ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(41), DOI: 10.17485/ijst/2016/v9i41/100981, November 2016* ISSN (Online) : 0974-5645

A Novel Technique for Analysis of Protein to Protein Interaction using EfficientMinimum Spanning Tree Techniques

S . Jeyabalan* and V. Cyril Raj

Department of Computer Science Engineering,
Dr. M.G.R. Educational and Research Institute University, Chennai - 600095,
Tamil Nadu, India; jeyabalan@rrgroupinsts.org, cyrilraj@drmgrdu.ac.in

Abstract

In this research article, the network concepts are proving extensive study of gene function, Protein–Protein Interaction and biochemical communication pathway. **Method/Analysis:** The understanding of huge-size protein network data is depending on skill to identify significant cluster in its data sets, which is a computationally precise task. This brings a new scope to carry out research work which helps for determining new paths in graph and assist to solve the problem for identifying pathways in protein interaction networks. **Findings:** This idea breaks through to implement new technique called efficient spanning tree algorithm for finding an efficient pathway with in networks under numerous biologically motivated constraints. This method helps to hunt for protein pathways over Protein-Protein Interaction network. **Application/Improvements:** The analysis results confirmed that the proposed algorithm is capable of restructuring the signal pathways and to identify well qualified paths in an unsupervised method.

Keywords: Algorithm, Cluster, E-MST, Protein

1. Introduction

The important elements of a cell is a protein and is capable for transferring signals and helping to control the function of enzymes and help to regulate the production and activities in the cell. In the protein, one of the activities is to communicate with other proteins, DNA and other molecules. There are two ways of communication permanent protein to protein interactions and it happens only during certain cellular processes. The protein complex is nothing but groups of proteins interact with certain cellular component. In a protein complex a domain is an important part which has its own function. There is a combination domain in a protein will determine their overall functions[1,2]. Protein to Protein Interaction (PPI) grasps information about the proteins coordination with other protein in the biological processes within the cell. The majority of proteins full sequence is already known but protein molecular function is not yet fully verified. There is a major problem for predicting protein function is still major problem in computational biology research. There are some genomic sequencing methods and computational techniques have been proposed to analysis and infer protein function from bio-molecules. The major problems are to need high-throughput techniques to help for identifying proteins interaction within an organism[3].

In the real scenario proteins are highly complex components in the cell activities. It is very much essential to understand of molecular functions and in biological process. In the biological organisms proteins and their functionalities are well classified and cooperatively carry out the specific biological functions[4]. For this there are some numerous ways to analysis the protein complexes. These techniques are toadstool two hybrid and tandem

* *Author for correspondence*

attraction sanitization with mass spectrometry and these actions have some limitations. Particularly in PPI during a high level testing is carried out an elevated false-positive and false-negative rates may be occurred and this result may affects the accuracy of the algorithm's experimental results[5]. Clustering is one kind of an unsupervised mechanism which is applied widely in various investigation analyses to find out the natural grouping in a specified set of entities. Varity of approaches are used to make clustering and be broadly grouped into two methods of hierarchical and partitioned approaches. Nested series of partitions in the structure are generated through the hierarchical clustering approaches like dendrogram tree. The partition clustering technique divides the dataset into k number of clusters straight away. The conventional approaches such as hierarchical and partitioned methods are lacked in their ability by detection of the built-in clusters due to the assorted nature of protein profiles[6]. The recent study suggested more novel techniques by applying graph theory and it was proposed to analysis the protein to protein pathway interaction. Several models were suggested out of those, one of the models was graph model. The models of graph are extensively considered in different problem fields caused by their skill which represents difficult data. Protein data sets are very often greatly connected. The clustering supports with graph connectivity algorithms are suitable for identifying the PPI by extremely interconnecting profiles[7]. From the results, it was observed that the performance of graph based clustering algorithms significantly better to recognize the signaling pathway in its inter-connectivity.

