Protein-Protein Interaction Network Clustering Using Particle Swarm Optimization

Iman Sharafuddin[1], Mehrdad Mirzaei[1], Masoud Rahgozar[1] and Ali Masoudi-Nejad[2]

[1] School of Electrical and Computer Engineering, Control and Intelligent Processing of excellence, Database Research Group, University of Tehran, Tehran, Iran
{i.sharafuddin|m.mirzaei|rahgozar}@ece.ut.ac.ir

[2] Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics and Center of Excellence in Biomathematics, University of Tehran, Tehran, Iran
amasoudin@ibb.ut.ac.ir

Abstract. The purpose of protein-protein interaction network clustering is to find major modules for cells function. Various clustering algorithms are used but only some of them take advantage of intelligent computational methods. In this work we introduced a particle swarm optimization method to find dense sub-graphs in protein-protein interaction networks. The experimental results proved that our method have advantages over previous methods upon modularity measure proposed by Grivan and Newman.

1 Introduction

Main functions within a cell are done by interactions between proteins. A collection of these interactions called protein-protein interaction network (PPIN) [15]. As an example, metabolic pathway consists of different proteins called enzymes, which produce a chain of chemical reactions to change a substance into another one named a product. Also protein interactions in signaling pathway is a set of proteins with a series of ordered interactions that convert a type of chemical signal to another. By this chain of interactions, cells can perceive environmental information.

In fact, interactions between proteins motivate them to function and grow [1]. In recent years, large progress of research in this context has occurred due to advances in mining protein-protein interaction networks [2, 3].

Clustering algorithm is the task of grouping set of items based on similarity between pairs of items. In other words clustering is the task of dividing a set of items into groups so that the items in the same group are more similar to each other than to those in other groups [1]. There are two main objectives in clustering; first one is “homogeneity”, which means that more similar items are placed in a same group. Second one “heterogeneity”, which means that different group’ elements have less similarity to each other. The main purpose of clustering protein-protein interaction network is to find dense sub-graphs showing significant functional modules in protein-protein interactions. Recognizing significant functional modules in protein-protein interaction network is first step to realize the structure and functional dynamic of cell [4, 14]. Significant functional module is a set of proteins that are actors for a specific cellular process .These dense sub-graphs that are mined from protein-protein interaction networks by clustering are an indicator of both functional modules and protein complexes. Also group of proteins that highly interact with themselves are known as protein complexes. In this paper both functional modules and protein complexes are considered as the same concepts.

Protein-protein interaction network clustering algorithms recently presented are based on identified semi-clique sub-graphs in the network [4, 5, 6, and 7]. Clique is a complete graph in which there is an edge between each two node. In [5] Hogue presented MCODE (Molecular Complex Prediction) algorithm that is an efficient clustering method for protein-protein interaction networks clustering. MCODE works in three stages, including weighting nodes, predicting complexes and an optional post-processing step for clustering protein-protein interaction networks. The MOCDE weakness is on modules with a large number of proteins. Adamcsek et al. [6] proposed algorithm called Cfinder that searches for cliques. However, this idea was not necessary correct because finding clique is not our main objective [7]. Ravaee et al. also presented an algorithm based on the body defense system [8]. In the current work, we introduce an algorithm based on particle swarm optimization. In this approach, each particle is considered as a solution for the problem and reaches the solution by distance between particles, which is defined later and swarm optimization facilities. Our algorithm has the ability to support diversity of population with suitable fitness [14].

2 Primary Definitions

2.1 Problem Definition

The first challenge in protein-protein interaction clustering is the mathematical representation of protein-protein interaction network. An ordinary approach is to use graph theory concepts. We can demonstrate protein-protein interaction network by graph G= (V, E). In G nodes V are corresponding to proteins and E to interaction and each edge connects two vertices meaning that two proteins have interaction with each other. Clusters in the graph also can be taking into account as dense sub-graphs that mean number of edges between clusters is the minimum number and number of edges in each sub-graph is the maximum number.

2.2 Swarm intelligence optimization algorithm

PSO was previously introduced by Eberhart and Kennedy in 1995 inspired by social behavior of swarm's [14, 16]; a particle could be referring to a bird in a bird flock, searching for a solution in a problem space. The location of a particle in multi-dimensional space shows a solution to the problem. When a particle moves in space, another solution for the problem would be obtained. A fitness function which quantitates the solution could help us to evaluate the solution quality. The velocity and direction of each particle moving in each dimension would change in each generation. In (1) and (2) we define two formulas to calculate the location of each particle. The term Pid in (1) is the particle’s personal experience and the term Pgd in (1) is its neighbors’ experience. Values of R1, R2 and R3 are produced randomly and used for the sake of the completeness of our algorithm. Values of C1, C2 and C3 specify the weighting effect of Pid and Pgd on velocity of particle. In each generation, particle’s position is calculated by adding current velocity to new position. These values and calculations in (1) and (2) are computed.

