ANALYSIS OF PRUNED PROTEIN PROTEIN INTERACTION NETWORK USING DIFFUSION METHOD

B Madhav Rao[1], V Srinivasa Rao[2] , K Srinivasa Rao[3]

[1] *Research Scholar, Rayalaseema University, Kurnool, India*

[2] *Department of CSE, V.R .Siddhartha Engineering College, Vijayawada 520007, India*

[3] *Department of CSE, V.R .Siddhartha Engineering College, Vijayawada 520007, India*

(E-mail: madhavraob@gmail.com, drvsrao9@gmail.com, kudipudi72@gmail.com)

Abstract—Protein protein interaction network analysis is one of most important research areas in molecular biology. This analysis can be used to detect the disease causing protein(s) in an organism. Further it will help discover the drug. Protein interactions can be studied through a huge number of experiments with high-throughput. These interactions are predicted by using computational methods that handle large scale of sequence data generated in the last decade. These methods have the capability of detecting the interaction between two proteins as well as thousands of proteins. Moreover, it also detects different proteins in the given cell. In the proposed approach, firstly diffusion method is used to analyze the interactions between proteins and purify the network based on rankings. Then the resultant protein sequences are analyzed with sequence analyzer to find the repeated sequence in the pruned protein dataset.

Keywords—Diffusion method, BRCA1, PPIs (Protein Protein Interactions), ORF

I. INTRODUCTION

In protein protein interaction network, amino acid sequences interact with the other proteins. The physical interaction of these proteins constructs a network [1]. These networks are either binary or complex structures. Binary structures are called as direct interaction and complex structures are called as indirect interaction. PPI network can be constructed by using the publicly available data sources.[6] The maintenance of these data sources is one of the tedious task based on the previous publication, the proposed approach uses the data sources[4] for the analysis of common sequence proteins which cause for the Breast Cancer. In the proposed approach, by using the query processing technique, interacting proteins can be retrieved through which one can identify direct and indirect edges. When the network is very large in size, it is difficult to analyze the interactions. The pruning of the proteins can be done by using two methods namely modular and topological methods [6]. A PPI network is a graphical representation of proteins. Where proteins are called as nodes and their interactions are called as edges. This method can reduce the network by identifying the driver nodes [11][5]. Taking into the consideration of some measuring features such as clustering, shortest path and coefficient matrix.[5]. This approach uses the diffusion method by considering edges and nodes which specify the occurrence of the target nodes based on the rankings it generates[7]. This method considers the existing database such as string and uniprot to investigate the proteins like BRCA1,RAD51 and their interacting proteins of homo sapiens category targeted considerations of breast cancer. These databases provide the physical association of protein sequences[4] and their functional activities.

There is a relation between one disease to another disease due to the similar molecules. The functionality of the protein family overlaps and which leads to the disease interacting with other molecules.[8] The protein functionality basically describes the healthiness of an organism. There are three approaches which can be taken to analyze the PPI called vivo, vitro and, silco[9] .

II. RELATED WORK

Breast cancer is one of the major diseases in women so there has been a large research in this area.PPI network analysis has been taken into consideration to study the interface which causes this disease since cancer proteins have the large constructed network based on target protein which generate 108 proteins and 912 interactions. The node centrally in a network contains the node betweenness for calculation of shortest path [3]. Purely identified cancers have the worst prognosis so the early diagnosis can be done which is important for therapy.

BIANA [7] is a database consists of interaction information of the proteins along with network management framework. Biologic Interactions and Network Analysis (BIANA) framework is implemented in python programming

language. It can grasp all biological information like individual entries and their relationships and also it supports external entities. For example a uniprot entry contains protein information, Gene Bank entry contains gene information or IntAct contains protein Interaction information are represented as external entities. Each external entity relation is defined by attributes like reliability, durability, detection method, role and cardinality. BIANA uses a unification protocol to unifies the external data inserted in the database. In this process, user can do back propagation to back into original state.

