Journal of Algorithms and Computation

journal homepage: http://jac.ut.ac.ir

Colorectal cancer driver gene detection in human gene regulatory network using independent cascade diffusion model

Mostafa Akhavan-Safar*[1], Babak Teimourpour†[2] and Mahboube Ayyoubi‡[3]

[1]Department of Computer and Information Technology, Payame Noor University (PNU), Tehran, Iran.
[2]Department of Information Technology Engineering, School of Systems and Industrial Engineering, Tarbiat Modares University (TMU), Tehran, Iran
[3]Department of Data Science, Tarbiat Modares University (TMU), Tehran, Iran

ABSTRACT

ABSTRACT

One of the important topics in oncology for treatment and prevention is the identification of genes that initiate cancer in cells. These genes are known as cancer driver genes (CDG). Identifying driver genes is important both for a basic understanding of cancer and for helping to find new therapeutic goals or biomarkers. Several computational methods for finding cancer driver genes have been developed from genome data. However, most of these methods find key mutations in genomic data to predict cancer driver genes. These

ARTICLE INFO

ARTICLE INFO

Article history:
Research paper
Received 09, September 2022
Received in revised form 11, November 2022
Accepted 10 December 2022
Available online 30, December 2022

Keyword: Gene regulatory network, Driver genes, Influence maximization, cancer, Independent Cascade.

AMS subject Classification: 05C78.

*akhavansaffar@pnu.ac.ir
†Corresponding author: B.Teimourpour. Email: b.teimourpour@modares.ac.ir
‡m.aubi_68@yahoo.com

1 Abstract continued

methods are dependent on mutation and genomic data and often have a high rate of false positives in the results. In this study, we proposed a network-based method, GeneIC, which can detect cancer driver genes without the need for mutation data. In this method, the concept of influence maximization and the independent cascade model is used. First, a cancer gene regulatory network was created using regulatory interactions and gene expression data. Then we implemented an independent cascade propagation algorithm on the network to calculate the coverage of each gene. Finally, the genes with the highest coverage were introduced as driver genes. The results of our proposed method were compared with 19 previous computational and network methods based on F-measure metric and the number of detected drivers. The results showed that the proposed method has a better outcome than other methods. In addition, more than 25.49% of the driver genes reported by GeneIC are new driver genes that have not been reported by any other computational method.

2 Introduction

Cancer is one of the diseases that is caused by oncogene activations such as genetic mutations, chromosomal rearrangements, etc., [6] and [15]. The disease is the second leading cause of death in the world and about 6.9 million people lost their lives in 2018 due to this disease, i.e. one out of every 6 people [29]. Lung cancer (2.09 million cases), Breast cancer (2.09 million cases), and Colorectal cancer (1.80 million cases) are the most common cancers [29]. During tumor progression, most of the altered genes identified are passenger-type, these genes do not contribute to the oncogene process, but a small portion of the altered genes are known to be driver genes that disrupt normal transcriptional processes and change the cell from normal to cancerous.

3 Literature review

Many computational methods have been proposed to find cancer-causing genes. In these methods, it is assumed that the genes that cause cancer are genes that are more prone to major changes in gene expression (mutation). Not all mutations that occur in the cancer genome lead to cancer. Therefore, most computational methods try to distinguish driver mutations from non-driver mutations. Most available methods for identifying CDG depend on transcriptomic or genomic data. On codrive CLUST, [25] is one of the computational methods proposed by Tamborero et al in 2013. This method identifies genes that have a significant tendency to cluster mutations in protein sequences. OncodriveCLUST creates a model for classifying genes by evaluating coding-silent mutations. Simon [30] is another computational method. This method has been proposed to improve the identification of cancer driver genes by estimating

the background mutation rate and can use the operational effect of mutations on proteins, changes in background mutations among tumors, and genetic code redundancy, predicting cancer driver genes. One of the features of this method is to differentiate between mutations that affect protein function and other mutations. It can also differentiate between the number of background mutations in different samples and patients. This method has been proposed to improve the identification of cancer driver genes by estimating the background mutation rate and can use the operational effect of mutations on proteins, changes in background mutations among tumors, and genetic code redundancy, predicting cancer driver genes. One of the features of this method is that it distinguishes between mutations that affect protein function and other mutations, as well as the difference between the number of background mutations in samples and different individuals. One of the challenges in interpreting DNA data is to distinguish driver mutations from passenger mutations. Dendrix [27] is a computational method that combines two characteristics of coverage: finding genes in different patient samples and exclusivity, meaning mutations that are rarely seen in certain patients Attempts to separate driver mutations from passenger mutations. The ActiveDriver [23] method was proposed by Reimand et al in 2013. This method uses phosphorylated protein sites that mutate by changing only one nucleotide to analyze and find cancer genes. This method identifies signaling sites where the mutation rate is significantly higher than the level of mutation in the entire gene sequence and shows the importance of those sites in cancer biology. The e-Driver [21] method extracts the internal distribution of malignant mutations between functional regions of proteins to find the mutation rate compared to other regions of the same protein. If the observations are positive, those genes could be cancer drivers.

Oncodrive-FM [11] is another computational method based on mutation data. One of the major challenges in cancer genomics is the identification of driver genes and pathways between different types of mutations. In this method, to overcome the limitations of traditional approaches, such as the difficulty of accurately estimating the mutation rate, and relying on increasing changes, a new criterion called FM bias is defined that does not rely on recurrence. In this way, it detects the cancer driver genes. The MDPFinder [31] method is another computational method that uses both mutation data and gene expression data. MDPFinder tries to solve the maximum weight matrix problem [27]] designed to identify mutant driver paths. To do this, it uses a random approach (genetic algorithm) and a combined approach (integration of gene expression and mutation data) to find the path of cancer mutations and then find the genes that cause cancer. The DriverML [13] method is a computational method that uses supervised machine learning and the Rao test score to identify cancer driver genes. This method uses mutation data and expression data. The weight parameters in the statistical test determine the functional effect of the mutations on the protein. The MutsigCV [18] method also uses mutation and expression data. This method tries to detect abnormal changes in genes by solving the problem of heterogeneity in mutation processes and mutation frequency of genes, thus identifying cancer genes. By examining transcriptional activity and comparing the number of mutations occurring

in different types of cancer and the number of mutations occurring in the human gene, the gene is diagnosed as drivers. iPAC [3] is an unsupervised approach in which, based on integrative analysis of the number of copies and gene expression data, it systematically performs a series of statistical tests on a list of genes to extract the driver list of genes.

Another category of driver gene recognition methods is computational methods that use part of the structure of biological networks in addition to mutation and genomic data. These methods are a combination of mutation and network data. The Netbox [5] method uses integrated network analysis to identify network modules, which change frequently. To do this, the network was created using protein-protein interactions and signaling pathways. The identified network modules and network modularity are then calculated. Finally, the significance of modularity is statistically evaluated. DawnRank [14] is a computational method that uses mutation data and focuses on each patient's cancer genes to discover rare and specific genes for each patient as cancer genes. This method uses the personal information of only one patient to diagnose the cancer gene. DawnRank ranks mutated genes in a patient according to their potential for delivery in the molecular interaction network. Mutated genes with higher rankings are likely to be drivers. MeMo [8] performs a systematic review of the oncogenic pathway module and uses mutation data and network structure. It uses correlation analysis, statistical tests, and three criteria to identify modules in the network. 1) The genes in the tumor samples have been altered, 2) the genes of each tumor have participated in the same biological process, and 3) the changes in the tumor genes have occurred exclusively. The MSEA [2] method has been proposed to provide an overall and integrated view of the disease mechanism rather than a separate review of the data. They developed a computational pipeline that could integrate multidimensional disease-related data with biological functions and molecular networks to retrieve biological pathways and gene networks, and then identify cancer-causing genes. The DriverNet [4] method is a computational framework for identifying driver mutations through miRNA expression networks. In this method, through gene interactions, the relationship between aberrations in the genome and transcription patterns are extracted. This method also relies on mutation data. Another category of methods that has recently been identified in cancer driver genes is network-based and Bioinformatics methods. These methods does not rely on mutation and genomic data and only uses biological network structures to identify driver genes. iMaxDriver-N and iMaxDriver-N [22] Rahimi, Majid, Babak Teimourpour, and Sayed-Amir Marashi. "Cancer driver gene discovery in transcriptional regulatory networks using influence maximization approach." *Comput. Biol. Med.* 114 (2019), 103362. [22] methods are among these methods that identify driver genes using gene expression data and transcriptional regulation network structure. In these methods, the Influence maximization and the linear threshold model are used. GenHITS [1] is another network-based approach. In this method, using the hyperlink induced topic search algorithm and modifying it based on the concept of diffusion, cancer driver genes are identified.

