International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

Motif Discovery of Protein-Protein Interaction using Minimum Spanning Tree

P. Lakshmi, D. Ramyachitra, E. Pavithravishalini

Abstract: ***In protein Interaction Networks, counting subgraph is a tedious task. From the list of non induced occurrence of the subgraph, motif topology calculated by using Combi Motif and Slider techniques. But, this approach was taken more time to execute. To reduce the execution time, the minimum weight value between the nodes, the Minimum spanning tree concept proposed. Prim's method implemented with the greedy technique (as Kruskal's algorithm) to calculate the minimum path between the nodes in the Protein interaction network. This technique uses to compare the similarity of the minimum spanning tree approach. Initially, this algorithm has discovered the path then calculated the weight matrix and found the minimum weight value. From the computational experiments, the proposed approach of MST providing better results in terms of time consumption and accuracy to count the motif pattern in the network of the interacted proteins.***

Keywords— Motif, Protein interaction Network, Minimum Spanning Tree, Graph, Sub tree.

I. INTRODUCTION

Proteins are a combination of amino acids, to carry out their function processes with other molecules, these biochemical activities in living cells are called Protein interactions [1]. Protein interaction prediction is a challenging task in bioinformatics. PPI predicted by various relevant biological information such as functions of proteins, protein sequences, gene expression, PIN, Gene interaction /Networks at the whole molecular level based on Amino acid sequences, structure information, physicochemical properties, etc.,[2.] Two classification aspects are addressed, first based on various attributed and features, second to predict whether the proteins have interacted or not interacted. These aspects depend upon the identification of the various sources of data, the information contains Gene ontology, Details of phylogenetic approach, and Synthesis of the gene, Genomic circumstances, Gene and Protein sequence conservation depends on the Protein interactions [3, 4]. Study about proteins known as proteomics, mainly applied for drug discovery [5, 6]. To reduce the cost and time, computational results are better than experimental methods, some computational methods of predicting PPI are 3D structural information through algorithm in human and yeast, Probabilistic decision tree with high throughput datasets to characterize the co-complex protein pair and [7], Incomplete primer extension method proposed for polymerase [8]. The existence of the feature can be categorized into 3 classes, based on structure, sequence, and hybrid for sequence and structural data [9]. Protein's amino acid sequence is determined by its physical and chemical properties, it contains mutation rate, hydrophobicity, structure acidity and alkalinity [10]. The problem of motif discovery can be evaluated with biological and electronic networks using DIP (Database of Interaction Proteins)[11]. In computational biology, similar feature topologies are discovered from PPIN to share degree distribution of power-law and huge cluster coefficient [12-16]. To predict PPIN, motif network discovery contains two levels: Subgraph count and statistical significance of subgraph and two approaches are: Enumeration of all subgraphs depends on the size and non-isomorphic class determination of subgraphs [17-19]. NAUTY tool decreases the process of the huge amount of isomorphism detection [20]. To detect isomorphism pattern, Slider, Combi motif and ACCMotif algorithms are used to reduce the time of execution [21] and for non-isomorphic detection G trees, Color coding and MODA algorithms are used [22].PPI information is quite often undirected; in this way the issue of orienting interaction edges for signal transmission in signaling network is expensive. This exhibits the fascination with finding an efficient algorithm for edge-orientation in PPI networks [24]. Following sections structured as given here: part 2 illustrates the methodology, part 3 represents about performance measurements finally part 4 concluded with future work.

Revised Manuscript Received on February 05, 2020.

* Correspondence Author
P. Lakshmi*, Ph.D. Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu.visalaks@gmail.
Dr. D. Ramyachitra, Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu. jaichitra1@yahoo.co.in
E. Pavithravishalini, Ph.D. Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu. pavishalin@gmail.com..

II. METHODOLOGY

A. Preliminaries

A Graph G contains a set of Edges and vertices represented in the form of Graph $G_{(V,E)} = (V_{pi}, E_{pi,pj})$ where P_i denotes the nodes as proteins and the P_i,P_j are the edges denotes the interaction between proteins. Graphs are categorized into Induced and NonInduced - Induced subgraph denotes, along with any edges, vertices and the endpoints are both in the subset, it's determined by vertices selected. Non-Induced Subgraph i.e., G` of G presented between node pairs in G` and it is determined by edges selected. The contribution of this paper is to identify and count the similar motif pattern in the protein interaction network.

