2014 3rd International Conference on Environment Energy and Biotechnology
IPCBEE vol.70 (2014) © (2014) IACSIT Press, Singapore
DOI: 10.7763/IPCBEE. 2014. V70. 1

Focal Structure Analysis in Large Biological Networks

Fatih Şen [1+], Rolf T. Wigand [2], Nitin Agarwal [3], Mutlu Mete [4] and Rafal Kasprzyk [5]

[1] Integrated Computing Program, EIT, and University of Arkansas at Little Rock

[2] Department of Information Science and Business Information Systems, University of Arkansas at Little Rock

[3] Department of Information Science, University of Arkansas at Little Rock

[4] Department of Computer Science & Information Systems, Texas A&M University-Commerce

[5] Military University of Technology, Faculty of Cybernetics

Abstract. After the completion of the Human Genome Project, identifying relevant protein structures became an important factor for detecting new disease-related structures. The newly available large-scale networks of molecular structures within the cell have made it possible to study protein function(s) in the context of a network. Those Protein-Protein Interaction (PPI) networks have been studied through the identification of clusters or communities by researchers. Proteins, however, interact in smaller and more pertinent groups. A new methodology, called Focal Structures Analysis (FSA), is presented to identify focal structures, i.e., smaller and more relevant structures. This research advances our understanding of the role and impact of the focal structures and can help researchers with discovering protein related diseases.

Keywords: PPI networks, focal structures, FSA, Louvain method

1. Introduction

After the completion of the Human Genome Project, researchers have been paying more attention to the study of the interactions of genes and proteins than only cataloguing and listing them to understand biological functions. Especially, protein-protein interactions play an important role to understand the architecture and function of the cellular organisms. Proteins interact with one another as networks and with DNA, RNA to form molecular functions. Studying the network of the proteins is very important, because the identification of relevant protein interactions can help researchers with identifying new disease-related structures, and can even elucidate the genetic basis of the related disease(s). Such studies can help people to improve methods for prevention, diagnosis, and treatment [1], [2]. To detect such biological complexes, many computational approaches have been proposed such as identifying clusters or communities in those Protein-Protein Interaction (PPI) networks. Most of those current methods detect larger structures in which non-relevant proteins most likely exist. However, proteins interact in smaller and more relevant groups. We propose a new methodology, called Focal Structures Analysis (FSA), addressing this challenge and identifying more relevant protein structures in a protein-protein interaction network.

We make the following contributions in this paper:

1. Defining a focal structure.
2. Developing a methodology to extract focal structures.
3. Distinguishing focal structures from communities by demonstrating biological relevance of focal structures.

[+] Correspondence author: Fatih Şen. Tel.: + 1 501-765-4057.
E-mail address: *fxsen@ualr.edu*

The paper is organized as follows: The next section 2 provides the related work. Focal structures are defined in section 3. Section 4 explains how to identify focal structures. The evaluation of the proposed approach with experimental results is presented in section 5. Lastly, the paper offers a conclusion and future work.

2. Related Work

In biological networks, the goal is to identify modular structures, such as protein complexes in a PPI network, which share common function(s). Some common clustering algorithms, such as MINE [3], LCMA [4], CODE [5], COACH [6] have been proposed for PPI networks over the past decade. Most of the studies of this field have been conducted through cliques or near-cliques, because protein complexes, for example, are usually densely connected sets of proteins. However, there might be some significant protein complexes in which the proteins are sparsely connected but important for causing certain diseases, such as breast cancer or Alzheimer disease. The algorithm we propose includes structures that are not densely connected structures but also structures with smaller density values.

Other algorithms have been proposed to identify communities in a complex network. The most well-known method is the Newman approach [7], which can identify communities in a biological network. SCAN [8] considers communities as functional modules in complex biological networks. The Louvain method [9] also detects communities in complex networks. It is known as the best partitioning method for large graphs [9]. However, those community identification algorithms suffer from the resolution problem [10]: it fails to identify communities smaller than a scale which depends on the total size of the network and on the degree of interconnectedness of the communities [11]. This challenge leads to identifying structures with irrelevant data and missing relevant protein complexes. In contrast to those community identification methods, our approach attempts to solve the resolution problem. It can identify smaller structures, named as focal structures, with more relevant protein complexes.