The proposed methods for identifying driver genes have limitations. Computational

methods rely on mutation data, and due to the noise in these data, these methods often have high false positives in the results. Also, most of the genes identified by these methods have overlap. The previous network-based methods, although do not have some of the problems of computational methods, can still be improved in terms of the number of genes identified and performance criteria. Due to the limitations, in this study, a network-based method without relying on mutation data has been proposed to identify cancer driver genes. In this method, the concept of network diffusion and the independent cascade model is used to rank genes. In this method, the coverage of each gene in terms of propagation power in the gene regulatory network is calculated. The Gene Regulatory Network (GRN) is a collection of DNA fragments in a cell, which interact indirectly with each other (via RNA and the expression of their protein products) and with other molecular regulators in the cell. As a result, they determine which genes in the network are transcribed into mRNA.

4 Background

With the rapid spread of the Internet around the world, social networks have become very popular. Information is spread on social networks, ideas and knowledge are shared through social networks. Hence, many subject are studied through the analysis of social networks, such as models of diffusion and social influence. Different models have been studied to model the behavior of social networks. One of them is the influence maximization problem, in which we look for the minimum k nodes that maximize diffusion in a social network. These nodes represent the people who have the most effect on other people in the network if they are active in the network. One of the most popular methods for modeling the diffusion process is the cascade model, which is inspired by particle motion models in physics [19]. Goldenberg et al. First studied diffusion maximization using cascading models [10]. In cascading models, starting with the seed nodes, in each step t, the active node v can try once to activate one of its adjacent active nodes with a p_v probability. If successful, the newly activated nodes will be activated in the next step $(t + 1)$, and will perform the same operation to activate the inactive nodes. Whether the attempt is successful or not, active nodes cannot try twice to activate the same node. This process continues until it is no more possible to activate a new node.

4.1 Independent Cascade Model

The independent cascade model is one of the two main models for modeling the information diffusion process at the social network level. The independent cascade model is one of the two main models for modeling the information diffusion process in a social network. The roots of this model go back to particle motion models in physics [16]. In general, this model and its derivatives have been used to model the acceptance and use of new and effective things in the context of social networks. Independent cascade

model (IC) [17] Kempe, David, Jon Kleinberg, and Ava Tardos. "Influential nodes in a diffusion model for social networks." *In International Colloquium on Automata, Languages, and Programming* , pp. 1127-1138. Springer, Berlin, Heidelberg, 2005. [17] is the simplest type of cascade model in which the probability that the active node v activates the adjacent active node u, p_u (v), is a constant value, independent of the previous propagation process. It has been claimed in [17] that the number of attempts to activate the nodes does not affect the output. In this model, the information spread platform is a static directional graph $G = (V, E)$ where V and E are a set of nodes and edges of G, respectively. Each connection means the existence of a directional edge from node n to node x in this network is defined as $e = (n, x)$ where $n \neq x$. In this model, each positive edge e is first assigned a positive $p_n(x)$ number with the condition $0 < p_n(x) < 1$. $p_n(x)$ is also called diffusion probability (n, x). The diffusion process begins by selecting an initial set A(0) of the nodes of the network. Assuming that each node can be active or inactive in one of two states, the nodes in A(0) are assumed to be active, and at each time step $t = \{0, 1, \ldots .. , w\}$ An active node like n can activate any node of its inactive child like x with a probability of $p_n(x)$. If several parent nodes, such as n, are active in step t, the order in which n is likely to be activated by one of them is optional, without regard to any particular priority, and the only important thing is that activation for n by the parent nodes must be done in unison in step t. In this model, apart from activating or not activating the child node in step t, each parent node has the opportunity to activate its child node only once. The independent diffusion process ends when a node can no longer be activated. The activation function and set θ for the arbitrary edge e are as follows:

$$Y_1(e_i) = p_{e_i} \; ; \quad \theta = \{p_{nx} \; ; (n, x) \epsilon E\} \tag{1}$$

Using set θ, the function of the degree of effectiveness of each node or the number of child nodes of the node that we assume will be active in the next step can be defined as follows:

$$Y_2(n) = \mu(n, \theta) \tag{2}$$

Figure 1 shows the process of propagating an independent cascade model for a small network with 10 nodes and three primary active nodes.
So in general the IC model works as follows:
The IC model starts with an initial set of active nodes (seed). The diffusion process is revealed in a discrete process according to a random rule:

1. When node n becomes active in step t, it is given a single chance to activate each currently inactive neighbor x; it succeeds with a probability $p(n, x)$
2. If x has several newly active neighbors, their efforts will be sorted as desired.
3. If n succeeds, x is activated in step t + 1. But whether v succeeds or not, it cannot make further effort to activate w in subsequent rounds.
4. This process runs until no more activation are possible

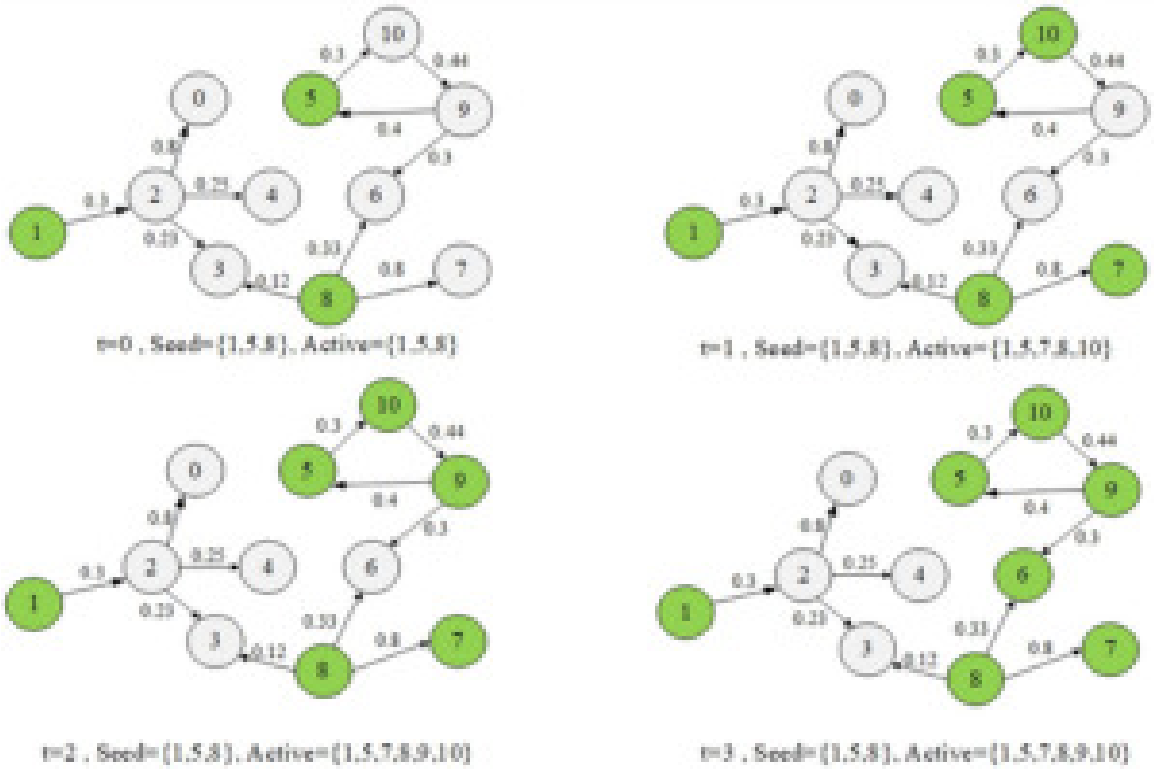


Figure 1:The IC model example in a network with 10 nodes and 11 edges. Active nodes in each step are shown in green. Nodes 1, 5 and 7 are the initial active nodes.

$$Y_1(e_i) = p_{e_i}\ ; \quad \theta = \{p_{nx}\ ; (n,x) \epsilon E\} \tag{1}$$

Using set θ, the function of the degree of effectiveness of each node or the number of child nodes of the node that we assume will be active in the next step can be defined as follows:

$$Y_2(n) = \mu(n,\theta) \tag{2}$$

Figure 1 shows the process of propagating an independent cascade model for a small network with 10 nodes and three primary active nodes.
So in general the IC model works as follows:
The IC model starts with an initial set of active nodes (seed). The diffusion process is revealed in a discrete process according to a random rule:

1. When node n becomes active in step t, it is given a single chance to activate each currently inactive neighbor x; it succeeds with a probability $p(n,x)$
2. If x has several newly active neighbors, their efforts will be sorted as desired.
3. If n succeeds, x is activated in step t + 1. But whether v succeeds or not, it cannot make further effort to activate w in subsequent rounds.
4. This process runs until no more activation are possible

5 Method

In this section, we first explain the framework of the proposed method, which consists of two parts, network construction and implementation of an independent cascading algorithm for identifying driver genes. This section describes how to modify the independent cascade algorithm to apply to the gene regulatory network. Also, the obtained results are compared with 19 computational and network-based methods. Overview of the proposed model is shown in Figure 2.

