Dijkstra’s based Identification of Lung Cancer Related Genes using PPI Networks

Praveen Tumuluru
M.Tech, (Ph.D)
Research Scholar,
Department of
Computer Science
and Engineering
GITAM University,
Visakhapatnam, India

Bhramaramba Ravi, PhD
B.Tech, MS, Ph.D
Associate Professor
Department of
Information Technology,
GITAM University,
Visakhapatnam, India

ABSTRACT

Biomedical area of research has grown dynamically for identification of various diseases and prediction of disease, among the most cancer is vital and critical disease caused from various sources of gene mutation. Cancer is one of the most common diseases in the developed world. Cancer arises from the mutation of a normal gene. Mutated genes that cause cancer are called oncogenes The Lung cancer is one of the leading causes of cancer mortality worldwide. The main types of lung cancer are small cell lung cancer (SCLC) and non small cell lung cancer (NSCLC). In this work, a computational method was proposed for identifying lung-cancer- related genes with a shortest path approach in a protein-protein interaction (PPI) network using R tool to set the computation. Based on the PPI data from STRING, a weighted PPI network was constructed. 54 NSCLC- and 84 SCLC-related genes were retrieved from associated KEGG pathways.Then the shortest paths between each pair of these 54 NSCLC genes and 84 SCLC genes were obtained with Dijkstra’s algorithm. Some of the shortest path genes have been reported to be related to lung cancer. Intriguingly, the candidate genes identified from the PPI community contained extra most cancers genes than the ones recognized from the gene expression profiles. Furthermore, these genes possessed greater purposeful similarity with the recognized cancer genes than those identified from the gene expression profiles. This work proved the performance of the proposed technique and showed promising consequences.

Keywords

Data related to Proteins, Weighted PPI Data, Computational algorithm, Resultant Protein information.

1. INTRODUCTION

Lung cancer, also known as lung carcinoma, is a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung. If left untreated, this growth can spread beyond the lung by the process of metastasis into nearby tissue or other parts of the body. Most cancers that start in the lung, known as primary lung cancers, are carcinomas. Two main types of lung cancer are non-small cell lung cancer (NSCLC), which accounts for 80%–85%, and small cell lung cancer (SCLC), which accounts for around 20% of all cases. However, the SCLC has an extraordinarily high degree of metastasis and a strong association with smoking. Diagnosis and treatment at the early stage of the disease process could reduce fatalities and increase the probability of d functions and thus may participate in the same pathways. This so-called “guilt by association” rule was initially proposed by Nabieva et al.. This rule could also be utilized to identify novel cancer-related genes[1]. The most common symptoms are coughing (including coughing up blood), weight loss, shortness of breath, and chest pains.

Search Tool for the Retrieval of Interacting Genes (STRING) is an online database resource that provides both predicted and experimental interaction information with a confidence score. It has been shown that proteins with short distances between each other in the PPI network tend to have the same biological functions , and interactive neighbors are prone to have the same biological functions as non interactive ones . The possible reason is that the query protein and its interactive proteins might form a protein complex to exert a particular function or might participate in the same pathways is ease-free survival. Therefore, it is meaningful.

Though great successes have been achieved for gene function prediction and identification of novel cancers related genes with the application of the high-throughput data, yet high-throughput data is not error free. This work, proposed a computational method for identifying lung-cancer-related genes based on PPI network constructed from STRING. 54 NSCLC and 84 SCLC related genes were retrieved from associated KEGG pathways. Then, Dijkstra’s algorithm was employed to obtain the shortest paths between each pair of the 54 NSCLC and 84 SCLC genes[2]. All the genes present on the shortest paths were extracted and analyzed. Several of these genes have been reported to be related to lung cancer. However, some of them were not previously reported. Therefore, there are probably novel lung cancer-related genes and have the potential to be biomarkers for diagnosis of lung cancer.