In the experiment section, we compare the performance of the proposed approach with other types of structures in network, such as communities.

3. Problem Statement

The main focus of this research is to identify the most significant focused structures in a complex network. A significant structure is a key set of vertices which may be responsible for some specific functions. A focal structure must have at least two vertices connected with an edge. A set of focal structures F needs to be identified in graph G with vertices V and edges E. Each focal structure is unique so that a focal structure cannot be a subset of any other focal structure. Formally, a focal structure can be defined as follows:

Given a social network G = (V, E), where V is the set of vertices and E is the set of edges, a focal structure can formally be defined as follows:

Focal structures in G are defined by F = {G′}, where G′ = (V′, E′) and V′ $\subseteq$ V and E′ $\subseteq$ E. For all i and j, i $\neq$ j, $G_i \in F$ and $G_j \in F$, such that no two focal structures can subsume each other, or $G_i \not\subset G_j$ and $G_j \not\subset G_i$.

4. Methodology

We present a methodology for FSA to identify focal structures from a complex network of interactions among individuals. First, we leverage the concept of modularity recursively to extract candidate focal structures from a complex network. The process of obtaining candidate focal structures is known as focal patterns presented in our previous research [12]. However, this approach tends to identify densely connected structures and may miss other focal yet sparsely connected structures. Therefore, in order to identify sparsely connected focal structures as well, we stitch the candidate focal structures together if there is sufficient similarity between them. We also define the similarity measure in the following discussion.

4.1. Stitching candidate focal structures

To stitch two candidate focal structures together, we propose to measure the degree of interconnectivity by the similarity of two subgraphs (candidate focal structures) i and j defined below:

$$Similarity(i, j) = \frac{\text{int}\, erconnectivity(i, j)}{unionsize(i, j)} \tag{1}$$

$$J = \frac{|X \cap Y|}{|X \cup Y|} \tag{2}$$

```
Data: Candidate Focal Structures
Result: Focal Structures

function StitchCandidateFocalStructures

    siblings ← Find all the siblings
    MaxPair: The candidate focal structure pair(i,j) with maximum
interconnections among all sibling pairs
    Similarity: The similarity value between i and j candidate focal structures
    Sort the siblings in descending order
    while siblings do
      MaxPair ← Find Maximum Pair
      while Similarity(MaxPair i and j) ≥ thresholdVal do
          newGraph ← Stitch candidate focal structures i and j
          Remove candidate focal structure i from siblings
          Remove candidate focal structure j from siblings
          Add newGraph to siblings
          Sort the siblings in descending order
          MaxPair ← Find Maximum Pair
      end
    end
    return siblings
end function
```

Fig. 1: Stitching the Candidate Focal Structures.

where interconnectivity(i,j) is the number of edges between subgraphs i and j, and unionsize(i,j) is the total edge number of subgraph i and j (Equation 1). The Similarity (i,j) between two subgraphs i and j is the ratio of the number of the edges between them to the total number of the edges of the two subgraphs. It is based on a well-known Jaccard's Coefficient in which the similarity is the ratio of the shared elements to the union size of the two sets X and Y (Equation 2). We stitch highly interconnected candidate focal structures based on their similarity values. The pair of candidate focal structures that have the highest similarity are stitched in each iteration and the stitching process iterates until the highest similarity of all sibling pairs is less than a given threshold (See Algorithm 1). Various threshold values such as 0.25, 0.3, 0.4, 0.5, 0.6, 0.7 and 0.8 are applied in the algorithm and 0.4 is taken since it is the best threshold value giving the highest performance results.

5. Experiments

To obtain focal structures and communities, we respectively apply FSA and Louvain methods for a PPI dataset and compare the performances of those two approaches.

5.1. Dataset & gene ontology

Protein-Protein Interaction Network Hand-curated databases of PPI in Saccharomyces cerevisiae have been studied earlier in the literature [13]-[15] and are demonstrated to be invaluable resources for bioinformatics research. For this study, the PPI network is downloaded from the Saccharomyces Genome Database (http://www.yeastgenome.org/). The network dataset contains 4,715 vertices (proteins) and 43,540 edges (interactions).

