International Journal of Computational Intelligence and Informatics, Vol. 3: No. 2, July - September 2013

A Study on Clustering the Protein Interaction Networks using Bio-Inspired Optimization

R Gowri
Department of Computer Science
Periyar University
Salem, Tamil Nadu, India
gowri.candy@gmail.com

R Rathipriya
Department of Computer Science
Periyar University
Salem, Tamil Nadu, India
rathi_priyar@yahoo.co.in

***Abstract*-The gist of the paper is to provide an insight about the various clustering using bio-inspired optimization for Protein Interaction Network. The major idea behind the clustering protein-protein interaction network is to identify dense sub-graphs that show significant functional modules in protein-protein interactions. A set of proteins that interact with each other are actors of a specific cellular process is termed as significant functional module which helps to recognize the structure and functional dynamics of the cell. In order to find those modules we have to cluster them based on their interaction. It is very difficult to define the boundaries among clusters and unable to identify the overlapping clusters. The bio-inspired optimization techniques are helpful in overcoming those difficulties.**

Keywords-Clustering, Protein Interaction Network (PIN), Ant Colony Optimization (ACO), Bacteria Foraging Optimization (BFO), Genetic Algorithm, Evolutionary Algorithm

I. INTRODUCTION

Clustering is the process of grouping data objects into sets (clusters) [1]. The objects in a cluster share some common characteristics. The similarity among objects in the same cluster is greater than in different clusters. Clustering differs from classification; in the latter, objects are assigned to predefined classes, while clustering defines the classes themselves. Thus, clustering is an unsupervised classification method, which means that it does not rely on training the data objects in predefined classes.

Protein-protein interactions are the fundamental to almost all biological processes [2, 3]. As advances in high-computational or hyper computational technologies, such as yeast-two-hybrid, mass spectrometry, and protein chip technologies, huge data sets of protein-protein interactions are available. Such protein-protein interaction data can be generally represented in the form of networks, which not only give us the initial global picture of protein interactions on a genomic scale but also help us understand the basic components and organization of cell machinery from the network level.

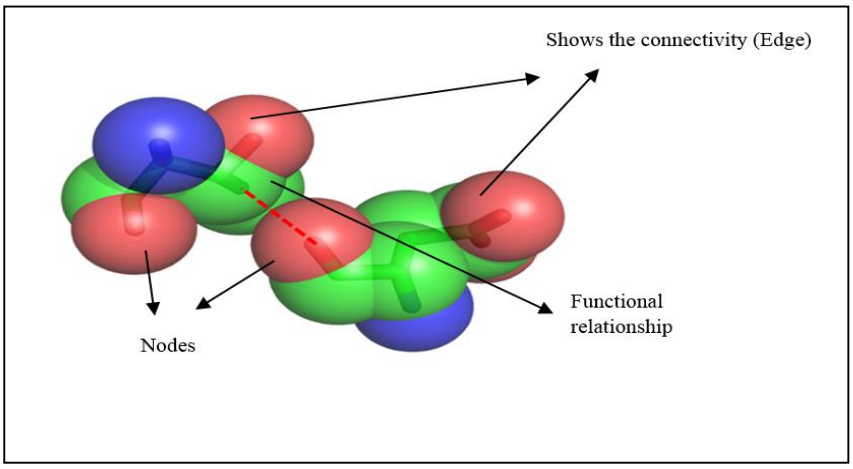

Figure 1. A Sample Protein Interaction Network

Generally, a protein interaction network is represented by an interaction graph with proteins as vertices (or nodes) and interactions as edges [4, 3]. Various topological properties of protein interaction networks have been studied frequently in the literature, such as the network diameter, the distribution of vertex degree, the clustering coefficient and etc. Here, in this paper gives a clear about the how the clustering techniques are applied for the

protein data to identify the protein-protein interactions. These interactions play a vital role in the any biological process.

The modules of PIN clusters are of two type protein complexes and functional modules [5]. Protein complexes are groups of proteins that interact with each other at the same time and place, forming a single multi-molecular machine. Functional modules consist of proteins that participate in a particular cellular process while binding to each other at a different time and place. Clustering in protein-protein interaction networks therefore involves identifying protein complexes and functional modules. This process has the following analytical benefits:

i. Clarification of PPI network structures and their component relationships.
ii. Inference of the principal function of each cluster from the functions of its members.

