (ARI) and the Rand Index (RI).
3. To identify which existing hierarchical community detection algorithm is the most effective in detecting sub-clusters on the *Saccharomyces cerevisiae* gene co-expression network. The current list of candidate algorithms can be found in Table 2.

4.4 Scope and Limitations of the Study

The study's primary focus was on identifying sub-clusters within the gene co-expression network of *Saccharomyces cerevisiae*, with a specific emphasis on cell-cycle regulated genes. The subset of 384 genes with available expression data was the center of this analysis. Our goal was to uncover the underlying structure and potential functional groupings within this network by employing various hierarchical community detection methods. The effectiveness of these methods in revealing biologically relevant sub-clusters was a key aspect of this study.

In specifying the scope of this research, it is important to note that we did not address overlapping nodes or communities within the gene co-expression network. Our analysis focused on distinct sub-clusters, and the methodologies we employed were tailored to identify non-overlapping hierarchical structures. This decision was made to maintain clarity and specificity in the interpretation of results.

In terms of methodology, we streamlined the process by adopting the most effective gene co-expression network generation method identified in the study by Clemente (2022). This method had been proven effective in constructing gene co-expression networks for various genomes, including the human genome (Lee et al., 2004). Notably, Clemente (2022) demonstrated its efficacy, showing that this method was the best algorithm for constructing graphs in the specific yeast genome dataset that we also utilized.

Furthermore, we did not venture into the development of new hierarchical community detection algorithms. Instead, we leveraged existing algorithms, recognized for their efficiency and accuracy in various contexts, including biological

networks. This approach ensured that we utilized proven methodologies while concentrating on the unique aspects of gene co-expression in yeast.

To evaluate the performance of these community detection algorithms, we relied on existing scoring metrics. The choice of these metrics was crucial, as they needed to be apt for assessing hierarchical community detection in a biological context. We carefully selected metrics that not only provided quantitative assessments of the algorithms' performances but were also relevant and interpretable in the realm of gene expression and cellular biology.

By focusing on established methods and tools, we aimed to efficiently utilize our resources and expertise to delve into the complex interactions and organization within the *Saccharomyces cerevisiae* gene network. This focused approach enabled a more in-depth understanding of the network's structure and the biological significance of its sub-clusters, potentially leading to valuable insights in the field of genomics and cellular biology.

5 Methodology

As we move forward from defining the precise aims and objectives of our research, we now step into the methodology. In this section, we outline the comprehensive theoretical framework, which harmonizes computational analysis with experimental validation alongside the flow of our research methodology.

5.1 Theoretical Framework

A comprehensive approach to gene function identification involves integrating in silico network analysis with wet lab experiments. This blend of computational and experimental methods allows for the development of robust hypotheses and testing strategies, ultimately leading to the updating of gene annotations in biological databases. By synthesizing these two approaches, we can achieve a more holistic understanding of gene functions and interactions, paving the way for significant advancements in the field of gene co-expression network research.

To visually represent the intricate interplay between wet lab experiments and computational analyses in our research, we present a theoretical framework diagram in Figure 1. This diagram serves as a visual roadmap, illustrating the connections between the outcomes of wet lab experiments and the subsequent application of community detection algorithms on gene co-expression networks. The integration of in silico network analysis (*e.g. process done by Clemente (2022)) with wet lab experiments is foundational to the comprehensive approach for gene function identification. This showcases how the collaborative efforts of wet lab experiments and computational analyses converge to advance our understanding of gene functions within co-expression networks.

Now, let's take a look at the specific steps in our methodology, which are outlined in Figure 2. We provide an overview of the processes involved, from data collection and pre-processing to network construction and community detection script implementation, highlighting the systematic approach we used in this study. Each step will be discussed in detail in the following sections.

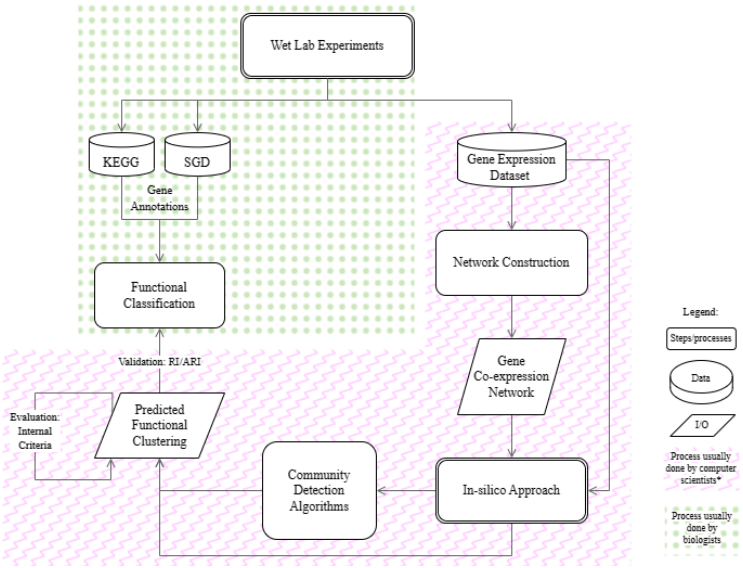

Fig. 1. Theoretical Framework Diagram

5.2 Data Collection and Organization

The first step in our methodology is to collect data from the Kyoto Encyclopedia of Genes and Genomes (KEGG) for each gene present in the dataset used by Cho et al. (1998). Our objective is to achieve a more precise grouping and classification for each gene, facilitating a comprehensive understanding of the dataset's hierarchical structure, essential for hierarchical community detection. This phase involves scraping relevant details for each gene. We have observed that there are three consistent levels of classification for every gene data that was scraped, each with a unique identification number. We will then further refine our scraped data by extracting only the specific ID for each classification at each level.

