

Identification of Key Modules of Lung Cancer in Gene Regulatory Network using Greedy Modularity Optimization Approach

Zahra Sadat Mirghadery ^a, Mehrdad Kargari^{a*}, Mostafa Akhavan-Safar^b

^a Department of Information Technology Engineering, School of Systems and Industrial Engineering, Tarbiat Modares University, Tehran, Iran; z_mirghaderi@modares.ac.ir, M_kargari@modares.ac.ir

^b Department of Computer and Information Technology Engineering, Payame Noor University, Tehran, Iran; akhavansaffar@pnu.ac.ir.

ABSTRACT

Cancer is a complex and dangerous disease in which cells uncontrollably begin to grow. Some cells, with mutated genes, cause abnormalities in the cell. These abnormalities are transferred to other genes through specific interactions between genes, leading to disruptions in the normal function of cells. The result of these cell abnormalities will be the occurrence of cancer. In cancer, modules are considered as clusters of genes and regulatory molecules that play a role in the processes of cancer initiation and progression. These modules usually have a specific gene sequence as a central unit that is important in controlling and regulating cellular processes related to cancer.

In this study, a novel network-based method called mdGRN is proposed for identifying modules effective in lung cancer occurrence in the gene regulatory network. In this method, first, using gene expression data and regulatory interactions, a lung cancer regulatory network is constructed. Then, using a greedy modularity optimization approach, communities related to lung cancer are identified. Subsequently, the obtained communities are ranked using influence diffusion metrics in the network. Finally, the top-ranked communities are introduced as effective modules.


To assess the efficacy of the proposed method, the standard Cancer Genome Atlas (TCGA) database and four classifiers including a decision tree, k-nearest neighbors, support vector machine, and random forest were utilized. The results obtained demonstrated that the proposed mdGRN method outperforms other methods in identifying cancer modules in terms of the average harmonic mean metric with the support vector machine classifier. Additionally, in terms of the AUC metric, the proposed method achieved a value of 0.997 using the random forest classifier, indicating better performance compared to other previous methods in identifying cancer modules. Furthermore, the number of genes identified by the top module is compared with other previous computational and network methods. The results show that the top-ranked module, besides containing a considerable number of driver genes, contains unique genes that have not been identified by other methods.

Keywords— Cancer-Effective Modules, Gene Regulatory Networks, Greedy Optimization Algorithm, Lung Cancer Driver Genes.

1. Introduction

Cancer is a serious health problem worldwide, resulting from genetic influences and environmental factors [1]. Cancer refers to diseases that occur due to uncontrolled growth and abnormal proliferation of cells. According to the World Health Organization, it is recognized as the second leading cause of death globally [2]. Essentially, the renewal, proliferation,

and death of each cell are tightly controlled by the cellular genetic combination. The precise genetic control is immediately lost, mutations occur, and clonal evolution begins irreversibly leading towards cancer. These genetic mutations play a role in all cancers [3]. Studies on vast volumes of cancer genomic data have shown that cancer is a systemic network phenomenon attributed to the accumulation of genetic or epigenetic changes in molecular

 <http://dx.doi.org/10.22133/ijwr.2024.465678.1229>

Citation Z. Mirghadery, M. Kargari, M. Akhavan-Safar, "Identification of Key Modules of Lung Cancer in Gene Regulatory Network using Greedy Modularity Optimization Approach", *International Journal of Web Research*, vol.7, no.3, pp.65-77, 2024, doi: <http://dx.doi.org/10.22133/ijwr.2024.465678.1229>.

*Corresponding Author

Article History: Received: 2 February 2024 ; Revised: 27 May 2024; Accepted: 6 June 2024.

Copyright © 2024 University of Science and Culture. Published by University of Science and Culture. This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 International license (<https://creativecommons.org/licenses/by-nc/4.0/>). Noncommercial uses of the work are permitted, provided the original work is properly cited.

network architecture. In the molecular networks within cells, certain parts or "nodes" play crucial roles in the initiation and growth of cancerous tumors. Therefore, New research employs network-based methods to identify genes that may trigger cancer. These studies are conducted by extracting crucial portions of molecular networks [4]. Network-based analysis helps us identify and understand various relationships and interactions between genes and molecules within cells that play important roles in cancer initiation or progression. Recently, evidence has shown that instead of focusing on specific mutations in the genome, we can explore relationships and interactions between different components within cells. In other words, by studying mutated networks or regulatory pathways occurring within cells, we can better understand cancer. Additionally, by examining the status of molecules within molecular networks, we can further understand the impact of drugs on cancer [5].

The spread of cancer has made it necessary to provide methods to control the disease and produce effective drugs. All cellular activities and behaviors are directly or indirectly related to interactions between biological components present within the cell, thus analyzing the networks of these interactions plays a crucial role in understanding cell function. Furthermore, one of the major reasons for cancer occurrence is attributed to abnormalities in regulatory networks within a cell, which cause the unconventional expression of a gene to impact the expression of other genes, leading to the proliferation of this abnormality to other genes and ultimately resulting in the cell deviating from its normal function and cancer initiation [6].

The gene regulatory network (GRN) and the examination of interactions within it are of paramount importance in research related to biological systems and particularly in understanding cancer. These networks represent mechanisms that illustrate how genes are regulated within cells in a specific network pattern [7]. The gene regulatory network consists of a group of genes that interact with each other inside the cell and are facilitated by RNA and proteins encoded by them [8]. So far, although many driver genes have been identified, cancer diagnosis remains challenging. Generally, cancer arises due to the interplay of genetic factors and environmental causes [9].

It has been reported that lung, breast, and colorectal cancers are the most common cancers in 2018 [10]. Genes or gene products often collaborate as functional modules in molecular interaction networks, playing pivotal roles in organizing complex biological processes [11]. Therefore, the identification of disordered gene sets or modules as biological markers in cancer research is essential. The existing challenges in cancer diagnosis and the

importance of identifying reliable biological markers for early detection and personalized treatment are crucial and vital.

So far, Various methods have been proposed for identifying gene modules. For instance, the mRank method focuses on the limitations of current biological markers and suggests the use of an extensive set of biological markers for cancer diagnosis as a solution. In this method, a module detection and ranking approach is introduced with the aim of identifying network modules as cancer diagnostic biomarkers. The authors evaluated the effectiveness of mRank using hepatocellular carcinoma (HCC) data and demonstrated its superiority over existing methods. Additionally, their proposed method is justified through Network Ontology Enrichment Analysis and comparison with known genes associated with HCC, and its advantage lies in discovering new biological markers [12].

The Gene Set Analysis (GSA) method is another approach that focuses on identifying groups of genes declared in microarray experiments. This method is based on the Gene Set Enrichment Analysis (GSEA) method proposed by Subramanian et al. The authors propose two potential advancements for GSEA: the use of the Maximum statistical average to define differential subset gene information and another involving re-standardization for obtaining more accurate inferences [13].

