Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

CinfuMax: An Influence Maximization-Based Model for Predicting Cancer Driver Genes in Gene Regulatory Networks

M. Akhavan-Safar [1], ***B. Teimourpour*** [2,*], ***M. Ayoubi*** [3]

[1] *Department of Computer and Information Technology Engineering, Payame Noor University (PNU), Tehran, Iran.*

[2] *Department of Information Technology, Faculty of Industrial and Systems Engineering, Tarbiat Modares University (TMU), Tehran, Iran.*

[3] *Department of Data Science, Tarbiat Modares University (TMU), Tehran, Iran.*

Article Info

Article History:
Received 18 July 2023
Reviewed 13 March 2024
Revised 05 April 2024
Accepted 08 April 2024

Keywords:
Cancer drivers
Influence maximization
Independent cascade model
Influence speed
Regulatory networks

*Corresponding Author's Email Address:
b.teimourpour@modares.ac.ir

Abstract

Background and Objectives: The identification of driver genes, which initiate cancer in cells, holds immense significance within the field of oncology. Discovering these genes is crucial for identifying markers that can indicate specific conditions or diseases, as well as for developing novel systemic and molecular treatment approaches for them. Several computational methods have been developed to identify the genes responsible for cancer based on genomic data. However, many of these methods find key mutations in genomic data to predict which genes are responsible for cancer. These methods rely on mutation and genomic data, but they often exhibit a high rate of false positives in the results. In this study, we propose an influence maximization-based approach, called CinfuMax, which can predict cancer-associated genes without relying on mutation information.

Methods: In this method, the concept of influence maximization and the independent cascade model is employed. Firstly, the gene regulatory network for breast, lung and colon cancers was constructed using regulatory interactions and gene expression data. Next, we implemented an independent cascade diffusion algorithm on the networks to calculate the coverage of each gene. Ultimately, genes with the highest coverage were identified and classified as drivers.

Results: The proposed method's results were compared to those of 19 other computational and network-based methods, utilizing the F-measure and the number of predicted driver genes as evaluation metricsThe results clearly indicate that the proposed method outperforms other methods. Furthermore, CinfuMax successfully identifies 18, 19, and 22 individual driver genes in breast, lung, and colon cancers, respectively, which were not previously identified by any other methods.

Conclusion: Corrected: The results indicate that independent cascading methods for identifying driver genes outperform linear threshold methods. Driver genes were also categorized based on their influence speed, and the genes with the highest diffusion rate in each type of cancer have been identified. The identification of these genes can be valuable for molecular therapies and drug development.

Introduction

Cancer is a disease caused by oncogene activations, such as genetic mutations, chromosome rearrangements, or the deactivation of tumor suppressor genes [1], [2]. It is the second leading cause of death globally, with approximately 6.9 million people losing their lives in 2018.

Lung cancer (2.09 million cases), Breast cancer (2.09 million cases), and Colon cancer (1.80 million cases) are among the most prevalent cancers [3]. During tumor progression, most of the altered genes identified are passenger-type, meaning they do not contribute to the oncogenic process. However, a small portion of the altered genes is believed to be driver genes that disrupt normal transcriptional processes and transform the cell from normal to cancerous.

Cancer Driver Genes prediction Problem

Many computational techniques had been proposed to discover CDGs. In those techniques, it is assumed that the genes that cause cancer are more susceptible to essential modifications in genes, such as mutations. Not all mutations that arise within the cancer genome led to cancer. Therefore, most computational methods attempt to distinguish driver mutations from non-driver mutations. Most of the methods available to identify CDGs depend on transcriptomic or genomic data. For instance, OncodriveCLUST, proposed by Tamboero et al in 2013 [4], identifies genes that exhibit a significant tendency to accumulate mutations in protein sequences. It creates a model to classify genes by evaluating the encoding of silent mutations. Another computational method, Simon [5], aims to improve the identification of cancer driver genes by estimating the background mutation rate and considering the operational effect of mutations on proteins, background mutation changes in tumors, and the redundancy of the genetic code. One of the features of this method is its capability to distinguish between mutations that affect protein function and other mutations. Furthermore, it can differentiate between the number of background mutations in various samples and patients. Dendrix [6] is a computational method that integrates two coverage characteristics: identifying genes in different patient samples and exclusivity. It aims to distinguish driver mutations from passenger mutations that are infrequently observed in certain patients. The ActiveDriver method, developed in 2013 by Reimand et al [7], identifies signal sites where the mutation rate is significantly higher than the mutation level in the entire gene sequence, highlighting the importance of these sites in cancer biology. Another method, e-Driver [8], extracts the internal distribution of malignant mutations between functional regions of proteins to determine the mutation rate relative to other regions of the same protein. If the observations are positive, those genes could be the CDGs. Oncodrive-FM is another computational method based on mutation data [9]. One of the major challenges in cancer genomics lies in identifying Cancer Driver Genes (CDGs) and their associated pathways amidst various types of mutations. The method calculates a functional influence metric using three established methods and evaluates the deviation in the functional influence of variants found in a gene across multiple tumor samples. To address the limitations of traditional approaches, a new criterion termed FM bias is introduced, aiming to overcome issues related to accurately estimating the mutation rate and dependence on incremental changes.

The MDPFinder method [10] uses both mutation data and gene expression data to identify the pathways of cancer mutations and genes that cause cancer. It aims to address the issue of identifying mutant driver paths by developing a maximum weight matrix [6]. To achieve this, it utilizes a genetic algorithm and integrates gene expression data and mutation data to identify cancer mutation pathways and the genes responsible for causing cancer. The DriverML method is another computational method [11] that utilizes machine learning and the Rao test to identify cancer-causing genes, relying on mutation data. The MutsigCV computational method also leverages mutation and expression data [12] to identify abnormal changes in genes and address the issue of heterogeneity in mutation processes and mutation frequency of genes. Additionally, iPAC is a computational method that systematically performs statistical tests on a list of genes to extract the CDGs [13], using gene expression and the number of copies as input data.

Another category of methods for identifying driver genes involves leveraging the structure of biological networks alongside mutation and genomic data. For instance, the Netbox method [14] utilizes protein-protein interaction network analysis to identify frequently changing modules. This method involves creating a network of PPI interactions and signaling pathways, identifying network modules, and statistically evaluating the significance of modularity. DawnRank is another method that utilizes mutation data [15] to focus on each patient's cancer genes, aiming to discover rare and specific driver genes for individual patients. DawnRank utilizes the individual patient's specific genetic information to identify cancer genes. It ranks mutated genes in a patient based on their potential for transmission in the Molecular Interaction Network, with higher-ranking genes being more likely to be drivers. Memo is a systematic approach to identifying cancer modules based on the concept of mutually exclusive events [16], utilizing mutation data and network structure. It searches for modules characterized by three key features: (1) frequent alterations of module genes in tumor samples, (2) involvement in known biological processes, and (3) mutually exclusive change events within the modules. Additionally, Memo conducts mutation enrichment analysis to examine mutational hotspots in genes [17], hypothesizing that genes with such hotspots could act as driver genes. The method integrates diverse data types, including multidimensional disease-related data, biological functional data, and

molecular networks. It employs two approaches: simulating a random walk in sequences to provide a quantitative measure of mutation location and clustering, and evaluating whether a protein domain exhibits a higher mutation rate than the rest of the protein. DriverNet is a computational framework designed to identify driver mutations within miRNA expression networks [18]. While leveraging a network structure, this method also relies on mutation data. It extracts the relationship between genome aberrations and transcription patterns through the gene network's structure.