At this point, we have acquired the classifications for 242 genes (out of the 384 genes from the dataset used by Cho et al. (1998)). However, the information for the remaining 142 genes was incomplete in the KEGG database, lacking essential classifications. Consequently, these genes were removed from the dataset.

To gain a clearer understanding of the groupings, we chose to generate a graph visualization. This visualization was represented as a forest, with each tree root corresponding to the level 1 classifications. The subsequent tree levels depicted the level 2 and level 3 classifications, while the leaves represented the genes themselves.

Figure 3 displays the generated graph for the 242 genes and their respective classifications at each level. The hierarchical structure of the classifications within KEGG is clearly depicted in the graph, as certain classifications are positioned "under" the classification at the level above them, and each classification

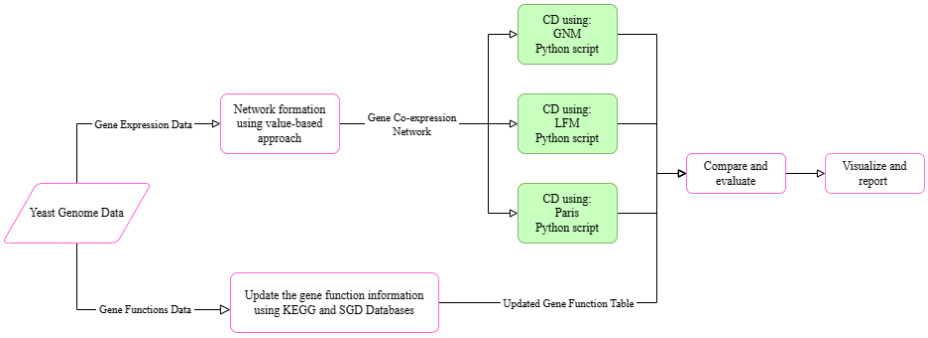

Fig. 2. Methodology Flowchart

in the lower levels belongs to a classification in the upper levels. Additionally, we observed from this graph that there are no overlapping nodes; therefore, we did not need to filter out anything, as overlapping nodes are not within the scope of this study.

The next step is to investigate potential relationships or correlations between the groupings at any level from the KEGG data and the two levels classified by Cho et al. (1998) in their study. To accomplish this, we utilized the Adjusted Rand Index (ARI). To proceed, we must filter the KEGG data to include only the genes present in Cho et al. (1998) more detailed classifications. This constitutes a smaller number (149 genes), as Cho et al. (1998) did not include all 384 genes in the more specific classifications.

We generated a graph for the groupings from Cho et al. (1998) which is shown in Figure 4. ARI was performed for each combination of levels, with the highest score being around 0.16. The scores for each combination are shown in Table 3.

Table 3. Relationships between KEGG and CHO classifications with ARI

KEGG Level	CHO Level	ARI
1	1	0.022699
1	2	0.083272
2	1	0.057357
2	2	0.118445
3	1	0.043423
3	2	0.156492

Collecting and organizing the data provided a foundational understanding of the hierarchical structures within the dataset and identified potential relationships between classifications from KEGG and classifications from Cho et al.

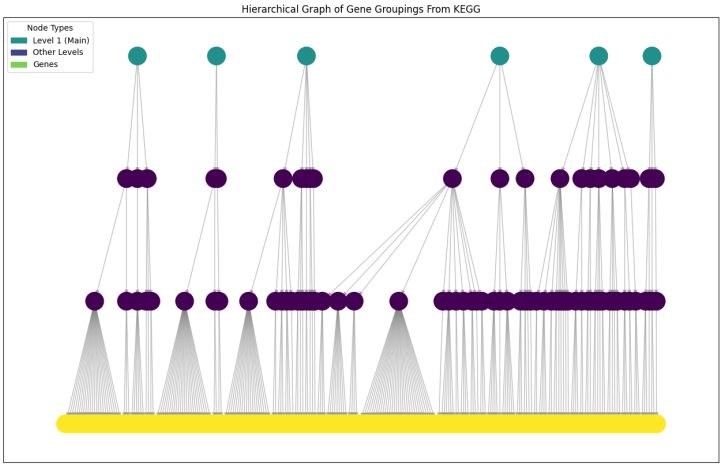

Fig. 3. Graph of Gene Groupings From KEGG

(1998). We will leverage this groundwork to construct gene co-expression networks.

5.3 Gene Co-expression Network Formation

Building upon the insights gained from the previous section, our next step is to delve into the construction of gene co-expression networks. Using the gene co-expression data provided by Yeung (2001), we employ a value-based algorithm for network construction, a method validated for its superior performance, as demonstrated in Clemente (2022).

We've organized our coding process in a Jupyter notebook, ensuring clarity and facilitating navigation across project stages. Utilizing the NetworkX library in Python, we aimed to construct a gene co-expression network by calculating pairwise gene similarities via Pearson's correlation coefficient. This method enables precise assessment of gene expression relationships.

Introducing the δ parameter, we established a crucial threshold for determining co-expression between gene pairs. This parameter's flexibility allowed us to fine-tune our network based on varying degrees of gene expression similarity.

In our analysis, we employed value-based graph construction, exploring thresholds from 0.70 to 0.95. This broad spectrum allowed us to understand how network sensitivity varies with different similarity thresholds, providing comprehensive insights into network dynamics.