Another network-based method for identifying modules is the Gene Set Enrichment Analysis (GSEA), which has been introduced as a powerful method for interpreting gene expression data, especially in the field of microarray analysis. GSEA focuses on gene sets rather than individual genes and utilizes prior biological knowledge to evaluate their correlation with phenotypic traits. This method addresses the limitations of single-gene analysis and has been shown to provide insights into various biological problems, including cancer-related datasets. The authors have demonstrated its application in various biological issues and introduced relevant software and databases for wider usage [14].

The Significance Analysis of Function and Expression (SAFE) method is another approach introduced by Barry et al. This method is two-staged and focuses on unknown information about gene correlations. It addresses issues arising from data collection in the analysis process and utilizes high-powered genomic and proteomic data for assessing the importance of gene sets. SAFE employs valid statistical analysis and uses permutations for error control. This framework utilizes gene categorization based on gene ontology and protein family databases to enhance its toolset and flexibility in discovering more biological insights [15].

In the field of bioinformatics, the network becomes a powerful tool for modeling functional interactions between genes/gene products, where nodes represent genes and edges denote their relationships [15].

The Crank method is another approach proposed for prioritizing network communities. Community detection methods often identify countless communities, but empirical validation of all of them is impractical. This method effectively evaluates the structural features of each community and combines them for community prioritization [16]. This method can be used with any community detection method and does not require additional metadata. Crank addresses the challenge of structurally prioritizing communities and finds applications in various domains, including those with specific domain knowledge.

As mentioned, detecting and describing community structures in networks is a fundamental subject in studying network systems in various scientific fields. Hence, the concept of modularity optimization-based approaches is a very effective approach for identifying community structures in networks. In fact, modularity can be expressed in terms of the eigenvectors of a specific matrix called the modularity matrix, leading to a spectral algorithm for community detection [17].

Graph clustering is a practical method that categorizes the vertices of a graph based on its edge structure. This method focuses on creating clusters with high internal connections and fewer inter-cluster connections [18].

In this study, a network-based method called mdGRN¹ is proposed to discover important and influential communities in the gene regulatory network. The goal is to identify communities containing the highest number of cancer genes. Identifying cancer communities can be effective in preventing further deviation in the network and also in therapeutic goals. Gene communities are the same as modules, referring to a group of genes with more interactions and interferences among them than with nodes outside the community, which is similar to the concept of protein complexes in protein-protein interaction networks.

The results obtained demonstrate that the efficiency of the proposed method is higher than other existing methods. Additionally, the importance of identified modules in terms of the number of cancer genes present in them was compared with other existing network and computational methods in the field of identifying cancer driver genes, showing significant results.

1.1. Community Detection Background

In a social network, people form groups based on their interests and commonalities, which are called communities in network science. From the perspective of network science, a community consists of a number of nodes and the edges between them, such that the relationships among members within a community are considerably stronger than the relationships between members of different communities. Figure 1 illustrates community detection in a social network and the types of relationships between communities. In some literature, the terms "group" or "cluster" are also used instead of "community".

Community detection is an active area of analysis in social network analysis aimed at understanding the flow of information and the dynamic nature of the network. In this approach, it is assumed that the set of interactions and information flow between communities determines the behavioral nature of the network. Therefore, community detection is the natural division of the network into densely interconnected nodes from other group nodes with fewer chances of connection [19].

Detecting communities in a network and determining their boundaries enable the classification of nodes based on their positions within the communities. In other words, nodes with central positions accommodate many nodes of the community and likely play a crucial and functional role in controlling and maintaining the stability of their community members. On the other hand, nodes positioned between communities play a vital role in communication and exchange between communities. In other words, these nodes act as intermediaries between communities. Consequently, large networks can be considered as networks where communities themselves are nodes, and the edges connecting these communities are considered as the main edges of the network. In this case, the relationships between the new network nodes, which are the communities themselves, can be examined. Ultimately, based on this new network, the base network can be evaluated [20].

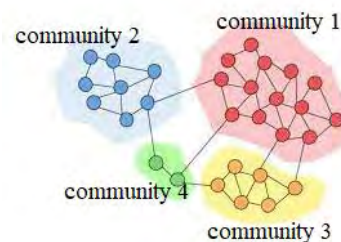


Figure 1. The community in the social network

¹ Module Detection in Gene Regulatory Network

Various methods have been proposed for community detection in networks, some of which are used in directed networks, and some are used in undirected networks. In this study, five algorithms including Louvain, Infomap, Walktrap, Floyd, and Greedy Modularity Optimization were examined, and ultimately the Greedy Modularity Optimization algorithm was selected due to the nature of the studied network. Additionally, the Greedy Modularity Optimization algorithm yielded the best results among the other algorithms in community detection and performance evaluation in the gene regulatory network.

1.2. Greedy Modularity Optimization Algorithm

Consider the graph $G = (V, E)$, where V is the set of vertices and E is the set of edges of the graph. Let n be the number of vertices in graph G and m be the number of its edges. Also, consider A as its adjacency matrix, where A_{ij} represents the number of edges connecting vertex i to vertex j . The greedy modularity optimization algorithm starts by creating n communities, each consisting of a single vertex. Let C_i be the community associated with vertex $i \in V$. In the next step, using Equ(1), merges two communities with the highest increase in modularity, where e_{ij} represents an edge in the network connecting a vertex in community i to a vertex in community j , and denotes the number of edges in the community i . This step continues until there are no more partitions in the network with higher modularity.

$$\Delta Q = 2(e_{ij} - a_i a_j) \quad (1)$$

Modularity is used to measure the quality of a community. The network modularity $Q(S)$ as the sum of modularity for each community, as expressed in Equ(2), is calculated, where l_c and k_c represent the number of edges and vertices in community $c \in S$. L represents the total number of edges and vertices in the network.

$$Q(S) = \sum_{c \in S} \left(\frac{l_c}{L} - \left(\frac{k_c}{2 \times L} \right)^2 \right) \quad (2)$$

According to the above explanation, the greedy modularity optimization algorithm is divided into two stages that are repeated sequentially. Suppose we start with a weighted network with N nodes. Initially, a distinct community is assigned to each node in the network. Therefore, in this initial partition, there exist as many communities as there are nodes. Then, for each node i , neighbors j from the side of I are considered, and the modularity gain obtained by removing I from its community and placing it in community j is evaluated. Then, node i is placed in the community where this gain is maximized. If no

positive gain is possible, i remains in its original community. This process is applied repeatedly and sequentially to all nodes until no further improvement is achieved, and then the first stage is completed. In this method, a node may be considered multiple times. This first stage stops when the local maximum modularity is achieved, meaning when no node movement can further improve modularity [21].

