Integrated Features Based on Graph Clustering and Gene Expression

Sura Ibrahim Mohammed Ali[1,2] and Sura Zaki Al Rashid[1]
[1]*Department of Software, College of Information Technology, University of Babylon, 51001 Babylon, Iraq*
[2]*Department of Mathematics and Computer Applications, College of Science, Al-Muthanna University, 66001 Sumaway, Iraq*
suraibraheem@mu.edu.iq, suraibrahimm.sw@student.uobabylon.edu.iq, sura.alrashid@uobabylon.edu.iq

Keywords: Topological Analysis, Multi-Feature Framework, Gene Expression, Graph Clustering, Gene Regulatory NetWork.

Abstract: Integrating different biological features – for example, the informativeness of topological features and gene expression – is challenging because each feature must be accounted for individually if the features are used to help forecast models. In this process, ensuring that the outcomes reflect the underlying biological structure of the network information while minimizing noise and irrelevant data is crucial. This study identifies the importance of rigorous pre-analyses in determining statistically significant correlations and joint effects among preprocess features before applying machine-learning techniques. Thus, when deploying multidimensional datasets, a systematic multi-feature methodology is presented in this paper to unify optimized graph clustering, weighted Jaccard similarity, and dimension reduction based on principal component analysis (PCA). Specifically, the objective was to identify novel uncharacterized gene associations in complex biological networks. Moreover, this study offers more refined insights into gene interactions within their networks, revealing patterns and relationships that might be hidden by broad data analysis. The method's performance was validated according to the benchmarks for a Dialogue on Reverse Engineering Assessment of Methods, fifth edition (DREAM5) challenge project, to determine its ability to analyze complex biological networks.

1 INTRODUCTION

Biological systems often involve complex interactions among numerous genes and chemicals within cellular networks, especially in multifactorial diseases [1], [2]. Advances in experimental and computational methods now enable precise mapping of physical (e.g., protein-protein, signaling, and regulatory) and functional (Yet, the scale of these networks poses challenges for extracting biological insights, making community detection a vital graph-clustering task for identifying functional modules [3].

While many algorithms have been tested on benchmark in silico networks, their utility in real molecular contexts remains uncertain. Nevertheless, identifying modules is critical for downstream analyses like link prediction and disease mechanism interpretation. Features from community structures and gene expression data can be combined to detect regulatory patterns and deeper functional associations. Dimensionality reduction techniques such as PCA help extract biologically relevant features from high-dimensional gene expression data. As early as 2010, PCA was shown to improve cancer subtype classification by incorporating biological structure [4], Moreover, integrating regulatory networks into statistical models allows for a richer interpretation of genetic architecture and biological function [5]. By combining community data with PCA loadings, researchers gain deeper insights into genetic network organization, surpassing traditional analytical methods [6]. This paper presents a novel multi-feature framework that integrates gene expression profiles, community structures, and principal components to identify highly variable genes within selected modules – prioritizing targets for experimental validation and enhancing biological understanding.

2 RELATED WORK

The graph structures are non-Euclidean, meaning they are characterized by irregular arrangements, make it hard to identify the nearby genes of a given point in the data, and vary in the number of surrounding nodes of various genes. Several recent studies have highlighted the importance of clustering

as a critical step in single-cell data integration and analysis. The method's primary innovation in this study [7] is its simultaneous consideration of the condition manifold's and gene manifold's internal geometric structures, this novel approach used DGPCA, its Laplacian embedding to approximate the cluster membership indicators and obtain principal components (PCs) to characterize the data. The condition manifold and gene manifold, two internal geometric structures that are appropriate for bi-clustering, are features of DGPCAA new interpretable framework for solving the single-cell RNA-Seq clustering challenge was proposed in [8]. The system could generate many clustering's of the same dataset at a low cost, in addition to retrieving results from a wide range of single-cell research. The algorithm was able to provide interpretable justifications for each of its choices and be used as a backend algorithm for interactive interpretation and analysis. In [9] to extract attribute information, the scDFN algorithm uses a dual mechanism that includes an autoencoder. This is the first study to use an enhanced graph network for single-cell topological data representation. Using the triple self-supervision technique and the cross-network information fusion mechanism, the information fusion module merges the retrieved attribute and topology data. To maximize the cell clustering representation throughout the entire model, quadruple joint losses are used. scGMAI, a novel Gaussian mixture clustering technique based on autoencoder networks and FastICA, was proposed in [10]. It is a powerful tool for precisely classifying and distinguishing cell types from scRNA-Seq data and demonstrates the wide range of applications it can have in scRNA-Seq data analysis. The defining genes of cell clusters are more precisely identified by this methodology. In scGMAI, FastICA selects the most important independent features to create a low-dimensional space that retains all of the data's fundamental properties. In order to assess how effectively biological variation is maintained during integration, the study in [11] has extensively benchmarked data integration techniques utilizing Louvain clustering. Researchers were able to statistically evaluate the quality of integration by comparing cluster assignments with reference cell type labels using measures such as NMI and ARI. This method has made clustering, when combined with strong validation metrics, a common technique for single-cell genomics batch correction evaluation and biological discovery.

