A METHOD FOR COMMUNITY DETECTION IN PROTEIN NETWORKS USING SPECTRAL OPTIMIZATION

Sminu Izudheen[1] and Sheena Mathew[2]

[1]Assistant Professor, Department of Computer Science, Rajagiri School of Engineering & Technology, Kerala, India
sminu_i@rajagirtech.ac.in
[2]Professor, Division of Computer Engineering, School of Engineering, Cochin University of Science & Technology, Kerala, India
sheenamathew@cusat.ac.in

ABSTRACT

Identification of community structures in complex networks has been a challenge in many domain and discipline. In protein networks these community interactions play a vital role in identifying the outcome of many cellular mechanisms. This paper reports the use of spectral optimization of triangular modularity as an effective method to identify these community structures. The algorithm has been carefully tested on real biological data and the results acknowledge that this is a powerful method for extracting community structures from protein networks.

KEYWORDS

Community Detection; Protein-Protein Interaction; Protein Networks; Spectral Optimization

1. INTRODUCTION

Protein-protein interactions are fundamental to virtually every cellular process [1]. It can inactive a protein, alter the kinetic properties of proteins, result in the formation of a new binding sites or change the specificity of a protein for its substrate. Protein complexes performing a specific biological function often contain highly connected protein modules [2]. These connected modules can be considered as community structures of protein networks, because they are more closely associated with themselves than with the rest of the network.

Within the networks literature a great many algorithms have been proposed to locate the communities in a network. In this paper we focus on spectral optimization of protein networks as a method to find closely communicating proteins. In this paper we improve our method [3] by introducing an error factor while optimizing the modularity value of protein networks. The rest of the paper is organized as follows. In Section 2, related work is described. Triangular modularity of protein networks is discussed in section 3 and spectral optimization on triangular modularity in

DOI: 10.5121/ijdms.2011.3412

section 4. The improved algorithm for community detection and the results are presented in section 5. This section also compares the performance of new algorithm with the previous method. Finally, we conclude our work.

2. RELATED WORK

A number of methods are proposed to detect community structures in complex networks. Many of these methods use principles of artificial intelligence and graph theory. While some of these algorithms use betweenness centrality, another category uses hierarchical clustering, k-clique percolation to identify the nodes that form a community [4]. Another class of algorithms views community as a evolution of social interaction. We adopted spectral optimization of triangular modularity as a method of community detection in protein networks [3]. We preferred to make use of the decomposition algorithm proposed by Newman and Girvan due to its ability not only to divide networks effectively, but also to refuse division when no good division exists.

3. IDENTIFYING TRIANGULAR MODULARITY IN PROTEIN NETWORKS

Many real networks contain groups of nodes in which the density of their internal connections is larger than the connections with the rest of nodes in the network. This lead to the concept of community structure in complex networks, and it was first pointed out by Girvan and Newman [5]. Within the networks literature a great many algorithms have been proposed that locate these community structures. One of the most successful approaches to identify this is through the quality function called *modularity* [5], [6]. In this paper, we propose to optimize the modularity value of a protein network. Once we have the optimum modularity value of the network, we can use it to partition the network into two communities such that the number of edges within each such community is larger than one expect to find at random. Or we get subset of nodes which are more connected among themselves than with the rest of the nodes.

The modularity for weighted directed networks [7] is calculated as:

$$Q(C) = \frac{1}{2w} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(w_{ij} - \frac{w_i^{\text{out}} w_j^{\text{in}}}{2w} \right) \delta(C_i, C_j) , \qquad (1)$$

where w_{ij} is the weight of the connection from the i[th] to the j[th] node, $w_i^{out} = \sum_j w_{ij}$ represents the output strength and $w_j^{in} = \sum_i w_{ij}$ the input strength, $2w = \sum_{ij} w_{ij}$ gives the total strength of the network, C_i is the index of the community to which the node i belongs, and the Kronecker δ is 1 if nodes i and j are in the same community, and 0 otherwise. For undirected networks, $w_i^{out} = w_i^{in} \equiv w_i$.