The possible functions of the cluster members are elucidated by comparing their functions with other members.

A. Various Clustering Approaches

Clustering approaches for PPI networks [1, 6] can be broadly characterized as follows

i. *Distance-based clustering:* It uses classic clustering techniques and focuses on the definition of the distance between proteins.
ii. *Graph-based clustering:* It includes approaches which consider the topology of the PPI network. Based on the structure of the network, the density of each sub graph is maximized or the cost of cut-off minimized while separating the graph.
iii. *Betweenness centrality-based clustering:* Betweenness centrality is an important metric for analyzing protein interaction network. There are two types of betweenness centrality: the vertex betweenness and the edge betweenness.
iv. *Hierarchical clustering:* It is one of the most common methods of classification used in biology and bioinformatics.

B. Organization of Paper

The organization of the paper is as follows. The section 1 deals with the optimization techniques that are applicable to the Protein-Protein Interaction Networks. The section 2 summarizes various optimized clustering algorithms that are applied so far to the PIN. The section 3 summarizes the process of detecting the PIN clusters in order to determine the functionality of protein using the optimized clustering techniques.

II. BIO-INSPIRED OPTIMIZATION

Biologically inspired algorithms are a category of algorithms that imitate the way nature performs [7]. This category has been quite popular, since numerous problems can be solved without rigorous mathematical approaches. Optimization is a commonly encountered mathematical problem in all engineering disciplines. It literally means finding the best possible/desirable solution. Optimization algorithms can be either deterministic or stochastic in nature. Former methods to solve optimization problems require enormous computational efforts, which tend to fail as the problem size increases. This is the motivation for employing bio inspired stochastic optimization algorithms as computationally efficient alternatives to deterministic approach. Meta-heuristics are based on the iterative improvement of either a population of solutions (as in Evolutionary algorithms, Swarm based algorithms) or a single solution (eg. Tabu Search) and mostly employ randomization and local search to solve a given optimization problem.

A. Various Bio-Inspired optimization Techniques:

1) Ant Colony Optimization (ACO)

ACO is inspired by the behavior of ants during the colony searching for the shortest path [8]. The pheromones deposited by ants attract other ants which then will increase the pheromones. ACO is a probabilistic technique which solves the computational problems that have ability to reduce finding good path through nodes in graph.

2) Bacteria Foraging Optimization

Bacteria Foraging Optimization (BFO) algorithm is a new class of biologically encouraged stochastic global search technique based on mimicking the foraging behavior of E. coli bacteria [9]. This method is used for locating, handling, and ingesting the food. During foraging, a bacterium can exhibit two different actions: tumbling or swimming [3]. The tumble action modifies the orientation of the bacterium. During swimming means the chemotaxis step, the bacterium will move in its current direction

3) Genetic Algorithm

Genetic algorithms (GAs) are adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetic [10]. The basic concept of GAs is designed to stimulate processes in natural system necessary for evolution, specifically those that follow the principles first laid down by Charles Darwin of survival of the

fittest. As such they represent an intelligent exploitation of a random search within a defined search space to solve problem.

4) Evolutionary Algorithm

Evolutionary Algorithm (EA) is a generic population-based Meta heuristic optimization algorithm [11]. An EA uses mechanisms inspired by biological evolution, such as reproduction, mutation, recombination, and selection. Candidate solutions to the optimization problem play the role of individuals in a population, and the fitness function determines the quality of the solutions (see also loss function). Evolution of the population then takes place after the repeated application of the above operators. Artificial evolution (AE) describes a process involving individual evolutionary algorithms; EAs are individual components that participate in an AE. Evolutionary algorithms often perform well approximating solutions to all types of problems because they ideally do not make any assumption about the underlying fitness landscape.