Minimum Spanning Tree (MST) was applied to the connected undirected graph to make bifurcation sub-graph within the graph. The MST used minimum weight to do sub-graph and used for combinatorial optimization problem to apply for diverse tasks[8]. The MST support clustering technique is used to classify and remove the incompatible edges for attaining a set of clusters. It is always a difficult task to get preferred structure of the cluster and this act as a basic problem by this move toward. Similarly an algorithm was developed to resolve the distribution network planning by grouping particle swarm optimization with MST and the results were revealed[9]. As per the literature many algorithms were proposed, but most of the algorithms not up to benchmarks when it applied on heterogeneous datasets due to diverse in shape, different size and densities. In PPI analysis these are the general setbacks which results in noise and outliers. To find an optimal robustness clustering algorithms are challenging problem in PPI analysis. The one of the approaches like MST spectral clustering method has outperformed in clustering performance for identifying arbitrary shaped clusters but it incurs delay. A new EMST technique algorithm was proposed in this research article and the objective of this method was to integrate above said methods and also supports verity of parameters for analysis. A methodical examination of topology and protocols used in the mode of sequential data transmission was carried out the comparison analysis for all the presented on-chip interconnection of the above methods of logical constituents[10].

The organization of our article as follows: In Section 2, MST based clustering algorithm were explained. Implementation setup was discussed in Section 3. In Section 4, the concluding remarks were pointed out and in Section 5 the references were listed. Understanding the protein clustering efficiency is a complex task in a complex signaling network. There is a method called as basto clustered protein sequences which is completely based on the similarity measures attained from BLAST. This method is used to scrutinize the proteins gene ontology in each of the clusters and learnt the centre of each cluster having the necessary information about protein cluster. Hierarchical clustered techniques are used to understand the protein sequences with new alignment independent resemblance measure called CLUSS and this was very effectual for both sequences and aligned. In continuation of year 2008 study a new algorithm was presented called as CLUSS2 which was useful for the analysis of protein clusters with multiple biological functionaries[11]. A PPI pathway networks using genetic algorithm was proposed and analyzed. This algorithm has played vital role for understanding cell activities and cell evolution. The objective of this algorithm mainly focused on orient Protein–Protein Interactions instead pathways. Normally the results were given by undirected networks or graphs and advance path findings algorithms specifies the orienting protein interactions, which can progress the process for discover pathway[12–14].

2. Efficient Minimum Spanning Tree Techniques

In this section the concept of the proposed algorithm of MST clustering algorithm was explained for the analysis

of protein pathway interaction. The MST is having set of points which used to replicate the similarity of the points and their region. The similarity graph is constructed cluster and corresponds of the graph is connected to sub graph. The protein cluster analyses are mainly grouped as an amino acid sequence based on euclidean distance algorithms. The objective of this work is to find out the disjoint subsets called as clusters. The construction of clustering method was applied based on minimum spanning tree cluster analysis. In MST, the N x N distance matrix was constructed where N is the number of proteins in the dataset and the entire graph was created using the distance matrix.

The proposed algorithm uses MST graph G (V, E), where V is the set of vertices and E is the set of edges. The first minimum spanning tree is defined as the acyclic subset T1 ϵ E that connects all vertices in V and whose total length $\sum_{i,j} CT_i^{d(vi,vj)}$ is minimal. The second MST is defined as the MST of reduced graph G (V, E -T1). Here the total length is measured as the number of edges connecting all vertices. Figure 1 shows protein(s) and its (their) interactions. Black nodes represents queried proteins, dark gray represents member of proteins in signaling pathways, light gray represents non pathway members and links represents the number of interactions. The signaling pathway memberships of each protein and the interactions are listed below. EMST techniques covered by constructing the dissimilarity matrix from protein sequence data set, constructing a complete weighted undirected graph G, obtaining MST of G, edges removing from MST based on the inconsistence measures and to create a sub-tree called clusters. The first algorithm explains the steps involved in construction of clusters.

Algorithm 1: Efficient Minimum Spanning Tree based Algorithm (EMST)

Input: Dataset D, Number of clusters k.

Output: Partition Set P = {P1, P_2 …P_k}

- Let G = (V, E) is the total chart acquired from the given dataset D.
- To construct EMST from G = (V, E)
- To make adjunct matrix Aj from G and find the degree
- To partition the points into k clusters apply a cluster algorithm.

After applying this algorithm, it constructs hierarchical clustering tree by using MST. In this cluster, the termination nodes are in the ultimate clusters. To inspection of the hierarchical cluster, the user has choice to choose higher levels degree of the tree as the final clusters.