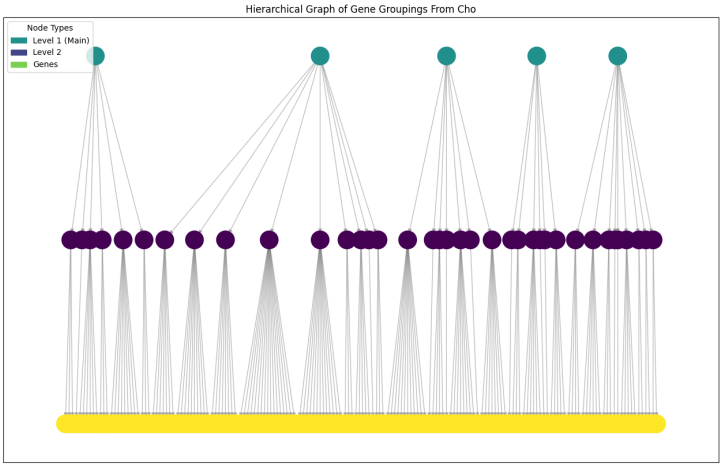

Fig. 4. Graph of Gene Groupings From Cho et al. (1998)

Our examination revealed that a small portion of nodes constitutes the largest connected component, with few singletons. Around 63% of nodes are integrated into non-trivial connected components, underscoring the network's complex and interconnected nature.

We produced a graph depicting the gene co-expression network, as depicted in Figure 5. This visualization facilitated interpretation of our findings and conveyed the complex relationships within the yeast gene expression data.

Our graph illustrates that setting δ to 0.85 results in a compelling structure, marked by the highest number of non-trivial connected components.

Remarkably, this mirrors the findings of the study by Clemente (2022) which we sought to build upon. This reveals analogous patterns of connectivity and functional grouping, affirming the credibility and significance of these co-expression relationships.

5.4 Script Implementation for Community Detection

For the script implementation of the automated community detection, the process began with a thorough review of available algorithms across different libraries, particularly NetworkX and CDLib. Notably, while NetworkX supports various algorithms, LFM and Paris were absent, emphasizing CDLib's unique capabilities.

We tested CDLib's community detection algorithm GN, which successfully identified meaningful communities within our network data. A pivotal parameter

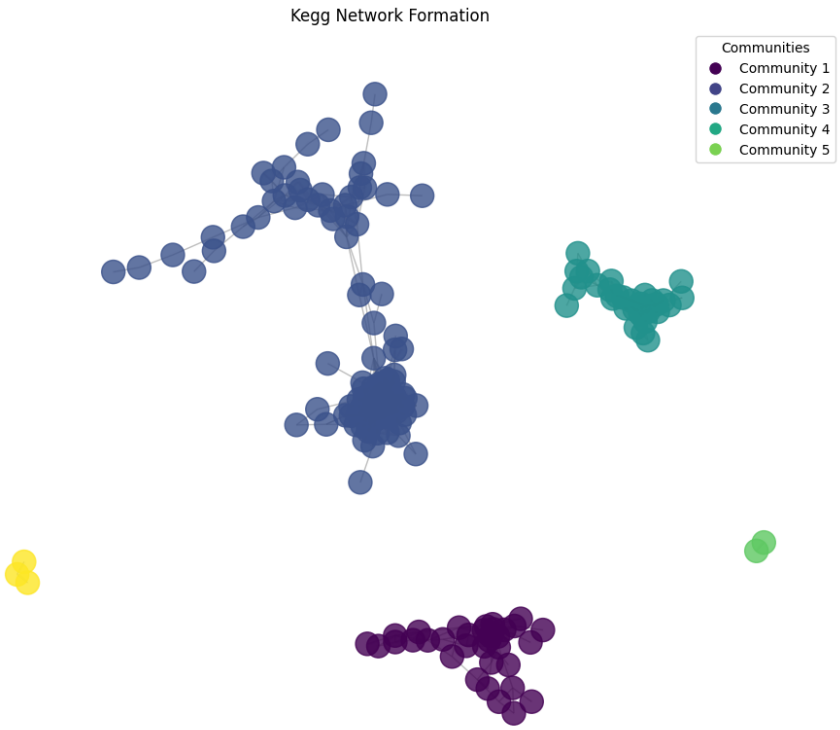

Fig. 5. Gene Co-expression Network

in this process is the "level," determining the depth of analysis and indicating the stage at which we opt to cut the dendrogram to observe the community structure.

As the level increases, we transition from a macro-view of the network's community structure (comprising broad, large communities) to a more micro-view (highlighting smaller, more specific communities). These insights will be elaborated upon in the following section, accompanied by detailed discussions of the generated graphs.

Similarly, CDLib is employed for the LFM algorithm, which also unveils a hierarchical structure of communities. The alpha parameter regulates the resolution of community detection, influencing both the size and the number of communities identified.

We experimented with varying alpha values from 0.5 to 1.0. The results indicate a significant impact of alpha on the outcome of community detection, with more communities emerging at higher alpha values. This underscores LFM's sensitivity to the alpha parameter and its capability to adapt to different scales of

community structure. Graphs illustrating different alpha values will be presented in the subsequent section.

Lastly, CDLib was used for the Paris algorithm. We have found that this specific method did not involve changing any parameters.

Now that we have finished implementing the scripts for the three algorithms, our next step is to evaluate each algorithm by comparing its performance (groups formed) against our dataset. This is discussed in the next section.

