Journal of Engineering and Applied Sciences 12 (22): 6088-6094, 2017
ISSN: 1816-949X

Shortest Path Enhancement using Improved Bellman Ford Algorithm in PPI

[1]Tumuluru Praveen and [2]Bhramaramba Ravi
[1]Department of Electronics and Computer Engineering,
Prasad V Potluri (PVP) Siddhartha Institute of Technology, Kanuru, Vijayawada, India
[2]Department of Information Technology,
Gandhi Institute of Technology and Management (GITAM) University, Visakhapatnam, India

Abstract: The chance that a man will develop lung cancer in his lifetime is about 1 in 14 and for a woman, the risk is about 1 in 17. The lung cancer can be cured only by the surgery alone in the early stage, no chemotherapy or radiation therapy is needed. There are cases where chemotherapy does not work in the stage 4. The shortest path in Protein-Protein Interaction (PPI) plays a vital role in identifying the cancer at an early stage. The identification of the cancer gene in the PPI is prolonged. This study proposed the improved bellman ford which is the optimized and efficient way to find the shortest path in the PPI network. PPI is constructed by obtaining the data from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) and Kyoto Encyclopedia of Genes and Genomes (KEGG) database.

Key words: Chemotherapy, improved Bellman Ford algorithm, lung cancer, protein-protein interaction, database, lifetime

INTRODUCTION

Lung cancer is a malignant lung tumor characterized by uncontrolled cell growth in tissues of the lung. Causes of lung cancer include smoking, second-hand smoke, exposure to certain toxins and family history (Wu *et al*., 2016; Zhang *et al*., 2013). Two major types of lung cancer are non-small cell lung cancer which accounts for 80-85% and small cell lung cancer which accounts for around 20% of all cases (Chen *et al*., 2011; Doungpan *et al*., 2016). Lung cancer is second most common cancer and it is by far the leading cause of cancer death among both men and women, about 1 out of 4 cancer deaths are from lung cancer (Zhang *et al*., 2012). Each year, more people die of lung cancer than of colon, breast and prostate cancers combined. Tobacco smoking is by far the main reason to lung cancer and radon which is from the decay product of uranium is the second most common factor for lung cancer. Protein to Protein Interactions (PPI) have to be analyzed to understand the change in molecular process related to different diseases mainly cancer (Qian *et al*., 2017).

Dijkstra's algorithm is one of the most famous algorithm for computing the shortest paths between nodes in PPI (Fontaine *et al*., 2015; Arkin, 2005; Tumuluru and Ravi, 2017). 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 find the shortest route between one city and all other cities. Dijkstra's algorithm doesn't work for the node with negative values and it consumes ot of time to collect the resources (Guda *et al*., 2009; Massanet-Vila *et al*., 2012). The Bellman Ford algorithm is another algorithm that finds shortest paths from a single source vertex to all of the other vertices in a weighted digraph. The Bellman Ford algorithm is capable of handling the graph which has edges of negative values. It wastes time for searching negative edges in the PPI which didn't have the negative edges.The improved Bellman Ford algorithm has high efficiency over the traditional Dijkstra's algorithm and Bellman Ford algorithm and also reduces the space requirement

Literature review: Yu *et al*. (2013) found the Pearson Correlation Coefficient (PCC) was determined to measure the correlation of expression between two proteins. The Support Vector Machine (SVM) Model was used to predict PPIs. By analyzing dynamic protein subnetworks, they found that the proteins of such subnetwork took part in cancer related biological processes. Hence, the researcher found that the PPI take part in the cancer but didn't identify the genes responsible for cancer.

Guda *et al*. (2009) and Sable and Jois (2015) have compared PPIs from different cancer types using the

Corresponding Author: Tumuluru Praveen, Department of Electronics and Computer Engineering,
Prasad V Potluri (PVP) Siddhartha Institute of Technology, Kanuru, Vijayawada, India

frequency of common GO terms between two partner proteins in an interaction. PPIs associated with differentially expressed cancer genes were used to create networks with the application of Cytoscape program. Comparison of protein interaction networks found the group of genes that are regulated uniformly across the cancer. These cancer types and those regulated only in a particular cancer indicate their importance in that specific cancer.