III. METHODOLOGY

Diffusion algorithm attempts to use a set of nodes and an entire interaction network to find the nodes most relevant to the original set. Conceptually, Diffusion applies heat to each node in the set, and lets the heat flow through connecting edges to adjacent nodes. It, then, produces a list of nodes ranked by the heat that accumulated. A node with many connections will tend to have a higher ranking, and an isolated node will tend to have low rank (and thus be excluded from the resulting node set).By default, Diffusion uses the set of selected nodes as the heat sources, with each node having the same initial heat. At the end of a diffusion, Cytoscape [10] leaves the top 90 percentile of hot nodes selected. It allows you to use the results panel to select a higher or lower percentile dynamically. It also stores the nodes' initial heat as a node attribute in the "diffusion_input" column, and returns the heat and ranking values in the "diffusion_output_heat" and "diffusion_ output _rank" columns.

In the proposed method constructs a sparse matrix for given input PPI network before applying the diffusion. Diffusion method considers inverse sparse matrix, time and diffusion_input as inputs. Calculate total number of interactions for each node. And sum the diffusion input values of all the interconnected nodes for each node, finds the determinate and calculate the mean. This will be the diffusion_output_value. Based on this, diffusion ranks are allotted for each protein node.

A. Algorithm

Step 1: Read the input as protein dataset form string db which is preprocessed by considering disease score=0.75 and correlation coefficient = 0.00004 .

Step 2: Apply diffusion method to given input dataset.

Step 3: Diffusion applies heat to each node in the dataset. Heat flows through connecting edges and its associated nodes.

Step 4: Calculate the ranking based on the heat that accumulated.

$$d=h * \exp(-Lt) \quad \text{---------- (1)}$$

Where resultant vector is represented as d and vector representation of query is denoted by h. Graph Laplacian ,L and fixed time for diffusion is denoted by 't', which is a scalar parameter. Time factor controls the extent to which the original signal is allowed to spread throughout the network. The matrix exponential is expressed as exp(*)[6]. Default value for time is 0.1.

Step 5: A node with many connections will tend to have a high rank, and an isolated node will tend to have low rank

Step 6: Eliminate the low rank nodes and output the resultant protein network.

Step 7:The resultant protein dataset is further analyzed by sequence analyzer

Step 8: Sequence analyzer finds the repeated sequence by comparing all the proteins in the network.

Step 9: It results the longest and shortest sequences and most repeated sequence.

The diffusion method takes the input as pruned proteins which are obtained after the preprocessing. This diffusion method sends the entire nodes and network to a web based REST service to calculate network propagation[2]. The REST service applies the heat to nodes and produces ranking based on nodes connected to each other. The resultant information will be stored in a node table.

This analysis method calculates the total sequences present in the given dataset and creates a dictionary to calculate each protein sequence length and produces output as largest sequence and smallest sequence protein.

In this method, "CPL"can be considered as starting sequence and "TAA", "TAG" or "TGA" is considered as ending sequence. ORF(Open Read Frame) can be identified by this method. For better analysis, it applies reverse complement mechanism to calculate the most accurate data. Another method Sequence analyzer used to find most repeated sequence in the given dataset, it results the top most proteins containing the same sequence.

IV. RESULTS

Figure 1 shows the input dataset to sequence analyzer method. The protein dataset is retrieved after pre processing the PPI network of 4000 nodes by applying filters like disease score =0.75 and correlation coefficient 0.00004 the resultant network of 108 proteins considered as input to diffusion method. These 108 proteins are most common proteins which cause breast cancer.

Figure 2 represents the network after diffusion method applied. Yellow colour nodes represents top fifteen proteins which cause breast cancer among the 108 proteins. Remaining

93 proteins are shown grey color circles which is also cause for the breast cancer but less priority then the top 15 proteins.