Another category of methods for recognizing cancer driver genes that has recently gained attention are network-based and bioinformatics methods. These approaches do not rely solely on mutational and genomic data, but also utilize biological network structures to identify CDGs. Notably, the iMaxDriver-N and iMaxDriver-W methods fall within this category [19], as they identify driver genes using gene expression data and the structure of the transcriptional regulatory network, employing the concept of influence maximization and the linear threshold model.

Previously proposed methods for identifying cancer driver genes (CDGs) have limitations. Computational methods rely on mutation data, which inherently contains a significant amount of noise, leading to high false positive results. Moreover, there is significant overlap in the genes identified by these computational methods. While previous network-based methods do not encounter the same issues as computational methods, there is room for improvement in terms of the number of identified CDGs and performance metrics. Addressing these limitations, a network-based method that does not rely on mutational data to identify cancer-causing genes was proposed in this study. This method uses the concept of influence maximization in the transcriptional regulatory network and an independent cascade model to prioritize genes. The data used include gene expression data and human regulatory interactions. In this method, the coverage of each gene is calculated in terms of diffusion power in the gene regulatory network (GRN). The Gene Regulatory Network consists of DNA fragments in a cell that interact indirectly with each other and with other molecular regulators, ultimately determining which genes in the network are transcribed into mRNA. The proposed method is capable of classifying genes based on network propagation speed, which holds significant potential for molecular therapy and drug development purposes.

Influence Maximization Problem

A social network is a social structure consisting of a set of individuals and the relationships or interactions between them. The rapid proliferation of the Internet has significantly increased the popularity of social networking among people. Consequently, social networks have emerged as a popular platform for product advertising and information dissemination [20].

Many topics are explored through the analysis of social networks, including models of diffusion and social influence. The influence maximization problem involves identifying the most influential nodes to achieve the maximum impact of diffusion in a social network, which is known as an NP-Hard problem.

The purpose of influence maximization is to leverage the network's capacity, such as social networks, to reach the widest audience and maximize the spread of information or influence. In general, inputs related to the influence maximizing problem include:

- A directed graph $G = (V, E)$, the network on which influence maximization is to be performed.
- A set of nodes as primary active nodes.
- A function $f : 2^V \rightarrow R$ that maps a set of nodes (S) to their diffusion values $(f(S))$. This shows how much it will affect the amount of diffusion on the network if we choose this set of nodes as seed (S) for the propagation process.
- a budget k

The goal is to find a set of seed(S) that

$$\max f(S) \quad |S| \leq k \tag{1}$$

Influence maximization involves identifying the minimum k nodes that can maximize diffusion in a social network, thereby enabling these nodes to exert the greatest impact on other nodes within the network. There are several models to solve the problem of influence maximization. One widely used method for modeling the diffusion process is the independent cascade model, which draws inspiration from particle movement models in physics [21]. The independent cascade model for influence maximization was initially studied by Goldenberg et al [22]. In this model, a set of nodes is chosen as initially active nodes (seeds).

At each step t, an active node v attempts to activate one of its neighboring inactive nodes with a given probability p_v. If successful, the newly activated nodes become active in the next step $(t + 1)$ and initiate a similar process to activate adjacent inactive nodes. Once a node's activation attempt succeeds or fails, it cannot attempt to activate the same node again. This process continues until it is no longer possible to activate a new node.

A. Independent Cascade Model

Cascading models for diffusion draw inspiration from particle motion and probability theory [23]. These models were initially studied in marketing, with one of the simplest being the independent cascade diffusion model

[24]. The impact of the influence maximization problem largely hinges on the selection of the diffusion model, specifically the diffusion function f. Understanding how a set of nodes (S) affects the entire network remains a significant challenge within the influence maximization problem. The concept involves establishing a random process on the network originating from S and spreading like a contagion. The expected number of infected nodes at the end of the diffusion propagation will indicate the influence f(S) through the set S. Consequently, the influence maximization problem is defined as a random transmission process in the network. The independent cascade model is a subset of cascade propagation models. In this model we have a directed graph G= (V, E). For each edge (u,v) in this graph, the weight $p_{u,v} \in [0,1]$ denotes the probability of diffusion for that edge.

At each time step $t \in N$, each node in the graph is in one of the following three states:

- ***Infected***: The node is newly infected and remains active for a period of time as it tries to infect its neighbors.
- ***Susceptible***: The nodes are not infected yet, but will probably become infected at this time step or in later step.
- ***Inactive***: These nodes were infected in the past but are now inactive. Infected nodes only become infected at a time step and then become inactive.

Fig. 1 shows the process of propagating an independent cascade model for a small network with 10 nodes and three primary active nodes.

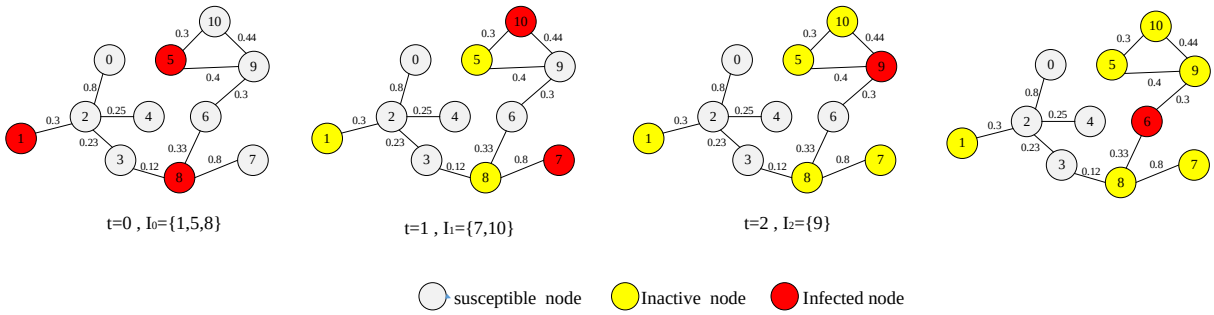

Fig. 1: The IC model example in a network with 10 nodes and 11 edges. Susceptible, infected and inactive nodes at each time step are shown in grey, red and yellow, respectively. Nodes 1, 5 and 7 are the initial active nodes in t=0.

As depicted, each infected node remains active for a time step to infect each of its susceptible neighbors. Infected node u will attempt to infect its neighbor's susceptible node v, with a success probability of $p_{u,v}$. Each attempt to activate susceptible nodes represents an independent random event. In cases where an infected node has multiple susceptible neighbors, attempts are made to activate them in a specified order. This process persists until an infected node successfully infects one of its neighbor's susceptible nodes.

Formally, for a susceptible node v in S_t, its probability of being infected at time step $t + 1$ is given by (2):

$$p(v, I_t) = 1 - \prod_{u \in P(v) \cap I_t} (1 - p_{u,v}) \quad (2)$$

where,

I_t: The infected nodes set

S_t: The susceptible nodes set

$P(v) \cap I_t$: infected parent sets of node v

Also,$P(v)$ are the parents of node v and are defined as (3):

$$P(v) = \{u \in V \mid (u, v) \in E\} \quad (3)$$

For a node $u \in (P(v) \cap I_t)$, the probability that an attempt will unsuccessful is equal to $(1 - p_{u,v})$, and v will be infected if not all attempts unsuccessful. The set of all nodes infected during a contagion process from S can define as (4):

$$I(S) = \bigcup_{t \geq 0} I_t \quad (4)$$

Finally, the penetration function $f(S)$ in the independent cascade model can be defined as (5):

$$f(S) = \mathbb{Z}[|I(S)|] \quad (5)$$

So, in general, the IC model works as follows:

The IC model starts with an initial set of infected nodes (seed). The influence process is revealed in a discrete process according to a random rule:

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

Methodology

In this section, we will outline the CinfuMax pipeline, which comprises two distinct steps: the construction of a gene regulatory network and the implementation of the independent cascade algorithm to identify driver genes and determine their influence rate. The driver genes identified through the independent cascade model will be clustered based on their influence rates. Finally, we compare the proposed method with 19 other computational and network-based methodss based on three cancer regulatory networks as benchmarks and several gold standard databases.