5.5 Visualization, Evaluation, and Analysis

To understand how the three community detection algorithms compare to one another, we employed several evaluation methods on the communities formed by each algorithm. Specifically, we used modularity, closeness centrality, and the Adjusted Rand Index (ARI) to assess the quality and effectiveness of the detected communities.

In addition to these quantitative measures, we created visual representations of the networks using graphs. These visualizations help provide a clearer and more intuitive understanding of the community structures and the overall network.

5.5.1 Girvan-Newman

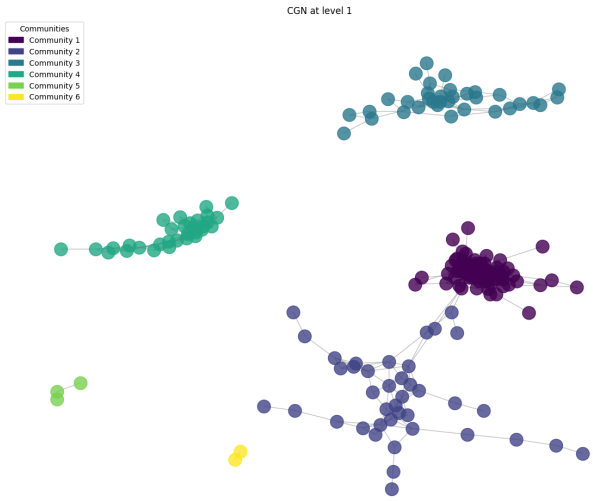

Fig. 6. Communities Detected Using GN at Level 1

At Level 1, the Girvan-Newman (GN) algorithm successfully identified six distinct communities within the gene co-expression network, as illustrated in

Figure 6. Each community is represented by a different color, highlighting the clear separation and clustering of nodes.

The visualization reveals that the communities are not only well-defined but also vary significantly in size. Community 1 and Community 2 are the largest, containing the majority of the nodes, while Community 6 is the smallest. This distribution of community sizes may reflect the inherent hierarchical nature of gene interactions, where certain groups of genes are more densely interconnected due to similar functions or regulatory mechanisms.

The identified communities are spread across the network, showing distinct clusters that are interconnected by fewer edges. This clear demarcation between communities indicates that the GN algorithm effectively captures the modular structure of the gene co-expression network at this level.

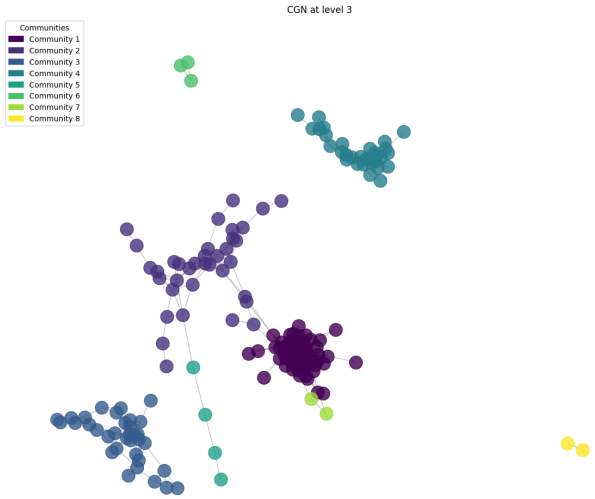

Fig. 7. Communities Detected Using GN at Level 3

At Level 3, the Girvan-Newman (GN) algorithm further refines the community structure within the gene co-expression network, as depicted in Figure 7. The algorithm successfully identifies eight distinct communities, each represented by a unique color. This level of analysis reveals more granularity compared to Level 1, allowing for a deeper exploration of the network's hierarchical structure.

The visualization shows a notable increase in the number of smaller communities, indicating that the GN algorithm is capable of detecting finer subdivisions within the larger clusters identified at earlier levels. Community 1 remains prominent, but it has split into several smaller groups, reflecting more detailed interactions among the genes. Similarly, Community 2 has fragmented into smaller,

more specific clusters, providing insights into sub-communities within the gene network.

The spatial distribution of the communities in the graph suggests that the GN algorithm can capture not only the broad divisions but also the subtle, intricate relationships among genes. The presence of smaller, isolated communities, such as Community 8, highlights the algorithm's ability to detect outliers or specialized clusters that may have unique functional roles.

This increased granularity at Level 3 offers a more comprehensive understanding of the gene co-expression network, revealing sub-structures that might be obscured at higher levels of abstraction. These detailed clusters can provide valuable insights into specific gene functions and interactions, facilitating targeted biological investigations and hypotheses generation.

5.5.2 Paris

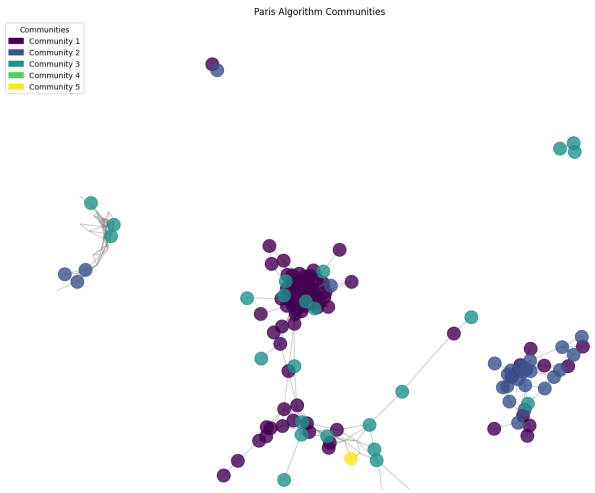

Fig. 8. Communities Detected Using Paris

The Paris algorithm, as illustrated in Figure 8, identifies five distinct communities within the gene co-expression network. Each community is marked by a different color, providing a clear visual representation of the clustering.

