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

Jaimielle Kyle Calderon[1], Princess Angel Ventures[2], and Jhoirene Clemente[3]

[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
[3] Algorithms and Complexity Lab,
University of the Philippines Diliman
jbclemente@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*. 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. 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

© The Author(s) 2025
J. Caro et al. (eds.), Proceedings of the Workshop on Computation: Theory and Practice (WCTP 2024) , Atlantis Highlights in Computer Sciences 23,
https://doi.org/10.2991/978-94-6463-684-0_14

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. Biological activities arise from interactions among multiple molecules, cells, and genes (Ma & Gao, 2012). Initially thought to be linear and isolated, biological processes are now understood as complex, interconnected networks (Bhalla & Iyengar, 1999).

Biological networks, such as protein-protein interactions, metabolic pathways, and gene co-expression networks, are analyzed for various topological properties. Key properties include degree and centrality for identifying hub genes or biomarkers, and motifs and modules for functional units and recurring patterns (Ma & Gao, 2012; Rahiminejad et al., 2019).

Schwikowski et al. (2000) studied a global yeast protein network with 1,548 proteins and 2,358 interactions, predicting functions for 72% of characterized proteins and clustering proteins with similar functions and locations.

Tang (2018) used weighted gene co-expression analysis on GEO data to identify prognostic biomarkers for breast cancer. They discovered five modules, one of which correlated with cancer severity. Key hub genes like CASC5 and RAD51 in this module were associated with poorer survival outcomes.

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). Two methods for network (or graph) construction are rank-based and value-based. In value-based construction, genes are represented as vertices, and edges are formed based on Pearson's correlation coefficient to create the gene co-expression network.

Central to understanding networks is the concept of communities—non-overlapping groups of nodes more densely connected internally than with the rest of the network (Fortunato & Newman, 2022). Identifying these communities is crucial for understanding gene functional clustering and can uncover new aspects of gene function and interaction.

Gene expression data provides insights into gene activity and co-expression patterns. Gene co-expression networks, which reflect collective up-regulation or down-regulation of genes, help in identifying functional relationships and can suggest annotations for genes with unknown functions based on known genes in the same cluster (Oyelade et al., 2016; Yeung, 2001).

2.2 Community Detection in Networks

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 aims to identify clusters where nodes (genes) have higher connectivity within their group than with nodes outside (Fortunato & Newman, 2022; Girvan & Newman, 2002). In gene co-expression networks, this approach helps reveal hidden structures and functional associations.

Fortunato categorizes community detection algorithms into:

- **Optimization approach:** Scores are assigned to potential network divisions, favoring those with higher scores. Examples include Modularity, Spectral Clustering, and Hierarchical Clustering.
- **Statistical inference approach:** Communities are viewed as key drivers of network structure, with nodes connecting based on shared affiliations. Techniques include Stochastic Block Model (SBM), Bayesian Inference, and Maximum Likelihood Estimation.
- **Dynamical process approach:** This method connects community structure to dynamical processes on networks, using techniques such as InfoMap and Random Walk Methods.

Application of 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

Table 1 helps researchers choose the best method for network analysis, particularly for biological data like gene co-expression networks in *Saccharomyces cerevisiae*. Each method offers unique strengths for different network structures and community detection needs.

We identified three key challenges in the literature: dynamic communities, hierarchical community structures, and overlapping communities. To address these, we compiled a list of algorithms specifically designed for these issues 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.).

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

3 Related work

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

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.

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.

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.

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,

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 examine 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.

Girvan-Newman Algorithm

The Girvan–Newman algorithm excels in community detection by systematically removing edges based on their betweenness centrality. This iterative process continues until distinct community structures emerge, with the graph fragmenting into tightly connected groups. The effectiveness of the algorithm is evident in its ability to reveal natural divisions within a network, visualized through a dendrogram. While it is particularly suited for undirected and weighted networks, it does not handle directed networks (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).

Paris Algorithm

The Paris algorithm combines modularity-based clustering with a hierarchical approach, using a unique distance metric based on the probability of sampling node pairs. Operating agglomeratively, Paris effectively reveals community structures at various levels of granularity. While it excels with undirected networks, it does not support directed or weighted graphs. Its simplicity and intuitive distance metric make it valuable for exploring complex gene co-expression networks (Girvan & Newman, 2002).

LFM Algorithm

The LFM algorithm is a powerful tool for detecting overlapping communities and hierarchical structures in networks. It uses local optimization of a fitness function and adjusts community granularity through the alpha parameter, with larger values yielding smaller communities. This adaptability makes LFM ideal for undirected networks, allowing for tailored community detection based on specific data needs (Lancichinetti et al., 2009).

3.3 Statistical Analysis of Hierarchical Community Detection

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 rigorous assessment of the algorithmic outputs.