GO database (http://www.geneontology.org/) provides controlled vocabularies for the description of the 1) molecular function (MF), 2) biological process (BP), and 3) cellular component (CC) of gene products. GO database is accepted as ground-truth and used for comparison and validation purposes. Figure 1 shows the size distribution of the annotations for the ground truth of BP. It is a heavy-tailed distribution with many annotations with few proteins and few annotations with many proteins. There are 383 annotations including

only one protein, 195 annotations of two proteins and 112 annotations of three proteins. The last three largest annotations, which occur only once, contain 111, 185, and 252 proteins. There are other annotations ranging between four and 100 proteins. Similar size distribution is observed for the categories of MF and CC.

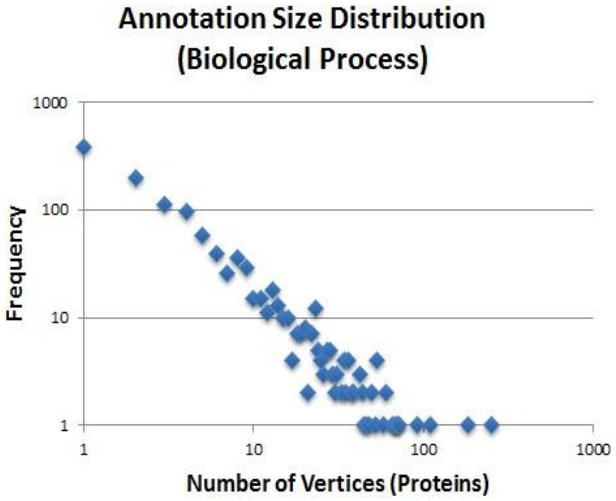

Fig. 2: Distribution of the size (number of proteins) of the annotation of "Biological Process".

5.2. Evaluation metrics

To evaluate the performance of our algorithm, we use precision and recall, which are widely used in machine leaning for evaluating the accuracy of pattern recognition algorithms. To calculate the accuracy of the structures, we utilize the Gene Ontology (GO) database as ground-truth. The definitions of precision and recall are

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \tag{3}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \tag{4}$$

where *True Positive (TP)* is the number of correctly detected proteins, i.e. with common annotation(s) in a structure, *False Positive (FP)* is number of identified proteins which do not exist in the GO database at all, and *False Negative (FN)* is the number of proteins in the related annotation, which could not be identified in a structure at all. According to these evaluation metrics, a higher value of precision indicates that proteins identified in a structure are more likely to share common annotation(s), and a higher value of recall shows that more proteins that share common annotation(s) in a structure are identified. Also, the F-Measure (Equation 5), i.e. the harmonic mean of precision and recall, is used to evaluate the overall performance. The F-Measure is defined as,

$$F - Measure = \frac{2 * \mathrm{Pr}ecision * \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call} \tag{5}$$

5.3. Results

As shown in Table 1, it is obvious that FSA generates more relevant structures than the Louvain method. For example, in BP, FSA detects 259 relevant structures out of 1123 focal structures with pre-existing 1,185 annotations while Louvain generates five relevant communities out of 36 communities with pre-existing 1,185 annotations. If all the proteins of a focal structure share at least one common annotation, then this structure is counted as a 100% biologically relevant structure. After calculating the F-Measure for each structure, we obtain the mean and standard deviation values as shown in Figure 2 and Figure 3. For all the

three GO categories, they clearly demonstrate that FSA outperforms Louvain and more relevant structures with less likelihood of noise are observed.

