IJICTR
International Journal of Information &
Communication Technology Research

Volume 14- Number 1 – 2022 (57 -68)

cMaxDriver: A Centrality Maximization Intersection Approach for Prediction of Cancer-Causing Genes in the Transcriptional Regulatory Network

Sajedeh Lashgari
Department of Data Science,
School of Mathematical
Sciences, Tarbiat Modares
University (TMU), Tehran, Iran
lashgarisajedeh@gmil.com

Babak Teimourpour *
Department of Information
Technology Engineering, School
of Systems and Industrial
Engineering, Tarbiat Modares
University (TMU), Tehran, Iran
b.teimourpour@modares.ac.ir

Mostafa Akhavan-Safar
Department of Computer and
Information Technology
Engineering, Payame Noor
University (PNU), Tehran, Iran.
akhavansaffar@pnu.ac.ir

Received: 3 October 2022 – Revised: 1 December 2021 - Accepted: 9 February 2022

Abstract—**Cancer-causing genes are genes in which mutations cause the onset and spread of cancer. These genes are called driver genes or cancer-causal genes. Several computational methods have been proposed so far to find them.** Most **of these methods are based on the genome sequencing of cancer tissues. They look for key mutations in genome data to predict cancer genes. This study proposes a new approach called centrality maximization intersection, cMaxDriver, as a network-based tool for predicting cancer-causing genes in the human transcriptional regulatory network. In this approach, we used degree, closeness, and betweenness centralities, without using genome data. We first constructed three cancer transcriptional regulatory networks using gene expression data and regulatory interactions as benchmarks. We then calculated the three mentioned centralities for the genes in the network and considered the nodes with the highest values in each of the centralities as important genes in the network. Finally, we identified the nodes with the highest value between at least two centralities as cancer causal genes. We compared the results with eighteen previous computational and network-based methods. The results show that the proposed approach has improved the efficiency and F-measure, significantly. In addition, the cMaxDriver approach has identified unique cancer driver genes, which other methods cannot identify.**

Keywords: Cancer-causing genes;Transcriptional regulatory network; Maximization; Centrality; Intersection.

Article type: Research Article

Publisher: ICT Research Institute

[DOI: 10.52547/itrc.14.1.57]

* Corresponding Author

I. Introduction

A. *The Importance of Cancer-Causing Genes Discovery*

Many studies have been done on cancer-causing genes detection. These genes are known as cancer driver genes (CDGs). CDGs are genes in which mutations cause cancer. The basic idea behind these methods, known as computational and statistical methods, is that repeated mutations in specific genes cause cancer. Not all mutations in a gene lead to cancer. As a result, in these methods, the detection and differentiation of cancer-causing mutations from normal mutations are essential for the identification of cancer genes. Existing methods for detecting CDGs rely heavily on genomic and transcriptomic data. Existing methods can be divided into three categories: computational-based, subnetwork-based and network-based. Computational methods using mutation data and transcriptomic data try to calculate the mutation frequency rate in genes. CoMDP [1], ActiveDriver [2], e-Driver [3], Simon [4], Oncodrive-Fm [5], OncodriverCLUST [6], Dendrix [7], iPAC [8] and MutSigCV [9] are among the computational methods. For example, Simon [4] calculates the effect of mutant function on proteins to find cancer-causing genes. OncodriveFM [5] and OncodriveCLUST [6] are approaches that categorize cancer-causing genes by evaluating the effect of cancer genome types on proteins. Dendrix [7], CoMDP [1] use mutation profiles to identify cancer signaling pathways. MutsigCV [9] uses exome sequences to detect heterogeneity in the cancer dataset and then identifies cancer-causing genes based on the frequency of mutations in different cancer. iPAC [8] also uses a combination of gene expression data and mutation data to identify cancer-causing genes. The ActiveDriver [2] uses information about changed post-transcription sites of proteins in mutant cancer genomes to identify cancer-causing genes. The e-Driver [3] method also tries to find the rate of biased mutations in the functional regions of a protein. Another group of methods for identifying cancer genes is known as sub-network methods. These methods are similar to network methods based on mutation data, but have also used part of the network structure. For example NetBox [10], DawnRank [11], MSEA [12], MeMo [13] and DriverNet [14] are among the sub-network methods. For example, DawnRank attempts to find cancer-causing genes using mutation and transcription data along with molecular interaction network information. Similarly, the NetBox [10] method finds cancer-causing genes from both protein-protein interactions and signaling pathways by finding cancer communities. The third category is network-based methods, which do not use mutation data and only use network structure analysis to identify driver genes. For example, iMaxDriver-N and iMaxDriver-W [15] are two network-based approaches that attempt to identify cancer-causing genes by influence maximizing approach. The characteristics of the methods compared to the proposed method in this study are shown in Table 1.

These methods have some limitations and shortcomings as follows:

- Most of these methods have a high rate of false positives in the results, which results in a decrease in precision and the F-measure.
- In addition, these methods rely heavily on mutation data. This data is naturally accompanied by noise and error. In addition, they may not always be available in the desired quality.
- Most of the genes identified by each of these methods overlap with the set of cancer-causing genes in other methods and are abundantly detected as unique cancer-causing genes.
- Some of these methods, such as the iMaxDriver approaches, are very time-consuming.

According to the limitations of existing methods, in this study, we proposed cMaxDriver as a new network-based approach to predicting cancer-causing genes. This approach identifies cancer-causing genes by analyzing the structure of the transcriptional regulatory network, without using mutation data. cMaxDriver uses an independent source of information. Transcriptional regulatory network (TRN) is one of the basic networks for controlling cellular processes. Transcription factors (TF) are key components of the cell and affect other genes, regulating their expression. In the other words, a transcriptional regulatory network shows how each transcription factor regulates the expression of other transcription factors and genes. Many diseases, including cancer, are caused by abnormalities in the function of transcription factors. This shows the importance of analyzing the structure of these networks in biomedical research.