A. Gene Regulatory Network

Biological networks represent the numerous of interactions within a cell, providing insight into how relationships between molecules regulate normal cellular behavior. Recent advances in bioinformatics and computational biology have facilitated the study of complex networks of transcriptional regulatory. These networks describe gene expression as a function of regulatory inputs characterized by interactions between proteins and DNA [24]. A gene regulatory network (GRN) is a directional graph in which gene expression regulators bind to target gene nodes through regulatory interactions. Gene expression regulators, including transcription factors (TFs) that can act as activators and repressors, RNA-binding proteins, and RNA regulators, constitute a collection of DNA fragments in a cell [25]. Analyzing and understanding the regulatory relationships between transcriptional regulators and their purposes is essential for comprehending biological phenomena, from cell growth and division to the identification and treatment of diseases, including cancer. Transcriptional regulatory networks (TRNs) are among the most important types of gene regulatory networks, playing a crucial role in the mechanism of diseases, especially cancer [26]. These types of networks are formed from the effect of a type of gene called a transcription factor on other genes. In this study, a list of confirmed regulatory interactions between the transcription factor and genes, described in the next section, was used to construct the cancer transcriptional regulatory networks.

B. Network Reconstruction

To construct transcriptional regulatory networks for each cancer, a list of regulatory interactions and gene expression data was necessary. The list of approved regulatory interactions was downloaded from the RegNetwork database [27], which is freely accessible[1]. This database reports five types of regulatory interactions related to pre-transcription and post-transcription for humans and mice. RegNetwork integrates regulatory interactions collected from various databases and extracts potential regulators based on TFBS[2]. In this study, interactions related to miRNA genes were filtered. Table 1 displays the information about the data used from the RegNetwork database.

Table 1: Characteristics of data taken from the RegNetwork

Element	Description	Number of elements
Gene	Gene regulatory network nodes	21175
interaction	Gene regulatory network edges	150202
TF	Transcription factors (a type network nodes)	1456
Gene	Target genes (a type network nodes)	19719
TF-gene	The TF-gene regulations (a type network edges)	149841
TF-TF	The TF'-TF regulations (a type network edges)	361

Gene expression data was also downloaded from the GEO database [28], which is freely available[3]. This database provides lung (GSE3268)[4], colon (GSE32323)[5], and breast (GSE15852)[6] gene expression data in the .CEL format. Expression data are reported separately in this database for normal tissue and adjacent tumour tissue for each tumour. Prior to use, these files must be processed using the Affy package in R and the RMA method. After processing, synonymous genes were isolated, and duplicate gene values were averaged. three text files corresponding to the three cancer tissues were generated, with each row containing the gene name and its expression values in normal tissue and its adjacent cancer tissue for different patients. The gene regulatory network for each cancer was then constructed by mapping the processed gene expression data to the list of regulatory interactions. This involved retaining interactions where both the source and destination genes were present in the list of related expression data, and filtering out the rest. The number of genes and regulatory interactions in each network is presented in Table 2.

Table 2: Number of nodes and edges in each cancer regulatory network

Network name	Number of nodes	Number of interactions
Breast cancer network	2499	7540
Lung cancer network	2782	8199
Colon cancer network	2500	7540

[1] http://www.regnetworkweb.org
[2] transcription factor binding sites
[3] https://www.ncbi.nlm.nih.gov/geo/
[4] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi
[5] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi
[6] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi

The CinfuMax Algorithm

Influence maximization models aim to select a minimal seed set that activates the largest number of nodes in the network. Given the time-intensive nature of influence maximization algorithms and the specific type of networks studied (TRNs), we focused solely on nodes of the transcription factor type as being "infected." Each node was individually considered as infected, and the influence score in the network was calculated. The algorithm was executed 300 times on each network, and the resulting coverage values were averaged , serving as the final coverage metric for each gene. The coverage score indicates the potential impact of infecting a gene on the other susceptible genes in the network. Genes with higher coverage are more likely to be associated with cancer driver genes. To optimize the algorithm's iterations for the best outcome, we tested iterations ranging from 10 to 500. In all three cancer networks, the optimal iteration value was obtained to be 300. Moreover, the proposed algorithm can determine the required iterations for each gene to achieve its maximum coverage, enabling the identification of genes with the potential to spread rapidly within the network. An overview of the proposed model is depicted in Fig. 2.

the CinfuMax method takes a cancer regulatory network as input and provides the coverage value and influence rate for each gene as output. Within the independent cascade model, a key parameter is the sensitivity of infected nodes in the network, which is typically set to 0.1 in the basic IC algorithm. To assess sensitivity, we implemented the algorithm using parameter values of 0.1, 0.2, 0.3, 0.4, and 0.5. Across all three networks, the best performance was observed when the parameter was set to 0.4.

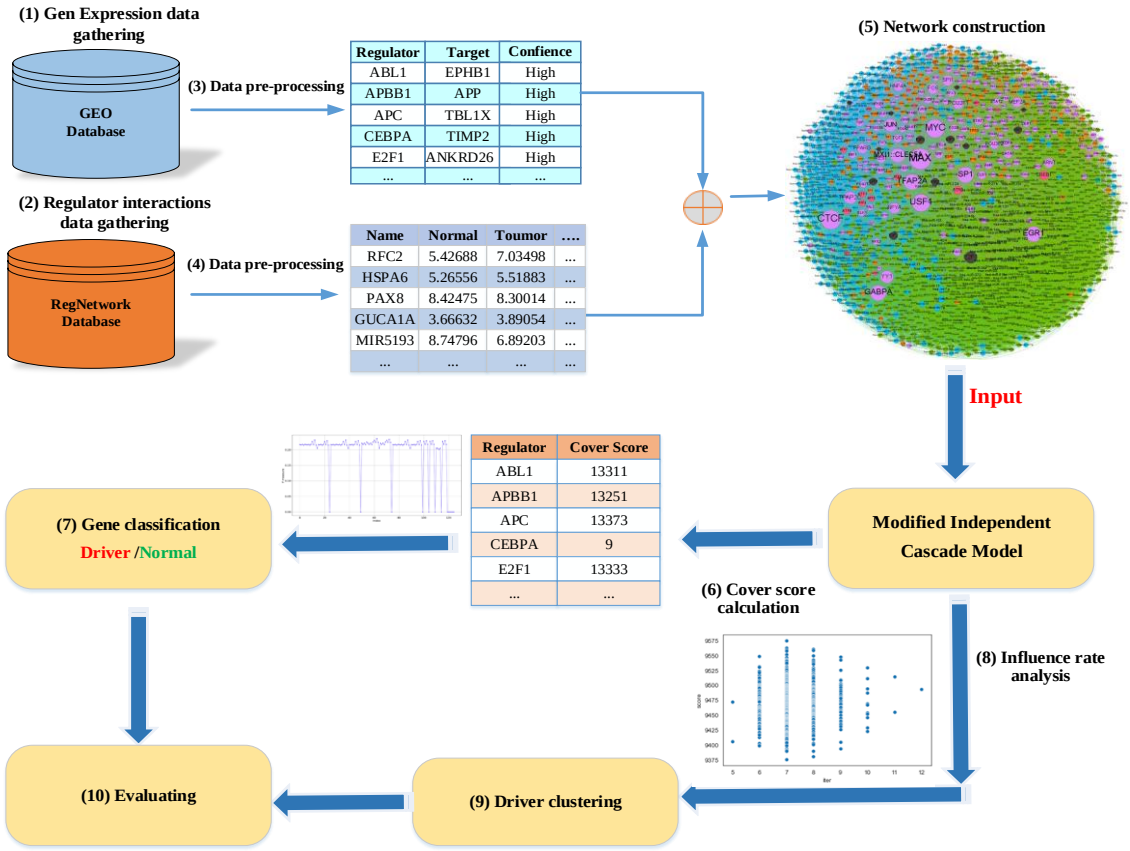

Fig. 2: (1 and 2) The CinfuMax pipeline. (3 and4) Collection of required data (gene expression and regulatory interactions) and pre-processing and data preparation. (5) Constructing transcriptional regulatory network. (6) Running the modified IC algorithm and calculating cover scores (7) Tuning of thresholds and classification of genes (8) Influence rate calculating (9) driver gene clustering (10) Evaluating.