2. MATERIALS AND METHODS

Cancer is a genetic disease; it results from mutations in somatic cells. To understand it at a molecular level, and need to identify the relevant mutations and to discover how they give rise to cancerous cell behavior. Finding the mutations is easy in one respect: the mutant cells are favored by natural selection and call attention to themselves by giving rise to tumors. The hard task then begins: how are the genes with the carcinogenic mutations to be identified among all the other genes in the cancerous cells? A similar needle-in-haystack problem arises in any search for a gene underlying a given mutant phenotype, but for cancer the task is particularly complex. A typical cancer depends on a whole set of mutations—usually a somewhat different set in each individual patient—and introduction of any single one of these into a normal cell is usually not enough to make it cancerous. This genetic cooperation makes it hard to test the significance of mutations on which suspicion falls. To make matters worse, most cancer cells will contain mutations that are accidental by-products of genetic instability, and it can be difficult to distinguish these from the mutations that have a causative role in the disease.

1. This may not provide truthful input data.
2. Identification of the genes is difficult in this method.

2.1 Proposed Method

Protein-Protein interaction (PPI) data has been widely utilized to annotate and predict the gene function assuming that interaction proteins possess the similar or identical functions and thus may participate in the same pathways. It has been shown that proteins with short distances between each other in the PPI network tend to have the same biological functions ,and interactive neighbors are prone to have the same biological functions as non-interactive ones" -guilt by association rule.

The initial weighted PPI network was constructed based on data from STRING .Shortest paths are calculated between every pair of proteins in the network by Dijkstra's algorithm. Finally, all the proteins present on the shortest paths were ranked according to their betweenness and extracted to analyze. Advantages of proposed system are that identification of genes becomes easier when compared to other methods. The shortest paths can be calculated easily using this method [3]. The plotting of graphs can be made easier by using the R Programming language. The truthful data is provided in this method with the lung cancer related genes one can identify whether there is any possibility of getting the disease.

First step is to extraction of genes from KEGG Database. With the genes list the Constructing of the PPI network by STRINGDB. Further applying the shortest path algorithm[4].

2.2 Implementation Steps:

1. Compile all 54 genes existing in the human non-small cell lung cancer (NSCLC) pathway and 84 genes in the small cell lung cancer (SCLC) pathway from KEGG.
2. Construct the protein-protein interaction (PPI) network by using the weights extracted from the STRING database
3. Then identify the shortest path between the every pair of proteins by using the Dijkstra's algorithm
4. And then calculate the betweenness at the nodes.
5. Finally, if the threshold value is less than the betweenness value then ignore it.
6. And if the threshold value is equal to the betweenness then the genes are ranked.
7. All the genes present on the shortest paths were extracted and analyzed.

2.3 Algorithm and Work flow

The algorithm exists in many variants; Dijkstra's original variant found the shortest path between two nodes, but a more common variant fixes a single node as the "source" node and finds shortest paths from the source to all other nodes in the graph, producing a shortest-path tree.

For a given source node in the graph, the algorithm finds the shortest path between that node and every other. It can also be used for finding the shortest paths from a single node to a single destination node by stopping the algorithm once the shortest path to the destination node has been determined. For example, if the nodes of the graph represent cities and edge path costs represent driving distances between pairs of cities connected by a direct road, Dijkstra's algorithm can be used to find the shortest route between one city and all other cities. As a result, the shortest path algorithm is widely used in network routing protocols, most notably IS-IS and Open Shortest Path First (OSPF)[5]. It is also employed as a subroutine in other algorithms such as Johnson's.

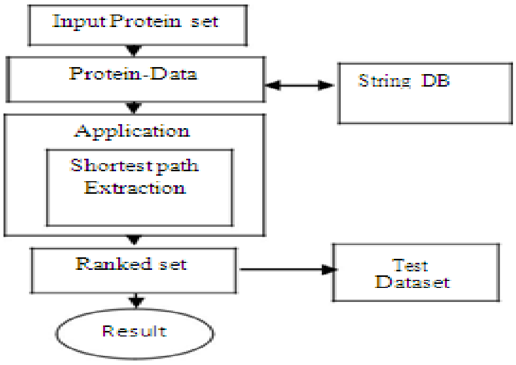

Fig 1: Flow of Operations to obtain Results

2.4 Workflow

The Protein set which was alleged to be cause to the cause of cancer was taken from the research done previously these protein set was an input to the system. The protein set then fed to the stringDB to get the required information about the proteins and their networks are taken in Teb-delimited Format. The taken format then given as input to the Application. The Processing was done on data that was given through the input[6-7].