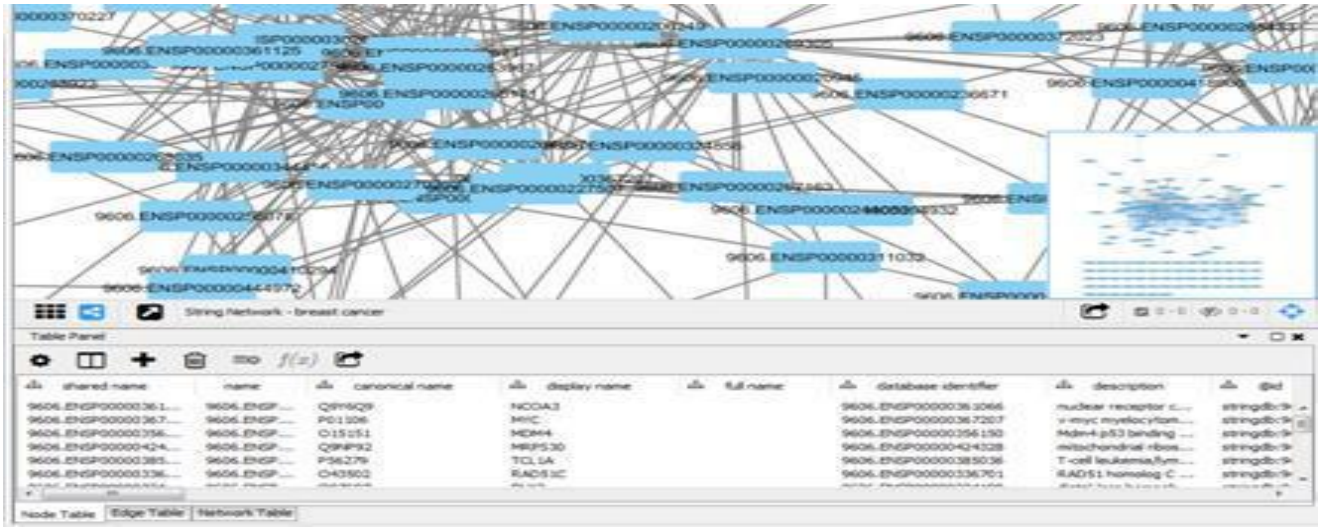

Figure 1: Breast Cancer dataset in cytoscape with 108 nodes 912 interactions.