Part of the efficiency of this algorithm arises from the fact that the gain in modularity ΔQ obtained by transferring a separated node i to community C can easily be calculated using Equ (3).

$$\Delta Q = \left[\frac{\sum in + k_{i,in}}{2 \times m} - \left(\frac{\sum tot + k_i}{2 \times m} \right)^2 \right] - \left[\frac{\sum in}{2 \times m} - \left(\frac{\sum tot}{2 \times m} \right)^2 - \left(\frac{k_i}{2 \times m} \right)^2 \right] \quad (3)$$

Where:

$\sum in$ is the sum of weights of links within community C .

$\sum tot$ is the sum of weights of links entering the nodes within community C .

k_i is the sum of weights of edge entering node i .

$k_{i,in}$ is the sum of weights of links exiting node i to other nodes within community C .

m is the sum of the weights of all links in the network.

Therefore, in this algorithm, the change in modularity by removing node i from its own community and then transferring it to a neighboring community is evaluated. The second phase of the algorithm involves constructing a new network whose nodes are now the communities found in the first stage [22].

This algorithm tends to generate super-communities that include a large portion of the nodes, even in networks where there is no significant community structure. This tendency to produce super-communities can slow down the algorithm and make it impractical for networks with more than a million nodes.

2. Methods and materials

2.1. Research methodology

As stated earlier, abnormality occurs in one or more genes inside the cell, and then its transfer to other genes disrupts the normal function of the cell and leads to cancer. The goal of systemic medicine is to identify the starting points of system disorders and prevent the entire system from failing, detecting effective modules in cancer occurrence in gene regulatory networks can aid in early identification, controlling proliferation, and Effective medicinal targets. Therefore, considering this approach, cancer

modules are more likely to disrupt the entire regulatory network and also have the highest number of cancer driver genes. Therefore, using community detection algorithms in networks can help identify effective modules in cancer progression and occurrence, as well as reveal communities that contain the most induced genes. The approach of identifying cancer modules in gene regulatory networks has not been used so far.

After constructing the lung cancer regulatory network, the greedy modularity optimization algorithm was applied to the network. Then, the gene modules obtained were ranked using propagation-based algorithms to identify the best plant in terms of impact on the incidence and spread of cancer. Two algorithms, HITS and PageRank, were used for this purpose, with the PageRank algorithm selected due to optimality of the results. Then, the performance of the proposed mRank model was compared in terms of efficiency with 6 module identification methods. Decision tree classifiers, k-nearest neighbors, support vector machines, and random forests were used for accurate performance calculation. Additionally, the number of driver genes identified by the top module was compared with 18 previous network and computational methods for identifying cancer driver genes. An overview of the proposed mdGRN approach is depicted in Figure 2.

2.2. Gene regulatory network

Changes in gene expression significantly impact various biological mechanisms within a cell, prompting numerous studies to characterize this process in both healthy and diseased states. Gene regulatory networks, particularly transcription regulation networks, are valuable tools for describing and investigating these complexities. The network analyzed in this research is a type of gene regulatory network known as transcription regulatory. In this network, the nodes represent transcription factors (TFs) and genes, while an edge indicates the regulatory effect of the source node on the destination node. This regulatory effect implies that changes in the expression of the source gene can influence the expression of the target gene. Transcription factors are crucial for the regulation of transcription and are key components in every cell that control gene expression. [23] Dysregulation of their function plays a significant role in the development of diseases, particularly cancer. Analyzing these networks and examining TF-target relationships can provide valuable insights into the effects of individual genes within a biological system and help identify complex characteristics associated with human diseases. These networks are directional, with nodes representing

genes and transcription factors, while the edges denote physical or regulatory interactions between them. This type of network is employed in studies aimed at identifying cancer driver genes.

2.3. Data set

In this study, the TCGA Lung Squamous Cell Carcinoma (LUSC) gene expression dataset downloaded from the University of California Santa Cruz database was used². This data set was collected by the AffyU133a array. Gene expression Specifications were experimentally measured using the Affymetrix HT U133a human genome microarray platform by MIT's Broad Institute and Harvard University's Cancer Genomic Characterization Center.

This dataset shows gene-level transcription estimates. Genes were mapped to the human genome coordinates using the UCSC Xena HUGO probeMap. To facilitate the observation of differential expression between samples, the default view was centered at zero by independently subtracting the mean of each gene or exon at the center of each gene or exon. These data contain gene expression values in cancerous tissues and adjacent healthy tissues. Each sample pair in this dataset corresponds to a lung cancer patient, one derived from cancer cells and the other from neighboring normal cells. The dataset includes expression values of 12,043 genes in both healthy and cancerous tissues of 133 patients. Additionally, the regulatory interaction dataset from the RegNetwork database was used for network construction [24], and the validated cancer driver dataset from the Cancer Gene Census (CGC) was employed as the gold standard for result evaluation³, as explained in the subsequent sections.

2.4. Network Construction

The gene regulatory network is responsible for monitoring the speed and amount of transcription of genes into mRNA. This network has the ability to enhance or suppress gene expression, which has a significant impact on protein production. Disruptions in this network can lead to the production of proteins outside of normal constraints and ultimately result in cellular abnormalities and cancer. Identifying genes that play a role in initiating abnormalities and cancer is of great importance.

To construct the network, a list of regulatory interactions was needed, which was downloaded from the RegNetwork database. RegNetwork is a database of transcriptional and post-transcriptional regulatory networks in humans and mice. TF and