In this study, a network-based approach called cMaxDriver was proposed to find cancer-causing genes. This approach uses degree, closeness, and betweenness centralities in the human transcriptional regulatory network. The results showed that cMaxDriver is able to improve the prediction precision of previous methods. In addition, cMaxDriver detects genes that other previous methods could not detect. Therefore, it can be used as a complementary method to other existing computational tools. The results show the proposed method performs better than many existing computational and network-based approaches.

B. *Theoretical Foundations*

Network centrality is a concept that is widely used in social network analysis to find the position and importance of each node in terms of communication with other nodes [16, 17]. Using centralities, noisy data from the network are reduced. In addition, the most important parts of the network are represented using them. There are different types of centralities, each with various definitions. Here, the three centralities of degree, closeness, and betweenness are used, which are defined below.

- Degree centrality: obtained by using the number of adjacent edges of a node through formula (1) [18].
- Closeness centrality: In connected networks, using the inverse calculation, the shortest path distance of each node from other nodes is obtained in the form of the formula (2) [18].

[DOI: 10.52547/itrc.14.1.57]

- Betweenness centrality: In connected networks, this centrality for each node is calculated by using the number of the shortest paths that pass through that node, as formula (3) [18].

$$C_D(p_k) = \sum_{i=1}^{n} a(p_i, p_k) \qquad (1)$$

$$C_c(p_k) = \left(\sum_{i=1}^{n} d(p_i, p_k)\right)^{-1} \qquad (2)$$

$$C_B(p_k) = \sum_{s \neq p_k \neq t} \frac{\sigma_{st}(p_k)}{\sigma_{st}} \qquad (3)$$

Where:

p represents the node of the network.

$a(p_i, p_k)$ is the value of the entry (i,j) of the adjacency matrix a. In other words the $a(p_i, p_k)$=1 if there is a directed edge from p_i to p_k are in the network and $a(p_i, p_k)$=0 otherwise.

$d(p_i, p_k)$ is the shortest path distance from p_i to p_k in the network.

σ_{st} is the total number of shortest paths from node s to node t.

$\sigma_{st}(p_k)$ is the number of paths that pass through the p_k node.

TABLE I. THE DETAILS OF COMPUTATIONAL AND NETWORK-BASED METHODS USED FOR COMPARISON.

Method name	Mutation data	Expression data	Network structure	Methodology
MeMo	✓	-	✓	correlation analysis and statistical tests
NetBox	✓	-	✓	sequence mutations and DNA copy number analysis
OncodriveCLUST	✓	-	-	clustering using mutations assessment
MDPFinder	✓	✓	-	Mutual exclusivity of gene modules
OncodriveFM	✓	-	-	The effect of mutation on genes
DriverML	✓	✓	-	machine learning approach
DawnRank	✓	✓	✓	The effect of downstream expression in molecular interaction networks
MeMo	✓	-	✓	gene correlation and statistical tests
Simon	✓	-	-	impact of mutations on proteins
Dendrix	✓	-	-	Classification of mutations by coverage and exclusivity
ActiveDriver	✓	-	-	identifies protein phosphorylation signaling sites
e-Driver	✓	-	-	Protein mutation rates by binomial test
MutsigCV	✓	✓	-	Calculation of mutations frequency
iPAC	✓	✓	-	Statistical methods
DriverNet	✓	-	✓	Effect of mutations on miRNA network
MSEA	✓	-	✓	combination of data associated with the disease development
iMaxDriver-N	-	✓	✓	Influence maximization approach
iMaxDriver-W	-	✓	✓	Influence maximization approach

II. METHODOLOGY

In this section, the cMaxDriver pipeline is described. It consists of three different steps:

1) Network construction

2) Cancer gene search algorithm based on the proposed model

3) Evaluation of results based on the existing gold standard

A. The Study Network

A gene is a specific region of a DNA[2] molecule of a specified length. Genes are found in every cell and carry the information needed to produce proteins, and by expressing these genes, different proteins are produced. Control of these processes plays a key role in determining the proteins present in the cell and their amounts [19]. That is a process that involves transcription on an RNA[3] molecule to translation into mRNA[4], which eventually leads to the production of new proteins. This process has a great effect on the rate of protein production. A transcriptional regulatory network is a type of biological network that comprises transcription factors and different

[2] Deoxyribonucleic Acid

[3] Ribonucleic Acid

[4] Messenger Ribonucleic Acid

[DOI: 10.52547/itrc.14.1.57]

[DOI: 10.52547/itrc.14.1.57]

genes and their interactions. The analysis of these networks is useful for examining the flow of information in a biological system and identifying different paths [20]. There are two types of modules in this network, gene module, and transcription factor module. In the first type, several genes are all regulated by one transcription factor, and in the second type, there are several transcription factors that all regulate common genes (see Fig. 1).

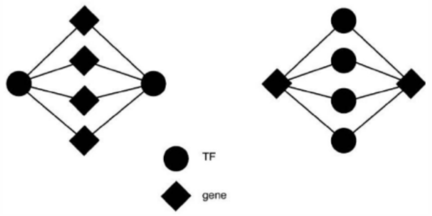

Fig. 1. Types of modules in the gene regulatory network [21].