According to equation (1) building block of community structure is the link between two nodes. But in many real time networks, functional structural entity of a graph is not a simple link but a small structure called motifs. Given a partition C of an unweighted network, motif modularity can be represented as the fraction of motifs inside the communities minus the fraction in a random network [8], given by

$$Q_M(C) = \frac{\sum_{i1i2\ldots iM} \prod_{(a,b)\in EM} w_{iaib}(C)}{\sum_{i1i2\ldots iM}^{n} \prod_{(a,b)\in EM} w_{iaib}} - \frac{\sum_{i1i2\ldots iM} \prod_{(a,b)\in EM} n_{iaib}(C)}{\sum_{i1i2\ldots iM}^{n} \prod_{(a,b)\in EM} n_{iaib}} \quad (2)$$

where

$$n^{ij} = w_i^{out} w_j^{in},$$
$$w_{ij}(C) = w_{ij}\delta(C_i, C_j)$$
$$n_{ij}(C) = n_{ij}\delta(C_i, C_j)$$

Among the different possible motifs, we used triangular motifs as method to detect the community structures in protein networks. Since we are considering an undirected graph, the triangle modularity [9] can be represented as

$$Q_{\triangle}(C) = \sum_i \sum_j \sum_k B_{ijk}\delta(C_i, C_j)\delta(C_j, C_k)\delta(C_k, C_i)\,. \quad (3)$$

Where

$$B_{ijk} = \frac{w_{ij}w_{jk}w_{ki}}{\sum_i \sum_j \sum_k w_{ij}w_{jk}w_{ki}} - \frac{(w_iw_j)(w_jw_k)(w_kw_i)}{\sum_i \sum_j \sum_k (w_iw_j)(w_jw_k)(w_kw_i)}\,.$$

4. SPECTRAL OPTIMATION OF TRIANGULAR MODULARITY IN PROTEIN NETWORKS

A proper optimization algorithm to calculate the triangular modularity value is demanding, due to the possibility of forming large number of traids. We have already developed an algorithm [3] by applying spectral optimization on the triangular modularity matrix. We have used the leading eigenvector of the modularity matrix to detect the community structure. Based on the sign of the elements the vertices are divided into two groups, with vertices whose corresponding elements are positive moves to one group and the rest moves to the other group. This process is repeated recursively, giving two partitions in each step until no new splits are possible. This is the case when there is no positive eigen value for the matrix. Then the leading eigenvector have all ones and all vertices fall in a single group. In this case, the algorithm is telling us that there is no division of the network that results in positive modularity. Hence the algorithm has the ability not only to divide networks effectively, but also to refuse to divide them when no good division possible.

To perform spectral optimization on the modularity value calculated in equation (3), we need to perform some transformations. In Belkacem Serrour et al [10], triangular modularity is reduced to standard spectral form as:

$$Q_{\triangle}(S) = \frac{3}{4}\sum_{i}\sum_{j} s_i M_{ij} s_j \cdot \qquad (4)$$

Where

$$M_{ij} = \sum_{k} B_{ijk} \cdot$$

5. ALGORITHM AND IMPLEMENTATION

We have already developed an algorithm [3] to identify the community structures in protein networks by the spectral optimization of triangular modularity as in equation (4). In this paper we are improving our algorithm by introducing an error factor ε while partitioning the vertices. In our previous algorithm, in order to divide the network into groups we computed the leading eigenvector of the modularity matrix. Then looking at the sign of the individual elements in this matrix we divided the network into two, with vertices whose corresponding elements are positive moves to one group and the rest moves to the other group. In this paper we modify the algorithm by introducing an error factor ±ε while separating the vertices. Instead of moving all positive elements to one group and remaining to other, we also included the elements falling within the error rate ε to join one among the two groups, whichever result in better modularity.

5.1. Algorithm

1. Initialize array s as 1
2. Create the adjacency matrix of protein network, w
3. Calculate $B_{ijk} = \dfrac{w_{ij}w_{jk}w_{ki}}{\sum_i \sum_j \sum_k w_{ij}w_{jk}w_{ki}} - \dfrac{(w_i w_j)(w_j w_k)(w_k w_i)}{\sum_i \sum_j \sum_k (w_i w_j)(w_j w_k)(w_k w_i)}$
4. Compute the modularity matrix $M_{ij} = \sum_k B_{ijk}$
5. Calculate modularity $Q(S) = \frac{3}{4}\sum_i \sum_j s_i M_{ij} s_j$
6. Calculate the leading eigenvector of M as V
7. Form vertex sets V_1 with +ve vertices of V and V_2 with -ve vertices of V within an error rate ±ε
8. If the vertex is in V_2 set the corresponding value in s as -1.
9. Compute the modularity matrix $M_{ij} = \sum_k B_{ijk}$

10. Calculate modularity $Q\_Sub(S) = \frac{3}{4}\sum_i \sum_j s_i M_{ij} s_j$

11. If ($Q\_Sub > Q$)

    11.1 Partition the graphs into two, one with vertices in V_1 and other with V_2

    11.2 Repeat from step 2 for V_1 and V_2

Else

    Stop

5.2. Dataset

For the present study protein interaction data is downloaded from MIPS [11] and MINT [12] databases.