Li *et al*. (2013) have used Dijkstra's algorithm to identify the shortest path between each protein pair corresponding to the 54 NSCLC and 84 SCLC genes in the PPI network, these shortest path genes indicate that some of these genes were related to lung cancer, namely ESR1, FDXR, ABCA1, IRS1, HSP90AA1, FOXM1 and IGBP1. The identification of the genes that are related to lung cancer in 158 genes takes more time to compile.

Yu *et al*. (2013) work clearly showed that the proteins of such subnetwork took part in cancer related biological processes. Guda *et al*. (2009) have proved that different genes are responsible for the different cancer. Li *et al*. (2013) have found the genes that are responsible for the small cell lung cancer and non-small cell lung cancer. To overcome this problem, the proposed system made a method efficient to implement.

MATERIALS AND METHODS

Dijkstra algorithm is a sequential access algorithm but poorly suited for parallel architecture whereas the Bellman Ford algorithm is suited for parallel execution but this feature comes at a slow process. The improved Bellman Ford algorithm is efficient over both Dijkstra algorithm and Bellman Ford algorithm and also reduces the space. In the improved Bellman Ford algorithm, flag f is introduced and it assigns to all vertices. The vertices which are to be processed next is set as 1 and all other vertices are set to zero. The vertices remain zero until its shortest distance is to be found. In this way, unwanted iteration and compression are removed. The graph coordinates format is used to store which is space efficient. It stores the value references to non-zero value which reduces the computational time. The few iterations value of the distance of the vertices is needed to be updated.

Dataset collection: The information about the PPI of the individual cancer gene is present in the Search Tool for the Retrieval of Interacting Genes (STRING) database. The KEGG pathway has been used to construct a PPI network of cancer gene. The STRING database has information from many sources like experimental data, computational prediction methods and public text collections. It is freely accessible and it is regularly updated and the resource also serves to highlight functional enrichments in user-provided list of proteins, using a number of functional classification systems such as GO, Pfam and KEGG.

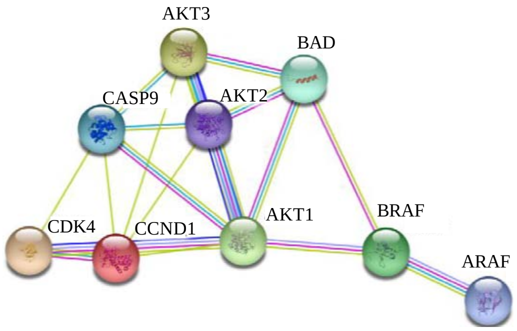

Fig. 1: Represent the PPI connection of genes

AKT1, AKT2, AKT3, ARAF, BAD, BRAF, CASP9, CCND1 and CDK4 are responsible for non-small lung cancer as shown in Fig. 1. AKT1, AKT2, AKT3, CASP9, CCND1 and CDK4 are responsible for small cell lung cancer. PPI networks are an important factor in the system-level understanding of cellular processes. That network can be used for filtering and assessing functional genomics data and for providing an insightful platform for annotating structural, functional and evolutionary properties of proteins. Exploring the predicted interaction networks can give new directions for further experimental research and provide cross-species predictions for efficient interaction mapping. It is like other databases that store protein association data STRING imports data from the experimentally derived PPI through literature curation. STRING also stores computationally predicted PPI from text mining of scientific texts, interactions computed from genomic features and interactions transferred from model organisms based on orthology. These predicted or imported PPI are benchmarked against a common reference of partnership as annotated by KEGG (Kyoto Encyclopedia of Genes and Genomes).

The data are weighted, integrated and a confidence score is obtained for all protein interactions. There are two types of STRING, they are the protein-mode and the COG-mode and predicted interactions are propagated to proteins in other organisms for which interaction has been described by inference of orthology. The results of the various computational predictions can be inspected from different designated views. A web interface is available to access the data and to give a quick overview of the proteins and their interactions. A plug-in for cytoscape to use STRING data is available. Another possibility to access data STRING is to use the Application Programming Interface (API) by constructing a URL that contains the request. KEGG (Kyoto Encyclopedia of

Genes and Genomes) is a collection of databases dealing with genomes, biological pathways, diseases, drugs and chemical substances. KEGG is utilized for bioinformatics research and education including data analysis in genomics, metagenomics, metabolomics and other omics studies, modeling and simulation in systems biology and translational research in drug development.