5.1 Gene Regulatory Network

One of the causes of cancer is the disruption of regulatory relationships between molecules in the cell. Therefore, studying the relationships between them, which can be examined in the form of a biological network, can help identify the causes of the disturbance. One of the most important biological networks, whose dysfunction leads to cancer, is gene regulation networks. The Gene Regulatory Network (GRN) is a collection of DNA fragments in a cell, which interact indirectly with each other (via RNA and the expression of their protein products) and with other molecular regulators in the cell. As a result, they determine which genes in the network are transcribed into mRNA.

5.2 Network Construction

Regulatory interactions and gene expression data were needed to construct the colorectal cancer gene regulatory network. The list of human regulatory interactions was downloaded from the RegNetwork [20] database, which is available at http://www.regnetworkweb.org. In this database, five types of regulatory interactions related to pre-transcription and post-transcription have been reported for humans and mouse. RegNetwork integrates regulatory interactions collected from different databases and extracts potential regulators based on transcription factor binding sites (TFBS). MiRNA interactions were filtered to construct the study gene regulatory network and human regulatory interactions were used. The final information about the data downloaded from this database is shown in Table 1.

Table1. Characteristics of data taken from the RegNetwork

Number	Description	Element
21175	All nodes used in the construction of the gene regulatory network	Node
150202	All regulatory interactions used in the construction of the gene regulatory network	Edge
1456	Transcription factors used in the construction of the gene regulation network	TF
19719	target genes used in the construction of the gene regulation network	Gene
149841	The 'TF-gene' regulations used in the construction of the gene regulation network	TF-gene
361	The 'TF'-'TF gene' self-regulations used in the construction of the gene regulation network	TF-TF

In addition to regulatory interactions, gene expression data were needed to construct the network. Gene expression data was downloaded from the GEO database [9], which is available for free (https://www.ncbi.nlm.nih.gov/geo/). Colorectal cancer gene expression data with GSE15852 ID are available as .CEL file. Expression data are reported in this database for each cancer separately for normal tissue and its adjacent cancer tissue. These files require pre-processing before use, which is done using the Affy package in R and the RMA method. In the obtained file, first, synonymous genes were isolated and the duplicate values of the genes were averaged. The final processed file included the name of a gene and its expression values in normal tissue and its adjacent cancer tissue for different patients. Colorectal cancer regulatory network was constructed by mapping processed gene expression data on the list of regulatory interactions. In this way, the source and destination genes were searched in the gene expression data. If both the source and the destination contained the gene expression data, the desired edge was preserved, otherwise it was removed from the final list.

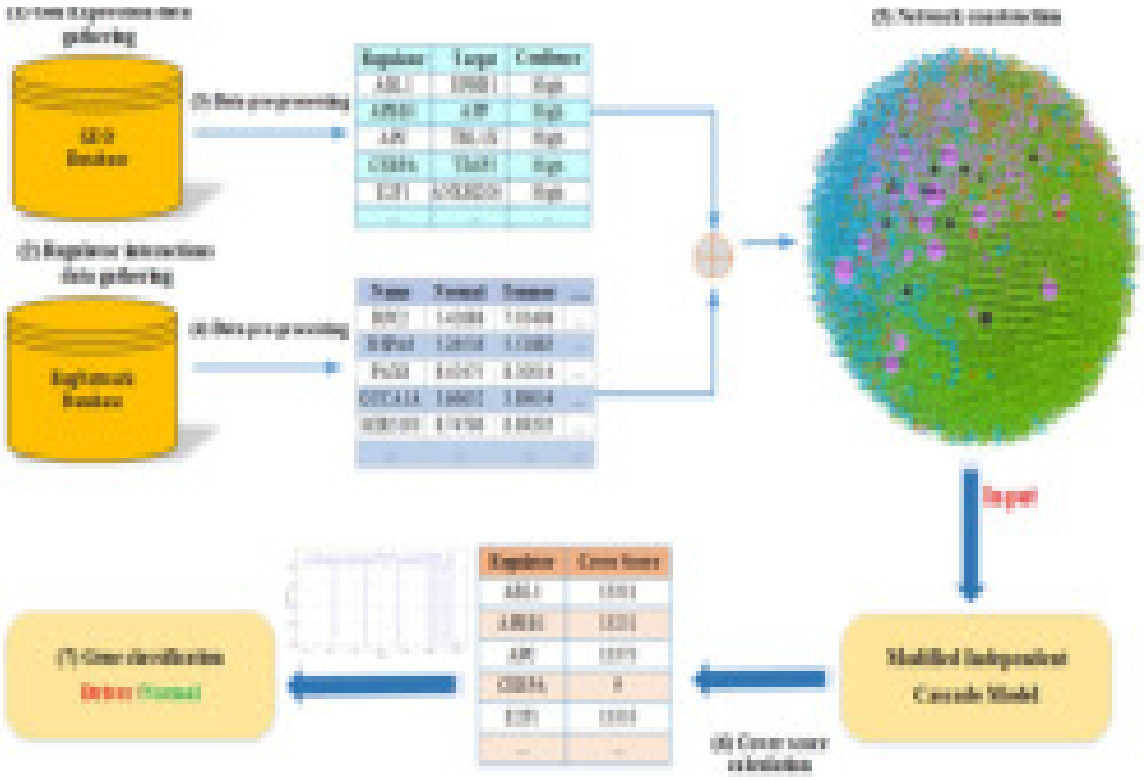


Figure 2: Overview of GeneIC method. (1 and 2) Colorectal gene expression and regulatory interactions data gathering (3 and4) Processing and preparation of raw data (**??**) colorectal gene regulatory network construction by mapping gene expression data and regulatory interactions (**??**) Running of modified independent cascade diffusion algorithm and cover scores calculating (**??**) Threshold fine tuning and gene classification

5.3 The GenIC algorithm

As mentioned in the independent cascade diffusion approach, this algorithm tries to select a minimum set as the seed, so that this set activates the most nodes in the network. According to the gene regulatory network in our study has 150329 genes, selecting the initial active genes set of all networks is very time consuming. Therefore, considering that the type of regulatory interactions of the studied network is TF-Target, we considered only TF nodes as active nodes and implemented the algorithm. Also, we considered only one node as active at a time. In this respect it is different from the basic IC algorithm. The algorithm was repeated 1000 times and the coverage values obtained for each gene were averaged and considered as the final coverage of each gene. The amount of coverage in the gene regulatory network means that if that gene is active, how many genes in the network can be affected and activated. So genes with higher coverage are more likely to be cancer drivers. The input of the proposed GeneIC method is a regulatory network related to colorectal cancer and the output is the amount of coverage of each gene. The obtained list, which shows the coverage of

each gene in the network, was sorted in descending order. One of the parameter in the IC model is the parameter that shows the sensitivity of activating nodes in the network. We set the value of this parameter to 0.4 for the gene regulatory network. In the basic algorithm, this value is set to 0.1 by default. Also, to analyze the sensitivity, we implemented the algorithm with values of 0.1, 0.2, 0.3 and 0.5, which the best result in terms of performance was 0.4.