Internal criteria assess community quality based on network structure alone. Key metrics include:

- **Modularity**: Measures the density of links within communities versus between them, indicating the strength of the community structure. A high modularity score signifies effective community detection, especially relevant for algorithms like Girvan-Newman (Hajibabaei et al., 2023).

- **Closeness Centrality**: Assesses how central each node is within its community, reflecting the coherence and accuracy of node allocation (Jarukasemratana et al., 2014).

These metrics validate the algorithms' ability to capture the network's intrinsic clustering and the complex relationships in the gene co-expression data.

For external validation, we use the **Adjusted Rand Index (ARI)** to compare our algorithmic clusterings with true clusterings based on prior biological knowledge. ARI adjusts for chance grouping, providing a more accurate similarity measure between predicted and true clusterings. The ARI is calculated as $\text{ARI} = \frac{\text{RI} - E(\text{RI})}{\max(\text{RI}) - E(\text{RI})}$. An ARI of 1 indicates perfect agreement, while a negative score suggests less than random alignment (Clemente, 2022; H. Zhang et al., 2018). ARI helps evaluate how well the algorithms categorize gene sets according to their biological functions, offering a comprehensive understanding of their performance.

4 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 primary objective is to assess the capability of predicting hierarchical subclusters within the *Saccharomyces cerevisiae* gene co-expression dataset using hierarchical community detection algorithms. This objective addresses the following key issues:

1. **Outdated Gene Function Classification:** To enhance our findings, we updated the dataset by integrating recent genomic data. The outdated classifications from Cho et al. (1998) were replaced with updated gene function classifications from resources like the Kyoto Encyclopedia of Genes and Genomes (KEGG).
2. **Limited Precision in Gene Function Identification:** The five major classifications from Clemente (2022) were too general compared to the more detailed groups and subgroups from Cho et al. (1998). Our research provided a more precise identification of gene roles within the *Saccharomyces cerevisiae* gene co-expression network, improving upon existing classifications.
3. **Absence of Network Analysis on Hierarchical Sub-clusters:** A previous study by Salido (2016) has investigated hierarchical clusters within our chosen dataset, it limited its analysis to traditional cluster methods, particularly Non-metric Multidimensional Scaling (nMDS). Our study applied a network-based analysis to identify hierarchical clusters, comparing the performance of various algorithms on the yeast cell-cycle regulated gene network for a more nuanced community detection analysis.

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 take a look at the specific steps in our methodology.

5.1 Data Collection and Organization

Our methodology starts with collecting data from the Kyoto Encyclopedia of Genes and Genomes (KEGG) for genes in the dataset used by Cho et al. (1998). We scraped gene details, identifying three consistent classification levels, each with a unique ID.

We acquired classifications for 242 of the 384 genes from Cho et al. (1998). The remaining 142 genes had incomplete data in KEGG and were removed from the dataset.

To clarify the groupings, we generated a graph visualization. This graph, structured as a forest, had level 1 classifications as tree roots, with level 2 and level 3 classifications as intermediate levels, and the genes as the leaves.

Figure 1 shows the graph of the 242 genes and their classifications at each level. The hierarchical structure is clearly depicted, with classifications at lower levels nested under those at higher levels.

The next step is to explore relationships between KEGG groupings and the two classification levels from Cho et al. (1998). We used the Adjusted Rand Index (ARI) for this comparison. KEGG data was filtered to include only the 149 genes with detailed classifications from Cho et al. (1998), as not all 384 genes were classified in detail.

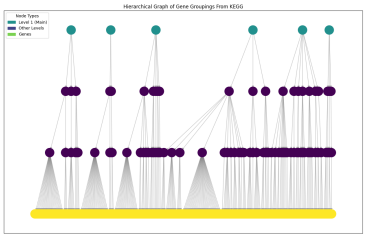

Fig. 1. Graph of Gene Groupings From KEGG

ARI was computed for each level combination, with the highest score around 0.16. The results are detailed 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

5.2 Gene Co-expression Network Formation

Building on previous insights, we proceed to construct gene co-expression networks using data from Yeung (2001). We employed a value-based algorithm for network construction, validated for its effectiveness in Clemente (2022).

Our process, organized in a Jupyter notebook for clarity, utilized NetworkX in Python to build the network by calculating pairwise gene similarities with Pearson's correlation coefficient. The δ parameter was introduced to set a threshold for co-expression, allowing fine-tuning based on gene similarity.

Figure 2 displays the gene co-expression network, illustrating that a δ value of 0.85 results in a structure with the most non-trivial connected components. This finding aligns with Clemente (2022), confirming similar connectivity patterns and functional groupings.

Fig. 2. Gene Co-expression Network

5.3 Script Implementation for Community Detection

For the automated community detection script, we reviewed algorithms from NetworkX and CDLib. NetworkX lacked LFM and Paris algorithms, highlighting CDLib's unique offerings.