B. Dataset

Three types of data collection in protein-protein interaction that is primary, secondary and prediction data from the databases.

C. Yeast's interaction network

To evaluate the algorithm protein interaction dataset taken from the yeast database placed in BioGRID. Online genetic interaction database, widely updated on time-based by researchers and biologists. It contains interactive information from the pair of proteins. This information combined with a confident score by experiment to determine the interaction edge weight. It depends on the type of experiment consistency and separate experiment interaction. In the pair of the protein interaction network, weight of the edge calculated by following the formula

$$P\ (\text{interact}\ (P_1, P2) = 1-\prod_{i\epsilon p1,p2} (1-c(i))\text{----------------}\ (1)$$

The set of I with P_1 and P_2 includes the member I, it contains interactive experiments individually, from the interacted database of protein.
To submit the protein interaction weight value to the equation, we can get interacted protein pair data and interaction weights. [23]

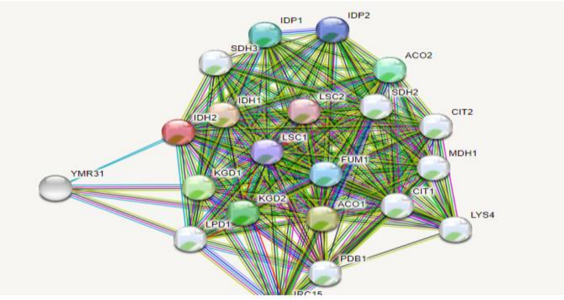


Figure -1 Protein Interaction Network

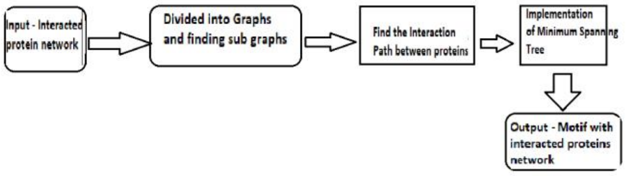
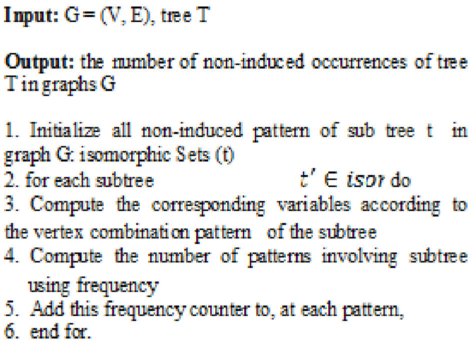

Input - Interacted protein network
Divided into Graphs and finding sub graphs
Find the Interaction Path between proteins
Implementation of Minimum Spanning Tree
Output - Motif with interacted proteins network

Figure -2 Framework of Finding Motif pattern in Protein interaction network

D. Counting Subgraphs

This algorithm gives an idea about, counting motif structure of tree T in Graph G.

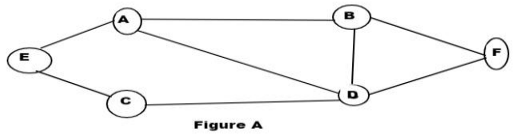


```
Input: G = (V, E), tree T

Output: the number of non-induced occurrences of tree
T in graphs G

1. Initialize all non-induced pattern of sub tree t in
graph G: isomorphic Sets (t)
2. for each subtree                    t′ ∈ isor do
3. Compute the corresponding variables according to
the vertex combination pattern  of the subtree
4. Compute the number of patterns involving subtree
   using frequency
5. Add this frequency counter to, at each pattern,
6. end for.
```

In an algorithm, calculating the pattern of isomorphism with the graph is completely tedious using combinatorial methods. Different sub-patterns used in algorithms, it achieved through the removal of particular vertices and edges from the pattern counting procedure of the subgraph. It requires necessary updating repeatedly.

E. Sub tree with no induced Sub tree t

Operations of the vertices, using combinatorial methods are demonstrated with non induced occurrences of the subtree pattern T with Graph G are listed.