6 Evaluated method

The results of GenIC were compared with 19 previous computational and network methods. The DriverDBv2 [7] database, which is available for free, was used to obtain output related to computational methods. It uses the Cancer Genome Atlas (TCGA) database, such as colorectal Cancer, as input for computing tools. The output of network-based methods was also taken from the relevant published articles. TCGA is a project for cataloging genetic mutations responsible for cancer, using genome sequencing and bioinformatics [26]. It is overseen by the Cancer Genomics Center of the National Cancer Institute [12] and is a central repository for TCGA data. In this study, we evaluated cancer driver genes identified by the proposed GenIC method and other methods using cancer driver genes (CGC) [28] as the gold standard. A list of cancer-related mutant genes in humans has been reported in the CGC (Table 2). We downloaded a list of colorectal cancer genes (identified as TCGA-COAD) from the free TCGA data portal (https://portal.gdc.cancergov), and CGC-approved driver genes was isolated and used as the gold standard of evaluation. In this dataset, 572 driver genes have been reported for colorectal cancer. In addition, we used two other standard driver gene databases to evaluate the proposed method. Mut-driver validated genes, introduced by Vogelstein et al [24], in this dataset 125 driver genes were reported. The MSKCC driver genes dataset; in this dataset for colorectal cancer 423 driver genes have been reported, and were downloaded from the cBioPortal database. (https://www.cbioportal.org).

Table2. Characteristics of standard driver gene databases

Number of driver	**ID**	**Name**
572	TCGA-COAD	CGC
125	-	Mut-Driver
423	MSKCC-COAD	MSKCC

We used the Recall, precision, and f-measure performance metrics common in binary classification problems to evaluate the proposed method. Recall represents the ratio of the number of genes correctly identified as drivers to the total number of genes reported as drivers. Also, precision indicate the accuracy of the prediction and shows how accurate the genes that are identified as drivers are. Recall and precision alone are not suitable for measuring the performance of a classification model, so the F-measure criterion, which is the harmonic mean of the two criteria Precision and Recall, is used.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$F{-}measure = 2{\times}\frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

7 Results

In this study, colorectal cancer gene regulatory network was constructed using gene expression data and regulatory interactions. Then, an independent cascade diffusion algorithm was implemented on the network to find the coverage of each gene. To do this, according to the network structure, to reduce the volume of calculations and execution time, we considered only the network regulators individually as the initial starting node and implemented the algorithm. Python language was used for implementation. The output was a list of genes along with their coverage rate. The genes were arranged in descending order of coverage. Then, based on a threshold value, they were classified into two categories: driver and normal. The precision_recall_curve and metric packages in the Python sklearn library were used to fine tuning the threshold value. Recall, precision and F-measure values for GenIC and other computational and network-based methods are shown in Figures 3.

As can be seen, the proposed method is higher than all computational methods in terms of F-measure criteria and has the highest value among network methods after GenHITS. In addition, GenIC has the highest recall value after iPac among all previous methods. We compared GenIC and other methods in number of predicted drivers. As shown in Figure 6, GenIC has identified 190 drivers, which is the highest number of drivers compared to the previous methods (after the iPAC calculation method). Although iPac was able to identify 286 drivers, it has a low F-measure (**??**). We compared GenIC and other previous methods for the amount of overlap of detecting driver genes. As shown in Figure 4 and 5, GenIC was able to identify 170 genes identified by other methods. In addition, GenIC identified 22 unique genes that were not identified by any of the previous computational and network-based methods. In addition, we compared the proposed method in terms of the degree of overlap of detected drivers separately with computational and network methods. As shown in the Venn diagram in Figure 6, GenIC has identified 63.5% of genes identified by other network-based methods. It also identified a significant number of 39 unique genes that were not detected by any of the network-based methods. In addition, compared to computational methods, GenIC identified 79 unique genes that were not detected by any of the previous computational methods.

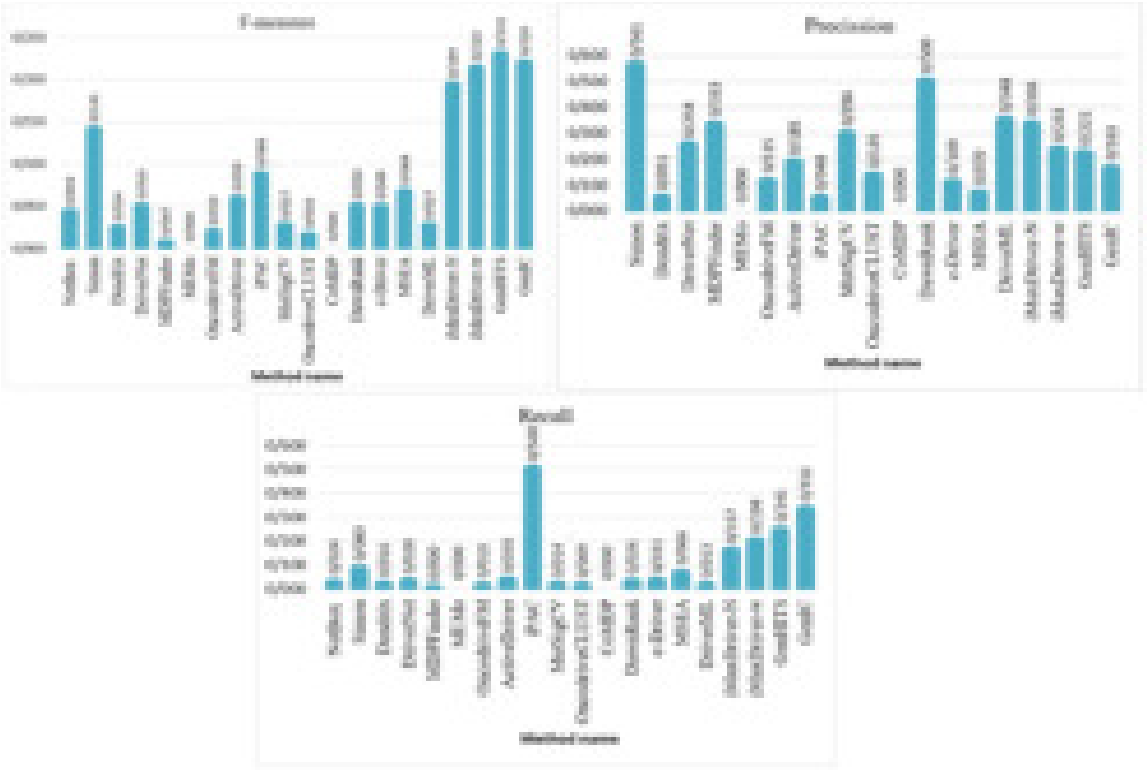


Figure 3. The F-measure, Recall and Precision of GenIC and other methods proposed for CDG prediction.

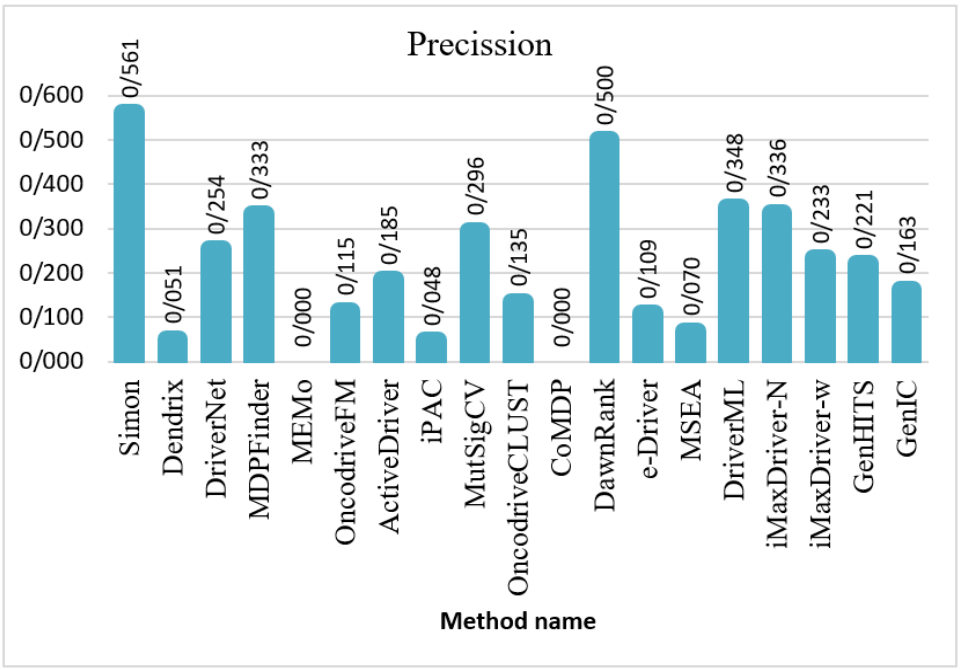


Figure 4. Number of detected drivers by GenIC and other methods.