The communities identified by the Paris algorithm exhibit a variety of sizes and structures. Community 1 is the largest, forming a dense cluster at the center of the network. This central cluster indicates a highly interconnected group of genes, possibly reflecting core functional relationships or regulatory modules.

Community 2 and Community 3 also form substantial clusters, albeit smaller than Community 1. These communities are relatively compact and appear to be

well-separated from each other, suggesting distinct functional groups within the network.

Community 4 and Community 5 are the smallest, with Community 5 consisting of only a few nodes. These smaller communities may represent more specialized or less connected genes within the network, highlighting the algorithm's ability to detect both large and small functional groupings.

The spatial distribution of the communities in the graph shows a clear separation between different clusters, with some nodes and edges connecting these clusters, indicating potential interactions or overlap between different gene groups. The Paris algorithm effectively captures these nuanced relationships, providing a detailed view of the gene co-expression network's modular structure.

5.5.3 LFM

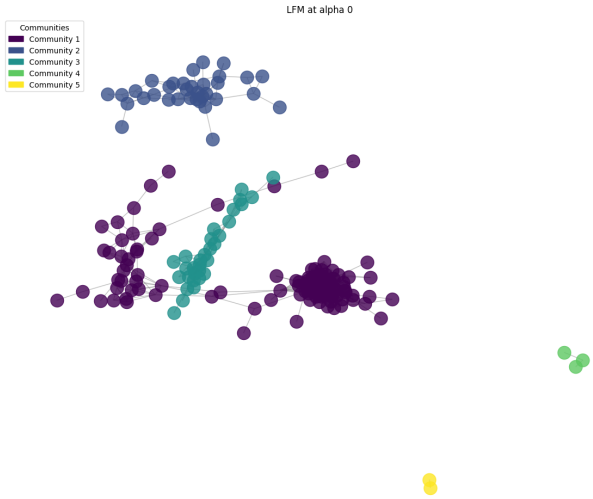

Fig. 9. Communities Detected Using LFM with Alpha=0

The visualization of the LFM algorithm at alpha 0, shown in Figure 9, identifies five distinct communities within the gene co-expression network. Each community is distinguished by a different color, illustrating the algorithm's ability to cluster nodes based on their connectivity.

Community 1 is the largest and most densely connected cluster, positioned centrally in the network. This suggests that Community 1 encompasses a significant portion of the network's core interactions, potentially representing a primary functional group or regulatory module.

Community 2 is another substantial cluster located towards the upper part of the network. Its size and density indicate a secondary but still prominent

grouping of genes, possibly related to a distinct set of functions or processes compared to Community 1.

Community 3 and Community 4 are smaller, more compact clusters. These communities are relatively isolated from the larger groups, highlighting their specialized roles or interactions within the gene network. Community 3, in particular, bridges between Community 1 and other parts of the network, suggesting a possible intermediary role in gene interactions.

Community 5 is the smallest cluster, consisting of only a few nodes. Its position on the periphery of the network indicates that it may represent a unique or less connected set of genes, potentially involved in specialized or less central functions.

The overall structure of the network under the LFM algorithm at alpha 0 shows a clear separation between the communities, with well-defined boundaries and varying sizes. This visualization underscores the algorithm's capability to identify both large, central clusters and smaller, peripheral groups, providing a comprehensive view of the hierarchical organization within the gene co-expression network.

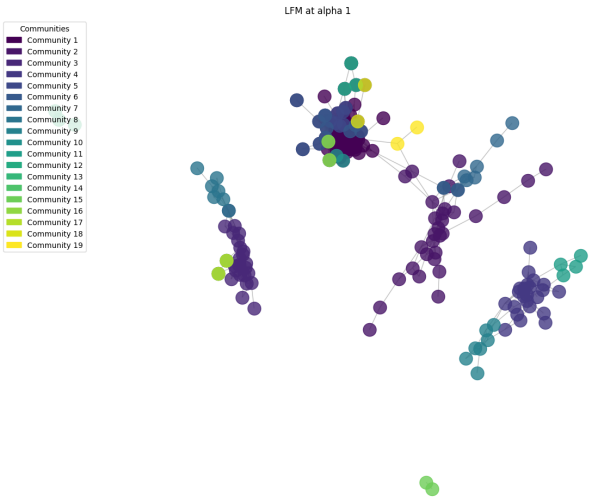

Fig. 10. Communities Detected Using LFM with Alpha=1

The visualization of the LFM algorithm at alpha 1, shown in Figure 10, identifies nineteen distinct communities within the gene co-expression network. Each community is represented by a different color, showcasing the algorithm's ability to detect a high level of granularity in the clustering.

At this alpha level, the network is divided into many smaller communities compared to the clustering observed at alpha 0. This increase in the number of

communities indicates the algorithm's sensitivity to the alpha parameter, which controls the resolution of the community detection process. As alpha increases, the LFM algorithm identifies more fine-grained structures within the network.

The largest clusters, such as Community 1 and Community 2, remain central in the network, but they have now split into several smaller sub-communities. This splitting reflects the detection of more specific functional groups within the larger clusters, highlighting intricate relationships among genes.

Other communities, such as Community 8 and Community 11, are relatively small and compact, indicating specialized groups of genes with strong internal connections but fewer links to other communities. The presence of numerous small communities, like Community 18 and Community 19, scattered throughout the network further emphasizes the detailed segmentation achieved at this alpha level.