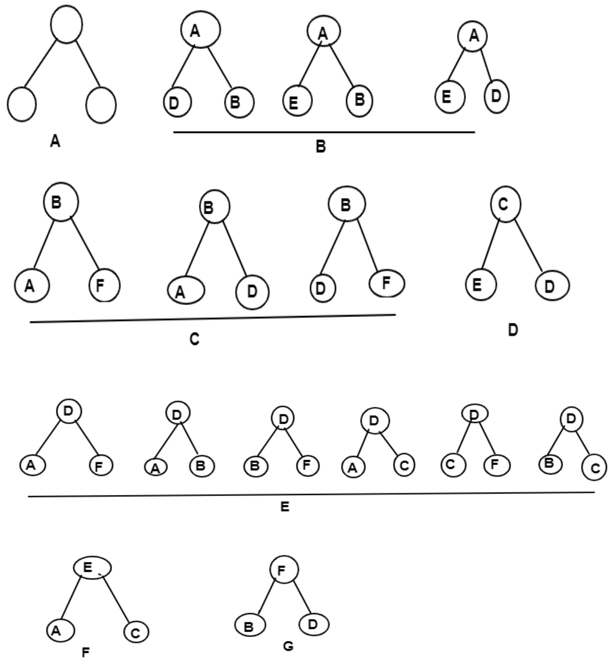
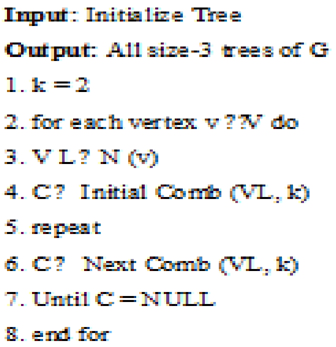


Figure – 3 represents the Tree Graph with sub trees motif pattern. A. Motif pattern - Tree T. Figure B, C, D, E, F, G shows the tree T includes 3, 1,6 no induced occurrences respectively.

By using the following equation, the weight value calculated between the nodes.

Weight w = Σ P1i + Σ P2i /2 --------------------- 2

Pseudo code of the algorithm to count the subtrees of G with size three.



```
Input: Initialize Tree
Output: All size-3 trees of G
1. k = 2
2. for each vertex v ??V do
3. V L? N (v)
4. C?  Initial Comb (VL, k)
5. repeat
6. C?  Next Comb (VL, k)
7. Until C = NULL
8. end for
```

Here we denote the number of interactions with relevant protein P.The graph, for each vertex, from the tree T, size 3 with two vertices selected from N (v). It considers various selections with vertices has (N (v), 2).

F. M-SLIDER: Sliding over motifs

With the starting neighborhood function N^{mot} , M-SLDIER (short for motif-SLIDER) method introduced.
In protein N^{mot} examined based on the comparison of (l, d) motif X ,which contains the sliding window with length l and l-d holes on the protein, till the matches of the protein holes with non wildcard of X. It's continued until get the motif with hole open and close with motif X'.If in case any non wild card motif X' found then it will be replaced by making (l, d) motif again.

G. SEQ-SLIDER: Sliding over sequences

The logic of counting motif pattern and the matching process is done based on the protein sequence with its sliding window, which is calculated by using the SEQ-SLIDER method. In case $G_{X,Y}$ & $G_{X',Y'}$ share interactions, motif pair $\{X, Y\}$ comes under N^{mot} selected sub network $G_{X',Y'}$ is also closed. It is also not guaranteed properly.
In the sequence level motif X with neighborhood N_u^{seq} ,consider all(l, d) motif. It will be similar area around the hits of the motif X in the particular protein sequence $u \in VX$. Fetching random pair $\{X, Y\}$, determines the interactions$\{u, v\}$. Based on motif pair, $\lambda(u)$ &$\lambda(v)$, are suggesting the given motif hits.

$$N_{u,v}^{seq}(\{X,Y\}) = \left\{\{X',Y\}|X' \in N_u^{seq}(X)\right\} \cup \left\{\{X,Y'\}|Y' \in N_v^{seq}(Y)\right\} \text{------------- } 3$$

In that way, $N_{u,v}^{seq}$ guarantees that the $G_{X',Y'}$ sub-network choose the $\{X', Y'\}$ neighbor pair of the $\{X, Y\}$ motif includes$\{u, v\}$. A motif(l, d), a protein u describes the position$pos(X, u)$. $\lambda(u)$ Substrings get the similarity with X.
If motif (l, d) with $X' \in N_u^{seq}(X)$, $p \in pos(X, u)$ and $p' \in pos(X', u)$ already existed, then $|p - p'| \leq \delta$, in which δ has short distance $(\delta = l/3)$. Hence, $N_{u,v}^{seq}(\{X, Y\})$ it defines the neighborhood $\{X, Y\}$ relative to$u \in V_X$ and $v \in V_Y$.