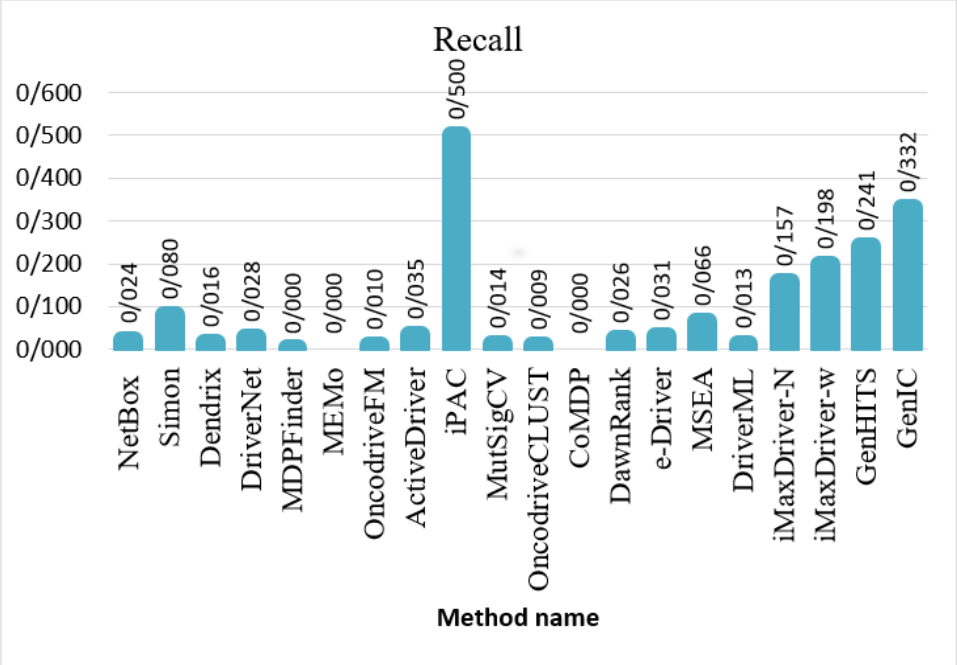


Figure 5. The Venn diagram for detected CDGs using GenIC and other computational and network-based methods.

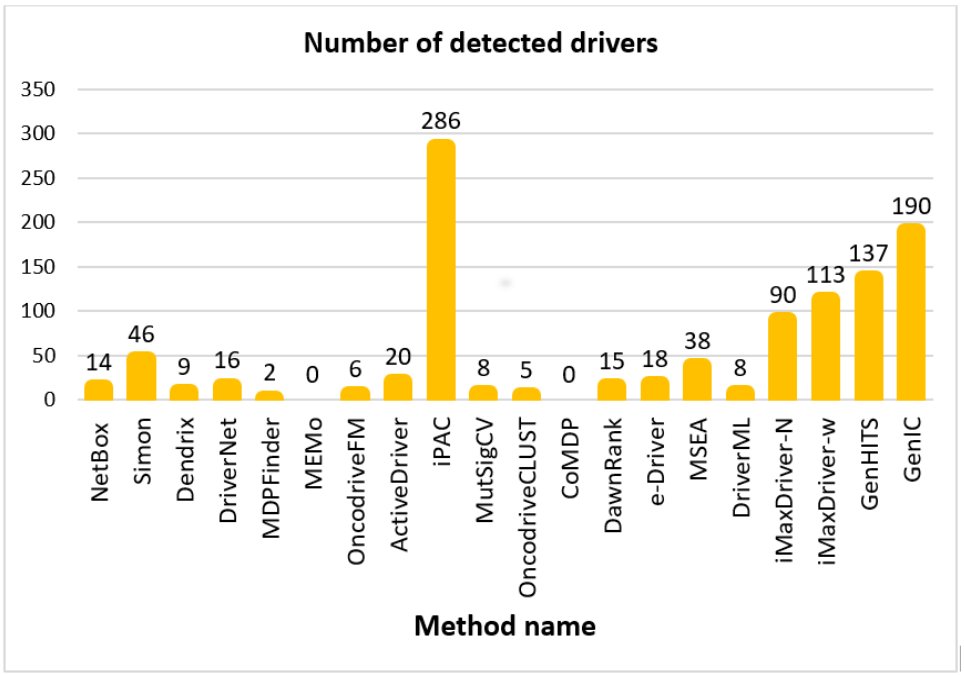
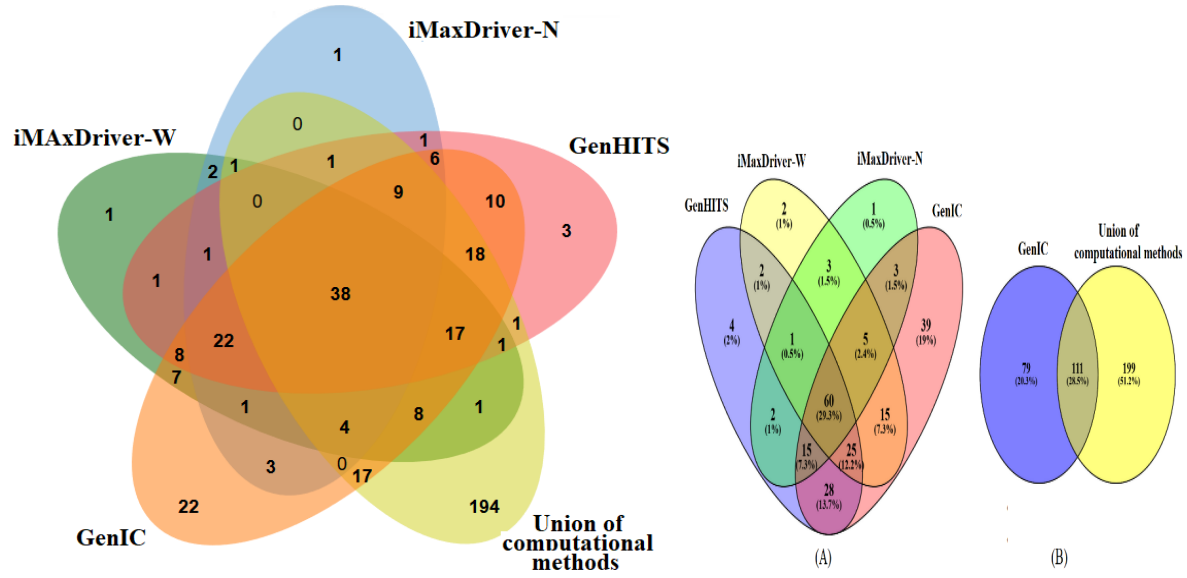


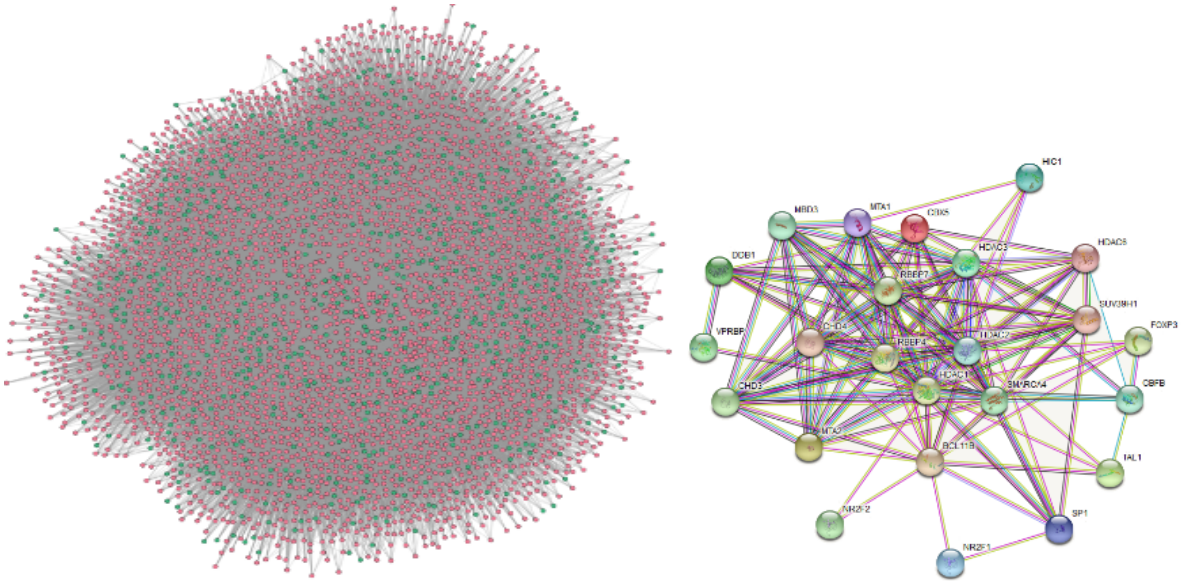
Figure 6. The Venn diagram for predicted CDGs using GenIC and (A). Other network-based methods, (B) the union of all other computational methods.