Evaluated Method

CinfuMax results were compared to 19 previous computational and network-based methods. The DriverDBv2 database was used to obtain results on the computational methods [29].

It uses the Cancer Genome Atlas database, such as colon, lung, and breast cancers, as input for computing tools. TCGA is a project aimed at cataloging genetic mutations responsible for cancer through genome sequencing and bioinformatics [30]. It serves as a central

repository for TCGA data and is licensed by the National Cancer Institute Genomic Cancer Center [31]. The results of network-based methods were obtained from their respective papers. The driver genes provided in the Cancer Gene Census (CGC) were used as the gold standard for evaluating CinfuMax and previous methods. The CGC reports a list of cancer driver genes, and we downloaded the lists for colon (TCGA-COAD), lung (TCGA-LUSC), and breast (TCGA-BRCA) cancers from the free TCGA data portal [7]. CGC-approved driver genes were then selected and used as the gold standard of evaluation. In this standard database, 572, 572 and 566 driver genes are reported for breast, colon and lung cancers, respectively.

The confusion matrix was used to calculate the evaluation criteria. Its various values are described in Table 2. To evaluate the performance of the proposed method and compare it with other methods, we used Recall, Precision and F-measure, which are common in binary classification problems.

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

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

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

Confusion matrix was used to calculate evaluation metrics. It shows the prediction results of a classification problem. The confusion matrix and its various values are described in Table 3.

Table 3: The confusion matrix and its various parts

		Actual class	
	Samples	**Positive**	**Negative**
Predicated class	**Positive**	TP (True positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

TP: It refers to the number of genes that have been cancer drivers and the algorithm has also identified them as drivers.
FP: It refers to the number of genes that have not been introduced as cancer drivers in the dataset used, but the algorithm has mistakenly categorized them as cancer drivers.
FN: It refers to the number of genes that have been actually cancer drivers but the algorithm mistakenly categorized them as normal.
TN: It refers to the number of genes that have not been carcinogens and the algorithm has also correctly identified them as non-carcinogens.

Results and Discussion

In this study, cancer regulatory networks were constructed using gene expression data and regulatory interactions. Subsequently, an independent cascade influence algorithm was applied to the network to determine the diffusion score of each gene. To streamline computations and reduce execution time, we focused solely on individual transcription factors as infected initial nodes, based on the network structure. The implementation of the algorithm and its evaluation were carried out using the Python language. The output comprised a list of genes with their respective coverage scores and influence rates, organized in descending order of coverage score. Subsequently, based on a specified threshold value, genes were categorized as either "driver" or "normal." Fine-tuning of the threshold value was performed using the precision_recall_curve and metric packages within the Python sklearn library. The confusion matrix of the proposed method for all three cancer networks is shown in Fig. 3.

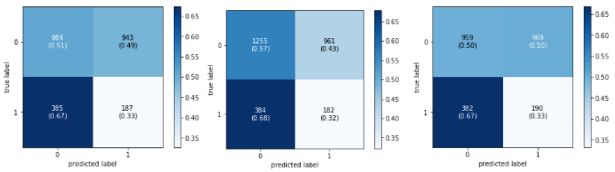

Fig. 3: Confusion matrixs for (a) Breast cancer, (b) Lung cancer (c) and colon cancer networks.

A. *Breast Cancer Network*

The F-measure values for CinfuMax and other computational and network-based methods in Breast cancer are shown in Fig. 4. It is evident that CinfuMax outperforms all other computational and network-based methods in terms of F-measure. Additionally, as illustrated in Fig. 4, CinfuMax has identified 187 drivers, representing the highest number of drivers compared to the previous computational and network-based methods (after iPac). Although iPac has predicted more drivers, its F-measure is notably lower.

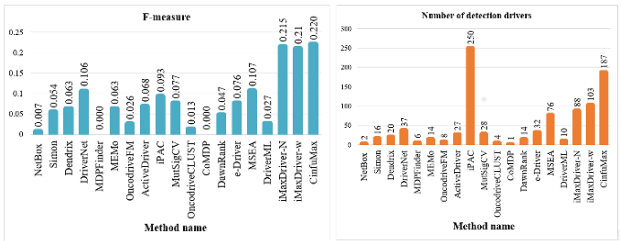

Fig. 4: Performance measures of CinfuMax and other pervious methods proposed in breast cancer.

As shown in Fig. 5, CinfuMax successfully identified 149 genes that were also identified by other methods. Furthermore, CinfuMax discovered 38 unique genes that were not predicted by any previous computational and network-based methods. Additionally, we conducted a comparative analysis of the proposed method in terms of

[7] https://portal.gdc.cancergov

the overlap of predicted drivers with computational and network methods. The Venn diagram in Fig. 6 illustrates that CinfuMax identified 64.9% (124) of the genes identified by other network-based methods, along with 63 unique genes not predicted by any of the network-based methods. Moreover, in comparison to computational methods, CinfuMax identified 93 unique genes that were not predicted by any previous computational methods.

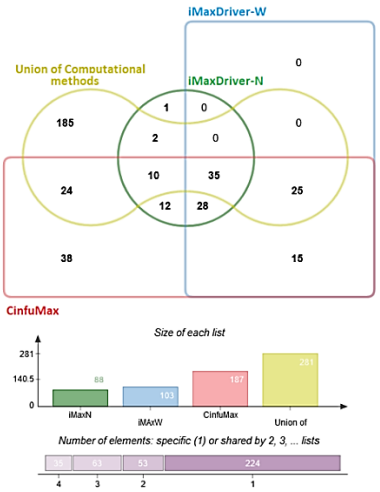

Fig. 5: The Venn diagram for predicted CDGs using CinfuMax and other computational and network-based methods in breast cancer.

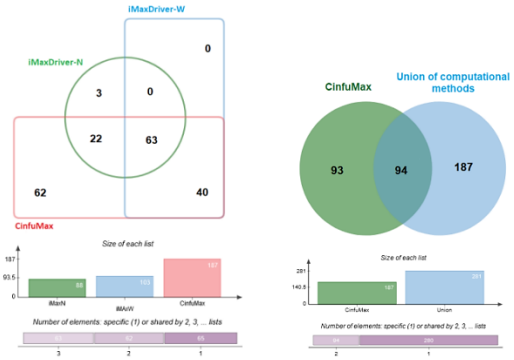

Fig. 6. The Venn diagram for predicted CDGs using CinfuMax and (A). Other network-based methods, (B) the union of all other computational methods in breast cancer

B. Colon Cancer Network

The F-measure values for CinfuMax and other computational and network-based methods in colon cancer network are shown in Fig. 6. It is evident that the proposed method outperforms all computational and network-based methods in terms of F-measure. Additionally, as illustrated in Fig. 7, CinfuMax has identified 190 drivers, representing the highest number of drivers compared to the previous methods (after the iPAC computational method). Although iPac was able to identify 286 drivers, its F-measure is notably low (0.088). We conducted a comparison between CinfuMax and 19 previous methods with respect to the overlap of CDGs predicted. As shown in Fig. 8, CinfuMax successfully identified 158 genes that were also identified by other methods. Additionally, CinfuMax discovered 32 unique genes that were not identified by any previous computational and network-based methods. Furthermore, we evaluated the proposed method in terms of the degree of overlap of predicted drivers separately with computational and network methods. The Venn diagram in Fig. 9 illustrates that CinfuMax identified 61.2% (123) of the genes identified by other network-based methods, along with 67 unique genes not predicted by any of the network-based methods. Moreover, in comparison to computational methods, CinfuMax identified 79 unique genes that were not predicted by any previous computational methods.