Minimum Spanning Tree (Proposed)

To identify the similar motif pattern of the protein interaction network, graph using minimum spanning tree applied. The contribution of the experiment here is, finding motif topology of the PPIN. To this purpose, improved path detection methods introduced from the subtree of the given graph. Based on the interaction, the weight is assigned to the edges. MST helps to count the number of similar patterns in the PPIN.

Kruskal's algorithm

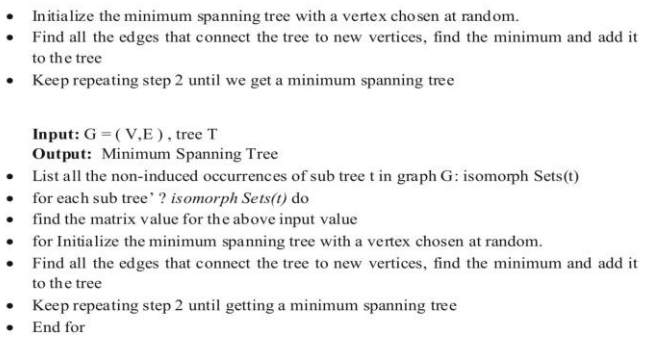
Kruskal's algorithm is the method of minimum spanning tree, which helps to find the edge of the smallest feasible weight; it connects the interaction between the proteins (nodes) to make a tree with the shortest path. It forms a tree with a subset of the edges using the greedy approach, which updates every vertex, to get the minimized tree of the total weight value of the edges. If the nodes are not connected then it starts to find the search of another MST in PPIN.

Algorithm Steps

- Start sorting with edges of their weights respectively.
- Add and update edges with the minimum to maximum weight value of the MST.
- We should build an acyclic graph with disconnected components.

H. Prim's algorithm

In PPIN, a particular segment of the network taken as a graph to apply Prim's algorithm. For the purpose of motif identification, counting the number of similar motif topology from the PPIN with Graph G, Minimum Spanning Tree algorithm utilized. A node (interacted protein) connected with the neighbor has the minimum weight value. This process repeated until all nodes connected with the shortest path without any cycle. It reduces the execution time, to find the minimum weight value with the shortest path of the Graph. MST created with the smallest path of the subtree pattern and updated every step until the target node is reached by using the greedy method. Steps of the Minimum Spanning Tree using Prim's algorithm as follows



- Initialize the minimum spanning tree with a vertex chosen at random.
- Find all the edges that connect the tree to new vertices, find the minimum and add it to the tree
- Keep repeating step 2 until we get a minimum spanning tree

Input: G = (V,E) , tree T
Output: Minimum Spanning Tree

- List all the non-induced occurrences of sub tree t in graph G: isomorph Sets(t)
- for each sub tree' ? *isomorph Sets(t)* do
- find the matrix value for the above input value
- for Initialize the minimum spanning tree with a vertex chosen at random.
- Find all the edges that connect the tree to new vertices, find the minimum and add it to the tree
- Keep repeating step 2 until getting a minimum spanning tree
- End for

III. RESULTS AND DISCUSSION

It shows the comparison between existing and proposed methods in terms of time consumption and Descent values are given below

Datasets used for Experimental study

To evaluate the algorithm with parameters, datasets are retrieved from the DIP (Database of Interacting Proteins) and results are compared.

TABLE -1 – Retrieval information of Data set from DIP database

Protein – Protein Interaction Network	Number of Proteins	Number of Interaction
DIP	8000	4000

Performance Measures

The comparison is made in terms of the performance metrics referred to as the time consumption that is defined in the following subsections.

Time Consumption

Time consumption is measured as calculating the difference between the start and end time of implementing the proposed and existing approaches. The proposed method MST has given better results than the existing approaches in terms of time consumption and accuracy for the DIP dataset. The tables and graphs are represented for the comparison of performance measures.