3 METHODOLOGY

The nature of molecular networks is modular, and the subsets of nodes are more interconnected than pure chance would suggest. These subsets often consist of genes or proteins that share biological functions. Therefore, community detection is an intrinsic step towards deriving findings from network data [12]. However, biological networks' sheer scale and complexity present an obstacle to their analysis. Consequently, the present researcher has focused on community structures to identify functionally related groups of genes or proteins involved in biological processes. Here, together with module identification and graph clustering, the traditional network science task of community detection was the objective of applying a number of proposed methods [13].

In particular, this paper [14] reports on detecting protein complexes based on gene expressions. However, the resulting framework addresses broader challenges, such as integrating multiple data sources and analyzing complex networks. Nevertheless, despite these advances, there is still very little understanding of how the different approaches identify biologically meaningful communities in molecular networks. For instance, although the Dialogue on Reverse Engineering Assessment of Methods (DREAM) project demonstrated robust methods of network inference [15], very little downstream analysis has been conducted of the reconstructed networks, for example, in predicting regulatory pairs or discovering new links.

Hence, a broad three-component framework is proposed in this study, deploying robust methods of community detection to identify groups of functionally related genes accurately. Consequently, additional biological insights were anticipated, together with a PCA-based feature selection strategy for reducing the dimensionality of gene expression data by selecting various pertinent principal components (PCs), according to cumulative variance. This framework is schematically represented in Figure 1, which illustrates the workflow and further clarifies the functions of the regulatory gene network. Here, highly diverse features are typically gathered from a variety of databases (see top panel).

A feature table or matrix is then produced by merging the feature data with the already identified gene regulatory network, sourced from open databases like DREAM5. From this gene regulatory network, appropriate algorithms are applied to the relevant genes, selected by graph clustering from, for example, an E. coli dataset.

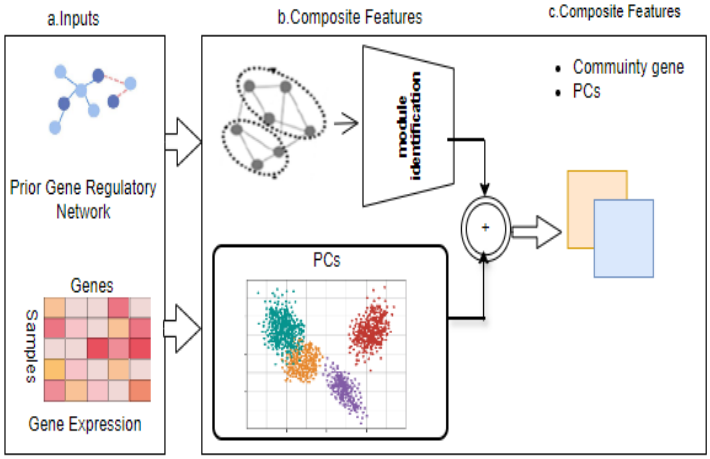

Figure 1: Overview of the proposed scheme for integrating biological data.

3.1 Feature Selection and Extraction

Feature selection enhances predictive accuracy and deepens understanding of biological networks, especially in the context of genomics and personalized medicine where accurate prediction is crucial [16]. To achieve this, the paper employs two main criteria: the "Predicting Communities" approach, which leverages prior knowledge of gene regulatory networks, and Principal Component Analysis (PCA), used to select the top 10 features (principal components) based on variance. Detailed descriptions of these criteria follow in the methodology section.

Features were extracted from both regulatory and inferred gene expression networks, focusing on dynamic elements such as transcription factors and co-expressed genes. Communities of highly interconnected genes and transcription factors were also incorporated. These were merged into a unified feature matrix, offering an integrated view of diverse biological data. Subsequent unsupervised clustering, based purely on network topology, identified non-overlapping communities ranging from 3 to 226 elements. This integrative approach supports robust exploration of molecular network structure.

3.2 Predicting Communities

In order to predict communities of related genes referred to as 'regulatory communities' or 'gene regulatory networks', it is crucial to cluster features using robust methods of community detection.

A framework of module predictions is proposed, consisting of four phases and represented in Figure 2. The architecture of the proposed framework consists of four key phases:

1) community detection to identify functional groups;
2) parameter optimization;
3) integration of transcription factors and target genes;
4) module validation to link biological functions with the detected communities.

In brief, Phase 1 consists of communities being used to seed or initialize groups of genes, in order to preprocess communities. The groups are refined by ranking and selecting those features that will best preserve the local structure, using mutual information.