B. *Network Construction*

Gene expression and regulatory interactions are needed to construct the study cancer networks. We used the RegNetwork[5] database [22], which is freely available[6], to obtain regulatory interactions. In this database, a list of gene regulatory interactions has been collected from various methods and multiple databases. It should be noted that RegNetwork, besides transcriptional interactions, also has regulatory interactions of microRNAs that have been omitted in this study. The retrieved dataset included 150,202 regulatory interactions between gene-TFs and TF-TFs. We also downloaded gene expression of three cancers: Breast (GSE15852), Colon (GSE32323), and Lung (GSE3268) from the GEO[7] database. In this database, gene expression data related to cancerous tissue and its adjacent normal tissue were reported for 10 patients. After initial processing, first, we deleted rows with missing gene names. Some rows had more than one gene name, which was separated. Finally, we computed the average expression values of rows that have the same gene name. Eventually, a file was obtained in which each row belonged to a unique gene and its expression values. Then we constructed separately, the regulatory network for breast, colon, and lung cancer using its gene expression data and regulatory interactions. In this way, for each network, the final list of gene expression values was mapped with the list of regulatory interactions. Thus, if a regulatory interaction of both origin and destination contained gene expression values, it was retained in the network and otherwise removed.

C. *Network Features*

The primary regulatory network for three types of cancer was constructed using the approach described in Section 2-2. These networks were disconnected and to analyze them in most cases, it is necessary to be connected. Therefore, we first converted the networks to connected networks and then performed the necessary analyzes. To do this, we used the largest weakly connected component. For example, in the lung cancer network, we had 11016 nodes, 87388 edges, 2 weakly connected components, and 9997 strongly connected components that the resulting network of lung cancer was constructed using the largest weakly connected component and the number of nodes and edges was 11015 and 87387, respectively. Information about the other two networks is also shown in Table 1. Also, the resulting networks are of directional and connected types. For example, the general view of the lung cancer network is illustrated using the force-directed algorithm in Fig. 2.

To identify and study the network more accurately, we calculated and examined the structural features of the networks. As shown in Fig. 3, the distribution and structure of the networks are more similar to a Scale-Free network.

According to the distribution of the network, it is expected that when a new node is added to the network, it will be connected to nodes with the highest connection and degree. Therefore, these features are investigated using more accurate indicators in the proposed algorithm described in the next section.

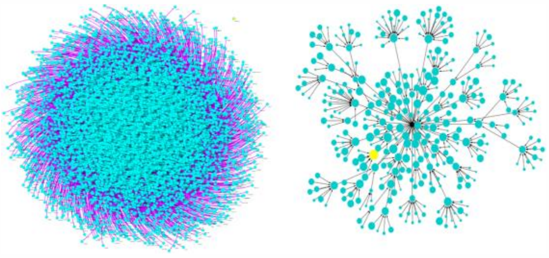

Fig. 2. From left to right: The primary disconnected lung cancer network using the force-directed algorithm, the disconnected component distinguished by its yellow color. Communities from the same network with the Louvain algorithm to better understand the schema and communications, this network has 349 communities, the largest consisting of 814 nodes and distinguished from the rest of the communities in yellow.

TABLE II. STRUCTURE INFORMATION OF THE RESULTING NETWORKS

Network type	Number of Nodes, Edges	Number of Strongly/ Weakly connected components	Criterion	Intended network	Random network
Breast cancer	10882, 86380	9870, 2	Average shortest distance	0.298	3.362
			Average clustering coefficient	0.222	1.459×10^{-3}
Colon cancer	15664, 117897	14517, 2	Average shortest distance	0.240	3.562
			Average clustering coefficient	0.214	9.610×10^{-4}
Lung cancer	11015, 87387	9997, 2	Average shortest distance	0.297	3.367
			Average clustering coefficient	0.223	1.440×10^{-3}

[5] Regulatory Network Repository

[6] http://www.regnetworkweb.org/

[7] Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/)