The KEGG pathway database, the wiring diagram database, is the core of the KEGG resource. It is a collection of pathway maps integrating many entities, including genes, proteins, RNAs, chemical compounds, glycans and chemical reactions as well as disease genes and drug targets which are stored as individual entries in the other databases of KEGG.

The metabolism section contains aesthetically drawn global maps to compare showing an overall picture of metabolism to regular metabolic pathway maps. The low-resolution global maps can also be used to compare metabolic capacities of different organisms in genomics studies and different environmental samples in metagenomics studies. KEGG Modules are defined as characteristic gene sets that can be linked to specific metabolic capacities and other phenotypic features, so that, they can be used for automatic interpretation of genome and metagenome data. KEGG modules in the KEGG Module database are higher-resolution, localized wiring diagrams, representing tighter functional units within a pathway map such as subpathways conserved among specific organism groups and molecular complexes.

Another database that contains KEGG pathway is the KEGG BRITE database. It is a database containing hierarchical classifications of various entities like proteins, including genes, diseases, organisms, drugs and chemical compounds. While KEGG pathway is limited to molecular interactions and reactions of these entities, KEGG BRITE incorporates many different types of relationships.

The genes that are responsible for non-small cell lung cancer is AKT1 and CDK4. Small cell lung cancer genes are AKT1 and CCND1.The basic interaction unit in STRING is to provide productive functional relationship between proteins, likely contributing to a common biological purpose. Interactions are derived from multiple sources: pathway knowledge is parsed from manually curated databases, known experimental interactions are imported from primary databases, interactions are predicted de novo by a number of algorithms using genomic information as well as by co-expression analysis, automated text-mining is applied to uncover statistical and/or semantic links between proteins based on medline abstracts and a large collection of full-text articles and interactions that are observed in one organism are systematically transferred to other organisms, via. pre-computed orthology relations. STRING centers on protein-coding gene loci-alternative splice isoforms or post-translationally modified forms are not resolved but instead collapsed at the level of the gene locus. All sources of interaction evidence are benchmarked and calibrated against previous knowledge using the high-level functional groupings provided by the manually curated Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway maps.

Algorithm: Improved Bellman Ford is based on the principle of relaxation, in which an infinity is applied to the distance is gradually replaced by more accurate values until eventually reaching the optimal solution. In this algorithm, the approximate distance to each vertex is always an overestimate of the true distance and is replaced by the minimum of its old value with the length of a newly found path. In each of these repetitions, the number of vertices with correctly calculated distances grows from which it follows that eventually all vertices will have their correct distances. This method allows the Bellman Ford algorithm to be applied to a wider class of inputs than Dijkstra.

The Bellman Ford algorithm may be improved in the process (although not in the worst case) by the observation that if an iteration of the main loop of the algorithm terminates without making any changes, the algorithm can be immediately terminated as subsequent iterations will not make any more changes. With this early termination condition, the main loop may in some cases use many fewer than |V|-1 iterations, even though the worst case of the algorithm remains unchanged. The number of iterations are given in Eq. 1:

$$|V|-1 \tag{1}$$

where, V is the vertices of the node. The Bellman Ford algorithm for process negative cycle of the node is given in Eq. 2:

$$O(|V|\times|E|) \tag{2}$$

where, E is the edge of the node (Algorithm 1).