² <https://xenabrowser.net/datapages/>

³ <https://portal.gdc.cancer.gov>

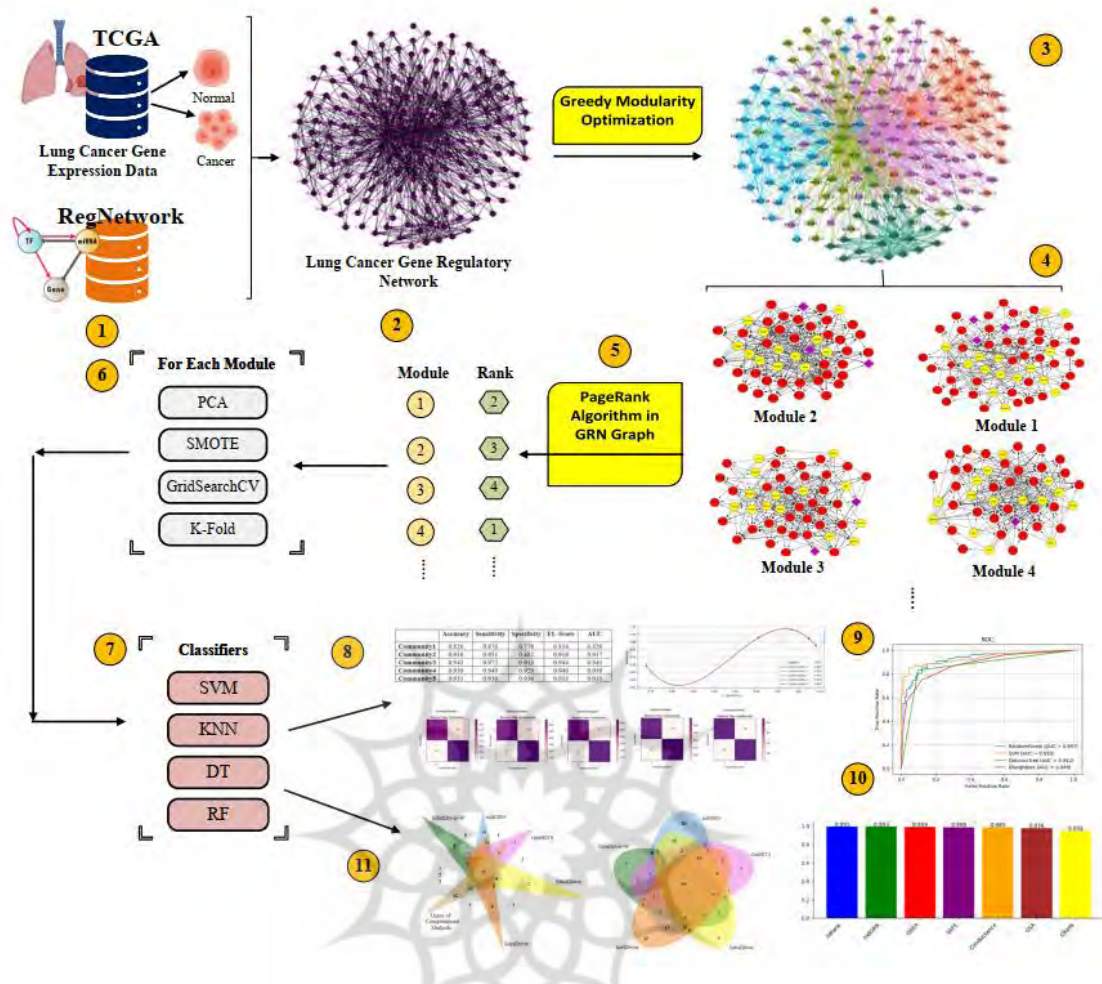


Figure 2. A view of the proposed mdGRN approach. 1.Data gathering,2. Network construction, 3. Module detection with GMA, 4.Module extraction, 5. Module ranking with PR, 6. Fine tuning methods for each module, 7. Classifier algorithms for each module, 8,9 and 10. Evaluation metrics, 11. Detected CDGs evaluating.

miRNA are the two main regulators that control gene expression. RegNetwork collects knowledge-based regulatory relationships as well as some potential regulatory relationships between two regulators and targets. It provides a platform to deposit known and predicted gene regulation at the transcriptional and post-transcriptional levels simultaneously [24]. Various interactions, including TF-gene, TF-TF, TF-miRNA, miRNA-gene, and miRNA-TF, are reported in this database. In this study, interactions related to miRNAs were ignored.

The gene expression dataset of lung cancer was mapped to the set of regulatory interactions to construct the network. Specifically, for each regulatory interaction, if there was a reported value in the gene expression dataset for both the source and target, the interaction was retained; otherwise, it was removed from the network. In this manner, the regulatory network related to lung cancer was built. Details of the constructed network are presented in Table 1.

Table 1. The details of the constructed network

Cancer network	Number of genes	Number of connections
Lung Cancer (LUSC)	18312	123224

The final image of the constructed regulatory network related to lung cancer, visualized using Gephi software version 0.10.0, is depicted in Figure 3. For a better understanding of the network, nodes with degrees greater than 10 (comprising 197 nodes and 1299 edges) are visualized.

2.5. Evaluation Metrics

An evaluation metric in data mining and machine learning is a quantitative measure used to assess the performance of trained models and algorithms on data. These metrics help us easily evaluate how accurately a model can predict data. Most metrics are derived from confusion matrices. The confusion matrix is shown in Table 2. In this matrix:

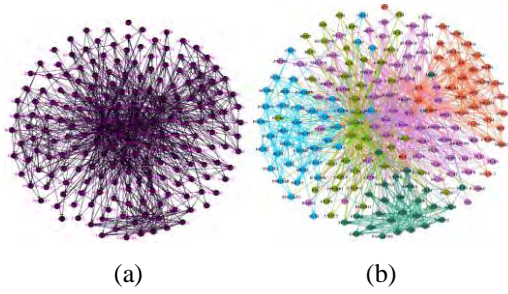


Figure 3. (a) Visualization of the lung cancer regulatory network, (b) visualization of network communities (for nodes with degrees higher than 10).

True Positive (TP): The number of positive instances that the classifier correctly predicted as positive.

False Positive (FP): The number of negative instances that the classifier incorrectly predicted as positive.

False Negative (FN): The number of positive instances that the classifier incorrectly predicted as negative.

True Negative (TN): The number of negative instances that the classifier correctly predicted as negative.

Using the elements of the confusion matrix, the following metrics are calculated and utilized:

Accuracy: Measures the proportion of instances that are correctly predicted (both positive and negative) out of the total number of instances. as shown

in Equ (4) this provides an overall evaluation of a model's performance.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (4)$$

Sensitivity: Sensitivity is also known as recall. It measures the ratio of true positive predictions out of all true positives. It is calculated using Equ(5).

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \quad (5)$$

Specificity: Specificity measures the ratio of true negative predictions to all true negative instances. It quantifies the model's ability to identify all negative instances. Specificity is calculated using Equ (6).

$$\text{Specificity} = \frac{TN}{(TN + FP)} \quad (6)$$

Precision: precision measures the ratio of true positive predictions to all positive predictions made by the model. It indicates how many of the positive

Table 2. Confusion Matrix for Binary Classification Problem

	<i>Positive Prediction</i>	<i>Negative Prediction</i>
<i>Positive Class</i>	True Positive (TP)	False Negative (FN)
<i>Negative Class</i>	False Positive (FP)	True Negative (TN)

predictions were correct. Precision is calculated using Equ (7).