Table 1: Gene symbols and their appropriate Ensemble protein ID's

Gene symbol	Ensemble protein ID
AKT1	ENSP00000270202
AKT2	ENSP00000375892
AKT3	ENSP00000263826
ARAF	ENSP00000366244
BAD	ENSP00000309103
BRAF	ENSP00000288602
CASP9	ENSP00000330237
CCND1	ENSP00000227507
CDK4	ENSP00000257904

By compiling the Dijkstra's algorithm in the application the shortest paths are compiled and the values are taken down. The complete application was developed under R programming which will provide a wide range of methods and interfaces to compute and represent the data to be displayable in a clear and straight view. The resulted protein list is then derived in to the David tool to generate their analogies and the remaining protein set is generalized to form the required protein data set. The resultant data is then made to analysis to obtain the Resultant protein information. This information is very helpful in finding the further more set of proteins that are closely capable of causing cancer. Which can be avoided to prevent the cause of cancer. The entire workflow will be compiled and executed by the application that is developed using the R programming.

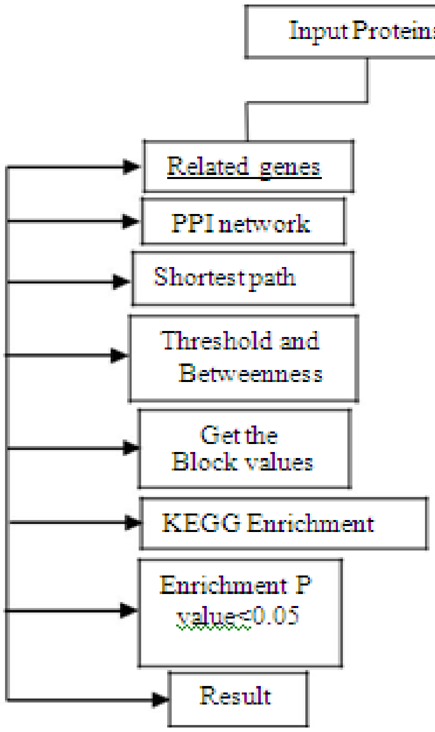

Fig 2 : Process of getting related genes

The initial stages of application are worked out to refine the data and to organize in a particular order to have complete utilization of the protein data. The phases then change their way to process the computed data and to produce the network with the data that was preprocessed in earlier phases of workflow. Constructing a network was completely made easy by the utilization of the packages in the R so that the task become easy to code the module to develop a network from the information.

The package that used to plot a network is igraph this is an package in R which help to construct and to address the information which was transformed in to Network. The parameter that are using to evaluate the shortest path is the weight between the two nodes which are equivalent and the approximation of the shortest path was done to last but one step. The final normalization of the values is quiet unavoidable the process comes up with the quantization of the results which are come under the computation of the shortest path. This Quantization done through the R package. Top go is an another package utilized to facilitate the automation in the process getting results through interactivity. Package Stringdb is an another package is used to get the results automatically through the single standalone application.

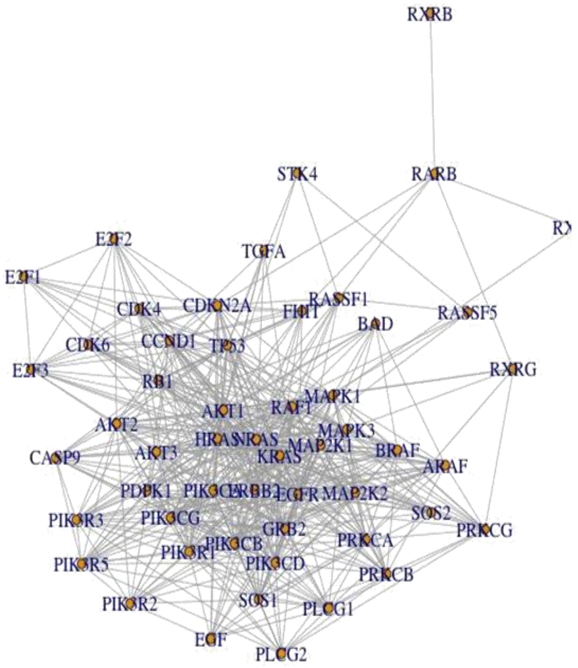

Fig 3: Network formed using R programming.