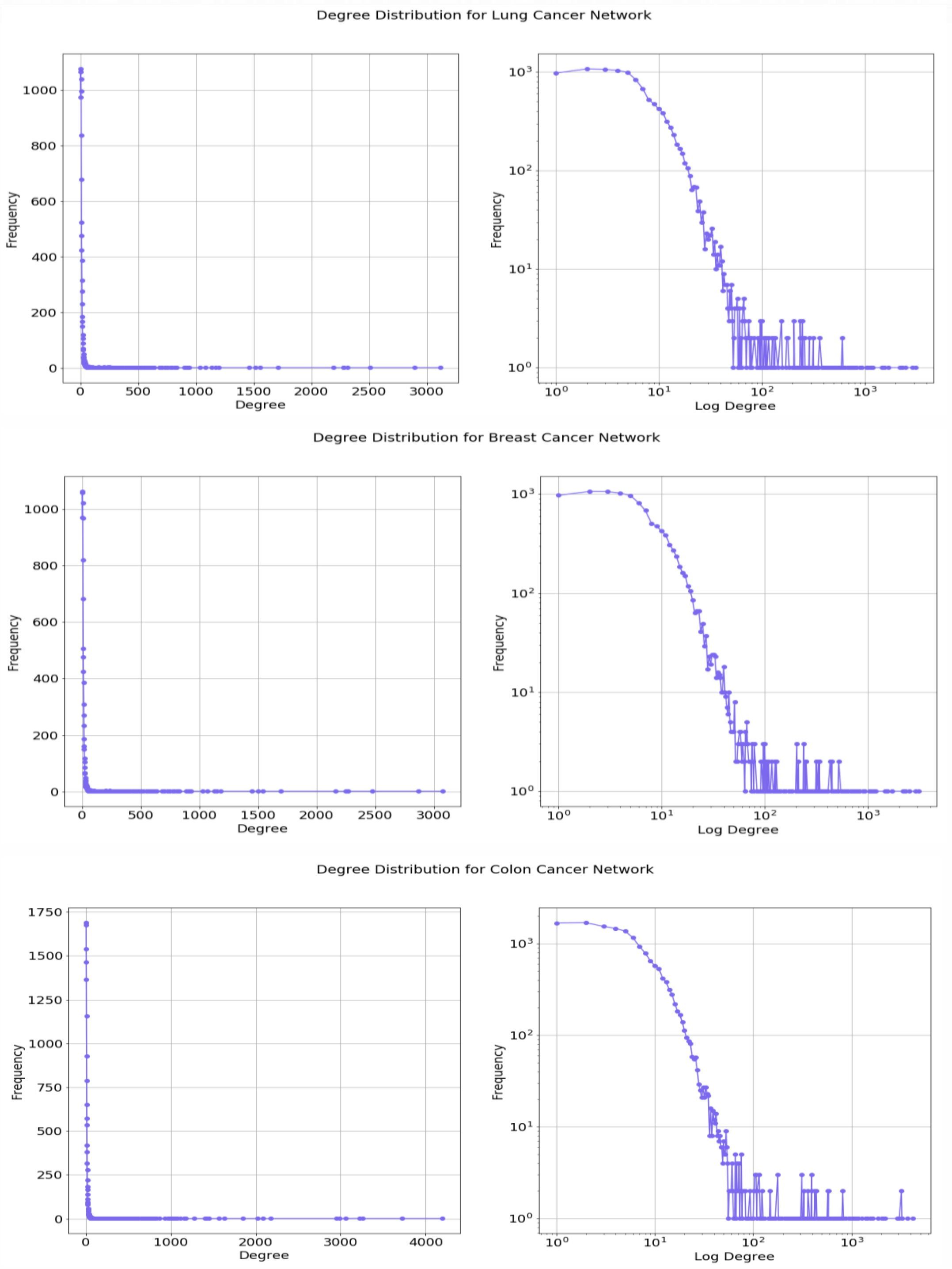

Fig. 3. From left to right: Degree distribution plot and distribution of network degrees in Log-Log scale (by ignoring the first and last nodes, a linear function is observed).

III. cMaxDriver: Centrality Maximization Intersection Proposed Approach

As mentioned earlier, cancer occurs because of abnormalities in some genes and their spread to other genes in the cell regulatory network. Thus, more important genes in network structure are more likely to be classified as driver genes. We proposed a new approach called centrality maximization intersection to predict cancer-causing genes. This algorithm tries to find a subset of nodes that are shared between at least two centralities. These sets of nodes are communities of genes that satisfy at least the following two conditions and indicate that they are more important in the network. This means that if a mutation occurs in them, it will affect a larger number of genes. Therefore, we considered these nodes as cancer-causing genes (drivers).

These conditions were defined as follows:

- They are more closely related to other genes. (Because of the greater degree centrality)

[DOI: 10.52547/itrc.14.1.57]

- They are shorter in distance from other genes and cost less to travel. (Because of the greater closeness centrality)
- They are on the path to more genes, so they affect a lot of genes. (Because of the greater betweenness centrality)

The proposed cMaxDriver algorithm based on the defined conditions comprises seven steps:

- Step 1- Calculation of degree, closeness, and betweenness centralities for all network genes.
- Step 2- Calculate the average of all three centralities.
- Step 3- Define the threshold value for each centrality according to the average value of each.
- Step 4- Find genes that have a value greater than the threshold, separately at each centrality.
- Step 5- Find and separate the data from step 4 that are common to at least two centralities.
- Step 6- Calculate the union of intersection data obtained in step 5.
- Step 7- Delete duplicate data.

The threshold values required in step 3 of the proposed algorithm for each centrality are obtained as described in Section 3-1.

The selected threshold values and other networks information are shown in Table 3. In the lung cancer network, for example, the MYC gene is known as an important node using the cMaxDriver algorithm, because its value in the two centralities has a value greater than the threshold associated with the relevant centrality. This is important, meaning that if a mutation (meaning a cancer-causing mutation) occurs in it, it will have a major impact on other things. Therefore, this gene is considered a driver gene in lung cancer.

TABLE III. INFORMATION ABOUT NETWORK CENTRALITIES.

Network Type	Criterion	Degree centrality	Closeness centrality	Betweenness centrality
Breast cancer	Minimum value	9.190×10^{-5}	0	0
	Average value	0.001	0.030	1.890×10^{-5}
	Maximum value	0.283	0.540	0.009
	The node corresponding to the maximum value	MAX	MYC	MYC
	Threshold	1.019×10^{-3}	2.547×10^{-1}	8.903×10^{-6}
Colon cancer	Minimum value	6.384×10^{-5}	0	0
	Average value	9.611×10^{-4}	0.023	1.068×10^{-5}
	Maximum value	0.268	0.529	0.007
	The node corresponding to the maximum value	MAX	SP1	SP1
	Threshold	1.011×10^{-3}	0.247	3.006×10^{-4}
Lung cancer	Minimum value	9.079×10^{-5}	0	0
	Average value	1.441×10^{-3}	0.029	1.858×10^{-5}
	Maximum value	0.283	0.540	0.009
	The node corresponding to the maximum value	MAX	MYC	MYC
	Threshold	1.491×10^{-3}	1.953×10^{-2}	1.458×10^{-5}

A. *Threshold tuning*

To select the best threshold, two stages were performed, in both of which the criterion for optimization was F-measure. In the first stage, five statistical indicators of minimum, average, median, mode, and maximum are used as threshold values. These five indicators summarize the information of all nodes in one value.

The best criterion among these five indicators for all three networks was the average. The best criterion is the index by which F-measure is maximized compared to the rest.

For example, Fig. 4 shows performance compared to different indices of the three centralities for the breast cancer network.

As shown in Fig. 4, the model is sensitive to the threshold value, and as the threshold changes, the performance of the model changes. Fig. 4 shows that the best index is 62 because it has the highest F-measure.