In Phase 2, weighted Jaccard similarity takes place. The key difference between normal and weighted Jaccard is that the latter does not operate with binary features alone; instead, the numerical values of each feature used to calculate similarity. Specifically, in weighted Jaccard, the numerical values of each feature considered to compute similarity, as in the proposal, using the numerical values of the genes, for example, PageRank, in_degree, etc. Genes that show very similar numerical values have high similarity, thereby increasing the accuracy of the gene grouping in populations.

In Phase 3, information about aggregation attributes, shared neighbours, and connectivity is used to build similarity matrices, thereby enhancing the performance of the existing community detection algorithms**.**

Finally, in Phase 4, communities are clustered on the similarity matrix to obtain the final clustering (module). All four of these phases are described in detail in the following sub-sections.

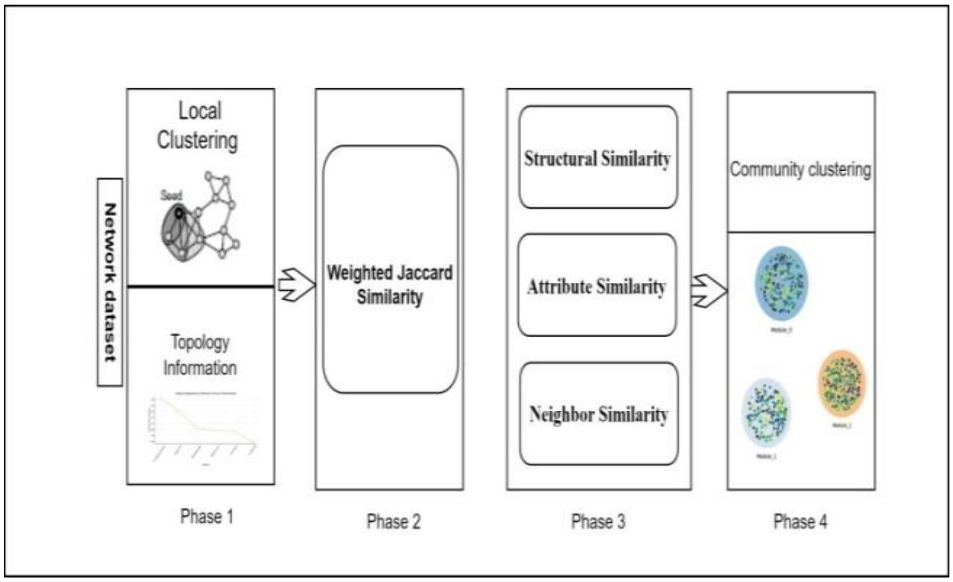

Figure 2: Framework of module predictions.

The communities illustrated in Figure 2 are made up of groups of genes and transcription factors that collectively perform specific biological functions, such as stress response or cell differentiation. Regulatory communities not only help shed light on the regulation of gene expression but also reveal the functional relationships between genes. This information helps to identify biological pathways and understand cellular processes. For instance, in stress response, a regulatory module could include stress resistance genes and their transcriptional regulators. Meanwhile, a cell differentiation network may contain differentiation genes and the transcription factors that control their expression. Therefore, this paper continues and extends the work reported in [17], introducing community detection algorithms that are enhanced by parameter learning. The latter has previously been applied in the literature to social networks but is now extended to biological networks. This parameter controls the tendency of nodes to form communities, consequently influencing community size by optimizing the modularity function.

The parameter would inversely influence the size of the detected communities, thereby increasing the value results in smaller, more tightly knit communities. This adjustment offers a scalable and flexible approach to analyzing large and complex biological networks. Unsupervised clustering algorithms were therefore employed, relying solely on network structure (link and attribute information) without additional biological annotations, consequently ensuring unbiased module identification.

3.2.1 Phase 1: Local Clustering and Topological Features

Phase 1 was subdivided into two stages: local clustering and topological feature extraction:

1) Local Clustering Phase. Initial clusters were formed using cluster density from the gene network, followed by attribute weighting and refinement via the Louvain modularity optimization algorithm [18], The DICCA algorithm was employed to group genes based on proximity and attributes, targeting densely connected subgroups. Jaccard similarity and modularity optimization further enhanced cluster identification and weighting.
2) Topological Feature Stage. Topological features – such as eigenvalue, closeness, and edge – were extracted to assess their influence on the target variable. Mutual information was used to rank these features, with the most informative ones selected for prediction tasks.