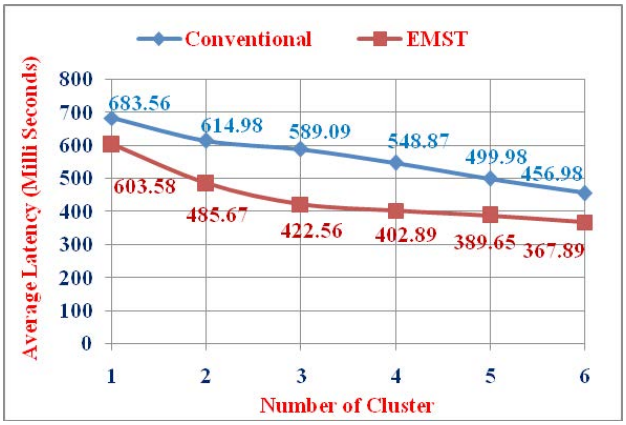

Figure 1. Number of cluster vs. average latency.

3. Implementation Setup

The proposed algorithm is standard and it can be applied to analysis the Protein to Protein Interaction pathway. The performance of this algorithm has been evaluated by using NS2 simulator. The results were compared between proposed algorithm and traditional methods for the following parameters such as speed, average linkage and accuracy rate.

3.1 Speed Analysis

The performance was analyzed to all the algorithms in terms of how many seconds was taken to bring the results.

3.2 Average Linkage

It is nothing but the distance among two clusters of shortest distance between two points in each cluster.

3.3 Accuracy Rate

Accuracy Rate = TP + TN /TP + FP + TN + FN

Where,

TP: The No. of cooperating proteins that are correctly classified.

FN: The No. of cooperating proteins that are wrongly classified non-interactive.

TN: The No. of non-cooperating protein pairs that are correctly classified.

FP: The No. of non-cooperating protein pairs that are incorrectly classified as interactive.

3.4 Speed Analysis

Figure 1 shows the performance of MST which compared against traditional algorithm in terms of speed by finding pathway. The simulations were run for about 10 times and average value was observed. From the Figure 1, it was also observed that, a diminution in an average latency can be attained through more number of clusters compared to the less number of clusters. Due to the repeated calculations conventional algorithm acquires more run time because of more distance at the time of calculations implicated with high dimensional data sets above the numerous iterations. The minimum time was taken by EMST as compared to conventional algorithm due to improve the adjacency matrix through rank transformation. The result reveals that the EMST out performance qualitative enhancement over existing method without compromise in terms of speed.

3.5 Average Linkage

From Figure 2, it was found that the performance of average linkage between the network load balance and in its links. The performance was evaluated in a single cluster through the networks in conditions with both average and maximum link stress. The domain was initially converted in to clusters using partitioning methods because of restricting the requests from traveling more distance through the network and also every state is responsible for more items. For this the EMST manage better way, when compared to conventional methods. Specifically the usage of the EMST clustering methods always permits the network to make use of its resources in superior. From Figure 2, it was also observed that the conventional algorithm perform poorer in terms of retrieval latency and hit ratio when compared with proposed algorithm EMST. Hence it was observed that the performance of EMST is always greater for more clusters during its move towards to load balancing metrics.

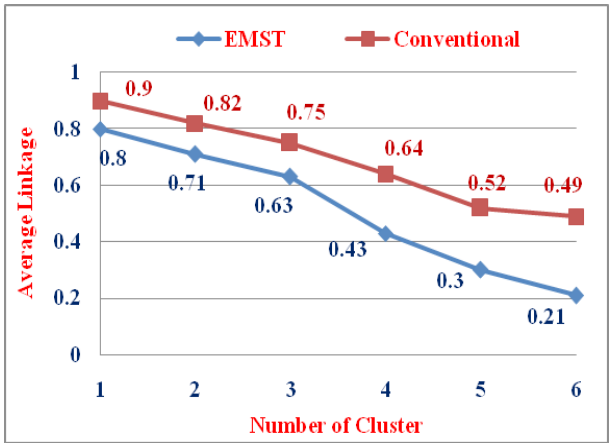

Figure 2. Cluster vs. average linkage.

3.6 Accuracy Rate

In Figure 3, the accuracy rate calculations between two algorithms by applying correlation method were discussed. Our proposed algorithm reasonably acts better than conventional methods with respect to PPI interactions. It was also observed that the effect of accuracy rate in MST is out performed than conventional. By comparing the results attained from measuring distance it was clearly indicates that the MST utilizes to calculate the distance effectively for the resemblance evaluations. Hence, it is concluded that we can obtain a better results through our proposed algorithm of EMST when compared to other conservative algorithms.