We tested CDLib's GN algorithm, which effectively identified communities in our network data. The key parameter, "level," determines the analysis depth and where to cut the dendrogram for community structure observation. Increasing the level shifts the view from broad communities to smaller, specific ones, which will be discussed further with corresponding graphs.

Similarly, CDLib's LFM algorithm revealed hierarchical community structures, with the alpha parameter adjusting resolution. Testing alpha values from 0.5 to 1.0 showed a significant impact on community detection, with higher alpha values yielding more communities. Graphs for different alpha values will be shown in the next section.

For the Paris algorithm, no parameters were adjusted.

With the scripts for all three algorithms implemented, we will now evaluate their performance by comparing the identified groups to our dataset. This analysis will be covered in the following section.

5.4 Visualization, Evaluation, and Analysis

To compare the three community detection algorithms, we evaluated the communities using modularity, closeness centrality, and the Adjusted Rand Index (ARI). We also created visual representations of the networks to provide a clearer, intuitive understanding of the community structures and overall network.

Girvan-Newman

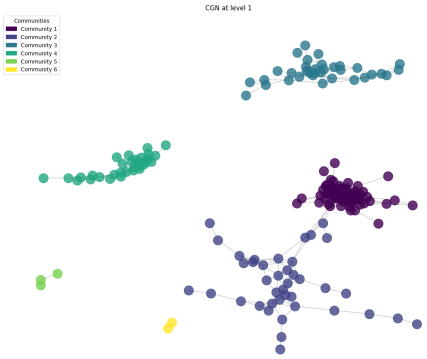

Fig. 3. Communities Detected Using GN at Level 1

At Level 1, the Girvan-Newman (GN) algorithm identified six distinct communities in the gene co-expression network, as shown in Figure 3. Each community is color-coded, illustrating clear separations and clustering of nodes.

The visualization reveals varying community sizes: Community 1 and Community 2 are the largest, while Community 6 is the smallest. This size distribution may reflect the hierarchical nature of gene interactions, where some gene groups are more tightly interconnected due to similar functions or regulatory roles.

The GN algorithm effectively captures the network's modular structure, with well-defined communities interconnected by fewer edges.

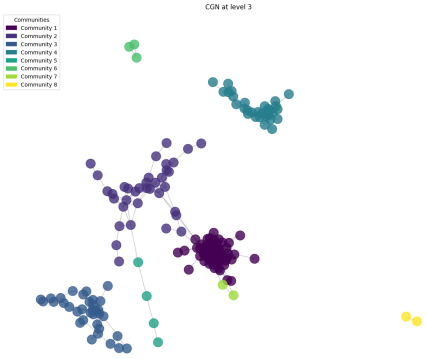

Fig. 4. Communities Detected Using GN at Level 3

At Level 3, the Girvan-Newman (GN) algorithm identifies eight distinct communities in the gene co-expression network, as shown in Figure 4. This level pro-

vides greater detail than Level 1, highlighting finer subdivisions within earlier identified clusters.

The visualization shows an increase in smaller communities, with Community 1 and Community 2 splitting into more specific clusters. This suggests that the GN algorithm effectively captures detailed interactions and sub-communities within the gene network.

The graph's layout indicates that GN can detect both broad and subtle relationships among genes, including isolated communities like Community 8. This finer granularity offers a deeper understanding of the network, revealing sub-structures that might be overlooked at higher levels, thus aiding targeted biological investigations.

Paris

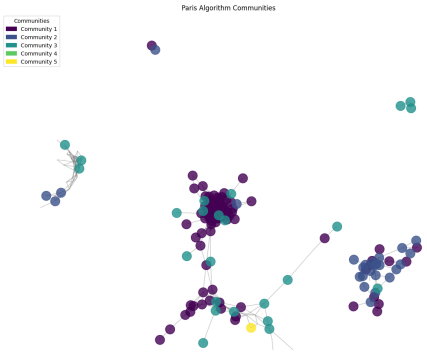

Fig. 5. Communities Detected Using Paris

The Paris algorithm, shown in Figure 5, identifies five distinct communities within the gene co-expression network. Each community is distinguished by color, revealing a range of sizes and structures.

Community 1, the largest and most central, forms a dense cluster, suggesting core functional relationships. Communities 2 and 3 are substantial but smaller, indicating distinct functional groups. Community 4 and Community 5 are smaller, with Community 5 comprising only a few nodes, possibly representing specialized or less connected genes.

The graph illustrates clear separations between communities, with some nodes linking different clusters, hinting at interactions or overlaps. The Paris algorithm provides a detailed view of the network's modular structure, capturing both large and small groupings effectively.