3.2.2 Phase 2: Weighted Jaccard Similarity Algorithm

Once the initial clusters are formed, the weighted Jaccard similarity algorithm can be applied as the key approach to detecting communities within biological networks. This algorithm is specifically designed to render similarity measures more relevant by accounting for the strength and importance of interactions. Other than the original Jaccard index,

which assigns equal weights to all attributes of interest, the weighted Jaccard similarity algorithm weights attributes according to their importance, minimizing the effect of weak/nonspecific interaction. In this paper, the weighted Jaccard similarity of two genes was calculated based on (1):

$$Similarity\ (i, j) = \frac{\sum min\ (atti, attj)}{\sum max\ (atti, attj)}. \quad (1)$$

In (1), *atti* and *attj* are the attribute vectors for the clustered *i*-th and *j*-th genes. This ensures that more importance is given to the features with higher values, while calculating the similarity score between the genes within a single cluster. It helps to identify biologically significant gene communities.

After constructing a Jaccard similarity matrix (see Definition 2) and calculating attribute weights (see Definition 3), the next logical steps consist of leveraging these definitions, in order to gain further insights into gene clustering and network analysis. On the basis of the attribute weights calculated therefrom, the algorithm emphasizes the importance of significant interactions and minimizes the influence of nonspecific interactions.

Definition 2: Jaccard similarity matrix construction: *J= [similarity (i, j)]*, where *J* is the Jaccard similarity matrix for a cluster, and *i*, *j* represent the index genes within that cluster. This matrix is populated by computing the Jaccard similarity for each pair of genes in the cluster.

Definition 3: Attribute weight calculation:

$$w_k = \frac{1}{n}\sum_{i=1}^{n} J_{ik}, \quad (2)$$

where w_k is the weight for the *k*-th attribute, calculated as the average Jaccard similarity involving that attribute across all pairs of genes within the cluster. Meanwhile, *n* denotes the number of genes in the cluster, and J_{ik} signifies the element of the Jaccard matrix for the *i*-th gene and *k*-th attributes.

Definition 4: Weighted gene similarity calculation: Using the attribute weight w_k derived in Definition 3, a weighted similarity score can be computed as (2) for any pair of *i* and *j* genes within a cluster. The weighted similarity measure adjusts the raw Jaccard similarity by amplifying the contribution of attributes with higher weights:

$$\text{Weighted Similarity (i,j)} = \sum_{k=1}^{m} w_k \ . J_{ik} \quad (3)$$

Where: w_k is the weight of the *k*-th attribute, J_{ik} is the Jaccard similarity matrix element for gene i and attribute k; *m* is the total number of attributes.

This step ensures that attributes with greater biological significance exert a stronger influence on the clustering process.

3.2.3 Phase 3: Aggregation of Information

Building on this, the algorithm incorporated a hybrid similarity measure (3) to integrate multiple aspects of similarity, including structural, attribute, and neighbour-based metrics:

$$\begin{aligned} Hyprid\ Similarity = \alpha \times Structural_{\ Similarity} + \\ + (1-\alpha) \times (\beta \times Attribute_{\ Similarity} + \\ + (1-\beta) \times Neighbour_{\ Similarity}) \end{aligned} \quad (3)$$

Structural similarity is captured by the adjacency matrix A while attribute similarity is represented by matrix W, and shared neighbour similarity by SNsim. The balance among these is controlled by parameters α and β, with α≥0.5 ensuring structural dominance. Jaccard similarity was integrated with these metrics to form biologically meaningful gene clusters. This method enhances network analysis accuracy and interpretability. Modularity , as defined in (4), was used to assess community quality by measuring intra-community density versus inter-community separation, thus validating the cohesiveness of detected structures. Using the following (4), modularity can be defined in formal terms, placing this evaluation in a mathematical framework:

$$Q = \frac{1}{2m}\sum \left[A_{ij} - \frac{K_i\ K_j}{2m}\right] \delta(C_i\ , C_j)\ . \quad (4)$$

Where:

- A_{ij} is entry into an adjacency matrix if the edge between the genes is i and j, otherwise valued as 0.
- $K_i\ and\ K_j$ represent the value of genes i and j, respectively. This refers to the number of edges that connect each gene.
- m is the total number of edges in the network.
- $\delta(C_i\ , C_j)$ represents the delta, equal to 1 when genes i and j are in the same module but otherwise equal to 0.

The Modularity formula evaluates the strength of communities by comparing actual and expected edge distributions within them – higher values indicate well-defined structures. This paper enhances modularity by optimizing weights α and β using network feature analysis rather than fixed values. Key features include the In-Degree Z-Score (which identifies influential nodes), PageRank (which assesses node importance via connectivity), and Betweenness Centrality (which highlights bridging nodes). This tuning aligns modularity with structural

and functional network properties, thereby enabling more precise detection of biologically and structurally significant communities.