Considering the three rings that are considered for each of the centralities and each of which has five defined values, index 62 is related to the values of 0.001, 0.030, and 1.890 × 10-5, which are related to the

[DOI: 10.52547/itrc.14.1.57]

[DOI: 10.52547/itrc.14.1.57]

average of the three centralities of degree, closeness, betweenness, respectively.

In the second stage, based on the output of the first stage (namely the average of each centrality), a search interval is found experimentally to find the improved values for the threshold. The results of this stage are shown in Fig. 5 for the breast cancer network.

Finally, the values that had the highest F-measure value at this stage are considered as the final threshold values, which are expressed in Table 2 as rounded values to three decimal places.

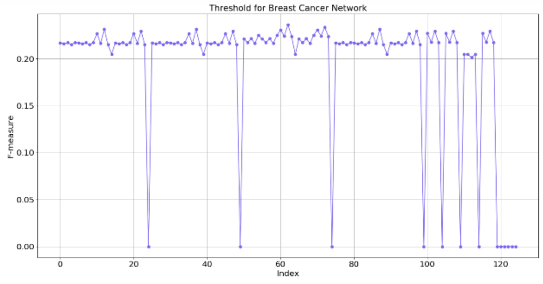

Fig. 4. Model performance at different threshold values in the first stage.

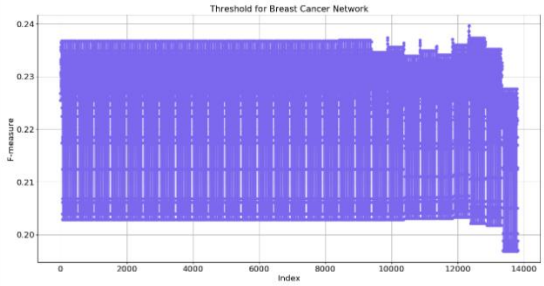

Fig. 5. Model performance at different threshold values in the second stage.

IV. Evaluation of The Proposed Algorithm

As described in Section 3, the algorithm in step 7 considers common unique nodes as cancer-causing genes. The algorithm can finally extract the list of genes as drivers based on the described approach. Then labeling is done for the actual and predicted data. So that the data that truly cause cancer and output data of step 7 of the algorithm, that means predicted driver, are labeled with 1 (negative: driver), and the rest of the data with 0 (positive: normal). We compared the results of the algorithm with eighteen previous computational and network-based methods. To obtain the results of the previous methods, we used the DriverDBv2 database [23]. In this database, the results of each cancer are reported based on the TCGA[8] dataset for each method. We also used a set of validated cancer-causing genes introduced by TCGA [24] to evaluate the results. TCGA is an evaluation database used in many bioinformatics studies such as [15, 20, and 21]. In this database, datasets are available[9] for three breast, colon, and lung cancers named TCGA-BRCA [10], TCGA-COAD [11], and TCGA-LUSC [12], respectively. That the number of genes as drivers introduced, the same order is 572, 572, and 566 for three different types of cancers.

To evaluate cMaxDriver, we used the criteria ^{of} precision, recall, accuracy, and F-measure that are common in binary classification approaches. The F-measure is a common and good criterion for evaluating classifiers, which obtained the percentage of correct positive predictions by calculating the harmonic mean of the two criteria of precision and recall, which is defined as follows:

$$\mathrm{F-measure} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \qquad (4)$$

While precision and recall are defined as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad (5)$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad (6)$$

Accuracy is also calculated by Equation (7).

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \qquad (7)$$

Details of the evaluation of the methods used are shown in Tables 3 and 4. For example, in breast cancer, cMaxDriver had 10249 true positives (TP) and 164 true negatives (TN), i.e. it correctly identified 164 genes as a driver and 10249 genes as

[8] The Cancer Genome Atlas

[9] https://cancer.sanger.ac.uk/census

[10] Breast invasive carcinoma

[11] Colon adenocarcinoma

[12] Lung squamous cell carcinoma

normal in breast cancer. It also has 408 false positives (FP) and 633 false negatives (FN), which are related to type I and type II errors, respectively. Which shows the number of normal genes that are incorrectly predicted as drivers by this algorithm, and the number of driver genes that are incorrectly predicted as normal by this algorithm, respectively. The complete results are shown in Table 3.

Considering the condition of at least two intersections will give better results than the intersection between all three centralities because the intersection between all three centralities imposes more restrictions on the data, so the results have less bias but more variance. Therefore, as shown in step 5, the proposed method uses the condition of at least two intersections. The results of running the cMaxDriver algorithm on three cancer networks are shown in Table 4. Comparing the results of cMaxDriver with other methods is given in Section 5.

TABLE IV. THE CONFUSION MATRIXES OF CMAXDRIVER

Network Type	Criterion			
	TP	TN	FP	FN
Breast cancer	10249	164	408	633
Colon cancer	15074	158	414	590
Lung cancer	10449	153	413	566

TABLE V. VALUES OBTAINED FROM CMAXDRIVER EVALUATION.

Network Type	Criterion			
	Accuracy	Precision	Recall	F-measure
Breast cancer	0.909	0.206	0.287	0.240
Colon cancer	0.938	0.211	0.276	0.239
Lung cancer	0.915	0.213	0.270	0.238

V. RESULTS

The proposed algorithm was run on three cancer networks. Then, based on the threshold values introduced in Section 3, the genes were classified into two classes: driver and normal. Afterward, using the performance criteria introduced in Section 4, we compared cMaxDriver with eighteen previous computational and network-based methods. The corresponding results for breast cancer are shown in Fig. 6. As seen, cMaxDriver with Recall = 0.287 is ranked first among network-based methods and ranked second among all computational and network-based methods. Although some computational methods have higher precision and recall they are not in a good position in terms of the F-measure and the number of cancer-casual genes they predict.