The overall structure shows a dense network with many inter-community edges, illustrating the complex and interconnected nature of the gene co-expression network. This high level of detail provides valuable insights into the modular organization of the network, potentially uncovering subtle functional relationships that are not visible at lower alpha levels.

The collected results from both the quantitative evaluations and visual analyses were then thoroughly examined. The insights gained from this analysis are discussed in detail in Section 6.

6 Results

	Level	Number of Communities	Nodes in Largest Community	Singletons	Modularity	Closeness Centrality	ARI (Level 1)	ARI (Level 2)
GN	1	4	64	0	**0.331094126**	0.4292961301	0.5947369851	0.15849833
	2	5	64	0	0.3305609729	0.5090427336	0.5999552912	0.1639007852
	...							
	14	17	58	6	0.3218033538	0.7451175912	0.574702	**0.1923712549**
	...							
	49	52	38	37	0.244454	**0.9030717181**	0.303390071	0.1659151548
	Alpha							
LFM	0	3	85	0	0.2577457511	0.1565445117	0.4436266995	0.091332
	0.5	4	65	0	**0.331094126**	0.1565445117	0.5947369851	0.15849833
	0.6	5	65	0	0.330336056	0.1565445117	0.5963115436	0.1600950956
	0.7	11	64	0	0.327002455	0.1565445117	0.6045236205	0.1753946801
	0.8	8	64	0	0.3298426235	0.1565445117	0.5940667658	0.1684366058
	0.9	9	64	0	0.3269775116	0.1565445117	0.6062407526	0.1704956699
	1	10	64	0	0.3237162358	0.1565445117	0.6038008742	0.1709489366
PARIS		3	85	0	0.1142872613	0.2106443044	**0.71500005**	0.1522614824
Highest Values:		52	85	37	**0.331094126**	**0.9030717181**	**0.71500005**	**0.1923712549**

Fig. 11. Evaluation Results on Cho Dataset

6.1 Cho Dataset Results

From our analysis of the Cho dataset evaluation results, which can be seen at Figure 11, we observed that the Local optimization Function Model (LFM) and

Girvan-Newman (GN) algorithms demonstrated a slightly higher average modularity compared to the Paris algorithm. This suggests that LFM and GN are more effective in forming tightly-knit communities within the gene co-expression network. Specifically, the LFM algorithm showed a significantly higher Adjusted Rand Index (ARI) at Level 1, indicating a better alignment with the ground truth clustering results. This highlights LFM's effectiveness in capturing the underlying biological structure in its initial clustering.

Interestingly, the Paris algorithm achieved the highest ARI at Level 1 among all algorithms, suggesting that Paris might also be highly effective at correctly identifying the true community structure at this level. However, GN excelled in closeness centrality, particularly at Level 49, where it achieved the highest value (0.9030717181). This implies that the clusters identified by GN are more central within the network, potentially indicating more meaningful or compact groupings of genes.

Moreover, at Level 2, GN's ARI slightly surpassed that of LFM and Paris. This indicates that GN maintains better consistency in its clustering across different hierarchical levels, making it a robust choice for multi-level community detection in gene co-expression networks. GN's balance between high modularity, closeness centrality, and consistent ARI scores across levels underscores its potential as a reliable method for uncovering hierarchical community structures in biological networks.

Overall, while LFM and Paris show strengths at specific levels, GN's performance across multiple evaluation metrics and levels makes it a strong contender for community detection in the Cho dataset. This comprehensive analysis emphasizes the importance of considering multiple metrics and hierarchical levels when evaluating community detection algorithms for biological applications.

6.2 KEGG Dataset Results

Moving on to the KEGG dataset evaluation results, which can be seen at Figure 11, our analysis revealed several key insights about the performance of the community detection algorithms. The Girvan-Newman (GN) algorithm consistently displayed higher modularity values compared to the Local optimization Function Model (LFM) and Paris algorithms, with the highest modularity observed at Level 3 (0.3438994678). This indicates that GN is more effective at identifying strong community structures within the dataset.

In terms of closeness centrality, GN again showed a significant advantage. At Level 49, GN achieved the highest closeness centrality value (0.8191581058), suggesting that the communities it identifies are more centrally positioned within the network, indicating tightly knit and meaningful clusters.

Additionally, GN outperformed both LFM and Paris in terms of average Adjusted Rand Index (ARI) values across various levels. At Level 1, GN and LFM shared the highest ARI (0.03375774012), but GN showed better consistency in ARI values at higher levels, with the highest ARI at Level 2 (0.081454). This indicates that GN provides more accurate clustering results that closely match the ground truth.

	Level	Number of Communities	Nodes in Largest Community	Singletons	Modularity	Closeness Centrality	ARI (Level 1)	ARI (Level 2)	ARI (Level 3)
GN	1	6	86	0	0.3434830981	0.5612222908	0.03375740132	0.063101	0.036326
	2	7	86	0	0.3430306206	0.5739693694	0.031901	0.061727	0.036023
	3	8	84	0	**0.3438994678**	0.6287934906	0.036905	0.065596	0.035593
	...								
	42	47	64	27	0.3084383113	0.803585	0.04781246601	**0.081454**	0.055733
	43	48	64	27	0.3049668155	0.8050001952	0.04798033944	0.080478	**0.056259**
	44	49	63	28	0.3038680595	0.8052317503	0.049356	0.08014	0.052442
	45	50	62	29	0.3023369761	0.8054931566	**0.051025**	0.080189	0.04908497428
	...								
	49	54	60	32	0.2941402547	**0.8191581058**	0.041532	0.078276	0.046109
	Alpha								
LFM	0	5	126	0	0.2560283436	0.1295114562	0.04922694429	0.069688	0.03168
	0.5	6	90	0	0.3434830981	0.1295114562	0.03375740132	0.063101	0.036326
	0.6	6	90	0	0.3434830981	0.1295114562	0.03375740132	0.063101	0.036326
	0.7	9	86	0	0.3436331397	0.1295114562	0.038579	0.065264	0.037519
	0.8	8	86	0	0.341914754	0.1295114562	0.03775841077	0.066443	0.03625
	0.9	15	86	0	0.3290094215	0.1295114562	0.035189	0.076349	0.042517
	1	19	86	0	0.3339341458	0.1295114562	0.02676239691	0.038042	0.034276
PARIS		5	126	0	0.1253258451	0.1749535266	0.030291	0.049484	0.02355150125
Highest Values:		54	126	32	**0.3438994678**	**0.8191581058**	**0.051025**	**0.081454**	**0.056259**