Algorithm 1:

Pseudo code

```
//Distance (v) store distance from the source vertex and weight (u,v) is the weight adjacency matrix. F(v) is a flag store's value of vertices that are processed next
T = sparse (Weight)
for each vertex v in parallel do
    Distance (v) = ∞, F(v) = 0
end for
```

```
Distance (S) = 0, F(S) =1; //Distance of source is set to zero
for i = 2 to V
    for each edge (u,v) in parallel do
        if (F(u) = 1)
            F(u) = 0
            if(T(u,v)>0 &&Distance(v)>(distance (u)+T (u,v)))
                Distance(v) = Distance (u)+T(u,v)
                F(v) = 1
            end if
        end if
    end for
end for
```

In its most basic form, the algorithm performs at most mn of these relaxation steps but this can be improved in two ways. Figure 2 processing the vertices in a FIFO sequence in order that avoids reprocessing vertices whose candidate distance has not changed in the previous step reduces the number of relaxation steps to <n 2/3, an improvement by a factor of two for dense graphs. The second improvement, published by Yen and since repeated in several textbooks, involves partitioning the input directed graph into two directed acyclic graphs and alternating between passes of the algorithm that relax the edges in one of these two DAGs. This method reduces the number of relaxation steps to mn/2/+m, an improvement by nearly a factor of two over the original algorithm even for sparse graphs.

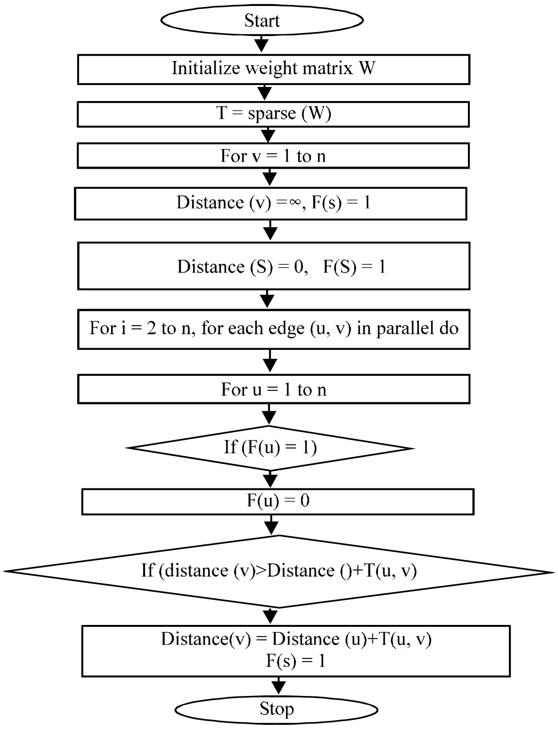
Start
Initialize weight matrix W
T = sparse (W)
For v = 1 to n
Distance (v) =∞, F(s) = 1
Distance (S) = 0,   F(S) = 1
For i = 2 to n, for each edge (u, v) in parallel do
For u = 1 to n
If (F(u) = 1)
F(u) = 0
If (distance (v)>Distance ()+T(u, v)
Distance(v) = Distance (u)+T(u, v)
F(s) = 1
Stop

Fig. 2: Flow char of improved Bellman Ford algorithm

RESULTS AND DISCUSSION

All experiments were implemented on PC with 1.8 GHz Pentium IV processor using MATLAB (Version 6.5). The different algorithm for the PPI was processed in the MATLAB. MATLAB worked on operating system of windows 7/8/10, Linux, Mac with processor of Intel and AMD x86 processor. The genes that were taken to process were AKT1, AKT2, AKT3, ARAF, BAD, BRAF, CASP9, CCND1 and CDK4.

Lung cancer is a leading cause of cancer death among men and women in industrialized countries. Figure 3 shows the Non-Small-Cell Lung Cancer (NSCLC) accounts for approximately 85% of lung cancer and represents a heterogeneous group of cancers, consisting mainly of squamous cell (SCC), Adeno (AC) and large-cell carcinoma.

Lung cancer is a leading cause of cancer death among men and women in industrialized countries. Figure 4 shows the Small Cell Lung Carcinoma (SCLC) is a highly aggressive neoplasm which accounts for approximately 25% of all lung cancer cases.