As mentioned, precision and recall alone cannot show the performance of a classification system. Therefore, the harmonic mean of these two criteria is used. As shown in the results, cMaxDriver with F-measure = 0.24 has the highest value among all computational and network-based methods and has significantly improved performance.

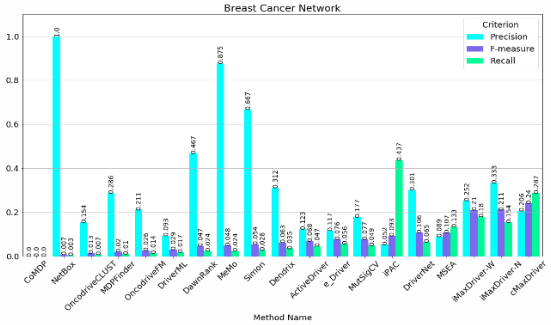

Fig. 6. Comparison of evaluation criteria of the cMaxDriver and other methods in breast cancer.

We also compared the cMaxDriver and other methods based on the number of driver genes predicted. The results are shown in Fig. 9. As seen, cMaxDriver reached ranks first among previous network-based methods and second among all methods by identifying 164 genes in breast cancer.

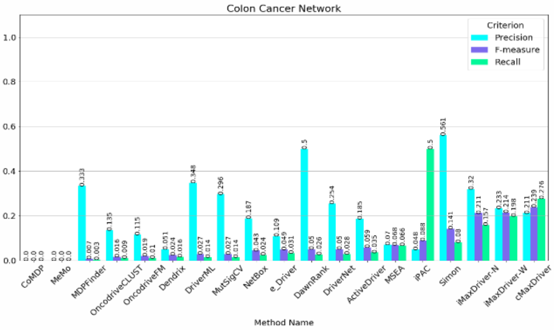

Fig. 7. Comparison of evaluation criteria of the cMaxDriver and other methods in colon cancer.

The results of cMaxDriver and other methods in colon cancer are shown in Fig. 7. As seen, cMaxDriver with F-measure =

0.239 has the highest value among all computational and network-based methods and has significantly improved performance. Also, in terms of the number of drivers detected, as shown in Fig. 9, cMaxDriver, with 158 drivers detected for colon cancer, ranks first among previous network-based methods and second among all methods.

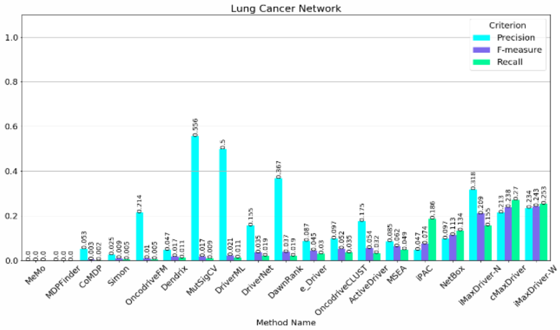

Fig. 8. Comparison of evaluation criteria of the cMaxDriver and other methods in lung cancer.

Also, based on Fig. 8 be seen, cMaxDriver for lung cancer by F-measure = 0.238 with a slight difference after iMaxDriver-W among all computational and network-based methods is ranked second, but based on Recall = 0.27, it's ranked first among all methods. In addition, based on the number of drivers detected, as seen in results are shown in Fig. 9, cMaxDriver reached ranked first among all computational and network-based methods by identifying 153 genes in lung cancer.

We also compared the overlap of genes identified by cMaxDriver and other methods. The results are shown as a Venn diagram in Fig. 10. As seen in Fig. 10, cMaxDriver was able to cover 140, 143, and 135 genes identified by other computational and network-based methods in breast, Colon, and lung cancer, respectively. In addition, it has identified 24, 15, and 18 unique genes in breast, colon, and lung cancer that have not been identified by any of the previous computational and network-based methods.

Also, compared to previous network-based methods, cMaxDriver identified 123, 123, and 129 cancer-causing genes detected by other network-based methods in the same order. In addition, cMaxDriver similarly identified 41, 35, and 24 unique genes in the three named cancers that were not detected by other network methods. Lists of unique cancer-causing genes correctly identified by cMaxDriver is given in Tables 5.

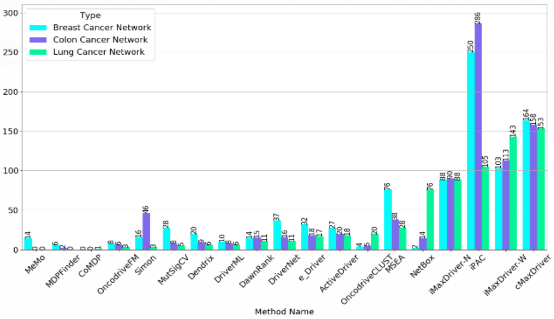

Fig. 9. Comparison of the number of cancer-causing genes identified by cMaxDriver and other methods in three types of cancer.