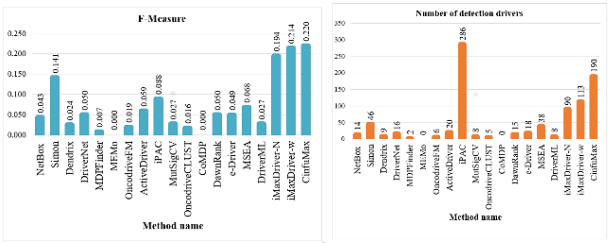

Fig. 7: Performance measures of CinfuMax and other pervious methods proposed in colon cancer.

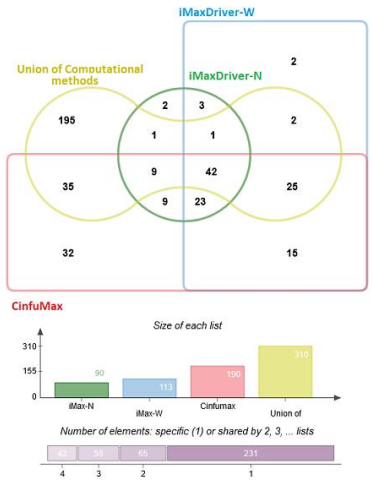

Fig. 8: The Venn diagram for predicted CDGs using CinfuMax and other computational and network-based methods in colon cancer.

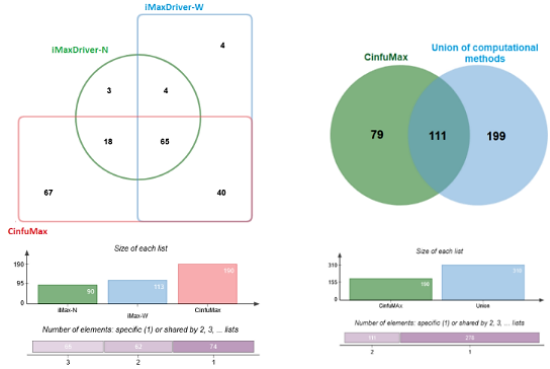

Fig. 9: The Venn diagram for predicted CDGs using CinfuMax and (A). Other network-based methods, (B) the union of all other computational methods in colon cancer.

C. *Lung Cancer Network*

F-measure values for CinfuMax and other computational and network-based methods in Lung cancer are shown in Fig. 10 . It is evident that CinfuMax surpasses all computational methods in terms of F-measure and has the highest value among network methods after iMaxDriver-W. Although its F-measure score is 0.03 less than that of the iMaxDriver-W method, it outperforms significantly in terms of the number of predicted drivers. As depicted in Fig. 10, CinfuMax has identified 182 drivers, representing the highest number of drivers compared to the previous computational and network-based methods.

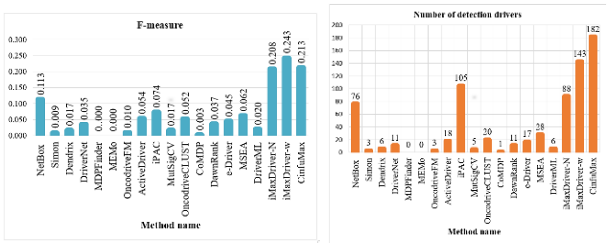

Fig. 10: Performance measures of CinfuMax and other pervious methods proposed in lung cancer.

CinfuMax and other previous methods were compared based on the overlap in predicting driver genes. As shown in Fig. 11, CinfuMax was able to identify 149 genes identified by other methods. In addition, CinfuMax identified 33 unique genes that were not identified by any of the previous computational and network-based methods. Furthermore, we assessed the proposed method's overlap in predicted drivers separately with computational and network methods.

As illustrated in the Venn diagram in Fig. 12, CinfuMax identified 58.44% (135) of genes predicted by other network-based methods and 47 unique genes not predicted by any network-based methods. Additionally, compared to computational methods, CinfuMax identified 130 unique genes not predicted by any previous computational methods. In addition to comparing the proposed method with other previous methods in terms of performance and number of predicted drivers, we also compared it with two previous methods based on linear threshold.

The ROC diagram for the CinfuMax method, which is based on the independent cascade model, and the two methods iMaxDriver-N and iMaxDriver-W, which are based on the linear threshold model, are depicted in Fig. 13.

The results show that in all three cancer networks, the ROC diagrams are almost the same, but the independent cascade model is significantly better than the linear threshold-based methods in terms of the number of predicted drivers as well as the number of unique drivers (Fig. 14).

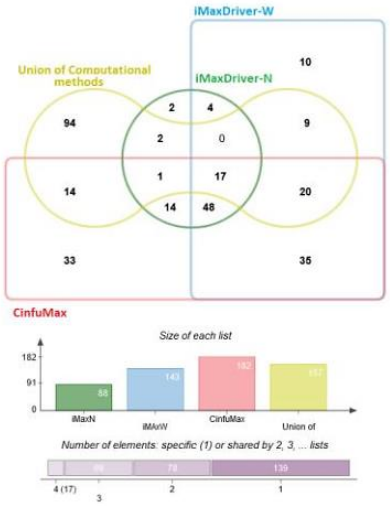

Fig. 11: The Venn diagram for predicted CDGs using CinfuMax and other computational and network-based methods in lung cancer.

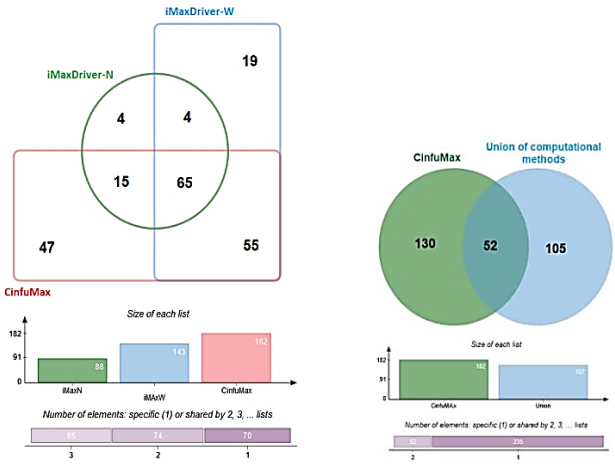

Fig. 12: The Venn diagram for predicted CDGs using CinfuMax and (A). Other network-based methods, (B) the union of all other computational methods in lung cancer.