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (7)$$

F1-Score (F-Measure): This metric combines Precision and Recall parameters to assess how well the model performs overall. It is also referred to as the "harmonic mean" of Precision and Recall. The F1-Score provides a more precise picture of the model's prediction performance for all classes in the data. The F1 criterion is one at best and zero at worst. It is calculated using Equ(8).

$$F - \text{Measure} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (8)$$

Area Under the ROC Curve (AUC-ROC): It is a measure that evaluates the ability of the model to distinguish between positive and negative classes by analyzing its performance at different probability thresholds. A higher value of this measure indicates a better ability to correctly classify samples, making it a useful measure when you need to evaluate the discrimination ability of a model. This criterion is calculated by drawing the receiver operating characteristic curve (ROC) and measuring the area under it. [25].

3. 3Results

The regulatory network of lung cancer was constructed using gene expression data and a list of regulatory interactions. Then, the greedy modularity optimization algorithm was executed to identify communities within the network. Subsequently, the obtained modules were ranked based on the PageRank metric to determine the most important modules. The modules were sorted according to the obtained scores. The module with the highest score was identified as the most effective module and in terms of performance metrics and the number of driver genes present in it, it was compared with other approaches in module detection and cancer driver identification methods. For all stages of the model execution and result evaluation, Python version 3.8.5 and various libraries (including NetworkX, ScikitLearn, etc.) were utilized.

In community detection, the fewer the number of communities and the higher the modularity of each community, the better the performance will be. Due

to the better performance of the greedy modularity optimization algorithm and the type of network studied, this algorithm was used as the final algorithm for identifying gene network modules. Table 3 shows the number of identified modules and the amount of normal and driver genes in each module.

Then, the identified modules were ranked using the PageRank algorithm. The PageRank algorithm is used to rank web pages in the World Wide Web. As shown in Figure 4, this algorithm assigns a numerical value (PageRank score) to each web page, indicating its importance in the web link structure. PR is a random algorithm that relies on the properties of random walks on web pages and Markov chain theory to calculate these scores [26]. This algorithm is used to rank web pages based on their connectivity and the quality and quantity of inbound links. Higher PR scores indicate greater importance, making it a valuable tool for search engines like Google to determine the relevance and credibility of web pages. The PR algorithm has also been employed in the context of identifying cancer driver genes. By doing this, genes are ranked based on their regulatory interactions and linking structures.

Results obtained from ranking gene modules are shown in Table 4.

Table 3. Identified modules using the greedy modularity optimization method and the number of normal and driver genes in each module.

Community Number	# Of Genes	# Of Driver Genes
Community1	3671	78
Community2	3449	57
Community3	3213	59
Community4	2792	249
Community5	2395	24
Community6	1811	46
Community7	700	12
Community8	277	2

Table 4. Ranking of communities obtained from the greedy modularity optimization method using PageRank.

Rank	Community	PageRank score
1	Community4	0.196918596278968
2	Community1	0.18882018480989746
3	Community2	0.17658317728813425
4	Community3	0.16594645263223817
5	Community5	0.12411289756502099
6	Community6	0.09597041227634573
7	Community7	0.036849357452636886
8	Community8	0.014150602771404456

Based on the ranks and PageRank scores obtained, the top 5 modules are described as follows: Community 4, Community 1, Community 2, Community 3, and Community 5. Each of the top 5 modules was visualized using Cytoscape version 3.7.1 based on nodes with high degrees and top genes in each of the 5 modules. As shown in Figure 5, normal genes are depicted in yellow, cancer genes in red, and top cancer genes in each community are drawn in diamond shapes and colored purple.

To evaluate the performance of the mdGRN model, genes present in it were labeled, with stimulatory genes labeled as 1 and normal genes labeled as 0. Then, according to the mRank model, evaluation metrics were calculated using decision tree classifiers, k-nearest neighbors, support vector machines, and random forests. Before running each classifier, one-sided encoding was applied to convert categorical data into numerical format, principal component analysis (PCA) was used for dimensionality reduction, the Synthetic Minority Over-sampling Technique (SMOTE) was used for class balancing, and a network search was conducted to calculate optimal hyperparameter values using the GridSearchCV technique.

Performance measures, confusion matrix, and ROC diagram of each of the top 5 communities are shown separately for each class in Figures 6 to 9.

To compare the performance metrics of the proposed model in each classifier, the averaged values were calculated. The final results of mdGRN and ROC curves are presented in Table 5 and Figure 10, respectively.



Figure 4. An example of PageRank ranking

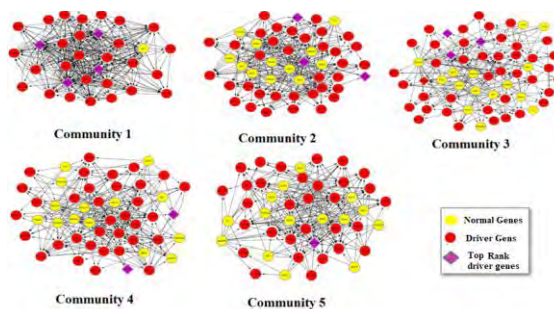


Figure 5. Visualization of the top modules in the lung cancer regulatory network.

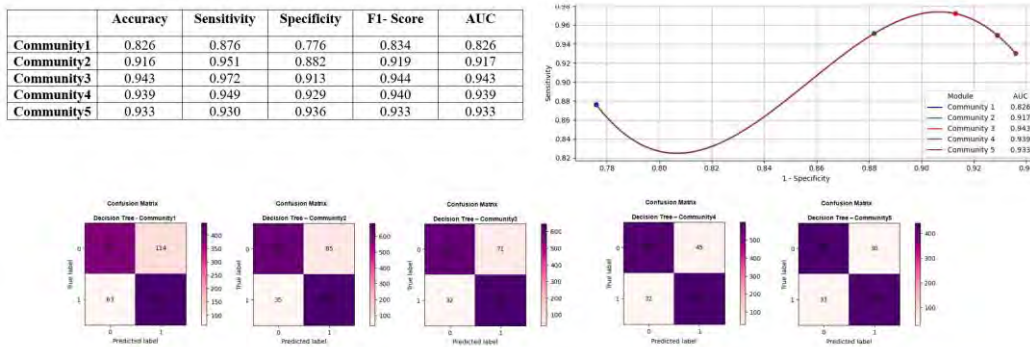


Figure 6. Performance metrics, ROC curve, and confusion matrix for the top 5 modules based on the decision tree classifier.