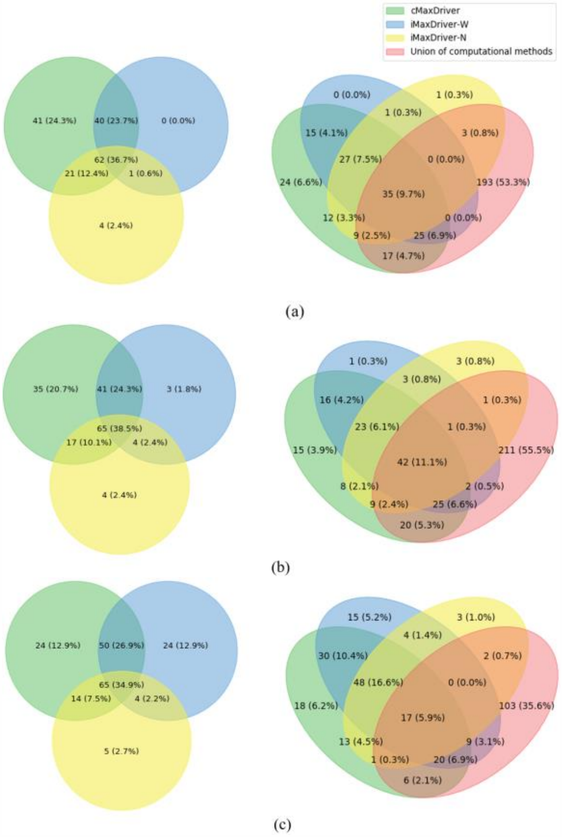

Fig. 10. Left to right: Overlap of genes identified by cMaxDriver with other network-based methods and with other methods. (a), (b), and (c) are related to breast, colon, and lung cancers, respectively.

The main aim of the study was not to examine time complexity. The primary objective was to improve the performance criteria and the number of identified driver genes. In addition the time complexity of the previous methods is not mentioned. However, program run time of our proposed method (step 1 to 7) in a system with CPU intel core i5 and Ram 8 are as follows: breast cancer=62ms, colon cancer= 61ms and lung cancer =75ms. These values do not include the time to set the optimal threshold.

VI. Conclusion

In this study, an algorithm called cMaxDriver was proposed to classification and detection cancer-causing genes in transcriptional regulatory networks. One of the advantages of proposed method is that it does not depend on mutation and genomic data. And

[DOI: 10.52547/itrc.14.1.57]

identifies driver genes only using the structure of gene interactions and the network approach. It is also much faster than previous networking methods. The cMaxDriver has the best performance compared to other computational and network-based methods in terms of f-measure and the number of diagnostic drivers. First, three cancer regulatory networks were constructed using gene expression and regulatory interactions data. Then the different steps of the algorithm were performed on the networks, as described in Section 3. Finally, genes were classified into cancer-casual and normal based on defined threshold values. The results were compared with eighteen computational and network-based methods in terms of efficiency criteria and the number of identified cancer-casual genes. The results in terms of efficiency and F-measure ranked first among all methods of detecting breast and colon cancer and second in terms of identifying the number of cancer-casual genes. Also, for lung cancer, the proposed algorithm ranks first among all computational and network-based methods in terms of the number of cancer-causing genes detected and second in terms of performance with a slight difference from the first rank. In addition, the proposed approach, while identifying a significant number of diagnostic genes by other methods, can identify genes that have not been identified by any of the other methods.

Data availability:

Data is available publicly at https://github.com/MASafar/cMaxDriver.

TABLE VI. THE LIST OF UNIQUE CANCER-CAUSING GENES PRODUCED BY CMAXDRIVER.

Breast Cancer Network		Colon Cancer Network		Lung Cancer Network	
Compared to all methods	Compared to network-based methods	Compared to all methods	Compared to network-based methods	Compared to all methods	Compared to network-based methods
SMARCB1	NONO	DDB2	SMARCB1	LMO1	ETV6
MLLT1	SMARCB1	IKZF1	CHD4	MLLT3	MLLT1
ETV6	MLLT1	SMARCE1	FOXP1	MLLT10	BCL11B
TAL1	CHD4	BTG1	APC	BCL11B	HOXC13
ERCC2	ETV6	BCL11B	SMARCE1	HOXC13	ETV1
FUBP1	TAL1	TRIM33	BCL11B	TRIM33	PAX3
SMARCE1	ERCC2	KDM5A	HOXC13	ETV6	TNFAIP3
MDM4	FUBP1	HMGA2	ETV1	MLLT1	AFF1
OLIG2	SMARCE1	XPC	PAX3	CIC	ERCC3
ETV1	PSIP1	POU2AF1	ZFHX3	DEK	SMARCD1
TNFAIP3	ZMYM2	NR4A3	HOXD13	ELL	MLLT10
SMARCD1	MDM4	HOXA13	ERCC3	AFF1	TRIM33
KAT6A	OLIG2	ETV5	ARID1B	ERCC3	NSD1
MAFB	ETV1	SRSF3	SMARCD1	ETV5	ELL
BTG1	TNFAIP3	SMARCD1	IKZF1	XPC	ETV5
ELL	HOXD13		BTG1	NFIB	XPC
ETV5	AFF1		TRIM33	TFE3	NFIB
DEK	ERCC3		HMGA2	SMARCD1	CIC
LMO1	PBRM1		PRRX1		TFE3
MLLT3	KAT6A		POU2AF1		DEK
DDB2	SMARCD1		NSD1		ATRX
NR4A3	MLLT10		ELL		LMO1
ELF4	MAFB		HOXA13		MLLT3
BCOR	BTG1		ETV5		MED12
	FUS		XPC		
	PRRX1		SRSF3		
	NSD1		NFIB		
	ELL		CIC		
	ETV5		DEK		
	XPC		ATRX		
	CIC		MLLT3		
	DEK		DDB2		
	ATRX		KAT6B		
	LMO1		KDM5A		
	MLLT3		NR4A3		
	DDB2				
	KAT6B				
	MED12				
	NR4A3				
	ELF4				
	BCOR				

[DOI: 10.52547/itrc.14.1.57]

[DOI: 10.52547/itrc.14.1.57]

References

[1] Zhang, J., Wu, L.-Y., Zhang, X. and Zhang, S. (2014), "Discovery of co-occurring driver pathways in cancer", *BMC Bioinformatics*, 15(1), 271. doi: 10.1186/1471-2105-15-271.

[2] Reimand, J., Wagih, O. and Bader, G.D. (2013), "The mutational landscape of phosphorylation signaling in cancer", *Scientific Reports*, 3, 2651.