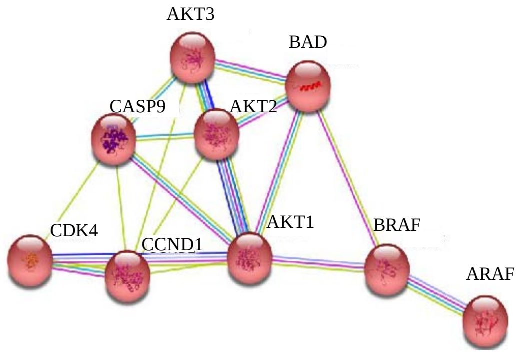

Fig. 3: Represent the gene responsible for non-small cell lung cancer

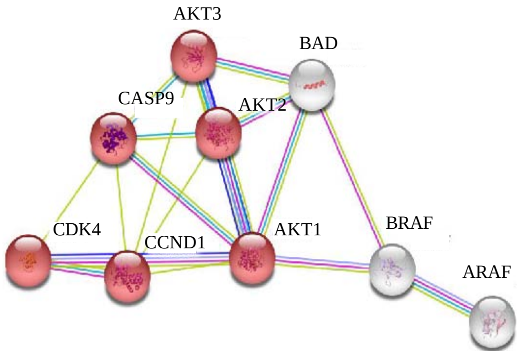

Fig. 4: Represent the gene responsible for small cell lung cancer

Table 1 shows the comparison of the Dijkstra, Bellman Ford and Improved Bellman Ford in time. Bellman algorithm is faster than Dijkstra but it checks negative cycle in PPI, it doesn't have the neg ative edges. Improved Bellman eliminates the negative cycle check and works faster than the Bellman Ford.

Table 1: Time comparison of Dijkstra, Bellman Ford, improved Bellman Ford

	Time (sec)		
Nodes	Dijkstra algorithm	Bellman Ford algorithm	Improved Bellman Ford algorithm
11	1.503017	0.589217	0.437878
22	1.984926	0.676689	0.542338
31	2.305561	0.874986	0.767801
40	2.703815	1.015421	0.920014
51	3.003022	1.282041	1.21345
60	3.913213	1.511156	1.492767
65	4.379276	1.672629	1.637205

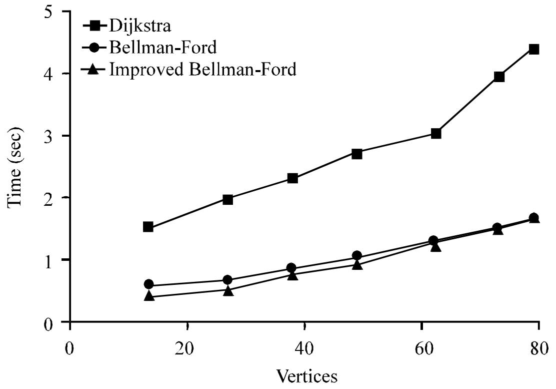

Fig. 5: Comparison graph of Dijkstra, Bellman Ford, improved Bellman Ford

All gene node data which are taken from the STRING dataset are fed into MATLAB and the genes were plotted. The shortest path between the genes of small cell lung cancer is plotted by the improved Bellman Ford algorithm. Then the shortest path between the gene non-small cell lung cancers were plotted. This can be obtained by calculating the shortest path between the AKT1 and CCND1 gene. The interaction between the genes is obtained by selecting the nodes and genes belong to the small cell lung cancer. Dijkstra's algorithm is used by LSR protocols like OSPF. LSR protocols are different from the DVR protocols as routers implementing these protocols store the entire topology of the network in their memory. There are 2 stages in building a routing table in LSR protocols. First, a map of the entire network should be stored in every router and then the shortest distance to each node must be calculated by each router. Figure 5 graph represents the Dijikstra algorithm, Bellman Ford algorithm and improved Bellman Ford algorithm is on avertices vs. time is shown in Fig. 5. When improved Bellman Ford is compared to the Dijkstra and Bellman Ford, the improved Bellman Ford is more than twice the time more efficient than Dijkstra and highly efficient than Bellman Ford algorithm.

All gene node data which are taken from the STRING dataset are fed into MATLAB and the genes were plotted. The shortest path between the gene of small cell lung cancer were plotted by the improved Bellman Ford algorithm. Then the shortest path between the gene small cell lung cancer were plotted. This can be obtained by calculating the shortest path between the AKT1 and CCND1 gene as shown in Fig. 6. The interaction between the genes are obtained by selecting the nodes and genes which belong to the small cell lung cancer.