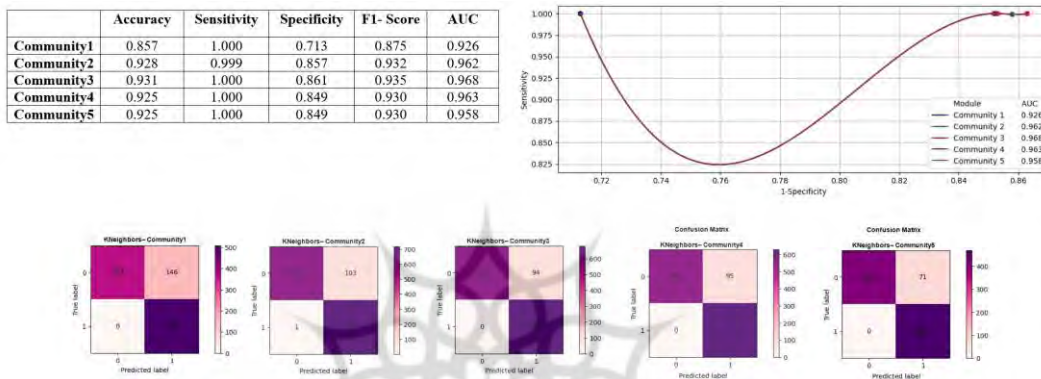


Figure 7. Performance metrics, ROC curve, and confusion matrix for the top 5 modules based on the k-nearest neighbors classifier

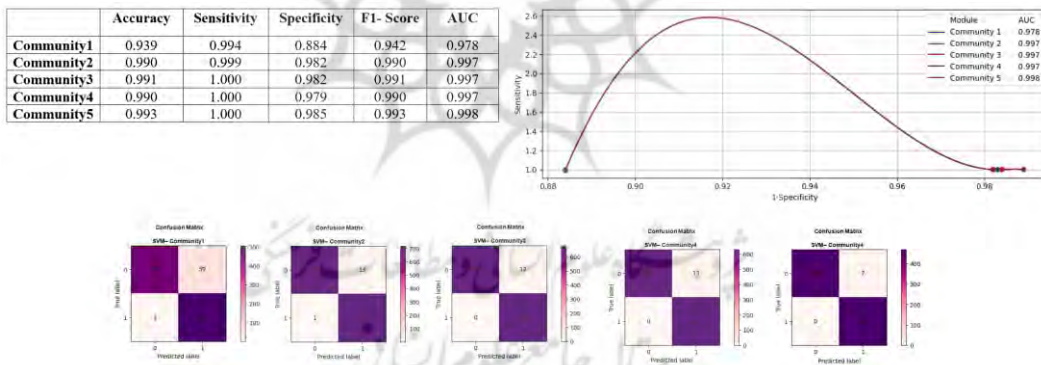


Figure 8. Performance metrics, ROC curve, and confusion matrix for the top 5 modules based on the support vector machine classifier.

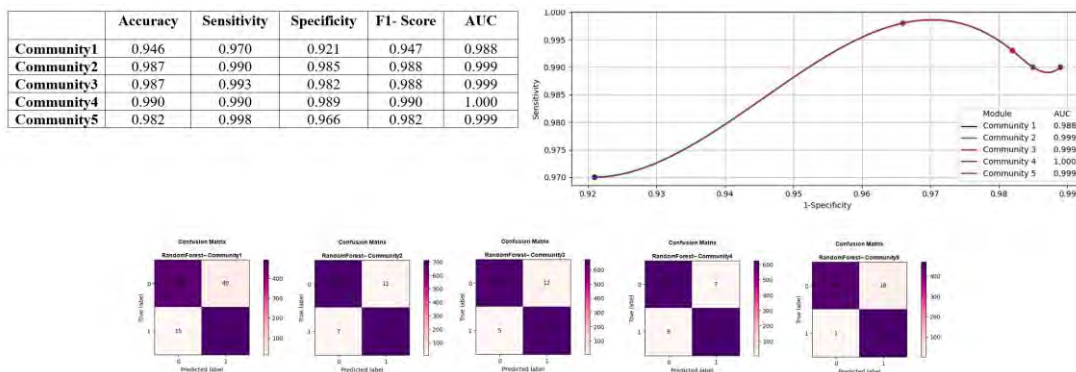


Figure 9. Performance metrics, ROC curve, and confusion matrix for the top 5 modules based on the random forest classifier.

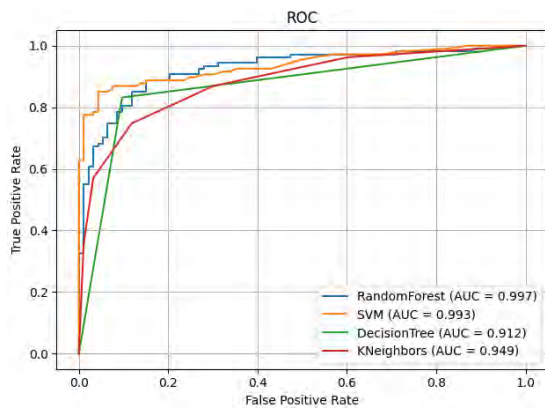


Figure 10. ROC curve of the mdGRN model with different classifiers.

Table 5. Comparison of performance metrics of the mdGRN model with different classifiers.

	Accuracy	Sensitivity	Specificity	F1-Score	AUC
DT	0.911	0.936	0.887	0.907	0.912
KNN	0.913	1.000	0.826	0.920	0.949
SVM	0.981	0.999	0.962	0.981	0.993
RF	0.978	0.988	0.969	0.979	0.997

Among the four classifiers, the support vector machine and random forest exhibit the best performance. The support vector machine shows AUC=0.993, a harmonic means of 0.981, and an accuracy of 0.981. also, the random forest shows AUC=0.997, a harmonic means of 0.979, and an accuracy of 0.978. Additionally, the ROC curves for all four classifiers demonstrate that the random forest has the highest area under the curve.

Furthermore, the performance of the proposed mdGRN model was compared with the baseline mRank model using two classifiers, namely, the support vector machine and random forest. As depicted in Figure 11, in the support vector machine classifier, the proposed mdGRN method shows a higher harmonic mean, sensitivity, and accuracy compared to the mRank method. also, in the random forest classifier, the proposed method demonstrates a higher AUC.

The mdGRN method has an AUC value of 0.993, indicating excellent class separation capability, and a strong F1 score of 0.981, which demonstrates a strong balance between precision and recall and is higher than the mRank method. Moreover, it has a specificity score of 0.962 and exceptional sensitivity (true positive rate) of 0.999, indicating its ability to correctly classify negative and positive instances. Its accuracy is also noteworthy at 0.981. Similarly, the mRank method performs with an AUC of 0.995, higher than mdGRN, an F1 score of 0.980, lower than mdGRN, and an accuracy of 0.981, which is the same

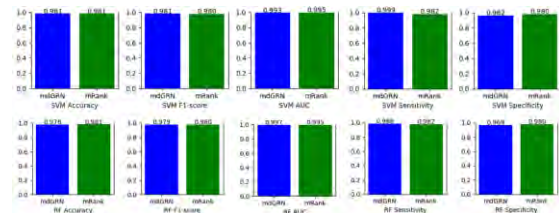


Figure 11. Comparison of the mdGRN model and the mRank model based on RF and SVM classifier.

as mdGRN, with a specificity of 0.980 and sensitivity of 0.982.