III. PIN CLUSTERING USING BIO-INSPIRED OPTIMIZATION

A. Ant Colony Optimization

ACO is inspired by the behavior of ants during the colony searching for the shortest path [12]. The pheromones deposited by ants attract other ants which then will increase the pheromones. ACO is a probabilistic technique which solves the computational problems that have ability to reduce finding good path through nodes in graph. ACOPIN works on the basis of the topological properties of PIN. The goal is to find an optimal path in a given PIN. As said earlier, the protein-protein interactions or PIN are represented using connectivity graph (G = V, E) where nodes (or vertices, V) correspond to proteins and edges, E correspond to the interactions. This PIN is represented by the interaction matrix aij, and the distances are calculated between nodes by transforming the interaction matrix aij to distance matrix dij using Bond Energy Algorithm (BEA). The TSP-problem approach best suited for the PIN where the ants just explore the PIN without exploiting the information of the nodes (proteins). The edge information (distances) is the input. In this case, we apply the ASDecision Rules when at junction to choose. The GO Terms is used to validate the initial predicted clusters and based on the probability only. The cut-off value is used to define the boundaries of protein clusters. If is too high, the clusters contain too many proteins even though some of which are not similar at all and vice versa.

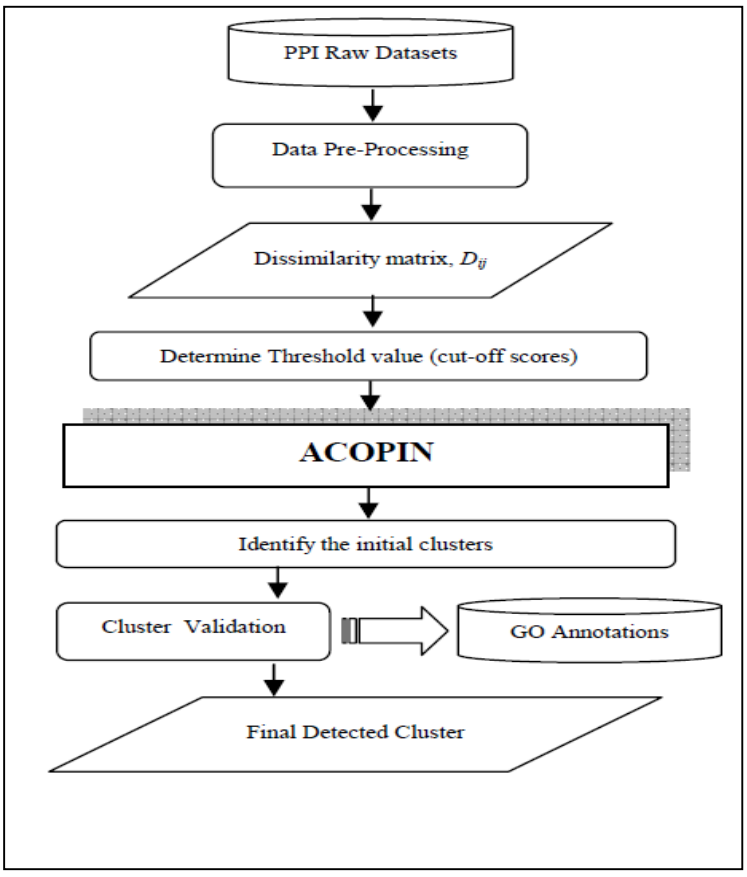

Figure 2. Flow chart for ACOPIN-v1

1) Bacteria Foraging Optimization

BFO algorithm is an evolutionary algorithm which consists of chemotactic, reproduction, and elimination dispersal operations [13]. The bacterium moves in two different ways to avoid noxious environment which is regarded as chemotactic behavior. In general, several bacteria which are becoming more and more incapable of searching food are obliged to be eliminated. In order to maintain the scale of population, the remained bacteria will reduplicate and generate new individual which is considered to be reproduction behavior. Because of the sudden change in the local environment, the bacteria population may be gradually inadaptable to the environment which leads to a fact that a group of bacteria are either killed or dispersed into a new location. This phenomenon is the elimination-dispersal behavior which can prevent the algorithm from trapping into the local optimal solution and search for a new individual which is much closer to the global optimal solution.