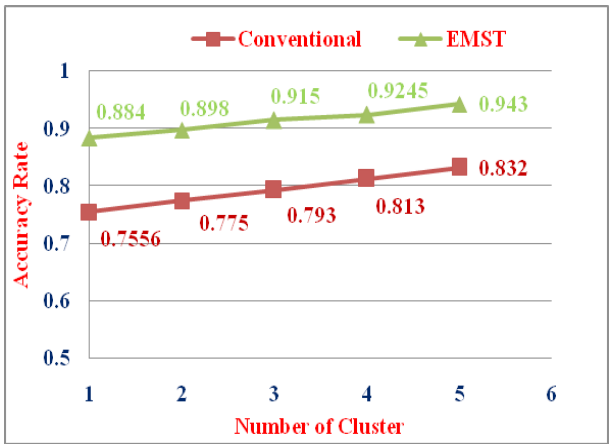

Figure 3. Cluster vs. accuracy rate.

4. Conclusions

In PPI networks to know about the cellular organizations and their biological functions the protein complex identification technique was used. We proposed EMST to dynamically construct cluster and find the pathway and compared the performance of our proposed EMST among conventional and without performed with respect to the results shown in implementation part. The experimental results showed that EMST gives maximum accuracy with less time when compared to conventional approach. It helps to analysis a protein complexes structures and accomplish to find the shortest signaling pathway effectively than MST. The overall results exhibits our algorithms was best when compared with conventional algorithm.

5. References

1. Bader GD, Hogue CW. Analyzing yeast Protein-Protein Interaction data obtained from different sources. Nature Biotechnology. 2002; 20:991–7.
2. Botlen E, Schliep A, Schneckener S, Schomburg D, Schrader R. Clustering protein sequences-structure prediction by transitive homology. Bioinformatics. 2001; 10:935–41.
3. Lodish H, et al. Molecular cell biology. New York and Basingstoke: W. H. Freeman and Co; 2005.
4. Zhang YJ, Lin HF, Yang ZH, Wang J, Li YP, Xu B. Protein complex prediction in large ontology attributed Protein-Protein Interaction networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2013; 10:729–41.
5. Sung HP, et al. Automatic protein structure clustering using secondary structure elements. Computational Science and its Applications - Part II. 2005; 277-279:324–30.
6. Golub TR, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science. 1999; 286:531–7.
7. Kelil A, Wang S, Brzezinski R, Fleury A. CLUSS: Clustering of protein sequences based on a new similarity measure. BMC Bioinformatics. 2007; 8:286–92.
8. Steffen M, Petti A, Aach J, D'haeseleer P, Church G. Automated modeling of signal transduction networks. BMC Bioinformatics. 2002; 3:34–44.
9. Hasan IJ, Gan CK, Shamshiri M, Ab Ghani MR, Omar RB. Optimum feeder routing and distribution substation placement and sizing using PSO and MST. Indian Journal of Science and Technology. 2014; 7(10):1682–9.
10. Roy S, Saha R, Bhunia CT. On efficient minimization techniques of logical constituents and sequential data transmission for digital IC. Indian Journal of Science and Technology. 2016; 9(9):1–9.
11. Kelil A, Wang S. CLUSS2: An alignment-independent algorithm for clustering protein families with multiple biological functions. International Journal of Computational Biology and Drug Design. 2008; 1(2):122–40.
12. Hoai AN, Cong LV, Minh PT, Thu LB. Discovery of pathways in Protein-Protein Interaction networks using a genetic algorithm. Data and Knowledge Engineering. 2015; 3:96–7.
13. Nallusamy S, Lakshmana Kumar DS, Balakannan K, Chakraborty PS. MCDM tools application for selection in manufacturing industries using AHP, Fuzzy Logic and ANN. International Journal of Engineering Research in Africa. 2015; 19:130–7.
14. Hoai AN. A multi-objective method for discovery of pathways in Protein-Protein Interaction networks. Proceedings of the 2014 7th IEEE Symposium on Computational Intelligence for Security and Defense Applications. 2015; 1:1–6.