3.2.4 Phase 4: Predicted Communities in Priors Network Analysis

Phase 4 consisted of constructing the hybrid similarity matrix, integrated with the Louvain algorithm to extract optimum community clusters. The predicted communities were compared with known biological databases to validate their biological relevance and interpret their genetic roles. In this study, the communities were grouped into three broad categories: (1) High Diversity, (2) High Potential, and (3) Moderate/Low (see Table 1). Meanwhile, some of the communities based on graph clusters and predicted in this study retained a considerable level of complexity and the potential for important biological activities, whereas others were expected to possess a wide range of biological roles and adaptabilities. Therefore, most of the communities were expected to have a less flexible or more specialized genetic configure ration, which could be concentrated in a small number of vital biological processes.

3.3 Principal Component Analysis Feature

Principal Component Analysis (PCA) was employed to reduce dimensionality while preserving key transcriptional signals [19]. In this study, PCA loadings were integrated with gene community information within a single interaction network to explore intra-community gene relationships and their biological implications. The normalized, log-transformed expression matrix was analyzed across all samples, retaining the top 10 principal components to capture maximum variance. To correct for batch effects, the Harmony algorithm was applied to the PCs [20]. ensuring the preservation of true biological signals. Focusing on the components with the highest contributions enhanced interpretability and supported more accurate modeling of gene relationships, ultimately improving feature matrices and biological prediction outcomes.

3.4 Biological Information Aggregation Using Multi-feature Reconstruction

Multi-feature reconstruction integrates outputs from PCA, community predictions, and gene expression into a unified feature matrix. This matrix captures key biological patterns by combining variance in gene expression, structural relationships, and community affiliation. Each sample is represented by a row, with columns encoding the top 10 principal components, expression values, and binary indicators of community membership. This integrative approach enhances the representation of latent factors influencing gene expression and improves predictive accuracy. By unifying multidimensional data, the matrix serves as a robust framework for identifying novel functional activities, structural links, and regulatory interactions – insights that would be missed using isolated data sources.

4 RESULT DISCUSSION

This study utilized gene expression microarray data from E. coli, sourced from the DREAM5 Challenge [21], comprising 4511 genes, 805 samples, and 2066 experimentally validated regulatory interactions. The network was constructed based on the RegulonDB database, which includes only interactions classified as strong evidence [22], [23] Given E. coli's status as a well-studied model organism, it serves as a reliable foundation for regulatory network analysis. In this context, validated interactions were treated as true positives, while others were considered negatives. The datasets used included E.coli_chip_features, E.coli_expression_ data, and the gold standard network.

For a clearer understanding of the data patterns, a smooth density curve is included with the histogram (see Figure 3), showing the distribution of average hybrid similarity among the clusters. Eight clusters are centred in the first noticeable peak, which is located at an average hybrid similarity of around 0.1. Another cluster grouping with eight clusters is indicated by the second peak, amounting to around 0.4. Based on the hybrid similarity metric, the distribution shows diverse levels of similarity in the nodes within clusters, ranging from 0.05 to 0.6. However, clusters close to 0.1 may have more varied features or weaker relationships.

The Louvain algorithm provided stability in the partitions. Section 3.3 lists a set of criteria besides community size that can be used to determine the likelihood of undiscovered genes and transcription factors. These criteria would help gain a better understanding of the regulatory network. Overall, the interactions between the gene communities were likely to be complex, involving shared regulators,

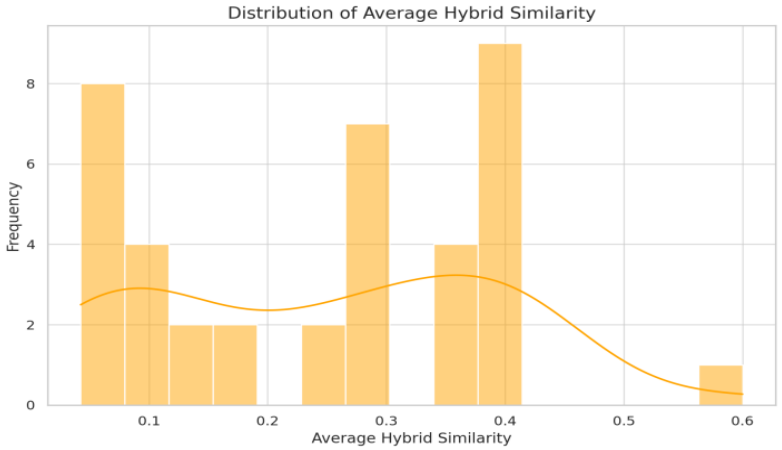

Figure 3: Distribution of hybrid similarity.

Table 1: Performance comparison of the proposed method with existing approaches.