3. RESULTS AND DISCUSSION

In this study, the proposed computational method is based on a protein-protein interaction network to identify cancer related genes. Next, this method was applied to lung cancer related genes, to find the shortest paths between 54 NSCLC and 84 SCLC genes in the protein-protein interaction network constructed based on STRING data and selected the 25 and 38 genes with a significant value for NSCLC and SCLC, respectively.

Analysis of these shortest path genes indicates that some of these genes, such as ESR1, FDXR, ABCA1, IRS1, HSP90AA1, FOXM1, and IGBP1 are related to lung cancer. In addition, the candidate genes of lung cancer identified in our study contain more cancer genes than those identified from gene expression profiles. Moreover, it is revealed that our candidate genes have greater functional similarity with the cancer genes than those identified from gene expression profiles. These candidate genes may be worth experiment validation and further research. It is expected that this method is useful in predicting novel cancer-related genes and has widespread use in cancer research. The Prevention measures can be done more elaborately by having this relative items which are derivatives of causing cancer. The probability function here used to quantize will also help in the guess of treating the patient with unique type[10]. As all know that, cancer will follow a stereotype it will form by of the other possible combinations[8-9]. Search Tool for the Retrieval of Interacting Genes (STRING) is an online database resource that provides both predicted and experimental interaction information with a confidence score. It has been shown that proteins with short distances between each other in the PPI network tend to have the same biological functions , and interactive neighbors are prone to have the same biological functions as non interactive ones has been developed.

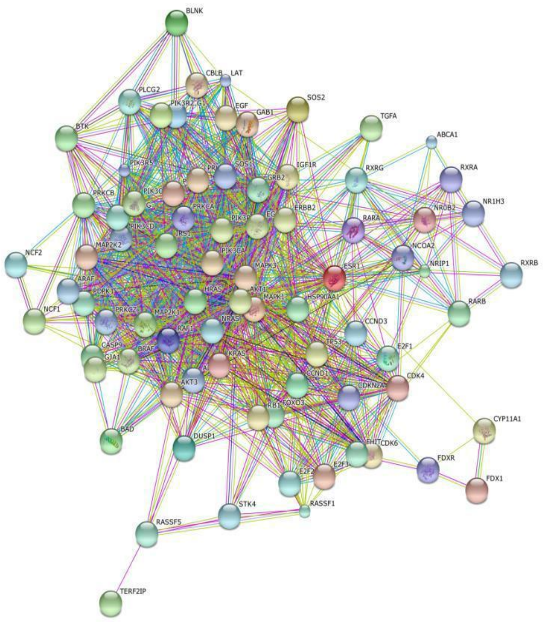

Fig 4: Network of 53 NSCLC Genes retrieved from stringDB

The structures in the figure are taken from the stringDB that which the input information was taken. The computational method was same as follows to different entities which can be realized to form a network and can able to produce the relation based on the network by which it was built on.The R tool also provide us high efficient packages which will provide more computation with less code and low space complexity as well as time complexity.