Fig2: Top 15 sequences generated using diffusion method

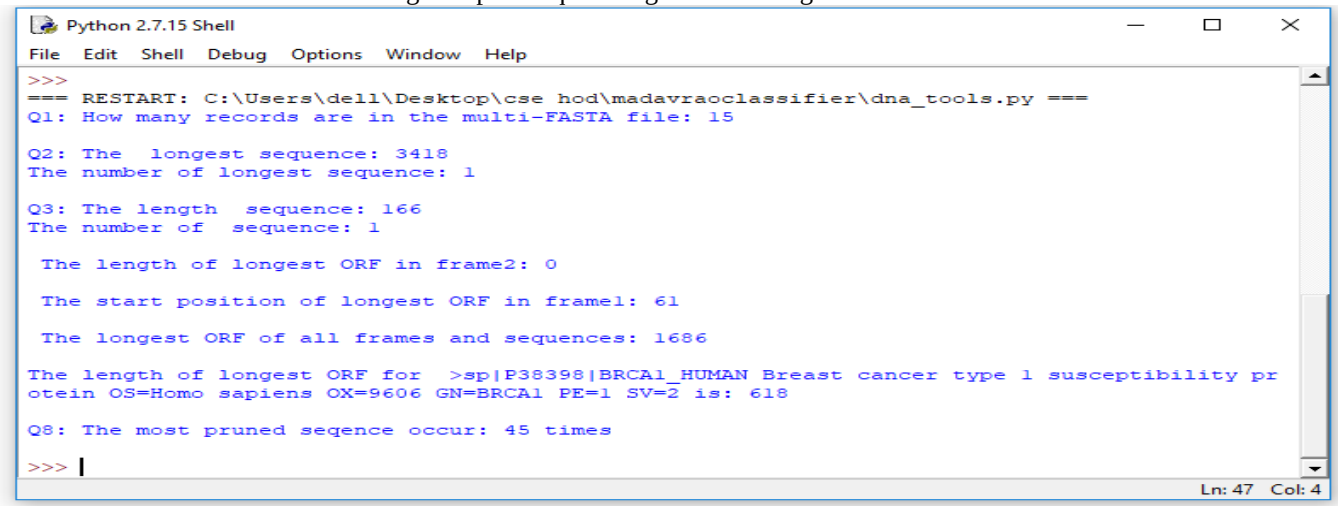

Figure 3:Analysis of protein sequence

Figure 3 shown that diffusion method yields the top 15 proteins which have high interactions. These proteins are analyzed by using sequence analyzer. The python script which used to generate information like longest protein sequence with length of 3418, shortest sequence with length of 166 and the most repeated sequence found in 45 times in given dataset. This analyzer method detects ORF (open read frame) in sequence frames. Start position of ORF found in line 61 and the longest ORF found in all frames was 1686. This method compares all the sequences with each other and results the most common sequence or part of the sequence which similar in 15 proteins.

Figure 4 shows the common sequence which is repeated in all the proteins in the given dataset. By comparing all the protein sequences in the given network with each other. BRCA1 protein sequence is more similar among the top 15 proteins.