Each of the identified driver genes can disrupt regulatory networks and initiate or

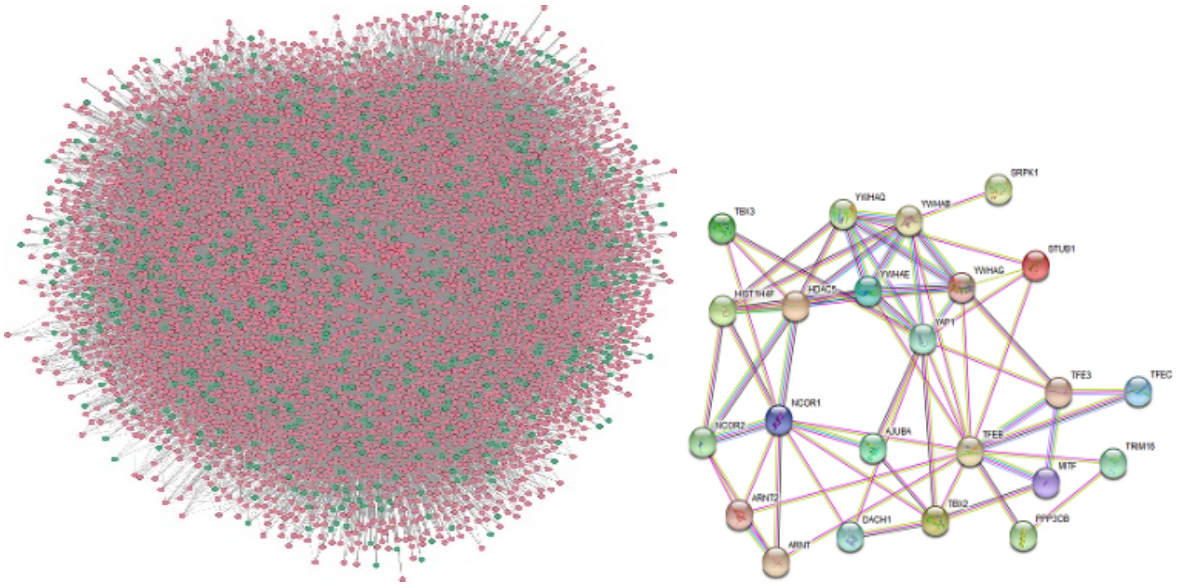
spread cancer. For example, as reported by Tomasetti et al. [32], only mutation in three driver genes could lead to the spread of colorectal cancer. For the top three unique genes identified by the GenIC, we showed which network genes may be abnormal if mutated (Indicated in red). It is also shown through which protein interactions they initiate the process of diffusion on other genes (Figure 7).



(a)



(b)



(c)

Figure 7. Disruption pathways and network of anomaly-spreading protein interactions for the top three drivers identified by GenIC ((a) BCL11B, (b) TFEB and (c) AFF4

We also evaluated the proposed method and other methods based on the mut-driver gold standard dataset. As shown in Figure 8, the proposed method identified the largest number of drivers compared to other network-based methods. In addition, as shown in Figure 9, GenIC was able to identify 45 genes identified by other methods. In addition, GenIC identified 65 unique genes that were not identified by any of the previous computational and network-based methods.

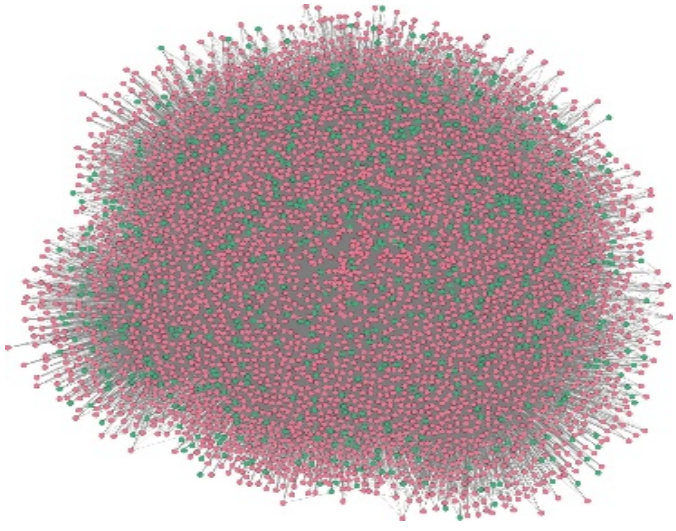


Figure 8. Number of detected drivers by GenIC and other methods (based on Mut-driver database).

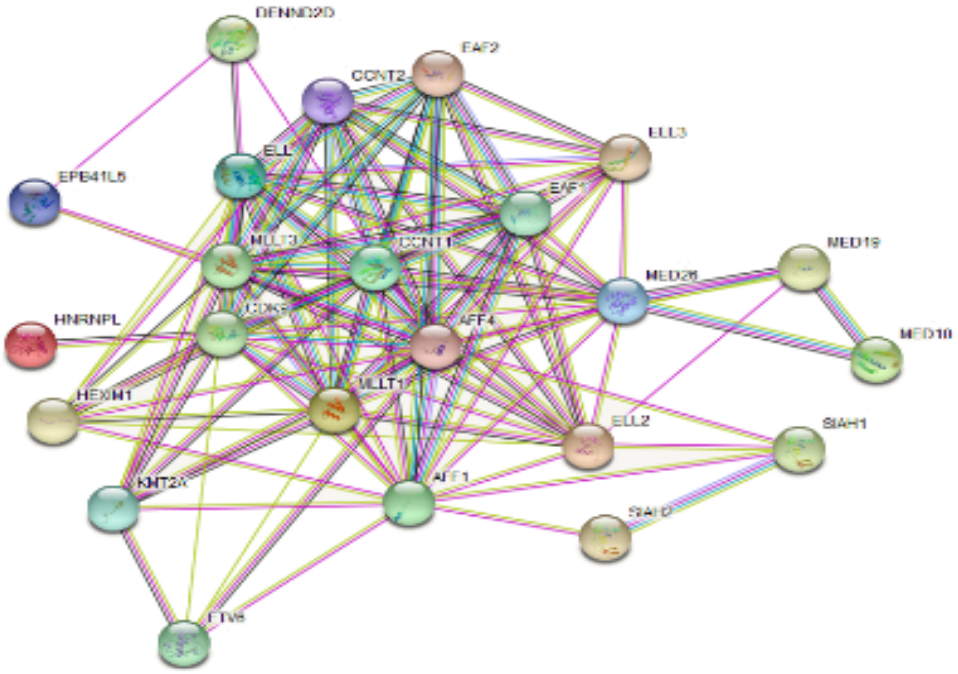


Figure 9. The Venn diagram for detected CDGs using GenIC and other computational and network-based methods (based Mut-driver database).

As shown in the Venn diagram in Figure 10, GenIC has identified 62.9% of genes identified by other network-based methods. In addition, compared to computational methods, GenIC identified 34.6% of genes identified by other computational methods.