TABLE – 2 – Subtree motif pattern count with subgraphs using existing algorithms

Existing algorithm	Dataset	Motif Structure	Number of classes	Sub tree / Sub graph	Execution time
Combi motif	S.cere.CR	0-0	9	5,160,508,293	0.25
Combi motif	S.cere.CR	0-0-0	13	20,468,265,322	7.59
Combi motif	S.cere.CR	0-0-0-0	9	25,726,058,159	59.78
Combi motif	C.eleg	0-0	9	209,394,077,970	0.29
Combi motif	C.eleg	0-0-0-0-0	4	230,623,915,351	957.83`
Combi motif	H.pylo	0-0-0	13	10,830,678,654	2.35
Combi motif	H.pylo	0-0-0-0	9	9,829,829,819	18.12
Slider	S.cere.CR	0-0-0-0	7	4,160,307,1982	0.24
Slider	S.cere.CR	0-0-0	14	21,431,453,311	8.50
Slider	S.cere.CR	0-0-0-0	8	35,829,321,148	60.18
Slider	C.eleg	0-0	4	108,269,517,286	0.75
Slider	C.eleg	0-0-0-0-0	8	254,687,842,978	843.69
Slider	H.pylo	0-0-0	13	11,235,248,479	3.07
Slider	H.pylo	0-0-0-0	9	9,759,654,277	19.01

TABLE – 3 Subtree motif pattern count with subgraphs using the proposed algorithm

Existing algorithm	Dataset	Motif Structure	Number of classes	Sub tree / Sub graph	Execution time
MST	S.cere.CR	0-0-0-0	2	286,432,255,0	52.85
MST	S.cere.CR	0-0-0-0-0-0-0	3	341,522,142,1	0.15
MST	S.cere.CR	0-0-0-0-0-0-0-0	4	397,135,45,2	1.25
MST	C.eleg	0-0-0-0-0-0-0	5	463,154,23,4	3.56
MST	C.eleg	0-0-0-0-0-0	6	530,123,54,75	25.84
MST	H.pylo	0-0-0-0-0	7	615,845,245,6	235.51
MST	H.pylo	0-0-0-0	9	732,235,24,57	728.02

TABLE – 4 Performance Comparison of existing with proposed algorithms in terms of Time consumption and Descent value

Database	Algorithms	Descent Value	Time Consumption
DIP	Combi Motif	78	82
	Slider	84	75
	MST	91	62

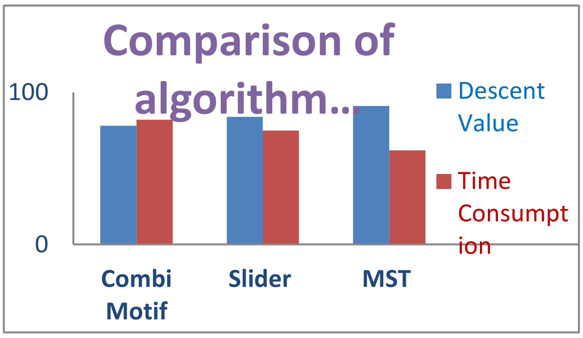


Figure 4 –Performance measures of algorithms for Descent value & Time Consumption

Figure 4 comparison results of the proposed approach with the existing method in terms of time consumption. Dataset is represented in X-axis and time consumption (ms) is denoted in Y-axis. From this analysis, it is decreased for the proposed approach compared to the existing approach.

IV. CONCLUSION

Nowadays, the discovery of the Motif pattern in PPIN is a tedious task for researchers. To know about the cellular activities and biological functions, path identification of the interacted protein network was used. For that, the graph with the subtree of the motif pattern counting helps to solve this method, to implement the Minimum spanning tree algorithm in PPIN. It helps to introduce new transaction factors for binding sites in the protein domain and it provides an idea to search for finding ungapped repeated sequence pattern occurrence in PPIN. Hence, the proposed approach of the MST using Prim’s and Kruskal’s algorithm, gives the maximum accuracy level, when compared with the existing methods to find the motif pattern of the PPIN. In the Future, in this research work could be extended as MST to determine lengthy motifs with combinatorial methods of each vertex for pattern similarity in PPIN.

REFERENCES

1. Leyi Wei, Pengwei Xing, JiancangZeng, JinXiu Chen, Ran Su, FeiGuo “Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier”, 2017 Elsevier B.V
2. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. CurrOpinStructBiol 2002;12:368–73.