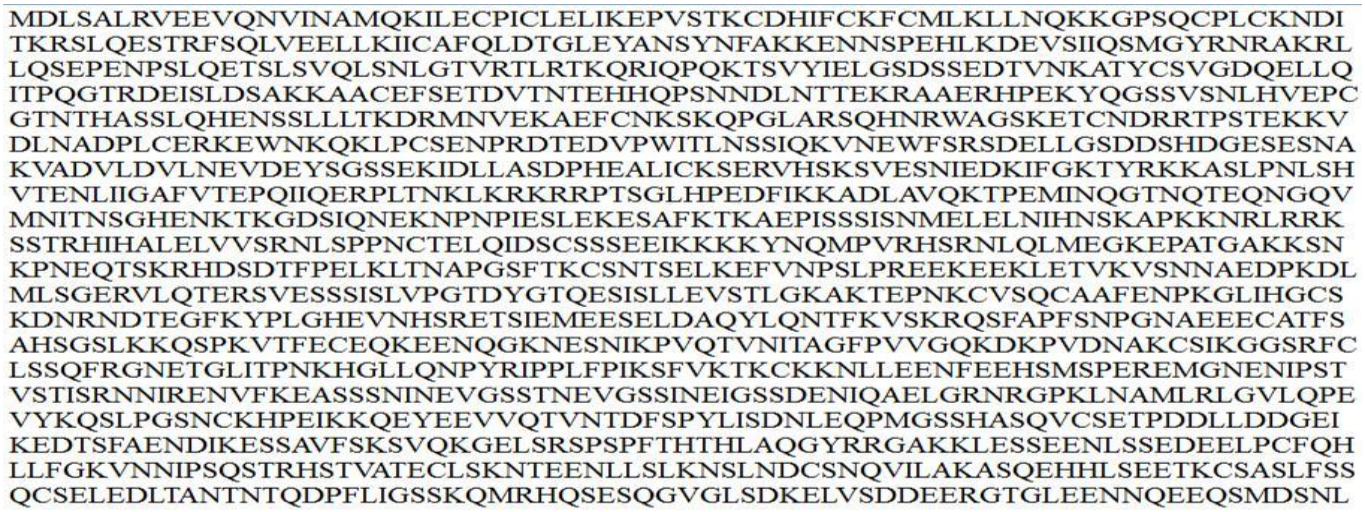

Figure 4: Most repeated sequence in given protein dataset

Table 4.1: Top pruned fifteen proteins after diffusion

SNO	CANONICAL NAME	PROTEIN NAME	DISEASE SCORE	DIFFUSION OUTPUT HEAT	DIFFUSION OUTPUT RANK
1	P08183	ABCB1	5	0.943590371	13
2	Q99728	BARD1	5	0.980142536	8
3	P38398	BRCA1	5	0.981078564	6
4	P51587	BRCA2	5	0.98125451	3
5	Q9BX63	BRIP1	3.509493	0.98125451	5
6	P11802	CDK4	5	0.980142536	7
7	P07339	CTSD	2.590049	0.993062461	2
8	O15151	MDM4	2.757025	0.967160887	12
9	Q86YC2	PALB2	5	0.973790986	11
10	E9PI54	RAD51	5	1.0	1
11	Q86U92	RAD51B	3.574687	0.979811108	9
12	O43502	RAD51C	5	0.976763781	10
13	Q15831	STK11	5	0.98125451	4
14	O15405	TOX3	2.546027	0.94279620	14
15	O43542	XRCC3	5	0.93644510	15

V. CONCLUSION

This proposed method analyzes the given protein dataset by using the diffusion method. It calculates the rankings for proteins in the network (network propagation). Diffusion method dynamically eliminates the less ranking proteins from the data set. Sequence analyzer calculates the most repeated sequence in the given protein dataset. It results the most repeated protein sequence which is similar in all the proteins in the dataset. This approach extracts the top protein sequences which cause the breast cancer.

REFERENCES

[1] Damian Szklarczyk et al , "STRING v10: protein–protein interaction networks, integrated over the tree of life", Nucleic Acids Research,2014, doi: 10.1093/nar/gku1003.

[2] Daniel E.Carlin,Barry Demchack,Dexter pratt ,"Network propagation in the cytoscape cyber infrastructure", PLOS Computational Biology | https://doi.org/10.1371/journal.pcbi. 1005598 October 12, 2017.

[3] Del Sol A, O'Meara P, "Small-world network approach to identify key residues in protein-protein interaction." PMID:15617065 DOI:10.1002/ prot.20348.

[4] Yassen Assenov,Fidel Ramirez, Sven Eric Schelhorn, Thomas Lengauer and Mario Albrecht, "Computing topological parameters of biological networks", Bioinformatics, vo 24(2), pg 282-284,2008.

[5] Pei Wang, Jinhu Lu and Xinghuo Yu, "Identification of important nodes in directed biological networks: A network motif approach", PLOS ONE, issue 8,e106132, 2014, pp 1-15.

[6] Jamie Snider et al "Fundamentals of protein interaction network mapping" DOI 0.15252/msb.20156351 Accepted 24 November 2015 Mol Syst Biol. (2015) , pp 11: 848 .

[7] Javier Garcia-Garcia, Emre Guney, Ramon Aragues, Joan Planas-Iglesias, Baldo Oliva," Biana: a software framework for compiling biological interactions and analyzing networks", BMC Bioinformatics 2010, https://doi.org/10.1186/1471-2105-11-56.

[8] K Srinivas ,R Kiran Kumar, M Mary Sujatha "A Study on Public Repositories of Human Protein Protein Interaction Data", IJIACS, ISSN 2347-8616 ,vol6-issue6, June 2017.

[9] Nicolas Kourtellis et al, "Identifying high betweenness centrality nodes in large social networks" DOI 10.1007/s13278-012-0076-6.

[10] Paul Shannon et al "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks" Genome Res. 2003 Nov; 13(11): 2498–2504.

[11] Xiao-Fei Zhang " Determining minimum set of driver nodes in protein protein interaction network" Zhang *et al. BMC Bioinformatics* (2015) 16:146, DOI 10.1186/s12859-015-0591-3.