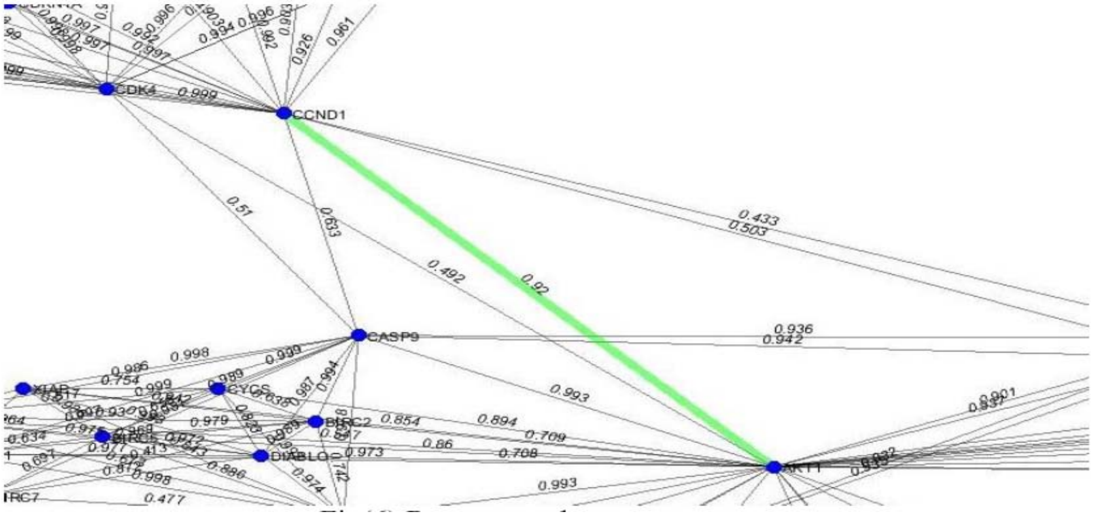

Fig. 6: Represents the gene structure

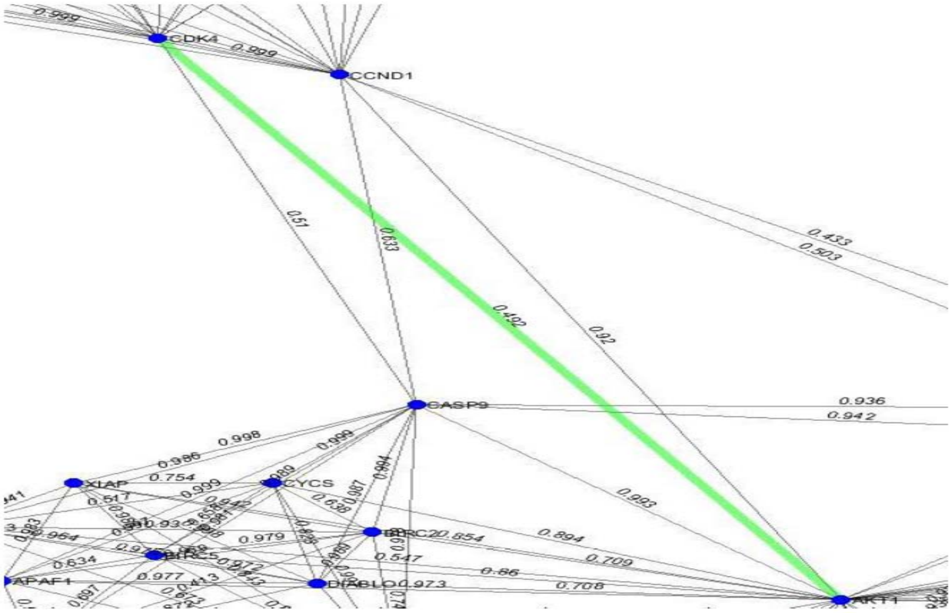

Fig. 7: The shortest path for small lung cancer gene

All gene node data which are taken from the STRING dataset are fed into MATLAB and the genes were plotted. The shortest path between the genes of non-small cell lung cancer was plotted using the improved Bellman Ford algorithm. Then the shortest path between the gene non-small cell lung cancers was plotted. This can be obtained by calculating the shortest path between the AKT1 and CDK4 gene as shown in Fig. 7. The interaction between the genes is obtained by selecting the nodes and genes which belong to the non-small cell lung cancer.