Studies	Algorithm Used	Dataset	Modular-ity Score	No. of Predicted Communi-ties
Choobdar et al. (2019)	Random Walk Clustering	DREAM (multiple species)	0.45~	60
Proposed method	Weighted Jaccard Similarity Louvain Clustering	DREAM (E.coli)	0.6488	38

metabolic pathway connections, stress response mechanisms, regulatory cascades, and transport/signalling processes. However, the evaluation of predicted modules is challenging because there is no ground truth of 'correct' modules in molecular networks. Therefore, in this study, a framework was introduced to empirically assess modules based on their association with complex traits using STRING[1]. The updated STRING version was evaluated and the prediction of E. coli genes in the communities of 1081 was assessed, in order to further ascertain the accuracy of community predictions using the current researcher's methods. All of the communities added to the most recent version of STRING had a true percentage of > 90%, with the exception of two communities that had < 90%, namely, Communities 1 and 9, which had respective Jaccard index values of 0.52 and 0.75. Average precision for the total number of given communities was approximately 0.97.

The modularity score, calculated as 0.6488 using (4), was used to assess community detection quality. Compared to the 0.45 score reported by Choobdar et al. [20], this reflects a 44% improvement, indicating enhanced clustering and network organization. The results demonstrate the method's effectiveness in identifying biologically meaningful structures and improving the interpretation of gene interactions.

Table 1 demonstrates the performance of this method and the method proposed in [24] to predict communities across DREAM datasets. As this method has significantly outperformed other competing methods, it indicates the formativeness of topological properties and content in predicting communities.

A set of genes was analyzed within specific communities. These genes were selected based on their highest PC variance of average gene expression across samples. Thus, the goal of the current approach was achieved by integrating biological information from different networks to enable more accurate prediction of new gene connections and to support a deeper understanding of complex biological systems. These communities prioritize novel candidate genes; reveal pathway-level rate was then applied to the genes with the highest variance for each component, to identify important genes. In Figure 4, horizontal axis reflects the amount of variance of each gene, showing how gene expression varied across the PCs. The vertical axis represents the mean expression of each gene for all samples. Each dot refers to a gene with the highest PC variance indicated. The orange dots denote the position of the genes, whereas the text accompanying each dot (in blue) presents the names of the corresponding genes.

[1] https://string-db.org/

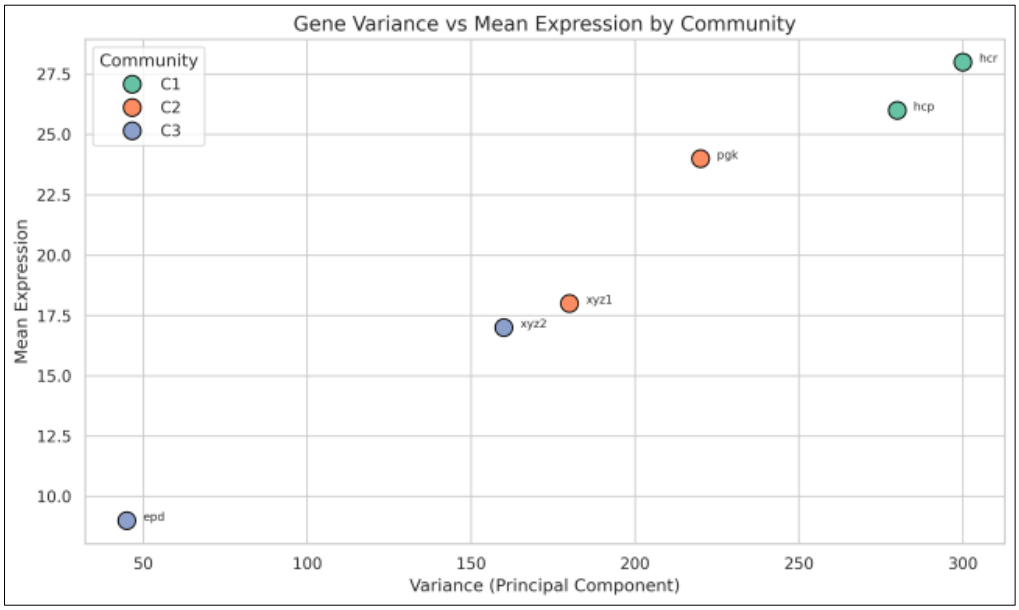

Figure 4: Relationships between variance and mean.

Table 2: Top genes selected by graph clustering dataset.

Genes		Biological Process	P-value
Community_11	ada, hns, lrp, tdh, …	Metabolic & Regulatory Networks	0.0283
Community_12	epd, aer, crr, crp,...	Cellular Regulation & Adaptation	0.00554
Community_25	fnr, hmp, hcp,…	Stress Response & Detoxification	**3.47×10^{-7}**
Community_2	hpt, acs, gcd, …	Energy Homeostasis & Metabolism	0.157
Community_16	lpd, icd, ssb	DNA Interaction & Metabolism	0.21
Community_23	uof, rob, slp, …	Environmental & Metabolic Adaptation	0.132

The genes to the upper right of the graphic (for example, hcr and hcp) exhibited high variance and high mean expression, indicating their potentially important role in a specific biological process. In contrast, the genes illustrated lower left (for example, epd) displayed low variance and low mean expression, and may therefore be less important in this context. The relationship between high variance and high mean expression can be interpreted as representing genes that are dynamically active and associated with a response or function.