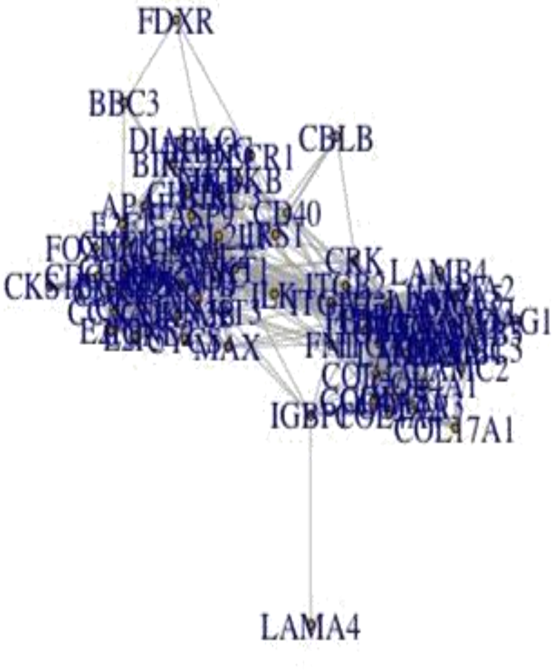

Fig 5: Graph in Extracting Nodes

4. CONCLUSION

The ESR1, FDXR, ABCA1, IRS1, HSP90AA1, FOXM1, and IGBP1 are related to lung cancer. In addition, the candidate genes of lung cancer identified in this study contain more cancer genes than those identified from gene expression profiles. Moreover, it is revealed that the candidate genes have greater functional similarity with the cancer genes than those identified from gene expression profiles. These candidate genes may be worth experiment validation and further research. Here by, this work done a method of extracting the additional related protein which was formed by the network that was built on the parameters that all ones commonly have. Shortest path approach is one which relied on to get the results. By this way of processing the entities one can able to form the additional related ones based on the betweenness which was evaluated through the weights between two entities. It is expected that this method is useful in predicting novel cancer-related genes and has widespread use in cancer research. The similar procedure may be applied for other cancer genes.

5. ACKNOWLEDGMENT

The first author published this paper under his Ph.D work. The first author wishes to thank the University grants commission (UGC) for extending financial support for this study, under the project – “Development of a software tool to identify Lung-cancer related genes using Protein Protein interaction Networks” with sanctioned Proposal number: 137, F.NO:4-4/2014-15 [MRP-SEM/UGC-SERO].

6. REFERENCES

[1] R. Siegel, D. Naishadham, and A. Jemal, “Cancer statistics,” A Cancer Journal for Clinicians, vol. 62,p.p10–29, 2012.

[2] J. P. van Meerbeeck, D. A. Fennell, and De Ruysscher, “DKM Small-cell lung cancer,” TheLancet, vol. 378, pp. 1741–1755.

[3] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh, “Whole-proteome prediction of protein function via graphtheoretic analysis of interaction maps,” Bioinformatics, vol. 21, supplement 1, pp. i302–i310, 2005.

[4] D. Szklarczyk, A. Franceschini, M. Kuhn et al.,“The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored,” Nucleic Acids Research, vol. 39, no. 1,pp D561–D568, 2011.

[5] R. Sharan, I. Ulitsky, and R. Shamir, “Network-based prediction of protein function,” MolecularSystems Biology, vol. 3, article 88, 2007.

[6] P. Bogdanov and A. K. Singh, “Molecular function prediction using neighborhood features,”IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 7, no. 2, pp. 208–217, 2010.

[7] Y. A. Kourmpetis, A. D. van Dijk, M. C. Bink, R. C. van Ham, and C. J. ter Braak, “BayesianMarkov Random Field analysis for protein function prediction based on network data,” PLoS One, vol. 5, article e9293, 2010.

[8] K. L. Ng, J. S. Ciou, and C. H. Huang, “Prediction of protein functions based on function-function correlation relations,” Computers in Biology and Medicine, vol. 40, no. 3, pp. 300–305, 2010.

[9] U. Karaoz, T. M. Murali, S. Letovsky et al., "Whole-genome annotation by using evidence integration in functional-linkage networks," Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. 9, pp. 2888–2893,

[10] S. Letovsky and S. Kasif, "Predicting protein function from protein/protein interaction data: a probabilistic approach," Bioinformatics, vol. 19, supplement 1, pp. i197–i204, 2003.

7. AUTHOR PROFILE

Praveen Tumuluru, received the M.Tech degree inComputer Science and Engineering from Koneru Lakshmaiah College of Engineering, Acharya Nagarjuna Univerisity, in 2008. He is a research Scholar in GITAM University working in Data Mining Techniques for Bioinformatics Protein-Protein Interaction. He has been working as Assistant Professor, in the Department of Electronics & Computer Engineering, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada since 2008.

Dr. Bhramaramba Ravi, presently working as Associate Professor in GITAM University. She has a total of 11 years of research experience and 16 years of teaching. She received her Ph.D from Jawaharlal Nehru Technological University, in 2011 and MS degree in Software Systems from BITS, in 1999. She has published 14 papers in various National and International Journals/ Conferences. Her current research interests are in the areas of Data Mining Techniques for Bioinformatics.