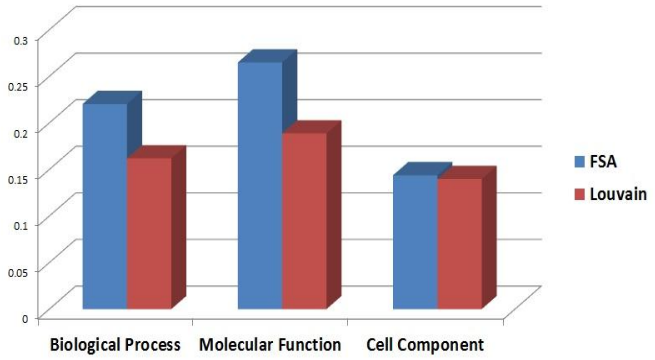

Fig. 3: F-Measures (According to Mean) of FSA and Louvain Algorithms.

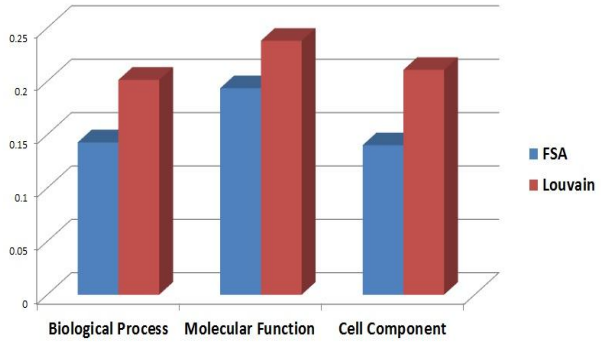

Fig. 4: F-Measures (According to Std. Dev.) of FSA and Louvain Algorithms.

Table 1. Number of Structures by FSA and Louvain Algorithms

	Annotation	Identified Relevant	Relevant (100%)
FSA - Biological Process	1185	1123	**259**
Louvain-Biological Process	1185	36	5
FSA-Molecular Function	1081	1123	**320**
Louvain-Molecular Function	1081	36	6
FSA - Cell Component	489	1123	**289**
Louvain-Cell Component	489	36	4

5.4. Analysis & discussion

We perform the analysis for the three GO categories and similar results are obtained. However, due to space limitation, we present only the size distribution of BP in Figure 4 which clearly shows that FSA identifies smaller structures as compared to Louvain, while Louvain identifies more larger structures as compared to FSA.

FSA identifies structures that are smaller and have higher F-Measure values, while Louvain generates much larger structures having lower F-Measure values. For example, for BP, FSA identifies one focal structure of two proteins, two focal structures of three proteins, and three focal structures of four proteins with F-Measures ranging between 0.6 and 0.8, whereas Louvain can identify only one community for the same range. The best F-Measure value of the Louvain method is 0.8 with only one community of three proteins. FSA's best F-Measure value is 1.0 with three structures of two proteins each.

For other two processes, i.e., MF and CC, similar results are obtained. For MF, the best F-Measure value is 1.0 for the eight focal structures, while the best F-Measure value is 0.838 for only one community with the

Louvain method. For CC, FSA identifies four focal structures with 1.0 F-Measure value, whereas Louvain identifies only one community with 1.0 F-Measure value. It is obvious that FSA generates smaller structures with higher accuracy values more frequently.