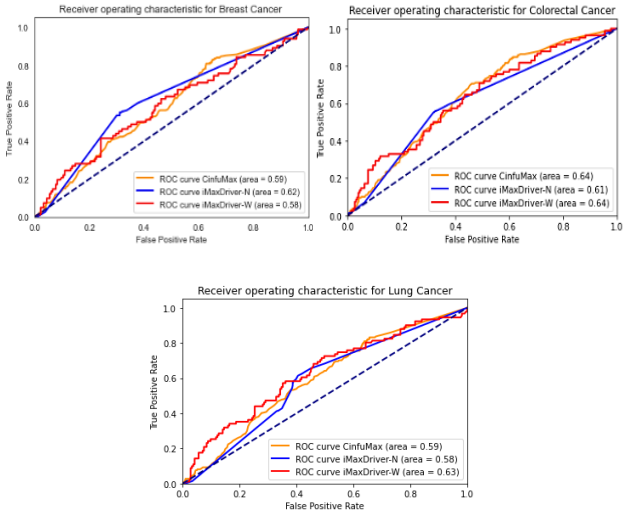

Fig. 13: The ROC diagram for the CinfuMax method and other LT based methods.

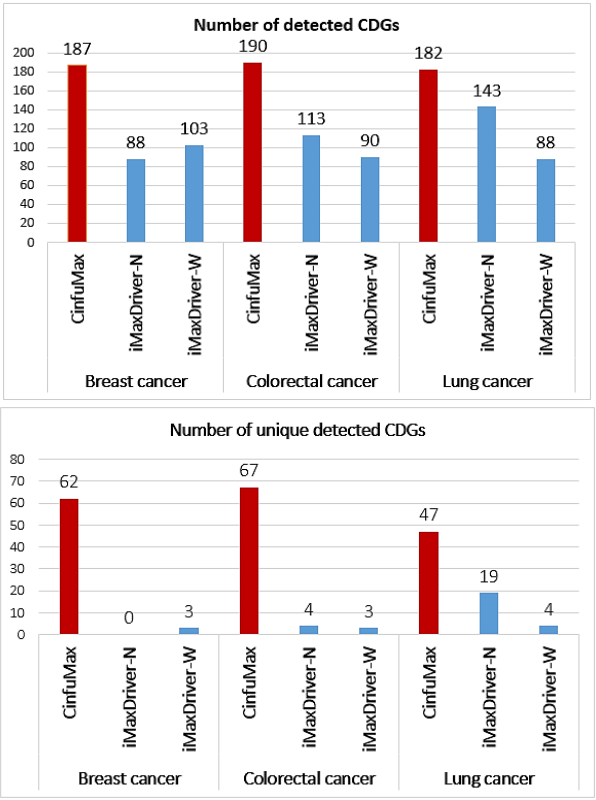

Fig. 14: Compare of the driver genes predicted and unique drivers predicted by CinfuMax and pervious LT based method.

Many researchers working in fields such as bioinformatics and biomathematics at some point face the well-known question of whether results need to be empirically confirmed [32]. Within the field of computational biology, 'experimental validation' refers to the process of replicating a scientific discovery obtained through computational methods by conducting investigations that do not heavily depend on computational resources. This procedure entails gathering additional evidence to bolster the conclusions drawn from the computational study. The integration of orthogonal sets of computational and experimental methods in a scientific study can enhance confidence in its results. However, the term 'experimental validation' may pose a hindrance to this collaborative effort [33]. As similar articles in this field also lack laboratory and experimental confirmation [5], [18], [34]-[36].

Prediction of Genes with The Fastest Speed of Influence

In addition to evaluating the proposed model in terms of the number of diagnostic drivers and comparing its performance, we also identified the driver genes with the highest influence speed compared to other drivers. Influence speed indicates which driver has achieved the highest spread rate earlier in the independent cascade model. This aspect has not yet been explored in diffusion-based methods for identifying cancer driver genes. Identifying driver genes with higher influence speed can be crucial for molecular therapy and targeted drug prescribing. The distribution diagram for iterations in all three tumor tissues is shown in Fig. 15. Nodes with low influence rates were excluded for better visualization. The diagram represents iteration values on the horizontal axis and influence values on the vertical axis, illustrating the nodes separately by iteration.

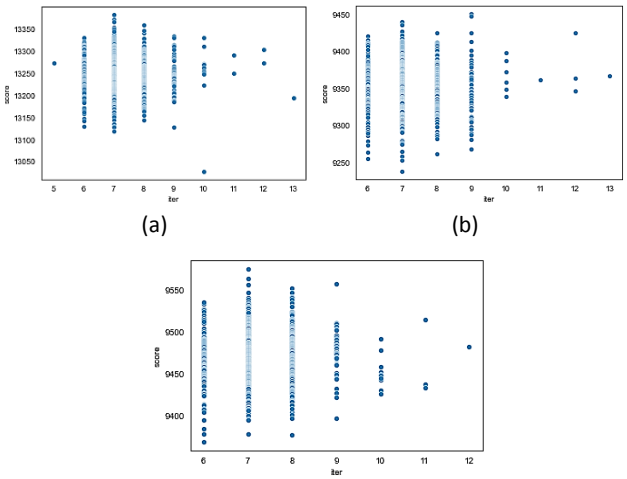

Fig. 15: Distribution diagrams related to the influence speed of each gene according to the algorithm iteration in CinfuMax algorithm. (a) colon cancer, (b) breast cancer, (c) lung cancer.

The cancer driver genes predicted by the CinfuMax method were classified into two categories based on their influence speed rates.

The first category includes genes classified as drivers by the proposed algorithm, which achieve their diffusion score faster than other drivers but do not have the highest diffusion score. These genes are cancer drivers and infect their target genes faster than other drivers, but their infection is less than that of genes with highest diffusion scores. These genes can be therapeutically significant in the early stages of the disease and in preventing the spread of cancer.

The list of these genes in the three cancerous tissues of the breast, lung, and colon is provided in Table 3. For instance, in lung cancer, 18 genes exhibit the highest influence speed in the network, with 17 of these genes not identified by any previous computational or network-based methods. In colon cancer, 15 genes have the highest penetration rate in the network, with 8 of these genes not identified as drivers by any previous computational or network-based methods. In breast cancer, for example, the AFF1 and ZNF384 genes, identified only by the CinfuMax method, are among the driver genes with the highest influence speed. The role of AFF1 in cancer metastasis has been confirmed by [32]. We have also depicted the EGO networks corresponding to

the two unique genes with the highest influence speed in breast cancer in Fig. 16 and Fig. 17. An ego network represents a set of regulatory relations from the perspective of a focal gene. As shown, the AFF1 driver gene, with the highest propagation rate, initiates propagation through only three genes The AFF1 Ego network comprises 1066 nodes and 85905 edges at a depth of 3, demonstrating that the spread speed of abnormality by a gene in the network depends on the structure and general position of that gene in the regulator network, not just its local connections. A similar result is observed for another gene with high influence speed in breast cancer, NF384.

The second category comprises driver genes with the fastest influence speed and the highest diffusion score (top 5%).

These genes are after the first category genes in terms of influence speed, but they have the most infection in the network.

The list of cancer driver genes belonging to the second category is provided in Table 4. These genes can be given special attention in order to prevent metastasis and treat cancer.

Some of these driver genes have not been identified in previous methods and have only been recognized by the proposed method. Information regarding these driver genes in the three cancer networks studied is presented in Table 5.

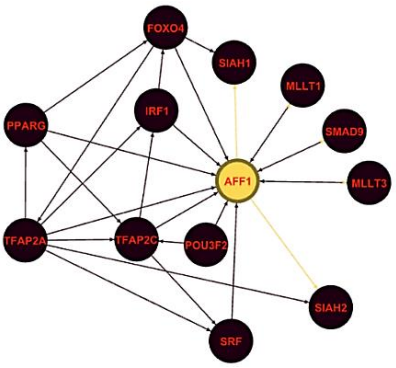

Fig. 16: AFF1 ego network up to 1 depth (13 nodes, 28 edges and Out-degree 5). This network up to depth 3 includes 1066 nodes and 85905 edges).