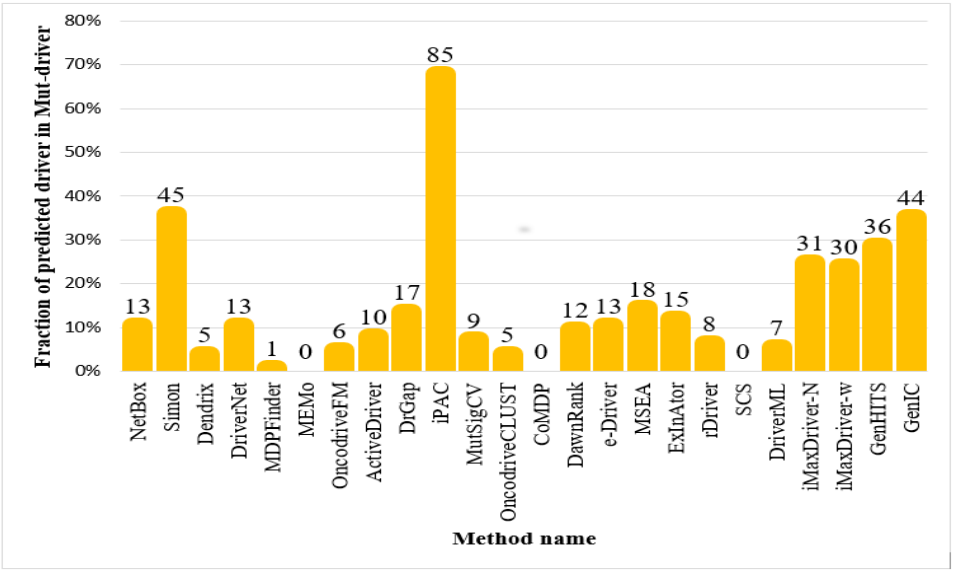


Figure 10. The Venn diagram for predicted CDGs using GenIC and (A). Other network-based methods, (B) the union of all other computational methods. (Based on Mut-driver database)

The comparison results of the proposed method and other methods based on the MSKCC gold standard are shown in Figure 11. As show, the proposed method has identified the highest number of driver genes among all previous network-based and computational methods (after iPac).

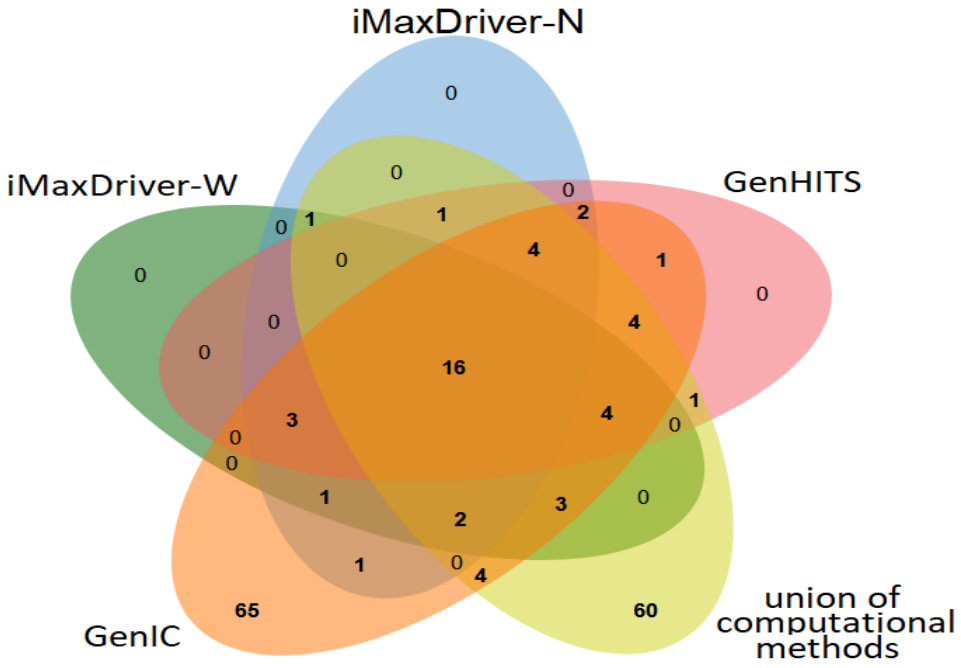


Figure 11. Number of detected drivers by GenIC and other methods (based on MSKCC database).

As shown in Figure 12, GenIC was able to identify 101 genes identified by other methods. In addition, GenIC identified 10 unique genes that were not identified by any of the previous computational and network-based methods. In addition, we compared the proposed method in terms of the degree of overlap of detected drivers separately with computational and network methods. As shown in the Venn diagram in Figure 13, GenIC has identified 56.9% of genes identified by other network-based methods. It also identified a significant number of 19 unique genes that were not detected by any of the network-based methods. In addition, compared to computational methods, GenIC identified 39 unique genes that were not detected by any of the previous computational methods.

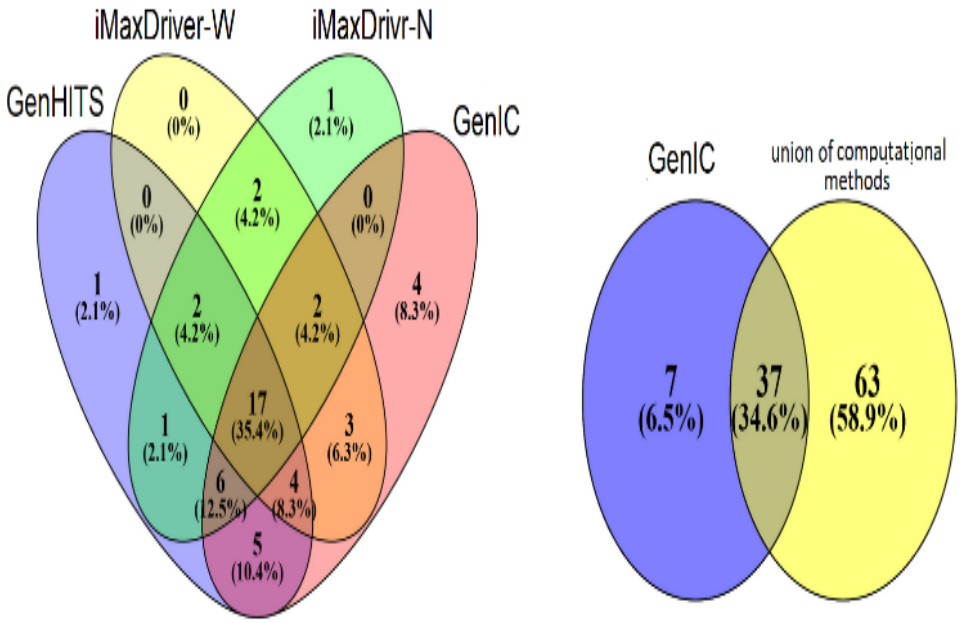


Figure 12. The Venn diagram for detected CDGs using GenIC and other computational and network-based methods (based MSKCC database).

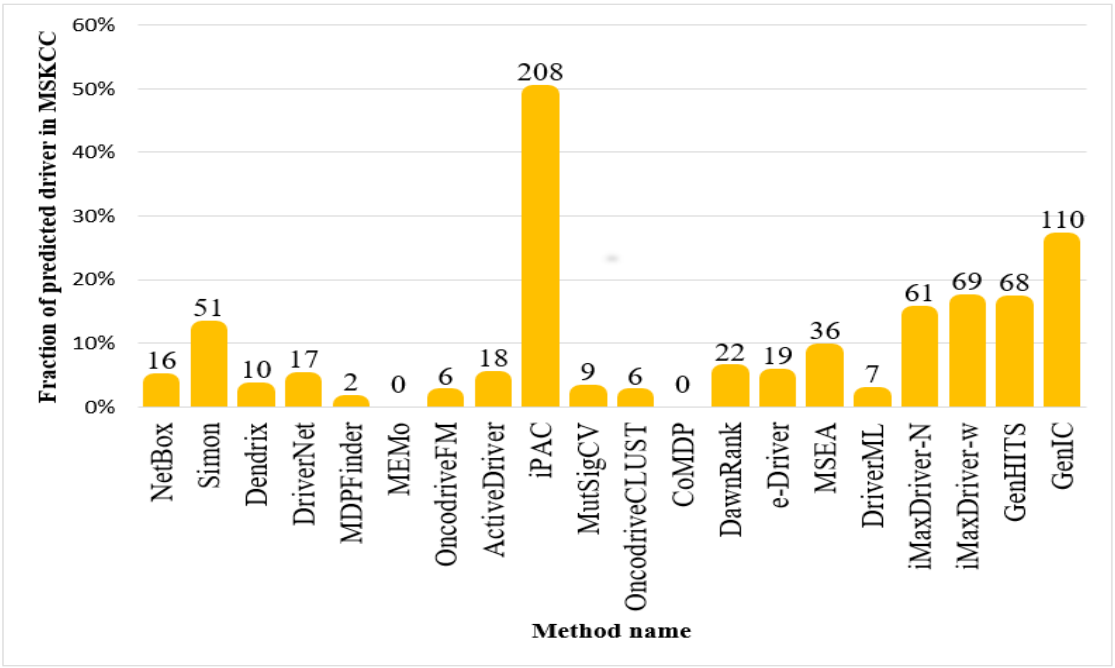


Figure 13. The Venn diagram for predicted CDGs using GenIC and (A). Other network-based methods, (B) the union of all other computational methods. (Based MSKCC database)