5.3. Results

In this section we present the results of the spectral optimization of triangular modularity applied to real protein interaction data from individually performed experiments. Figure 1 shows the result of the algorithm when applied on actual protein interaction data downloaded from MINT. It represents a plot showing the value of the elements in the leading eigen vector during various iterations of the algorithm. From figures 1*d*,1*e* and 1*f* it is clear that we can get a better result when we set ε as -0.1, i.e., vertices with value greater than -0.1 moves to one group and the rest to the other. Similarly, Figure 2 shows the result for data downloaded from MIPS. Here, again figure 2*f* shows that we can get a better result when we set ε as -0.1.

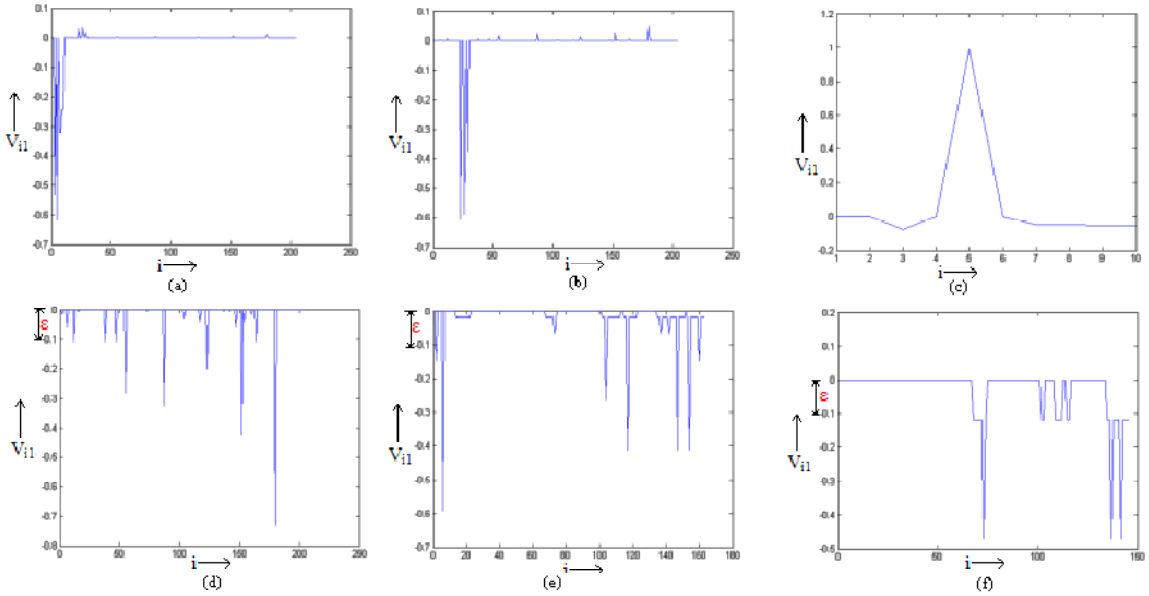

Figure 1: Result for protein interaction network extracted from MINT database. Figures 1*a* -1*f* shows the elements of the leading eigenvector during various iterations of the algorithm.