LFM

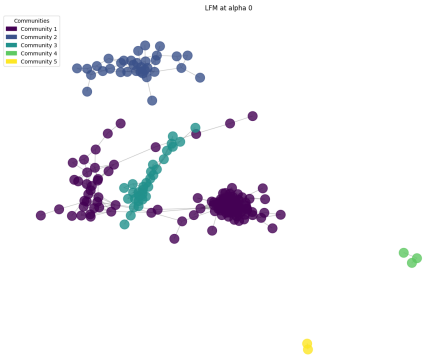

Fig. 6. Communities Detected Using LFM with Alpha=0

The LFM algorithm at alpha 0, illustrated in Figure 6, identifies five distinct communities within the gene co-expression network. Each community is color-coded to show clustering based on node connectivity.

Community 1, the largest and most central, indicates a core functional group or regulatory module. Community 2, another significant cluster, is located towards the network's upper part, suggesting a secondary functional grouping.

Communities 3 and 4 are smaller and more isolated, with Community 3 bridging between larger groups, implying an intermediary role. Community 5, the smallest and peripheral, may represent specialized or less central genes.

The network structure at alpha 0 shows clear community boundaries and varying sizes, demonstrating the LFM algorithm's ability to reveal both central and peripheral clusters in the gene co-expression network.

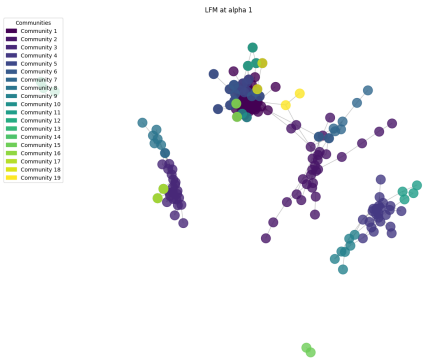

Fig. 7. Communities Detected Using LFM with Alpha=1

The LFM algorithm at alpha 1, shown in Figure 7, identifies nineteen distinct communities in the gene co-expression network, each color-coded for clarity.

At this alpha level, the network is divided into many smaller communities compared to alpha 0, demonstrating the algorithm's sensitivity to the alpha parameter, which enhances the resolution of community detection. As alpha increases, the LFM algorithm uncovers more detailed structures.

While Communities 1 and 2 remain central, they have fragmented into smaller sub-communities, revealing finer functional groupings. Smaller, compact communities like Community 8 and Community 11 indicate specialized gene groups with strong internal connections. The presence of many small communities, such as Community 18 and Community 19, highlights the detailed segmentation achieved.

The dense network with numerous inter-community edges illustrates the gene co-expression network's complex, interconnected nature, providing deeper insights into its modular organization and subtle functional relationships.

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. 8. 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 8, 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

	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	0.0337574013	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	0.04922694425	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. 9. Evaluation Results on KEGG Dataset

The evaluation of community detection algorithms on the KEGG dataset, shown in Figure 8, reveals several key insights:

- **Girvan-Newman (GN) Algorithm:** GN consistently outperforms the Local Optimization Function Model (LFM) and Paris algorithms. It achieved the highest modularity at Level 3 (0.3439), indicating superior identification of strong community structures. GN also excelled in closeness centrality, with the highest value at Level 49 (0.8192), suggesting that its identified communities are more centrally positioned. In terms of the Adjusted Rand Index (ARI), GN showed the best performance across levels, with the highest ARI at Level 2 (0.0815), indicating accurate clustering that closely aligns with the ground truth.

- **Local Optimization Function Model (LFM):** LFM demonstrated flexibility with different alpha values, adapting well to varying granularity of community structures. However, it did not achieve the highest values in modularity or ARI compared to GN.

- **Paris Algorithm:** Paris provided a baseline comparison with consistently lower performance across metrics. Its highest ARI at Level 1 was 0.0492, which was lower than those achieved by GN and LFM.

Overall, GN is identified as the most effective algorithm for detecting hierarchical community structures within the gene co-expression network, based on its superior modularity, closeness centrality, and consistent ARI values.

7 Recommendations

Based on our findings, we propose the following recommendations to advance gene co-expression network analysis and community detection:

1. Explore dynamic community detection algorithms to capture time-varying changes in gene co-expression networks. This could provide deeper insights into gene interactions over time.
2. Develop standardized benchmarking frameworks to systematically compare the performance of community detection algorithms, assessing their robustness and accuracy for various biological datasets.
3. Apply the methodologies from this study to gene co-expression networks in other organisms to assess the generalizability of our findings and validate the effectiveness of the algorithms across different species.
4. Collaborate with biologists and geneticists to enhance the interpretation of results, leveraging their expertise to guide computational approaches and accurately interpret detected gene groups.

These recommendations aim to enhance our understanding of gene co-expression networks and improve the application of community detection algorithms in biological research.

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

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

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

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

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.