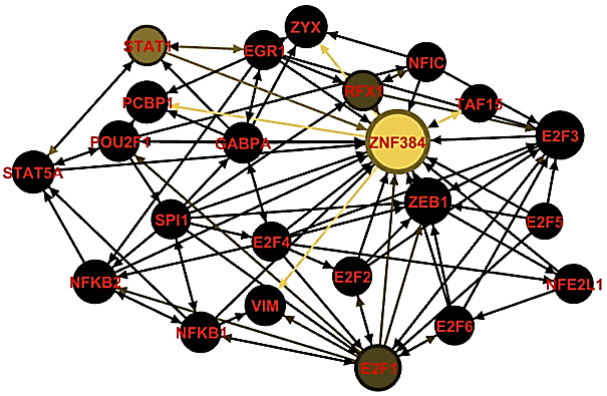

Fig. 17: ZNF384 ego network up to 1 depth (23 nodes, 91 edges and Out-degree 4) This network up to depth 3 includes 10852nodes and 86021edges.

Table 3: List and number of driver genes with the highest influence speed identified by CinfuMax

Tumor tissue	# CDGs	# Unique CDGs (Compare with other methods)	# Unique CDGs (Compare with network-based methods)	# Unique CDGs (Compare with computational methods)	Name of unique CDGs
Colon	15	8	13	8	SRSF3, FUBP1, RAD51B, AFF4, HOXA13, PLAG1, TAL2, THRAP3
Breast	15	5	11	3	HOXC11, ELF4, RAD51B, FUBP1, TNFAIP3
Lung	18	17	17	18	ERCC3, RAD51B, TMF1, HCLS1, HAX1, HMG20B, PDLIM4, ZFP36, FOXP3, SNAI1, HMGN2, CNOT8, GTF3C4, ARNTL2, CTDSPL, BAZ1B, DOT1L,

Table 4: List of genes with the highest diffusion speed and highest influence score identified by CinfuMax

Tumor tissue	# CDGs	# Unique CDGs (Compare with all other methods)	# Unique CDGs (Compare with network-based methods)	# Unique CDGs (Compare with computational methods)	Name of unique CDGs
Colon	38	0	1	12	SMARCB1
Lung	35	0	0	31	-
Breast	42	1	1	17	TAL1

Table 5: Name of genes with the highest diffusion speed and highest influence score identified by CinfuMax

Tumor tissue	Cancer driver genes
Breast	ESR1, FOXO4, FLI1, STAT3, NCOA1, TP53, DAXX, NCOR2, TAL1, RUNX1T1, ERG, KLF6, CTCF, LEF1, NCOR1, RARA, GATA1, MAX, MYOD1, RB1, FOXO1, TCF12, ARNT, HMGA1, MYB, SMAD3, NFE2L2, BCL6, WT1, CREB1, JUN, NFKB2, HMGA2, PML, EP300, SMAD4, ATF1, BRCA1, STAT6, SMARCA4, PAX5, AR
Colon	FOXO1, ABL1, CREB1, SMAD2, SMAD3, SMAD4, MITF, RB1, CTNNB1, ATF1, JUN, REL, PATZ1, CREBBP, ZBTB16, MAF, MYOD1, BRCA1, ESR1, TP53, DDX5, TCF7L2, CTCF, MAX, HNF1A, GATA2, STAT5B, CUX1, ARNT, NCOR2, SMARCB1, LMO2, CDX2, PAX5, STAT3, EP300, PPARG, CEBPA
Lung	EP300, CEBPA, REL, CREB1, HNF1A, FOXO4, PPARG, MYOD1, ABL1, GATA2, JUN, MYB, FOXO1, CREBBP, BCOR, GATA3, STAT5B, RB1, STAT3, CTCF, AR, NFKB2, PML, SMAD2, ZBTB16, RARA, ARNT, LMO2, BCL6, NCOR2, TCF7L2, MYC, SMARCA4, ATF1, MAX, LEF1, GATA1, TP53, CUX1, TAL1, BRCA1, NCOA2, SMARCB1, SMAD4, STAT6, TRRAP, SMAD3, ESR1, CTNNB1

Conclusion and Future Work

A method based on influence maximization was introduced to identify cancer driver genes in human gene regulatory networks, utilizing the independent cascade diffusion model. The independent cascade is one of the popular models in the problem of maximizing influence. This method is also able to classify the identified driver genes based on the influence speed. In this method, regulatory networks associated with breast, lung, and colon cancers are initially constructed using gene expression data and regulatory interactions. Subsequently, the modified independent cascade algorithm is independently executed on each cancer network. Genes are then sorted in descending order of influence scores. Based on a predefined threshold value, the genes are classified into two categories: driver and normal. Moreover, cancer drivers are further categorized into two classes based on their influence speed. The results showed that the CinfuMax has better performance in terms of F-measure and number of predicted drivers than other computational and network-based methods. Additionally, CinfuMax successfully identifies a considerable number of unique drivers that were not previously predicted by other computational and network-based methods. Therefore, it can be utilized as a complementary tool alongside other computational approaches. The proposed method exhibits superior performance compared to iMaxDriver methods, which are based on the linear threshold model. This result shows the use of independent cascade model is more appropriate than linear threshold model in the gene regulatory network to identify driver genes. One of the limitations of influence maximization models is the computational time and the selection of the initial active set (seed). To address this, an effective technique was proposed in this study to reduce execution times. The running time of the proposed algorithm was 35 minutes on a computer equipped with an Intel CORE i7 microprocessor and 8 GB of RAM, which is reasonable for influence maximization algorithms. However, future research could focus on providing methods to further reduce the execution time of the algorithm and ensuring the proper selection of seed nodes. Furthermore, it is noted that driver genes have not been previously classified based on influence speed in diffusion-based methods. Prioritizing genes based on infection and influence speed can be crucial for therapeutic purposes and the development of targeted drugs.

Author Contributions

B. Teimourpour and M. Akhavan-Safar contributed to the study design. M. Akhavan-Safar conducted the writing of the article and data processing. M. Ayyoubi performed data analysis and model implementation. B. Teimourpour critically revised the article and provided final approval of the version to be submitted. All authors have reviewed and approved the final manuscript.

Acknowledgment

The authors would like to thank the appreciate the anonymous reviewers and the editor of JECEI for their useful comments and suggestions.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Data availability

Data is available publicly at:
https://github.com/makhsafar/CinfuMax

Abbreviations

CDG	Cancer Driver Gene
IM	Influence Maximization
IC	Independent Cascade
LT	Linear Threshold

TF	Transcription Factor
TRN	Transcriptional Regulatory Network
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic acid
TP	True Positive
FP	False Positive
FN	False Negative
TN	True Negative

References

[1] F. Cheng, J. Zhao, Z. Zhao, "Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes," Briefings Bioinf., 17(4): 642-56, 2016.

[2] H. S. Jang, N. M. Shah, A.Y. Du, Z. Z. Dailey, E. C. Pehrsson, P. M. Godoy, D. Zhang, D. Li, X. Xing, S. Kim, D. O'Donnell "Transposable elements drive widespread expression of oncogenes in human cancers," Nat. Genet., 51(4): 611-617, 2019.

[3] World Health Organization, Cancers, 12 September 2018.

[4] D. Tamborero, A. Gonzalez-Perez, N. Lopez-Bigas, "OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes," Bioinf., 29(18): 2238-2244, 2013.

[5] A. Youn, R. Simon, "Identifying cancer driver genes in tumor genome sequencing studies," Bioinf., 27(2): 175-181, 2011.

[6] F. Vandin, E. Upfal, B. J. Raphael, "De novo discovery of mutated driver pathways in cancer," Genome Res., 22(2): 375-385, 2012.