[3] Porta-Pardo, E. and Godzik, A. (2014), "e-Driver: a novel method to identify protein regions driving cancer", *Bioinformatics*, 30(21), 3109–3114. doi: 10.1093/bioinformatics/btu499.

[4] Youn, A. and Simon, R. (2011), "Identifying cancer driver genes in tumor genome sequencing studies", *Bioinformatics*, 27(2), 175–181. doi: 10.1093/bioinformatics/btq630.

[5] Gonzalez-Perez, A. and Lopez-Bigas, N. (2012), "Functional impact bias reveals cancer drivers", *Nucleic Acids Res*, 40(21), e169. doi: 10.1093/nar/gks743.

[6] Tamborero, D., Gonzalez-Perez, A. and Lopez-Bigas, N. (2013), "OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes", *Bioinformatics*, 29(18), 2238–2244. doi: 10.1093/bioinformatics/btt395.

[7] Vandin, F., Upfal, E. and Raphael, B.J. (2011), "De novo discovery of mutated driver pathways in cancer", *Genome Research*, 22(2), 375–385. doi: 10.1101/gr.120477.111.

[8] Aure, M.R., Steinfeld, I., Baumbusch, L.O., Liestøl, K. and et al. (2013), "Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data", *Europe PMC*, 8(1). doi: 10.1371/journal.pone.0053014.

[9] [9] Lawrence, M., Stojanov, P., Polak, P. and et al. (2013), "Mutational heterogeneity in cancer and the search for new cancer-associated genes", *Nature*, 499(7457), 214–218. doi: 10.1038/nature12213.

[10] [10] Cerami, E., Demir, E., Schultz, N., Taylor, B.S. and Sander, C. (2010), "Automated Network Analysis Identifies Core Pathways in Glioblastoma", *PLoS ONE*, 5(2), e8918. doi: 10.1371/journal.pone.0008918.

[11] Hou, J.P. and Ma, J. (2014), "DawnRank: discovering personalized driver genes in cancer", *Genome Medicine*, 6(7), 56. doi: 10.1186/s13073-014-0056-8.

[12] Arneson, D., Bhattacharya, A., Shu, L., Mäkinen, V.-P. And Yang, X. (2016), "Mergeomics: a web server for identifying pathological pathways, networks, and key regulators via multidimensional data integration", *BMC Genomics*, 17, 722.

[13] G. Ciriello, E. Cerami, C. Sander, N. Schultz. (2012), Mutual exclusivity analysis identifies oncogenic network modules, Genome Res. 22 (2) 398–406.

[14] A. Bashashati, et al., (2012). DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer, Genome Biol. 13 (12).

[15] Rahimi, M., Teimourpour, B. and Marashi, S.-A. (2019), "Cancer driver gene discovery in transcriptional regulatory networks using influence maximization approach", *Computers in Biology and Medicine*, 114, 103362. doi: 10.1016/j.compbiomed.2019.103362.

[16] Freeman, L.C. (1978–1979), "Centrality in Social Networks Conceptual Clarification", *Social Networks*, 1(3), 215-239. doi: 10.1016/0378-8733(78)90021-7.

[17] Nieminen, J. (1974), "On the centrality in a graph", *Scandinavian Journal of Psychology*, 15(1), 332–336. doi: 10.1111/j.1467-9450.1974.tb00598.x.

[18] Wasserman, S. and Faust, K. (1994), "Social Network Analysis: Methods and Applications", 8, Cambridge: *Cambridge University Press*. doi: 10.1017/CBO9780511815478.

[19] Akhavan-Safar, M., Teimourpour, B. and Kargari, M. (2021), "GenHITS: A network science approach to driver gene detection in human regulatory network using gene's influence evaluation", *Journal of Biomedical Informatics*, 114(2), 103661. doi: 10.1016/j.jbi.2020.103661

[20] Akhavan-Safar, M. and Teimourpour, B. (2021), "KatzDriver: A network based method to cancer causal genes discovery in gene regulatory network", *Biosystems*, 201, 104326. doi: 10.1016/j.biosystems.2020.104326.

[21] MacNeil, L.T. and Walhout, A.J.M. (2011), "Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression", *Genome Research*, 21(5), 645-657. doi: 10.1101/gr.097378.109.

[22] Liu, Z.P., Wu, C., Miao, H. and Wu, H. (2015), "RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse", *Database: The Journal of Biological Databases and Curation*. doi: 10.1093/database/bav095.

[23] Chung, I.-F., Chen, C.-Y., Su, S-.C. And et al. (2016), "DriverDBv2: a database for human cancer driver gene research", *Nucleic Acids Research*, 44(D1), D975–D979. doi: 10.1093/nar/gkv1314.

[24] Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M. and et al. (2013), "The Cancer Genome Atlas Pan-Cancer analysis project", *Nature Genetics*, 45(10), 1113–1120. doi: 10.1038/ng.2764.

Sajedeh Lashgari is a M.Sc. student in Data Science at Tarbiat Modares University and she received her B.Sc. degree in Statistics from Imam Khomeini International University. Her research interests include Data Mining, Deep Learning, Graph Learning, and Social Network Analysis. And her favorite fields in scientific study include Neuroscience, Clean Energy, Strategy, and Genetics.

Babak Teimourpour obtained his Ph.D. in Industrial Engineering from Department of Industrial Engineering, Tarbiat Modares University (TMU), Tehran, Iran. He teaches Ph.D. and M.Sc. level courses. His research interests include Data Mining and Social Network Analysis. His team won the Iran Data Mining Cup in 2010.

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 (PNU). He received his M.Sc. in Information Technology Engineering form 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.

[DOI: 10.52547/itrc.14.1.57]