Focusing on the results in Table 2, which correspond to the integration of expression data with topological properties, it may be seen that the method of integrating the expression data played an essential role in performance. Meanwhile, the p-values in the various groups denote the levels of statistical significance and stand to offer new insights into functional integration among the genes in each group. By way of illustration, Communi-ty_25 was topped with genes such as fnr, hmp, hcr, and hcp and had a p-value of 3.47×10-7, exhibiting highly significant interactions and integrations as physiological responses to oxidative conditions and nitrosative stress. Here, some communities such as 29 and 0 had p-values at 1, reflecting comparatively lower significance in their interactions. In contrast, high p-values in groups such as Community_29 and Community_0 could be the starting point for further, more detailed investigation into new relations and interactions, which could be relevant to deeper insights into biological networks.

Moreover, advanced bioinformatics tools and techniques may give way to new horizons opening up for possible discoveries in molecular biology and genomics. Conversely, Community_11 is composed

of ada, hns, lrp, tdh, and kbl genes with a p-value of 0.0283, thereby evidencing significant functional cooperation in terms of the impact on cellular regulatory networks and metabolic pathways. This difference in p-values enabled highly interactive groups to be identified, namely, high functional integration, which suggests the important biological functions of these genes in the face of environmental and physio-logical change. The identified genes were associated with critical biological functions such as gene regulation, cell signalling, and basic biological processes, consequently enhancing their biological significance.

5 CONCLUSIONS

This paper presents a comprehensive framework that leverages multiple types of features and advanced graph clustering techniques to predict gene interactions with notable precision. By integrating information from diverse biological data sources, the approach not only increased the accuracy of link prediction but also allowed for a nuanced exploration of both dynamic and static aspects within gene networks. The two-step clustering strategy proved effective in disentangling these facets, ultimately shedding light on key gene relationships that might otherwise remain obscured by traditional analysis methods. Importantly, the findings highlight the value of combining different data perspectives – such as gene expression profiles, topological network properties, and community structures – to gain a more holistic understanding of genetic connectivity. The resulting framework has enabled a more detailed mapping of gene regulatory mechanisms and offered new insights into the fundamental organization of cellular processes. This work not only enriches our understanding of gene networks but also establishes a methodological foundation for future studies aiming to interpret complex molecular interactions.

Moreover, the methodology has practical implications beyond basic research. Its ability to reliably predict novel gene interactions positions it as a valuable tool for applications in fields such as precision medicine and drug discovery, where understanding the intricacies of gene regulation is critical. The structured, data-driven approach laid out in this research provides a reproducible pathway for future exploration and validation in both experimental and clinical settings. Ultimately, this study contributes meaningfully to the ongoing effort to decode the complexity of biological systems through computational innovation.

REFERENCES

[1] D. Marbach et al., "Perturbations Across Complex Diseases," Genome Biology, vol. 18, no. 1, p. 236, 2016, [Online]. Available: https://doi.org/10.1038/nmeth.3799.

[2] M. J. Bonder, R. Luijk, D. V. Zhernakova, and M. Moed, "Disease variants alter transcription factor levels and methylation of their binding sites," 2015.

[3] S. S. Ahmed, S. Roy, and J. Kalita, "Assessing the Effectiveness of Causality Inference Methods for Gene Regulatory Networks," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 17, no. 1, pp. 56–70, 2020, [Online]. Available: https://doi.org/10.1109/TCBB.2018.2853728.

[4] B. A. Logsdon and J. Mezey, "Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations," PLoS Computational Biology, vol. 6, no. 12, 2010, [Online]. Available: https://doi.org/10.1371/journal.pcbi.1001014.

[5] S. Z. AlRashid, M. H. Dosh, and A. J. Obaid, "Classification of the Senescence-Accelerated Mouse (SAM) Strains With Its Behaviour Using Deep Learning," International Journal of e-Collaboration, vol. 18, no. 2, pp. 1–13, 2022, [Online]. Available: https://doi.org/10.4018/IJeC.304035.

[6] F. Wagner, "GO-PCA: An unsupervised method to explore gene expression data using prior knowledge," PLoS One, vol. 10, no. 11, pp. 1–26, 2015, [Online]. Available: https://doi.org/10.1371/journal.pone.0143196.

[7] J. X. Liu, C. M. Feng, X. Z. Kong, and Y. Xu, "Dual Graph-Laplacian PCA: A Closed-Form Solution for Bi-Clustering to Find 'Checkerboard' Structures on Gene Expression Data," IEEE Access, vol. 7, pp. 151329–151338, 2019, [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2941227.