Fig. 12. Evaluation Results on KEGG Dataset

The LFM algorithm, while slightly behind GN in modularity and ARI, showed significant adaptability with different alpha values, highlighting its flexibility in detecting community structures of varying granularity. Despite this, LFM did not reach the highest values in any metric compared to GN.

The Paris algorithm, although it provided a baseline comparison, consistently had lower performance across the metrics, with its highest ARI at Level 1 being 0.0492484, which is still lower compared to the ARI values achieved by GN and LFM.

Overall, the results from the KEGG dataset strongly support the superior performance of the GN algorithm in detecting hierarchical community structures within the gene co-expression network. Its high modularity, significant closeness centrality, and consistent ARI values make it a robust choice for uncovering complex biological interactions and functional groupings in the KEGG dataset.

7 Recommendations

Based on our findings and the insights gained from this study, we propose the following recommendations to further advance the field of gene co-expression network analysis and community detection:

1. Future research could investigate the application of dynamic community detection algorithms. These algorithms can handle time-varying data, capturing dynamic changes in gene co-expression networks over various time points. This approach could provide deeper insights into gene interactions.
2. There is a need to develop standardized benchmarking frameworks to systematically compare the performance of different community detection algorithms. Such frameworks would help in assessing the robustness and accuracy of different methods, facilitating the identification of the most suitable algorithms for specific biological datasets.

3. The methodologies developed in this study can be utilized for gene co-expression networks of other organisms. This approach would assist in assessing the generalizability of our findings and confirm the effectiveness of the community detection algorithms across various species and datasets.

4. Collaboration with domain experts, including biologists and geneticists, is essential for the successful interpretation of results. The experts can provide valuable biological insights that can guide the computational approaches and help in the accurate interpretation of the detected gene groups.

With these recommendations, future research can build upon this study to achieve a more comprehensive understanding of gene co-expression networks with community detection algorithms and their applications in biological mechanisms.

References

Bhalla, U. S., & Iyengar, R. (1999). Emergent properties of networks of biological signaling pathways. *SCIENCE*, *283*, 381–387. https://doi.org/10.1126/science.283.5400.381

Bonald, T., Charpentier, B., Galland, A., & Hollocou, A. (2018, June 22). Hierarchical graph clustering using node pair sampling. Retrieved November 27, 2023, from http://arxiv.org/abs/1806.01664

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., & Wagner, D. (2008). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, *20*(2), 172–188. https://doi.org/10.1109/TKDE.2007.190689

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., & Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, *2*(1), 65–73. https://doi.org/10.1016/S1097-2765(00)80114-8

Clemente, J. (2022). Predicting the biological classification of cell-cycle regulated genes of saccharomyces cerevisiae using community detection algorithms on gene co-expression networks. *Philippine Computing Journal*, *16*. https://doi.org/10.48550/arXiv.2208.10119

Coscia, M., Rossetti, G., Giannotti, F., & Pedreschi, D. (2014). Uncovering hierarchical and overlapping communities with a local-first approach. *ACM Transactions on Knowledge Discovery from Data*, *9*(1), 1–27. https://doi.org/10.1145/2629511

DasGupta, B., & Desai, D. (2013). On the complexity of newmans community finding approach for biological and social networks. *Journal of Computer and System Sciences*, *79*(1), 50–67. https://doi.org/10.1016/j.jcss.2012.04.003

Fortunato, S., & Newman, M. E. J. (2022). 20 years of network community detection. *Nature Physics*, *18*(8), 848–850. https://doi.org/10.1038/s41567-022-01716-7

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826. https://doi.org/10.1073/pnas.122653799

Gregory, S. (2010). Finding overlapping communities in networks by label propagation. *New Journal of Physics*, *12*(10), 103018. https://doi.org/10.1088/1367-2630/12/10/103018

Hajibabaei, H., Seydi, V., & Koochari, A. (2023). Community detection in weighted networks using probabilistic generative model. *Journal of Intelligent Information Systems*, *60*(1), 119–136. https://doi.org/10.1007/s10844-022-00740-6

Hanneman, R. A., & Riddle, M. (2005). Introduction to social network methods.