[7] J. Reimand, O. Wagih, G. D. Bader, "The mutational landscape of phosphorylation signaling in cancer," Sci. Rep., 3(1): 2651, 2013.

[8] E. Porta-Pardo, A. Godzik, "e-Driver: a novel method to identify protein regions driving cancer," Bioinf. 30(21): 3109-3014, 2014.

[9] A. Gonzalez-Perez, N. Lopez-Bigas, "Functional impact bias reveals cancer drivers," Nucleic Acids Res., 40(21): e169, 2012.

[10] J. Zhao, S. Zhang, L. Y. Wu, X. S. Zhang, "Efficient methods for identifying mutated driver pathways in cancer," Bioinf., 28(22): 2940-2947, 2012.

[11] Y. Han, J. Yang, X. Qian, W. C. Cheng, S. H. Liu, X. Hua, L. Zhou, Y. Yang, Q. Wu, P. Liu, Y. Lu, "DriverML: a machine learning algorithm for identifying driver genes in cancer sequencing studies," Nucleic Acids Res., 47(8): e45, 2019.

[12] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, A. Kiezun, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," Nature, 499(7457): 214-218, 2018.

[13] M. R. Aure, I. Steinfeld, L. O. Baumbusch, K. Liestøl, D. Lipson, S. Nyberg, B. Naume, K. K. Sahlberg, V. N. Kristensen, A. L. Børresen-Dale, O. C. Lingjærde, "Identifying in-trans process associated genes in breast cancer by integrated analysis of copy number and expression data," PloS one, 8(1): e53014, 2013.

[14] E. Cerami, E. Demir, N. Schultz, B. S. Taylor, C. Sander, "Automated network analysis identifies core pathways in glioblastoma," PloS one, 5(2): e8918, 2010.

[15] H. P. Hou, J. Ma, "DawnRank: discovering personalized driver genes in cancer," Genome Med., 6: 1-6, 2014.

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

[17] D. Arneson, A. Bhattacharya, L. Shu, V. P. Mäkinen, X. Yang, "Mergeomics: a web server for identifying pathological pathways, networks, and key regulators via multidimensional data integration," BMC Genomics, 17: 1-9, 2016.

[18] A. Bashashati, G. Haffari, J. Ding, G. Ha, K. Lui, J. Rosner, D. G. Huntsman, C. Caldas, S. A. Aparicio, S. P. Shah, "DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer," Genome biol., 13(12): 1-4, 2012.

[19] M. Rahimi, B. Teimourpour, S. A. Marashi, "Cancer driver gene discovery in transcriptional regulatory networks using influence maximization approach," Comput. Biol. Med., 114: 103362, 2019.

[20] T. M. Liggett, Interacting particle systems. New York: Springer; 1985 Feb 13.

[21] J. Goldenberg, B. Libai, E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," Mark. Lett., 12: 211-223, 2001.

[22] W. O. Kermack, A. G. McKendrick, "Contributions to the mathematical theory of epidemics--I. 1927," Bull. Math. Biol., 53(1-2): 33-55, 1991.

[23] D. Kempe, J. Kleinberg, É. Tardos, "Influential nodes in a diffusion model for social networks," in Proc. Automata, Languages and Programming (ICALP 2005): 1127-1138, 2005.

[24] A. Blais, B. D. Dynlacht, "Constructing transcriptional regulatory networks," Gene. Dev., 19(13): 1499, 2005.

[25] C. A. Jackson, D. M. Castro, G. A. Saldi, R. Bonneau, D. Gresham, "Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments," elife, 9: e51254, 2020.

[26] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, N. M. Luscombe, "A census of human transcription factors: function, expression and evolution," Nat. Rev. Genet., 10(4): 252-263, 2009.

[27] Z. P. Liu, C. Wu, H. Miao, H. Wu, "RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse," Database, bav095, 2015.

[28] E. Clough, T. Barrett, "The gene expression omnibus database," Statistical Genomics: Methods and Protocols: 93-110, 2016.

[29] I. F. Chung, C. Y. Chen, S. C. Su, C. Y. Li, K. J. Wu, H. W. Wang, W. C. Cheng, "DriverDBv2: a database for human cancer driver gene research," Nucleic Acids Res., 44(D1): D975-D979, 2016.

[30] K. Tomczak, P. Czerwińska, M. Wiznerowicz, "Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," Contemp. Oncol., 19(1A): 68-77, 2015.

[31] M. R. Grever, S. A. Schepartz, B. A. Chabner, "The national cancer institute: cancer drug discovery and development program," in Proc. InSeminars in Oncology, 19(6): 622-638, 1992.

[32] Q. X. Meng, K. N. Wang, J. H. Li, H. Zhang, Z. H. Chen, X. J. Zhou, X. C. Cao, P. Wang, Y. Yu, ZNF384–ZEB1 feedback loop regulates breast cancer metastasis," Mol. Med., 28(1): 111, 2022.

[33] M. Jafari, Y. Guan, D. C. Wedge, N. Ansari-Pour, "Re-evaluating experimental validation in the Big Data Era: a conceptual argument," Genome Biol., 22(1): 1-6, 2021.

[34] Y. Lu, Y. Wang, N. Sheng, H. Wang, Y. Fu, Y. Tian, "RDDriver: A novel method based on multi-layer heterogeneous transcriptional regulation network for identifying pancreatic cancer biomarker," in Proc. 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): 497-502, 2022.

[35] F. Dietlein, D. Weghorn, A. Taylor-Weiner, A. Richters, B. Reardon, D. Liu, E. S. Lander, E. M. Van Allen, S. R. Sunyaev, "Identification of cancer driver genes based on nucleotide context," Nat. Genet., 52(2): 208-218, 2020.

[36] R. Gillman, M. A. Field, U. Schmitz, R. Karamatic, L. Hebbard, "Identifying cancer driver genes in individual tumours," Comput. Struct. Biotechnol. J., 21: 5028-5038, 2023.

Biographies

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 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, Machine learning, Information systems and Social Network Analysis.

- Email: akhavansaffar@pnu.ac.ir
- ORCID: 0000-0002-7337-712X
- Web of Science Researcher ID: NA
- Scopus Author ID: 57221234257
- Homepage: https://cv.pnu.ac.ir/HomePage/akhavansaffar

Babak Teimourpour is an Associate professor of Information Technology Engineering at the School of Industrial and Systems Engineering of Tarbiat Modares University (TMU). He obtained his Ph.D. in Industrial Engineering from Department of Industrial Engineering, Tarbiat Modares University (TMU), Tehran, Iran. He teaches Ph.D. and M.S. level courses. His research interests include 'Data Mining' and 'Social Network Analysis'. His team won the Iran Data Mining Cup in 2010.

- Email: b.teimourpour@modares.ac.ir
- ORCID: 0000-0002-9286-2286
- Web of Science Researcher ID: NA
- Scopus Author ID: 36115355600
- Homepage: https://www.modares.ac.ir/~b.teimourpour

Mahboube Ayoubi is a master student of data science in Tarbiat modares University, Tehran, Iran. She has a M.Sc. In biostatistics from Isfahan medical University, Isfahan, Iran. She's research interests include Social Network analysis, Machine Learning, Deep Learning and Survival Analysis.

- Email: m.aubi_68@yahoo.com
- ORCID: 0000-0001-6773-7996
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
M. Akhavan-Safar, B. Teimourpour, M. Ayoubi, “CinfuMax: An influence maximization-based model for predicting cancer driver genes in gene regulatory networks,” J. Electr. Comput. Eng. Innovations, 12(2): 373-386, 2024.

DOI: 10.22061/jecei.2024.10026.673

URL: https://jecei.sru.ac.ir/article_2095.html