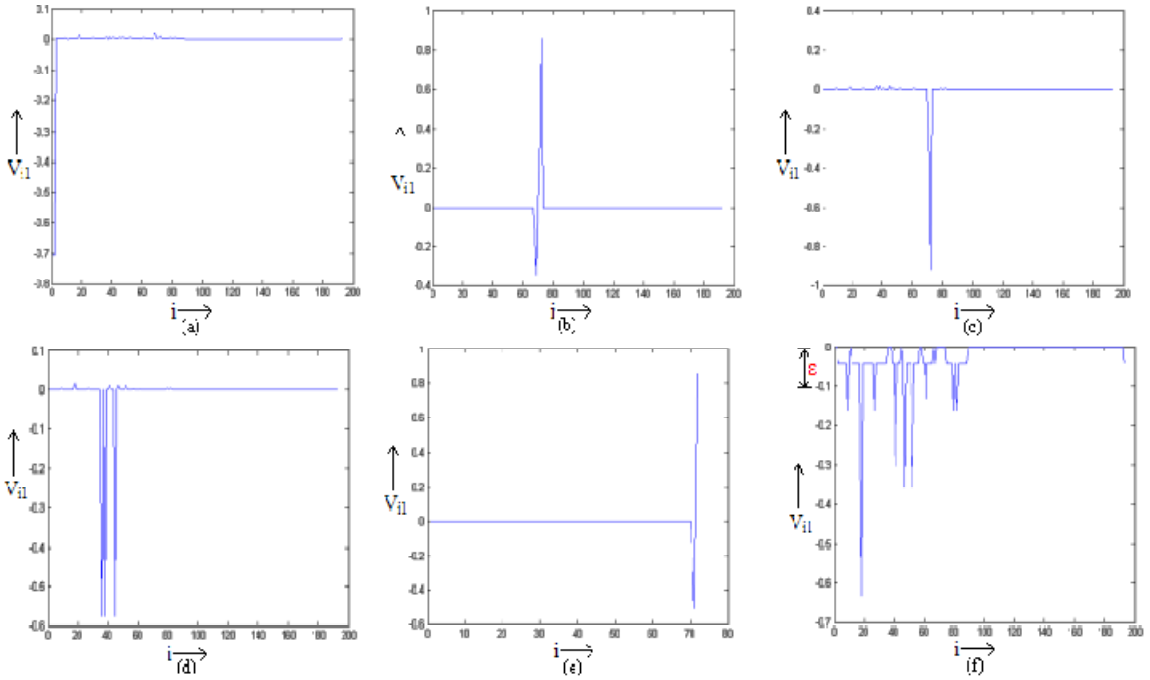

Figure 2: Result for protein interaction network extracted from MIPS database. Figures 1*a* -1*f* shows the elements of the leading eigenvector during various iterations of the algorithm.

The above figures show that the improved algorithm gives a better performance when tested on real protein interaction data extracted from MINT and MIPS. Performance comparison showing the percentage of communities identified by the previous algorithm and the proposed modified algorithm by introducing error factor ε is shown in Figure 3. The figure shows that the modified algorithm gives an average improvement of 10.5 percent over the previous algorithm when tested on data extracted from MINT and MIPS.

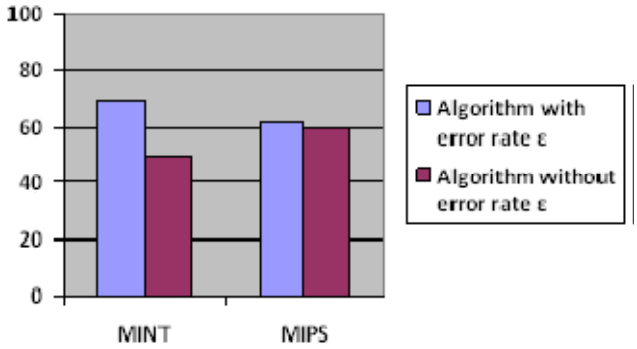

Figure 3: Performance comparison showing the percentage of community structures identified by the algorithm with error factor ε with the algorithm without error factor ε.

6. CONCLUSIONS

Identification of community structures in complex networks is a challenging issue in many domains and disciplines. Here, we are suggesting spectral optimization on triangular modularity as a promising method to analyze protein interactions. In this paper we have improved our previous work by introducing an error factor while partitioning the vertices into different communities. The modified algorithm have been tested on protein interaction data retrieved from databases like MIPS and MINT and the results shows this is a powerful method in extracting community structures from protein networks.

REFERENCES

[1] Hakes L, Lovell SC, Oliver SG, et al (2007) Specificity in protein interactions and its relationship with sequence diversity and coevolution. PNAS 104(19), 7999-8004

[2] Harwell LH, Hopfield JJ, Leibler S, Murray AW, (1999) From molecular to modular cell biology Nature. 402, c47-c52

[3] Sminu Izudheen, Sheena Mathew, (2012) A Heuristic Approach for Community Detection in Protein Networks, (Accepted in CCSIT 2012 – will be published in LNICST 85,479-485).

[4] D. Bader et al., (2007) Approximating betweenness centrality. Georgia Institute of Technology,

[5] Girvan M, Newman ME., (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA, 99:7821-7826

[6] Newman M E J and Girvan, (2004) Finding and evaluating community structure in networks Phys. Rev. E 69 026113

[7] Arenas A, Duch J, Fern´andez A and G´omez S., (2007) Size reduction of complex networks preserving modularity New J. Phys. 9, 176

[8] A. Arenas, A. Fern´andez, S. Fortunato, S. G´omez, (2008) Motif-based communities in complex networks, Journal of Physics A: Mathematical and Theoretical 41, 224001

[9] M. E. J. Newman, (2006) Modularity and community structure in networks, Proceedings of the National Academy of Sciences USA (103), 8577

[10] Belkacem Serrour, Alex Arenas, Sergio G´omez, (2010) Detecting communities of triangles in complex networks using spectral optimization, Computer Communications, May 11

[11] Munich Information Center for Protein Sequences, http://www.helmholtz- muenchen.de/en/mips/home/index.html

[12] MINT: the Molecular INTeraction database, http://mint.bio.uniroma2.it/mint/Welcome.do