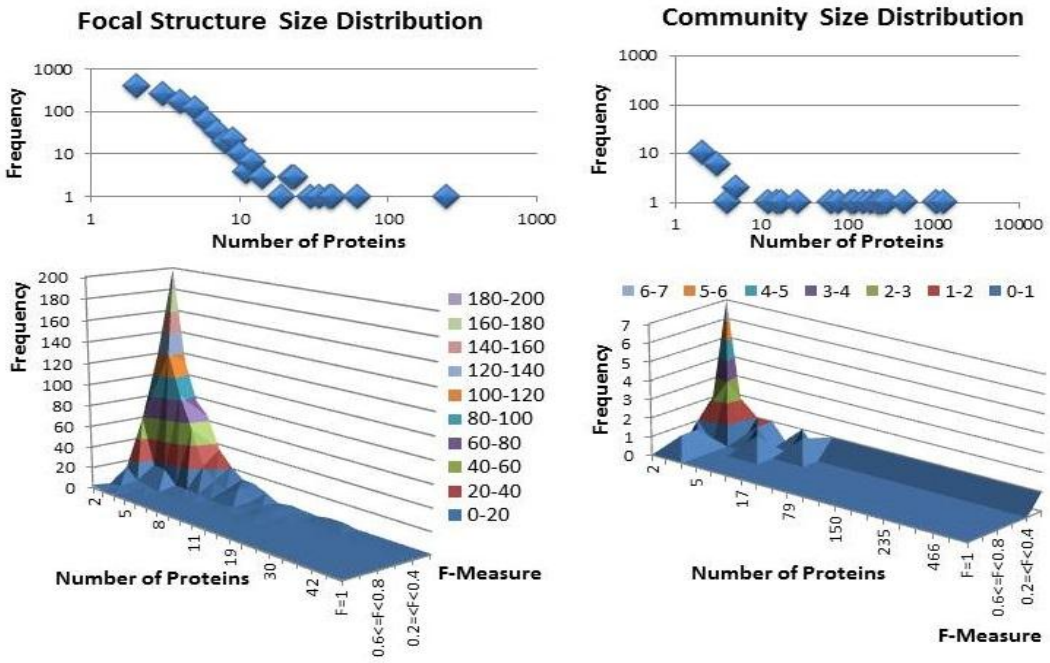

Fig. 5: The size distribution and accuracy values of the focal structures and communities for the "Biological Process".

6. Conclusion and Future Work

We develop a methodology, called FSA, to extract focal structures of a complex network. Significant protein complexes are identified for a biological network. The performance between FSA and Louvain community identification method is compared. We demonstrate that focal structures are smaller and more relevant groups of proteins as compared to communities identified by the Louvain method. For our future work, we plan to compare FSA with other community identification methods. This research can help identify disease-related structures and improve methods for prevention, diagnosis, and treatment.

7. References

[1] M. G. Kann. Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Briefings in Bioinformatics*, 2007 **8** (5): 333-346.

[2] T. Ideker and R. Sharan. Protein networks in disease. *Genome Research*, 2008, 18(4): 644-652.

[3] K. Rhrissorrakrai and K.C. Gunsalus. Mine: module identification in networks. *BMC Bioinformatics* 12 (2011) 192.

[4] X. L. Li, S. H. Tan, C. S. Foo, S. K. Ng, et al. Interaction graph mining for protein complexes using local clique merging. *Genome Informatics Series*, 2005, **16** (260).

[5] G. D. Bader and C.W. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 2003, **4** (2).

[6] M. Wu, X. Li, C. K. Kwoh and S. K. Ng. A core-attachment based method to detect protein complexes in ppi networks. *BMC Bioinformatics*, 2009, **10** (169).

[7] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 2002, **99** (12): 7821-7826.

[8] M. Mete, F. Tang, X. Xu and N. Yuruk. A structural approach for finding functional modules from large biological networks. *BMC Bioinformatics, 2008,* 9 (Suppl 9) S19.

[9] V. D. Blondel, J. L. Guillaume, R. Lambiotte and E. Lefebvre. Fast unfolding of communities in large networks.

Journal of Statistical Mechanics: Theory and Experiment, 2008, (10) P10008.

[10] H. Kwak, Y. Choi, Y. H. Eom, H. Jeong and S. Moon. Mining communities in networks: a solution for consistency and its evaluation. *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement, ACM*, 2009, 301-314.

[11] J. Chen, O. R. Zaiane and R. Goebel. Detecting communities in social networks using max-min modularity. *SDM*, 2009, Volume 3: 20-24.

[12] F. Sen, R.T. Wigand, N. Agarwal, D. Mahata and H. Bisgin. Identifying focal patterns in social networks. *Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on, IEEE*, 2012, 105-108.

[13] S. H. Yook, Z. N. Oltvai and A. L. Barabasi. Functional and topological characterization of protein interaction networks. *Proteomics*, 2004, 4(4): 928-942.

[14] V. Arnau, S. Mars and I. Marins. Iterative cluster analysis of protein interaction data. *Bioinformatics* 2005, **21** (3): 364-378.

[15] L. F. Wu, T.R. Hughes, A. P. Davierwala, M. D. Robinson, R. Stoughton, S. J. Altschuler. Large-scale prediction of saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nature Genetics*, 2002, **31**(3): 255-265.