Additionally, the AUC of the proposed method was compared with five previous module detection methods. As shown in Figure 12, mdGRN achieved the highest AUC value among the previous methods with the random forest classifier. It also ranks second among the previous methods with the support vector machine classifier.

In addition to the above evaluations, the overlap of driver genes in the top module of the proposed method was compared with other methods for cancer driver gene detection. Considering the ranking of identified modules, only the driver genes identified in module number 4, which has the highest rank, were compared with other network-based and computational methods. Previous computational methods include the set of methods introduced in the study [1] (15 methods), and network-based methods include methods introduced in [1], [6], and [9]. The results of computational methods were extracted similarly from the DriverDB v2 database [27], and network-based methods were extracted from relevant articles. The accuracy of predicted drivers by mdGRN and other methods is evaluated by comparing each list with the list of standard genes introduced by the Cancer Gene Census (CGC) as the gold standard.

The top module of the mdGRN method identifies 249 driver genes. The overlap of identified genes in the Venn diagram in Figure 13 shows that mdGRN identified 71 driver genes identified by computational methods and additionally identified 178 driver genes not detected by computational methods. Furthermore, in comparison with network-based methods, mdGRN identified 150 genes identified by other methods and managed to identify 99 unique driver genes. Moreover, the proposed method successfully identified 75 driver genes that were not identified by any of the previous network-based and computational methods.

4. Conclusion

In this study, a method for identifying effective modules in lung cancer occurrence in the gene regulatory network called mdGRN, was proposed. The greedy module optimization algorithm was used for module aggregation. The use of this method for

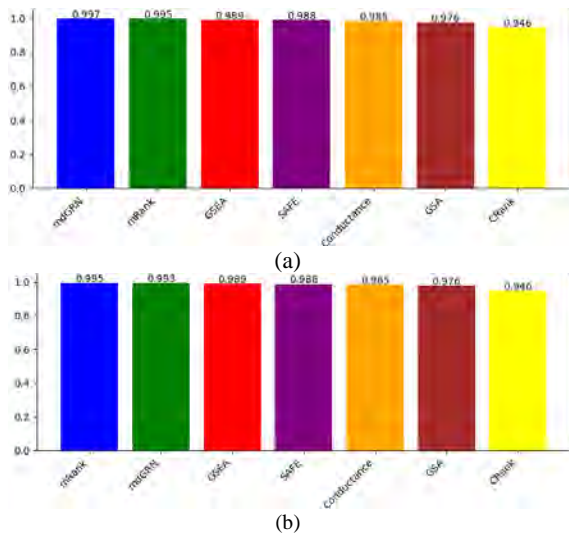


Figure 12. AUC comparison in the mdGRN method and other methods with (a) the random forest classifier and (b) based on the support vector machine classifier.

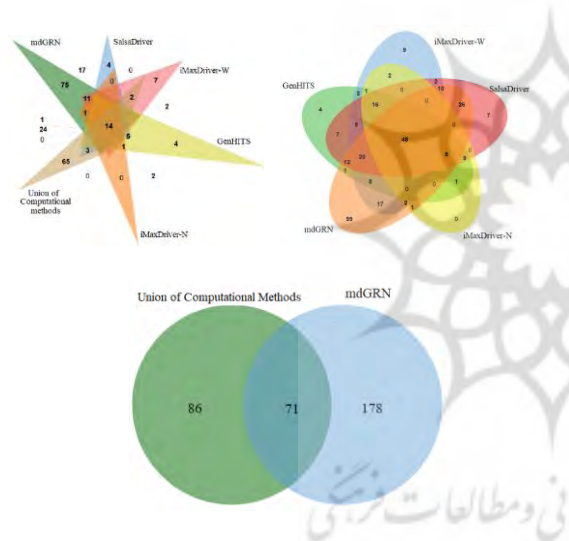


Figure 13. The level of gene overlap identified by mdGRN and other network-based and computational methods

identifying cancer modules in the gene regulatory network has not been used before. Then, the obtained modules were prioritized based on the PageRank algorithm for importance ranking. The performance of the proposed method was compared with six previous module identification methods using the standard Cancer Genome Atlas (TCGA) database and four classifiers: decision tree, k-nearest neighbors, support vector machine, and random forest. The results showed that the proposed method mdGRN outperforms other methods in terms of the average harmonic mean with the support vector machine classifier. Additionally, in terms of the AUC metric, the proposed method with the random forest classifier with a value of 0.997 also outperformed other previous methods in identifying cancer modules. Furthermore, the number of genes identified by the

top module was compared with 18 previous computational and network-based methods. The results show that the top module in the proposed method not only contains a significant number of driver genes but also contains unique genes that were not identified by other methods. This approach can be used in other cancers and biological networks as well. One of the limitations pertains to the lack of data during the formation of gene regulatory networks for each disease. Since expression values were not reported for certain genes, some driver genes were excluded from the final structure of the constructed networks. Addressing this issue in future research could enhance the performance of the proposed methods.

5. Future works

Based on the findings of this study, several avenues for future research can be explored. The methods employed to identify and rank communities can be applied to other biological networks. The dynamics of gene regulatory networks, including how rankings change over time or under different conditions, can be investigated. In fact, dynamic network analysis can uncover the evolving impact of potential cancer genes. Additionally, more advanced machine learning techniques and other network centrality measures can be explored to enhance ranking accuracy. Furthermore, the proposed method can be applied to other cancers or diseases.

Declarations

Funding

This research did not receive any grant from funding agencies in the public, commercial, or non-profit sectors.

Authors' contributions

Z.S. Mirghadery: acquisition of data, interpretation of the results, statistical analysis, drafting the manuscript.

M. Kargari: Study design, interpretation of the results, revision of the manuscript

M. Akhavan-Safar: Study design, interpretation of the results, revision of the manuscript.

Conflict of interest

The authors declare that no conflicts of interest exist.