2) Relevant concepts of PPI networks

The weighted degree of node is defined as the summation of the weight value of edges between nodes i and its neighbors. The clustering coefficient of node which is used to assess the quality of cluster result is calculated by the following equation.

$$C_i = 2n_i/k_i(k_i - 1) \quad (1)$$

Where k_i represents the degree of node i, n_i refers to the number of edges connecting all the neighbor nodes of i with each other. Recently the concept of clustering coefficient of node is extended to edge and the accumulation coefficient of edge is defined as follows:

$$WCu,v=\frac{\sum_{k\in I_{u,v}} w(u,k)\cdot\sum_{k\in I_{u,v}} w(u,k)}{\sum_{s\in N_u} w(u,s)\cdot\sum_{t\in N_v} w(v,t)} \quad (2)$$

Where the sets N_u and N_v represent the sets of directly adjacent nodes of node u and node v respectively. The symbol w (u,s) refers to the weight value of edge linking nodes u with s. The set $I_{u,v}$ stands for the set of common nodes between the adjacent nodes of nodes u and v. The Comprehensive Network Feature Value (CNFV) of node can reveal the joint strength among this node and other nodes. The CNFV of node i is defined as follows:

$$CNFV_i = \beta * C_i + (1 - \beta) * w(i)/n \quad (3)$$

The parameter _ is a random number within 0 and 1, w(i) refers to the weighted degree of node i, and n stands for the number of protein nodes in PPI network.

3) Object function

The clustering coefficient of a cluster module is defined as the average clustering coefficient of all the protein nodes belonging to this cluster module [14]. The equation is as follows:

$$C_B= \frac{\sum_{j=1}^{h} C_{Bj}}{h} \quad (4)$$

In Eq. (4), the parameter C_{Bj} represents the cluster coefficient of node *j* and *h* stands for the number of nodes which belong to cluster *B*.

4) Evaluation criteria of cluster result

In general, a large number of studies on clustering analysis [14] adopt precision and recall values to evaluate cluster result. Suppose that X represents one cluster module in the cluster results, Fi stands for the matched cluster module in the standard PPI dataset.

$$recision(X,F_i) = \frac{|X\cap F_i|}{|X|} \quad (5)$$

$$Recall(X,F_i) = \frac{|X\cap F_i|}{|F_i|} \quad (6)$$

where the expression |X∩Fi| stands for the number of common proteins between cluster modules X and F_i. However, these two criteria have drawbacks in facing with the unexpected circumstances of larger and smaller cluster modules. Therefore, we assess the accuracy of modules with the f-measure value which balances precision, recall, and running time:

$$f\text{- Measure} = \frac{3}{\frac{1}{precision}+\frac{1}{recall}+time} \quad (7)$$

B. Evolutionary Algorithm

Differential Evolution (DE) [11] is a stochastic, population-based search strategy. While DE shares similarities with other evolutionary algorithms (EA), it differs significantly in the sense that distance and direction information from the current population is used to guide the search process. Furthermore, the original DE strategies were developed to be applied to continuous-valued landscapes.

DE differs from these evolutionary algorithms in that

- Mutation is applied first to generate a trial vector, which is then used within the crossover operator to produce one offspring, and
- Mutation step sizes are not sampled from a prior known probability distribution function.

```
Set the generation counter, t = 0;
Initialize the control parameters, β and pr;
Create and initialize the population, C(0), of ns individuals;
while stopping condition(s) not true do
  for each individual, x_i(t) ∈ C(t) do
     Evaluate the fitness, f(x_i(t));
     Create the trial vector, u_i(t) by applying the mutation operator;
     Create an offspring, x'_i(t), by applying the crossover operator;
      if f(x'_i(t)) is better than f(x_i(t))  then
         Add x'_i(t) to C(t + 1);
     end
     else
        Add x_i(t) to C(t + 1);
     end
   end
 end
Return the individual with the best fitness as the solution;
```

Figure 3. Steps in Differential Evolution Algorithm

Memetic Algorithm is a class of stochastic global search heuristics in which Evolutionary Algorithms based approaches are combined with problem-specific solvers [15]. The later might be implemented as local search heuristics techniques, approximation algorithms or, sometimes, even (partial) exact methods. The term MA is now widely used as a synergy of evolutionary or any population-based approach with separate individual learning or local improvement procedures for problem search. Quite often, MA is also referred to in the literature as Baldwinian evolutionary algorithms (EA), Lamarckian EAs, cultural algorithms, or genetic local search.

```
Procedure Memetic Algorithm
Initialize: Generate an initial population;
 while Stopping conditions are not satisfied do
    Evaluate all individuals in the population.
    Evolve a new population using stochastic search operators.
    Select the subset of individuals, Ω_il, that should undergo the individual improvement procedure.
    for each individual in Ω_il do
    Perform individual learning using meme(s) with frequency or probability of f_il,
     for a period of t_il.
     Proceed with Lamarckian or Baldwinian learning.
 end for
 end while
```