8 Conclusion and future work

In this study, we proposed a new network-based method for identifying colorectal cancer driver genes. In this method, the independent cascade diffusion model is used. Independent cascade is one of the popular models in the Influence maximization problem. This approach is the first gene regulation network method used to identify colon cancer genes. This approach has not been used in the gene regulatory network to identify genes for colorectal cancer drivers. The results showed that the proposed method has a higher performance in terms of F-measure and the number of detected drivers compared to other computational and network-based methods. GenIC was also able to identify a significant number of unique drivers that were not detected in any of the previous computational and network-based methods. Therefore, it can be used as a complementary tool along with other computational methods. GenIC performance was significantly higher than iMaxDriver network methods. This suggests that the use of independent cascade diffusion models is more appropriate than linear threshold models in the gene regulatory network for identifying driver genes.
One of the limitations of methods based on influence maximization is the computational time and selection of the initial active set (seed). In this study, an effective technique was proposed to reduce the execution time. The execution time of the proposed algorithm was 35 minutes on a computer with an Intel CORE i7 microprocessor and 8 GB of RAM. Which is a reasonable time in influence maximization algorithms. However, providing methods to reduce the execution time of the algorithm and the proper selection of seed nodes can be one of the future research topics.

References

[1] Akhavan-Safar, M.,Teimourpour, B., and Kargari, M.. "GenHITS: A network science approach to driver gene detection in human regulatory network using gene's influence evaluation." *J. Biomed. Inf.* 114 (2021), 103661.

[2] Arneson, D., Bhattacharya, A.,Shu, L., Mkinen, V. P., and Yang, X., "Mergeomics: a web server for identifying pathological pathways, networks, and key regulators via multidimensional data integration." *BMC genomics* 17, no. 1 (2016), 1-9.

[3] Aure, M. R., Israel Steinfeld, Lars Oliver Baumbusch, Knut Liest , Doron Lipson, Sandra Nyberg, Bjrn Naume et al. "Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data." *PloS one* 8, no. 1 (2013), e53014.

[4] Bashashati, A., Haffari, G., Ding, J., Ha, G., Lui, K., Rosner, J., Huntsman, D. G., Caldas, C., Aparicio, S. A., and Shah, S. P., "DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer." *Genome Biol.* 13, no. 12 (2012), 1-14.

[5] Cerami, E., Demir, E., Schultz, N., Taylor, B. S.,, and Sander, C.. "Automated network analysis identifies core pathways in glioblastoma." *PloS one* 5, no. 2 (2010), e8918.

[6] Cheng, F., Zhao, J and Zhao, Z.. "Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes." *Briefings Bioinf*, 17, no. 4 (2016), 642-656.

[7] Chung, I. F., Chen, C. Y., Su, S. C., Li, C. Y., Wu, K. J., Wang, H. W.,, and Cheng, W.C., "DriverDBv2: a database for human cancer driver gene research." *Nucleic Acids Res.* 44, no. D1 (2016), D975-D979.

[8] Ciriello, G., Cerami, E., Sander, C, and Nikolaus Schultz, N., "Mutual exclusivity analysis identifies oncogenic network modules." *Genome Res.* 22, no. 2 (2012), 398-406.

[9] Clough, E., and Barrett, T.. "The gene expression omnibus database." In Statistical genomics, pp. 93-110. *Humana Press*, New York, NY, 2016.

[10] Goldenberg, J., Liba, B., and Muller, E.,"Talk of the network: A complex systems look at the underlying process of word-of-mouth." *Marketing letters* 12, no. 3 (2001), 211-223.

[11] Gonzalez-Perez, A., and Lopez-Bigas, N.. "Functional impact bias reveals cancer drivers." *Nucleic Acids Res.* 40, no. 21 (2012), e169-e169.

[12] Grever, M. R., Schepartz, S. A., and Bruce A. Chabner. "The National Cancer Institute: cancer drug discovery and development program." *In Seminars in oncology*, vol. 19, no. 6, pp. 622-638. 1992.

[13] Han, Y., Yang, J., Qian, X., Cheng, W. C., Liu, S. H., Hua, X., Zhou, L. et al. "DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies." *Nucleic Acids Res.* 47, no. 8 (2019), e45-e45.

[14] Hou, J. P., and Ma, J.. "DawnRank: discovering personalized driver genes in cancer." *Genome Med.* 6, no. 7 (2014), 1-16.

[15] Jang, H. S., Shah, N. M., Du, A. D., Dailey, Z. Z.,Pehrsson, E. C., Godoy, P. M., Zhang, D. et al. "Transposable elements drive widespread expression of oncogenes in human cancers." *Nat. Genet.* 51, no. 4 (2019), 611-617.

[16] Kermack, W. O., and McKendrick, A. G.. "Contributions to the mathematical theory of epidemics. II.—The problem of endemicity." *Proceedings of the Royal Society of London*. Series A, containing papers of a mathematical and physical character 138, no. 834 (1932), 55-83.

[17] Kempe, D., Kleinberg, J., and Tardos, A., "Influential nodes in a diffusion model for social networks." *In International Colloquium on Automata, Languages, and Programming*, pp. 1127-1138. Springer, Berlin, Heidelberg, 2005.

[18] Lawrence, M.l S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K. C., Sivachenko, A., Carter, S. L. et al. "Mutational heterogeneity in cancer and the search for new cancer-associated genes." *Nature* 499, no. 7457 (2013), 214-218.

[19] Liggett T., Interacting particle systems. *Springer Science & Business Media*. (2012) Dec 6.

[20] Liu, Z. P., Wu, C., Miao, H., and Wu, H.. "RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse." *Database 2015* (2015).

[21] Porta-Pardo, E., and Godzik, A.. "e-Driver: a novel method to identify protein regions driving cancer." *Bioinformatics* 30, no. 21 (2014), 3109-3114.

[22] Rahimi, M.,Teimourpour, B., and Marash, S. A., "Cancer driver gene discovery in transcriptional regulatory networks using influence maximization approach." *Comput. Biol. Med.* 114 (2019), 103362.

[23] Reimand, J.i, Wagih, O., and Bader, G. D.. "The mutational landscape of phosphorylation signaling in cancer." *Sci. Rep.* 3, no. 1 (2013), 1-9.

[24] Tomasetti, C., Luigi Marchionni, Martin A. Nowak, Giovanni Parmigiani, and Bert Vogelstein. "Only three driver gene mutations are required for the development of lung and colorectal cancers." *Proceedings of the National Academy of Sciences* 112, no. 1 (2015), 118-123.

[25] Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N.. "OncodriveCLUST: exploiting the positional clustering of so4matic mutations to identify cancer genes." *Bioinformatics* 29, no. 18 (2013), 2238-2244.

[26] Tomczak, K., P. , Czerwinska, P., and Wiznerowicz, M., "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, *Współczesna Onkologia*, vol. 19, no. 1A, pp." A68-A77 (2015).

[27] Vandin, F., Upfa, E.l, and Raphael, B. J,. "De novo discovery of mutated driver pathways in cancer." *Genome Res.* 22, no. 2 (2012): 375-385.

[28] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr, L. A., and Kinzler, K. W., "Cancer genome landscapes." *science* 339, no. 6127 (2013), 1546-1558.

[29] World Health Organization, Cancers, 12 September 2018. (https://www.who.int/en/news-room/fact-sheets/detail/cancer).

[30] Youn, A. , and Simon, R.. "Identifying cancer driver genes in tumor genome sequencing studies." *Bioinformatics* 27, no. 2 (2011), 175-181.

[31] Zhao, J., Zhang, S., Wu. L. Y., and Zhang, X. S.. "Efficient methods for identifying mutated driver pathways in cancer." *Bioinformatics* 28, no. 22 (2012), 2940-2947.