References

[1] M. Akhavan-Safar, B. Teimourpour, and M. Kargari, "GenHITS: A network science approach to driver gene detection in human regulatory network using gene's influence evaluation", *Journal of Biomedical Informatics*, vol. 114, p. 103661, 2021. <https://doi.org/10.1016/j.jbi.2020.103661>

- [2] [2] World Health Organization, Cancers, 12 September 2018. <https://www.who.int/en/news-room/fact-sheets/detail/cancer>
- [3] [3] A. S. Nath, A. Pal, S. Mukhopadhyay, and K. C. Mondal, "A survey on cancer prediction and detection with data analysis", *Innov. Syst. Softw. Eng.*, vol. 16, pp. 231–243, 2020. <https://doi.org/10.1007/s11334-019-00350-6>
- [4] [4] J. Q. Liu, X. R. Li, and J. C. Dong, "A survey on network node ranking algorithms: Representative methods, extensions, and applications", *Science China Technological Sciences*, vol. 64, no. 3, pp. 451–461, 2021. <https://doi.org/10.1007/s11431-020-1683-2>
- [5] [5] W. Zhang, J. Chien, J. Yong, R. Kuang, "Network-based machine learning and graph theory algorithms for precision oncology", *Npj Precision Onc.*, vol. 1, no. 1, p. 25, 2017. <https://doi.org/10.1038/s41698-017-0029-7>
- [6] [6] M. Akhavan-Safar, B. Teimourpour, and A. Nowzari-Dalini, "A Network-Based Method for the Detection of Cancer Driver Genes in Transcriptional Regulatory Networks Using the Structural Analysis of Weighted Regulatory Interactions", *Current Bioinformatics*, vol. 17, no. 4, pp. 327–343, 2022. <https://doi.org/10.2174/1574893617666220127094224>
- [7] [7] K. Feng, H. Jiang, C. Yin, and H. Sun, "Gene regulatory network inference based on causal discovery integrating with graph neural network", *Quantitative Biology*, vol. 11, no. 4, pp. 434–450, 2023. <https://doi.org/10.1002/qub.2.26>
- [8] [8] M. Shruti, D. Mishra, and S. Kumar Satapathy, "Integration and visualization of gene selection and gene regulatory networks for cancer genome. Academic Press, 2018.
- [9] [9] M. Rahimi, B. Teimourpour, and S. A. Marashi, "Cancer driver gene discovery in transcriptional regulatory networks using influence maximization approach", *Computers in Biology and Medicine*, vol. 114, p. 103362, 2019. <https://doi.org/10.1016/j.combiomed.2019.103362>
- [10] [10] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018. <https://doi.org/10.3322/caac.21492>
- [11] [11] A. L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization", *Nature reviews genetics*, vol. 5, no. 2, pp.101-113, 2004. <https://doi.org/10.1038/nrg1272>
- [12] [12] H. Shang and Z. P. Liu, "Network-based prioritization of cancer biomarkers by phenotype-driven module detection and ranking", *Computational and Structural Biotechnology Journal*, vol. 20, pp. 206–217, 2022. <https://doi.org/10.1016/j.csbj.2021.12.005>
- [13] [13] B. Efron and R. Tibshirani, "On testing the significance of sets of genes", *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 107–129, 2007. <https://doi.org/10.1214/07-aos101>
- [14] [14] A. Subramanian, et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15545-15550, 2005. <https://doi.org/10.1073/pnas.0506580102>
- [15] [15] W. T. Barry, A. B. Nobel, and F. A. Wright, "Significance analysis of functional categories in gene expression studies: A structured permutation approach", *Bioinformatics*, vol. 21, no. 9, pp. 1943–1949, 2005. <https://doi.org/10.1093/bioinformatics/bti260>
- [16] [16] M. Zitnik, R. Sosič, and J. Leskovec, "Prioritizing network communities", *Nature Communications*, vol. 9, no. 1, p. 2544 2018. <https://doi.org/10.1038/s41467-018-04948-5>
- [17] [17] M. E. J. Newman, "Modularity and community structure in networks", *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006. <https://doi.org/10.1073/pnas.0601602103>
- [18] [18] S. E. Schaeffer, "Graph clustering", *Computer Science Review*, vol. 1, no. 1, pp. 27–64, 2007. <https://doi.org/10.1016/j.cosrev.2007.05.001>
- [19] [19] R. George, K. Shujaee, M. Kerwat, Z. Felfli, D. Gelenbe, and K. Ukuwu, "A Comparative Evaluation of Community Detection Algorithms in Social Networks", *Procedia Computer Science*, vol. 171, no. 2019, pp. 1157–1165, 2020. <https://doi.org/10.1016/j.procs.2020.04.124>
- [20] [20] S. Fortunato, S. "Community detection in graphs", *Physics reports. Elsevier*, vol. 486, no. 3–5, pp. 75–174, 2010. <https://doi.org/10.1016/j.physrep.2009.11.002>
- [21] [21] H. C. Rustamaji, W. A. Kusuma, S. Nurdianti, and I. Batubara, "Community detection with greedy modularity disassembly strategy", *Scientific Reports*, vol. 14, no. 1, pp. 4694, 2024. <https://doi.org/10.1038/s41598-024-55190-7>
- [22] [22] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks", *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, P10008, 2008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- [23] [23] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, "A census of human transcription factors: function, expression and evolution", *Nature Reviews Genetics*, vol. 10, no. 4, pp. 252–263, 2009. <https://doi.org/10.1038/nrg2538>
- [24] [24] Z. P. Liu, C. Wu, H. Miao, and H. Wu, "RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse", *Database*, vol. 2015, p. bav095, 2015. <https://doi.org/10.1093/database/bav095>
- [25] [25] F. B. Gereme and W. Zhu, "Early Detection of Fake News 'Before It Flies High'", *Proceedings of the 2nd International Conference on Big Data Technologies - ICBDT2019*, presented at the the 2nd International Conference, ACM Press, Jinan, China, 2019, pp. 142–148. <https://doi.org/10.1145/3358528.3358567>
- [26] [26] N. Ma, J. Guan, and Y. Zhao, "Bringing PageRank to the citation analysis", *Information Processing & Management*, vol. 44, no. 2, pp. 800–810, 2008. <https://doi.org/10.1016/j.ipm.2007.06.006>
- [27] [27] I. F., Chung, et al., "DriverDBv2: a database for human cancer driver gene research", *Nucleic acids research*, vol. 44, no. D1, pp. D975–D979, 2016. <https://doi.org/10.1093/nar/gkv1314>



Zahra Sadat Mirghaderi was born in 1994. She received her B.Sc. in Computer Engineering in 2018. Currently, she is an M.Sc. student in Information Technology at Tarbiat Modares University. Her research interests include bioinformatics, genetics, information systems, and social network analysis.



Mehrdad Kargari received his PhD degree in industrial engineering from Tarbiat Modares University of Iran. He is currently an Associate professor at the department of Information Engineering, in Tarbiat Modares University. His research interests are in the fields of machine learning, artificial intelligence, IoT and their applications in health or Banking.



Mostafa Akhavan-Safar is an Assistant Professor of Information Technology Engineering currently at the School of Computer and Information Technology Engineering of Payame Noor University. He received his M.Sc. in Information Technology Engineering from Iran University of Science and Technology (IUST), and Ph.D. in Information Technology Engineering from Tarbiat Modares University (TMU), Tehran, Iran. His research interests include Bioinformatics, Data Mining, Machine learning, Information systems and Social Network Analysis.