3. Zhu-Hong You, Member, IEEE, MengChu Zhou, Fellow, IEEE, XinLuo, Member, IEEE, and Shuai Li, Member, IEEE," Highly Efficient Framework for Predicting
Interactions between Proteins" IEEE Transactions on Cybernetics (Volume: 47, Issue: 3, March 2017).
4. Quraishi M, Koytiger G, Jenney A, MacBeath G, Sorger PK. A multiscale statistical mechanical framework integrates biophysical and genomic data to assemble cancer networks, Nat Genet 2014;46:1363–71.
5. Califano A. Predicting protein networks in cancer. Nat Genet 2014; 46:1252–3.
6. Jiang M, Chen Y, Zhang Y, Chen L, Zhang N, Huang T, et al. Identification of hepatocellular carcinoma related genes with k-th shortest paths in a protein-protein interaction network. MolBiosyst 2013; 9:2720–8.
7. Lan VZ, Wong SL, King OD, Roth FP. Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinf 2004;5:1–15.
8. Pitre S, Dehne F, Chan A, Cheetham J, Duong A, Emili A, et al. PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. BMC Bioinf2006; 7:365.
9. H.X. Zhou, Y.Shan, Prediction of protein interaction sites from sequence profile and residue neighbor list, Proteins: Struct.Funct.Bioinform.44 (2001) 336–343.
10. Wu, G.Nov 1 2010. Functional Amino Acids its growth, reproduction, and health. Adv.Nutr: Int.Rev. J.1,31-37.
11. I. Xenarios, L. Salwinski, X. Duan, P. Higney, S. Kim, D. Eisenberg, Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions, Nucleic Acids Res. 30 (1) (2002) 303–305.
12. G.C.K.W. Koh, P. Porras, B. Aranda, H. Hermjakob, S.E. Orchard, Analyzing protein-protein interaction networks, J. Proteome Res. 11 (2012) 2014–2031.
13. D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (6684) (1998) 440–442.
14. A.L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
15. G. Bebek, P. Berenbrink, C. Cooper, T. Friedetzky, J. Nadeau, S.C. Sahinalp, The degree distribution of the generalized duplication model, Theoret.Comput. Sci. 369 (1) (2006) 239–249.
16. B. Bollobás, O. Riordan, J. Spencer, G. Tusnády, The degree sequence of a scale-free random graph process, Random Structures Algorithms 18 (3) (2001) 279–290.
17. S. Wernicke, Efficient detection of network motifs, IEEE/ACM Trans. Comput. Biol. Bioinform. 3 (4) (2006) 347–359.
18. J. Chen, W. Hsu, M. Lee, S. Ng, Nemofinder: dissecting genome-wide protein-protein interactions with mesoscale network motifs, in Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2006, pp. 106–115.
19. Z. Kashani, H. Arabian, E. Elahi, A. Nowzari-Dalini, E. Ansari, S. Asadi, S. Mohammadi, F. Schreiber, A. Masoudi-Nejad, Kavosh: a new algorithm for finding network motifs, BMC Bioinformatics 10 (1) (2009) article 318.
20. B. McKay, Practical graph isomorphism, Congr. Numer. 30 (1981) 45–87.
21. S. Khakabimamaghani, I. Sharafuddin, N. Dichter, I. Koch, A. Masoudi-Nejad, QuateXelero: an accelerated exact network motif detection algorithm, PLoS ONE 8 (7) (2013) e68073.
22. P. Ribeiro, F. Silva, G-Tries: an efficient data structure for discovering network motifs, in Proceedings of the 2010 ACM Symposium on Applied Computing, ACM, 2010, pp. 1559–1566.
23. L.A.N. Amaral, A. Scala, M. Barthelemy, H.E. Stanley, Classes of small-world networks, Proc. Natl. Acad. Sci. USA 97 (21) (2000) 11149–11152.
24. Fischer E, Sauer U, "Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism", Nat. Genet. 37 (2005) 636–640.

AUTHORS PROFILE

P.Lakshmi Ph.D, Research Scholar, Computer Science, Bharathiar University. Coimbatore, Tamilnadu, India.

Dr.D.Ramyachitra, Assistant professor, Bharathiar University, Coimbatore. She has more than 50 National, International Conferences, Journal Publications in multi aspects such as SCI, SCIE, Science Direct, Web of Science, Scopus, and UGC Care respectively. She is one of a member of CSI. She presented and participated in various workshops, seminars, etc., she has done an UGC-Minor Research Project, 2009-2011, Title: An efficient scheduling strategy for protein sequence analysis on the grid. She had research experience with more than 10 years of research experience and 19 years of teaching experience. She guided M.Phil and Ph.D. Research Scholars.

PavithraVishalini, M.Phil Research Scholar, Computer Science from Bharathiar University, Coimbatore, Tamilnadu, India.