[8] S. R. Datasets, J. M. Zhang, J. Fan, H. C. Fan, D. Rosenfeld, and D. N. Tse, "An Interpretable Framework for Clustering," bioRxiv, pp. 1–15, 2017.

[9] T. Liu and C. Jia, "scDFN: enhancing single-cell RNA-seq clustering with," vol. 25, no. 6, 2024.

[10] B. Yu et al., "ScGMAI: A Gaussian mixture model for clustering single-cell RNA-Seq data based on deep autoencoder," Briefings in Bioinformatics, vol. 22, no. 4, pp. 1–10, 2021, [Online]. Available: https://doi.org/10.1093/bib/bbaa316.

[11] M. D. Luecken et al., "Benchmarking atlas-level data integration in single-cell genomics," Nature Methods, vol. 19, no. 1, pp. 41–50, 2022, [Online]. Available: https://doi.org/10.1038/s41592-021-01336-8.

[12] N. A. A. Shanan, H. A. Lafta, and S. Z. Al Rashid, "Using alignment-free methods as preprocessing stage to classification whole genomes," International Journal of Nonlinear Analysis and Applications, vol. 12, no. 2, pp. 1531–1539, 2021, [Online]. Available: https://doi.org/10.22075/ijnaa.2021.5281.

[13] S. Fortunato and D. Hric, "Community detection in networks: A user guide," Physics Reports, vol. 659, pp. 1–44, 2016, [Online]. Available: https://doi.org/10.1016/j.physrep.2016.09.002.

[14] S. Noori, N. Al-A'araji, and E. Al-Shamery, "Construction of dynamic protein interaction network based on gene expression data and quartile one principle," Proteins: Structure, Function, and Bioinformatics, vol. 90, no. 5, pp. 1219–1228, 2022, [Online]. Available: https://doi.org/10.1002/prot.26304.

[15] S. M. Hill et al., "Inferring causal molecular networks: Empirical assessment through a community-based effort," Nature Methods, vol. 13, no. 4, pp. 310–322, 2016, [Online]. Available: https://doi.org/10.1038/nmeth.3773.

[16] C. G. Urzúa-Traslaviña et al., "Improving gene function predictions using independent transcriptional components," Nature Communications, vol. 12, no. 1, 2021, [Online]. Available: https://doi.org/10.1038/s41467-021-21671-w.

[17] A. Bhih, P. Johnson, and M. Randles, "An optimisation tool for robust community detection algorithms using content and topology information," The Journal of Supercomputing, vol. 76, no. 1, pp. 226–254, 2020, [Online]. Available: https://doi.org/10.1007/s11227-019-03018-x.

[18] A. Bhih, P. Johnson, T. Nguyen, and M. Randles, "Decentralized iterative community clustering approach (DICCA)," IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), vol. 2017-Octob, pp. 1–7, 2017, [Online]. Available: https://doi.org/10.1109/PIMRC.2017.8292677.

[19] N. Alrefaai and S. Z. Alrashid, "Classification of gene expression dataset for type 1 diabetes using machine learning methods," Bulletin of Electrical Engineering and Informatics, vol. 12, no. 5, pp. 2986–2992, 2023, [Online]. Available: https://doi.org/10.11591/eei.v12i5.4322.

[20] I. Korsunsky et al., "Harmony 2," Nature Methods, vol. 16, no. 12, pp. 1289–1296, 2019, [Online]. Available: https://doi.org/10.1038/s41592-019-0619-0.

[21] D. Marbach et al., "Wisdom of crowds for robust gene network inference," Nature Methods, vol. 9, no. 8, pp. 796–804, 2012, [Online]. Available: https://doi.org/10.1038/nmeth.2016.

[22] H. Salgado, A. Santos, U. Garza-Ramos, J. Van Helden, E. Díaz, and J. Collado-Vides, "RegulonDB (version 2.0): A database on transcriptional regulation in Escherichia coli," Nucleic Acids Research, vol. 27, no. 1, pp. 59–60, 1999, [Online]. Available: https://doi.org/10.1093/nar/27.1.59.

[23] G. J. Kang, S. R. Ewing-Nelson, L. Mackey, J. T. Schlitt, A. Marathe, and K. M. Abbas, "Neonatal Rat Ventricular Myocyte Isolation: HHS Public Access," Physiology & Behavior, vol. 176, no. 1, pp. 139–148, 2018. https://doi.org/10.1002/cpbi.43.

[24] S. Choobdar et al., "Assessment of network module identification across complex diseases," Nature Methods, vol. 16, no. 9, pp. 843–852, 2019, [Online]. Available: https://doi.org/10.1038/s41592-019-0509-5.