Hevey, D. (2018). Network analysis: A brief overview and tutorial. *Health Psychology and Behavioral Medicine*, *6*(1), 301–328. https://doi.org/10.1080/21642850.2018.1521283

Jaguzovic, M., Grbic, M., Dukanovic, M., & Matic, D. (2022). Identification of protein complexes by overlapping community detection algorithms: A comparative study. *2022 21st International Symposium INFOTEH-JAHORINA (INFOTEH)*, 1–6. https://doi.org/10.1109/INFOTEH53737.2022.9751314

Jarukasemratana, S., Murata, T., & Liu, X. (2014). Community detection algorithm based on centrality and node closeness in scale-free networks. *Transactions of the Japanese Society for Artificial Intelligence*, *29*(2), 234–244. https://doi.org/10.1527/tjsai.29.234

Lancichinetti, A., Fortunato, S., & Kertesz, J. (2009). Detecting the overlapping and hierarchical community structure of complex networks. *New Journal of Physics*, *11*(3), 033015. https://doi.org/10.1088/1367-2630/11/3/033015

Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J., & Pavlidis, P. (2004). Coexpression analysis of human genes across many microarray data sets. *Genome Research*, *14*(6), 1085–1094. https://doi.org/10.1101/gr.1910904

Liu, X., Cheng, H.-M., & Zhang, Z.-Y. (2019, February 18). Evaluation of community detection methods. Retrieved November 27, 2023, from http://arxiv.org/abs/1807.01130

Ma, X., & Gao, L. (2012). Biological network analysis: Insights into structure and functions. *Briefings in Functional Genomics*, *11*(6), 434–442. https://doi.org/10.1093/bfgp/els045

Montenegro, J. D. (2022, January 16). Gene co-expression network analysis. In *Plant bioinformatics* (Vol. 2443). https://doi.org/10.1007/978-1-0716-2067-0_19

Moore, C. (2017, August 23). The computer science and physics of community detection: Landscapes, phase transitions, and hardness. Retrieved November 27, 2023, from http://arxiv.org/abs/1702.00467

Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M., & Adebiyi, E. (2016). Clustering algorithms: Their application to gene expression data. *Bioinformatics and Biology Insights*, *10*, BBI.S38316. https://doi.org/10.4137/BBI.S38316

Palla, G., Derenyi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814–818. https://doi.org/10.1038/nature03607

Pons, P., & Latapy, M. (n.d.). Computing communities in large networks using random walks.

Rahiminejad, S., Maurya, M. R., & Subramaniam, S. (2019). Topological and functional comparison of community detection algorithms in biological networks. *BMC Bioinformatics*, *20*(1), 212. https://doi.org/10.1186/s12859-019-2746-0

Redekar, S. S., & Varma, S. L. (2022). A survey on community detection methods and its application in biological network. *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 1030–1037. https://doi.org/10.1109/ICAAIC53929.2022.9792913

Reichardt, J., & Bornholdt, S. (2006, July 18). Statistical mechanics of community detection. In *PHYSICAL REVIEW e* (Vol. 74). https://doi.org/10.1103/PhysRevE.74.016110

Rossetti, G. (2019). *Algorithms' table*. https://cdlib.readthedocs.io/en/latest/reference/cd_algorithms/algorithms.html

Rossetti, G., Pappalardo, L., Pedreschi, D., & Giannotti, F. (2017). Tiles: An online algorithm for community discovery in dynamic social networks. *Machine Learning*, *106*(8), 1213–1241. https://doi.org/10.1007/s10994-016-5582-8

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123. https://doi.org/10.1073/pnas.0706851105

Salido, J. A. (2016). Identification of candidate gene function group of yeast cell cycle genes using gene expression data.

Schwikowski, B., Uetz, P., & Fields, S. (2000). A network of protein–protein interactions in yeast. *Nature Biotechnology*, *18*(12), 1257–1261. https://doi.org/10.1038/82360

Tang, J. (2018). Prognostic genes of breast cancer identified by gene co-expression network analysis. *Frontiers in oncology*, *8*(374). https://doi.org/10.3389/fonc.2018.00374

Vieira, V. D. F., Xavier, C. R., & Evsukoff, A. G. (2020). A comparative study of overlapping community detection methods from the perspective of the structural properties. *Applied Network Science*, *5*(1), 51. https://doi.org/10.1007/s41109-020-00289-9

Wagenseller III, P., & Wang, F. (2017, December 2). Size matters: A comparative analysis of community detection algorithms. Retrieved November 3, 2023, from http://arxiv.org/abs/1712.01690

Xie, J., Szymanski, B. K., & Liu, X. (2011, November 10). SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. Retrieved November 27, 2023, from http://arxiv.org/abs/1109.5720

Yang, J., & Leskovec, J. (n.d.). Overlapping community detection at scale: A nonnegative matrix factorization approach.

Yeung, K. Y., Haynor, D. R., & Ruzzo, W. L. (2001). Validating clustering for gene expression data. *Bioinformatics*, *17*(4), 309–318. https://doi.org/10.1093/bioinformatics/17.4.309

Yeung, K. Y. (2001). Cluster analysis of gene expression data. https://api.semanticscholar.org/CorpusID:61503158

Yu, D., Kim, M., Xiao, G., & Hwang, T. H. (2013). Review of biological network data and its applications. *Genomics & Informatics*, *11*(4), 200. https://doi.org/10.5808/GI.2013.11.4.200

Zhang, H., Zeng, J., Tan, Y., Lu, L., Sun, C., Liang, Y., Zou, H., Yang, X., & Tan, Y. (2018). Subgroup analysis reveals molecular heterogeneity and provides potential precise treatment for pancreatic cancers. *OncoTargets*

and Therapy, *Volume 11*, 5811–5819. https://doi.org/10.2147/OTT.S163139

Zhang, X., Miller-Hooks, E., & Denny, K. (2015). Assessing the role of network topology in transportation network resilience. *Journal of Transport Geography*, *46*, 35–45. https://doi.org/10.1016/j.jtrangeo.2015.05.006