CONCLUSION

Recent technology helps in understanding the cellular mechanism of diseases in the PPI interconnection. Over the past decades, PPI helps to overcome the two major disadvantages in association with developing cyclic peptide drugs, developing cyclic peptide drugs, target engagement and membrane permeability. In this study, the improved Bellman Ford algorithm was used to find the shortest path effectively. Cancer can be effectively treated if it is found in the early stage. By finding the shortest path between the cancer genes, we can identify the cancer in the early stage.

REFERENCES

Arkin, M., 2005. Protein-protein interactions and cancer: Small molecules going in for the kill. Curr. Opin. Chem. Biol., 9: 317-324.

Chen, Q.L., Z.J. Qiao, H.J. Yang, T.F. Ni and X.Y. Chen *et al*., 2011. Bioinformatics analysis the hub-proteins and co-expression proteins of lung squamous carcinoma and adenocarcinoma based on protein interaction network. Proceedings of the IEEE International Conference on Computer Science and Automation Engineering (CSAE) Vol. 3, June 10-12, 2011, IEEE, Shanghai, China, ISBN:978-1-4244-8727-1, pp: 256-261.

Doungpan, N., W. Engchuan, J.H. Chan and A. Meechai, 2016. GSNFS: Gene subnetwork biomarker identification of lung cancer expression data. BMC. Med. Genomics, 9: 241-250.

Fontaine, F., J. Overman and M. Francois, 2015. Pharmacological manipulation of transcription factor protein-protein interactions: Opportunities and obstacles. Cell Regeneration, 4: 1-12.

Guda, P., V.S. Chittur and C. Guda, 2009. Comparative analysis of protein-protein interactions in cancer-associated genes. Genomics, Proteomics Bioinform., 7: 25-36.

Li, B.Q., J. You, L. Chen, J. Zhang and N. Zhang *et al*., 2013. Identification of lung-cancer-related genes with the shortest path approach in a protein-protein interaction network. BioMed. Res. Intl., 2013: 1-8.

Massanet-Vila, R., F.F. Albert, P. Caminal and A. Perera, 2012. Network-based enrichment analysis of gene expression through protein-protein interaction data. Proceedings of the 2012 IEEE Annual International Conference on Engineering in Medicine and Biology Society (EMBC), August 28 - 1, September 2012, IEEE, San Diego, California, USA., ISBN:978-1-4244-4119-8, pp: 6317-6320.

Qian, Z., P.G. Dougherty and D. Pei, 2017. Targeting intracellular protein-protein interactions with cell-permeable cyclic peptides. Curr. Opin. Chem. Biol., 38: 80-86.

Sable, R. and S. Jois, 2015. Surfing the protein-protein interaction surface using docking methods: Application to the design of PPI inhibitors. Mol., 20: 11569-11603.

Tumuluru, P. and B. Ravi, 2017. Dijkstra's based identification of lung cancer related genes using PPI networks. Intl. J. Comput. Appl., 163: 1-5.

Wu, C.H., C.L. Hsu, P.C. Lu, W.C Lin and H.F. Juan *et al*., 2016. Identification of lncRNA functions in lung cancer based on associated protein-protein interaction modules. Sci. Rep., 6: 1-11.

Yu, W., L.R. He, Y.C. Zhao, M.H. Chan and M. Zhang *et al*., 2013. Dynamic protein-protein interaction subnetworks of lung cancer in cases with smoking history. Chin. J. Cancer, 32: 84-90.

Zhang, M., M.H. Chan, W.J. Tu, L.R. He and C.M. Lee *et al*., 2013. Using the theory of coevolution to predict protein-protein interactions in non-small cell lung cancer. Chin. J. Cancer, 32: 91-98.

Zhang, S., C.C. Liu, W. Li, H. Shen and P.W. Laird *et al*., 2012. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res., 40: 9379-9391.