Figure 4. Steps in Differential Memetic Algorithm

IV. INFERENCE

This work has provided a review of a set of clustering approaches and optimization algorithms that are applied so far to the PIN which has yielded promising results in application of protein-protein interaction networks. Clustering a PPI network permits a better understanding of its structure and the interrelationship of constituent components. More significantly, it also becomes possible to predict the potential functions of unannotated proteins by comparison with other members of the same cluster.

TABLE I. VARIOUS WORKS ON CLUSTERING PROTEIN INTERACTION NETWORKS IN THE LITERATURE

Author Name	Title	Proposed Work
Jamaluddin,et.al	An Improved Ant Colony Optimization Algorithm for Clustering Protein Interaction Network [12]	ACOPIN
J. Sallim, et.,al	ACOPIN: An ACO Algorithm with TSP Approach for Clustering Proteins from Protein Interaction Network [13]	ACOPIN with TSP for clustering
LEI Xiujuan,WU Shuang,GE Liang et al.,	Clustering PPI Data Based on Ant Colony Optimization Algorithm [16]	ACO for clustering
Xiujuan Lei, et.al.	Clustering PPI Data Based on Bacteria Foraging Optimization Algorithm [14]	BFO for clustering PIN
Lei X, et.al.	Clustering and overlapping modules detection in PPI network based on IBFO [17]	IBFO for detecting functional modules by clustering
Iman Sharafuddin, et.al.	Protein-Protein Interaction Network Clustering Using Particle Swarm Optimization [18, 19]	PSO for clustering PIN
Pratyusha Rakshit	Protein-Ligand Docking And Protein-Protein Interaction Using Evolutionary Algorithm[15]	Evolution with Temporal Difference Q-Learning (DE-TDQL)
Pratyusha Rakshit	Protein-Ligand Docking And Protein-Protein Interaction Using Evolutionary Algorithm[15]	Memetic algorithm for clustering PIN
Jos´E Juan Tapia Valenzuela	A clustering genetic algorithm for inferring protein-protein Functional interaction sites [10]	Genetic algorithm for finding PIN cluster boundaries
TIAN Jian-Fang, et.al.	PPI Network Clustering Based on Artificial Bee Colony and Breadth First Traverse Algorithm [20]	Artificial Bee Colony for clustering

The inference from the different sort of papers is as follows.

- The functions of unidentified proteins could be evaluated with the help of functions of the other known proteins present in same clusters.
- Even though the intergenetic distances between the proteins in a cluster is minimum, sometimes it may not provide the functional relationship among the species.
- The Development of flexibility in the receptor structure is required
- Optimization algorithms like Bees Colony Optimization and other Swarm Intelligence algorithms are well suited and applied for Protein Interaction Network.

V. CONCLUSION

This paper began with an overview of clustering approaches and followed by a discussion of optimization techniques have been used for PIN. Various optimized clustering algorithms based on different approaches have been reviewed and summarized. This review study could help us understand and clarify some of the unnoticed issues that need further study in clustering approaches. It is evident that the limitations that have been identified and discussed above have been overcome to a certain extent. So it shows that optimized clustering approaches also provide better results as the normal clustering approaches.

VI. ACKNOWLEDGEMENT

The second author acknowledges the UGC, New Delhi for financial assistance under minor research project under grant no. 41-1354/2012(SR).

REFERENCES

[1] Jianxin Wang, Min Li, Youping Deng and Yi Pan, “Recent advances in clustering methods for protein interaction networks”, The ISIBM International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, December 2010, pp-1471-2164.

[2] Hunter B. Fraser, Aaron E. Hirsh, Lars M. Steinmetz, Curt Scharfe and Marcus W. Feldman, “Evolutionary Rate in the Protein Interaction Network”, Science, April 2002, Vol. 296, Issue- 5568, pp-750-752.

[3] H. Jeong, S. P. Mason, A.-L. Barabási and Z. N. Oltvai, "Lethality and centrality in protein networks", Nature, May 2001, Vol. 411, pp-41-42.

[4] V. Srinivasa Rao, K. Srinivas, G. N. Sujini, and G. N. Sunand Kumar, "Protein-Protein Interaction Detection: Methods and Analysis", International Journal of Proteomics, 2014, pp-1-12.

[5] Mohd Saberi Mohamad, Nor Farhah Binti Saidin, Chuii Khim Chong, Yee Wen Choon, Lian En hai, Safaai Deris, Rosli M. Illias and Mohd Shahir Shamsir, "Using Ant Colony Optimization (ACO) on Kinetic Modeling of the Acetoin Production in Lactococcus Lactis C7", Advances in Biomedical Infrastructure 2013, Vol. 477, pp- 25–35.

[6] Chuan Lin, Young-rae Cho, Woo-chang Hwang, Pengjun Pei and Aidong Zhang, "Clustering Methods In Protein-Protein Interaction Network", Knowledge Discovery in Bioinformatics: Techniques, Methods and Application 2006 (book).

[7] Binitha S and S Siva Sathya, "A Survey of Bio inspired Optimization Algorithms", International Journal of Soft Computing and Engineering, May 2012 , Vol.2, Issue-2, pp-137-151.

[8] Nada M. A. Al Salami, "Ant Colony Optimization Algorithm", UBICC Journal, August 2009, Vol. 4, Issue-3, pp-823-826

[9] Riva Mary Thomas, "Survey of Bacterial Foraging Optimization Algorithm", International Journal of Science and Modern Engineering, March 2013, Vol.1, Issue-4, pp-11-12.

[10] Jos´E Juan Tapia Valenzuela, "A clustering genetic algorithm for inferring protein-protein Functional interaction sites", July 2009, (thesis).

[11] Eduardo R. Hruschka, Ricardo J. G. B. Campello, Alex A. Freitas and André C. P. L. F. de Carvalho, "A Survey of Evolutionary Algorithms for Clustering", To appear in IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews.

[12] Jamaluddin, Rosni Abdullah and Ahamad Tajudin Khader "An Improved Ant Colony Optimization Algorithm for Clustering Protein Interaction Network", International Conference on Software Engineering & Computer Systems 2009, pp-19-21.

[13] J. Sallim, Rosni Abdullah Ahamad Tajudin Khader "ACOPIN: An ACO Algorithm with TSP Approach for Clustering Proteins from Protein Interaction Network", Computer Modeling and Simulation, September-2010, pp-203-208.

[14] Xiujuan Lei Shuang Wu, Liang Ge and Aidong Zhang, "Clustering PPI Data Based on Bacteria Foraging Optimization Algorithm", IEEE International Conference on Bioinformatics and Biomedicine, November-2011, pp-96-99.

[15] Pratyusha Rakshit, "Protein-Ligand Docking And Protein-Protein Interaction Using Evolutionary Algorithm", May 2012 (Thesis).

[16] Jamaluddin, Rosni Abdullah, Ahamad Tajudin Khader "An Improved Ant Colony Optimization Algorithm for Clustering Protein Interaction Network", International Conference on Software Engineering & Computer Systems 2009, pp-19-21.

[17] Min Li, Xuehong Wu, Jianxin Wang, Yi Pan, "A New Measurement for Evaluating Clusters in Protein Interaction Networks", IEEE International Conference on Bioinformatics and Biomedicine, 2011, pp-63-68.

[18] Iman Sharafuddin, Mehrdad Mirzaei, Masoud Rahgozar and Ali Masoudi-Nejad, "Protein-Protein Interaction Network Clustering Using Particle Swarm Optimization", IWBBIO proceedings, March 2013, pp-317-324.

[19] TIAN Jian-Fang, LEI Xiu-Juan, " PPI Network Clustering Based on Artificial Bee Colony and Breadth First Traverse Algorithm", Pattern Recognition and Artificial Intelligence, 2012, Vol. 25, pp- 481-490.