#

$$v_{i,j} \leftarrow c_0 v_{i,j} + c_1 r_1 \left(Pgd - x_{i,j}\right) + c_2 r_2 \left(local\ best - x_{i,j}\right) + c_3 r_3 \left(Pid - x_{i,j}\right) \qquad (1)$$

$$x_{i,j} \leftarrow x_{i,j} + v_{i,j} \qquad (2)$$

It is possible to consider clustering problem as an optimization problem; hence we proposed a PSO clustering algorithm in this paper.

2.3 Encoding Solution in a Particle

For coding the solution in a particle, we use a one dimensional array of integers [12, 8] filled with vertexes number; each two vertices are divided by separation bits. In example with a graph G(V,E) with |V| vertices and |E| edges, a particle is a collection of subsets of vertexes, each of these subsets represents a cluster in our solution.

Given an undirected connected graph with N vertex, each particle p is an array of integers with length of 2N-1 in which the N integers that are in odd position denote the vertices of graph. There are N-1 zero or one that are in even positions, these integers act as a separator between items of a cluster or separation between clusters. An example of a particle for a given graph with nine vertices is illustrated in Fig.1.
As you see in Fig. 1 nodes 1,2,3,4 are in a cluster and node 5, 6,7,8,9 in other cluster.

1	0	2	0	3	0	4	1	5	0	6	0	7	0	8	0	9

Fig. 1. Separator value to distinguish vertices in a sub-graph is 0, and 1 means separation between two dense sub-graphs . a) graph G with 9 vertices and b) particle corresponding graph G.

2.4 Fitness function

An accurate and suitable fitness function is essential since this function is used for discriminating the candidate solution and optimal solution. Fitness function examines each particle in respect of all existing sub-graphs in that particle and scores it. We evaluate each sub-graph with “clustering score” [7]. Clustering score of each sub-graph is the product of number of nodes and its density. Density of sub-graph is proportion of edges connecting that sub-graph E and maximum possible edges Emax. Equation (3) demonstrates Density.

$$Density(S) = 2 * E/(V * V - 1) \tag{3}$$

Final score of sub-graph is calculated by (4).

$$Score(S) = Density(S) * V \tag{4}$$

At the end, the final fitness of that particle is equal to sum of each sub-graph multiply by coefficient in (5).

$$N = \frac{\sum_{i=1}^{m} E_i}{E} \tag{5}$$

In (5) E means number of edges and Ei is number of edges in sub-graph ith and m is number of sub-graphs.

2.5 Proximity between two particles

In our algorithm, we need a similarity function based on structure to investigate best nearest neighbor of each particle. To calculate similarity (6) is used.

$$D(a, b) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{S_i^a \cap S_j^b}{S_i^a \cup S_j^b} \tag{6}$$

Here D(a,b) is the distance between particles a and b. S_i^a is the set of sub-graph i nodes existing in a and S_j^b is so on. m and n are the number of sub-graphs in a and b. Based on this distance function, to move the worse particle near the better particle, best particle sub-graph is transferred to worst sub-graph of worse particle.

3 Swarm Intelligence Optimization Operators

3.1 Initialize operator

This operation for creating initial particles is performed at first step of algorithm. For all initial particles, we use a well-designed heuristic to bootstrap our solution in order to achieve better results; first of all ,to make clusters we select highest degrees as seeds; then select the adjacent nodes of seed nodes, next select nodes which is connected to two or more nodes of selected cluster.

3.2 Local random movement

In each particle substitute a node of a sub-graph with a node from another sub-graph by a predefined probability.

3.3 Global random movement

Complement one of selected bits in each particle with a predefined probability.

3.4 Movement function

Sometimes we want to move a worse particle close to a better particle with respect to proximity function explained in the previous section.

4 The Algorithm

Amount of particles in our algorithms is constant and specified by value n. Alike other algorithms our algorithms has input and output. The input of our algorithms is protein-protein interaction network in form of adjacency matrix. In Initialize population phase our algorithm uses adjacency matrix to generate initial particles. One of the particles with the best fitness in last generation presents the approximation of optimum solution for the graph clustering problem. Our algorithm would terminate when it converge; it is defined by two termination condition, the first one is the maximum number of iterations that will be define by user and the second one occurs when the fitness of the best particle between two generations be less than a specified threshold. Our algorithm flowchart is shown in Fig. 2.

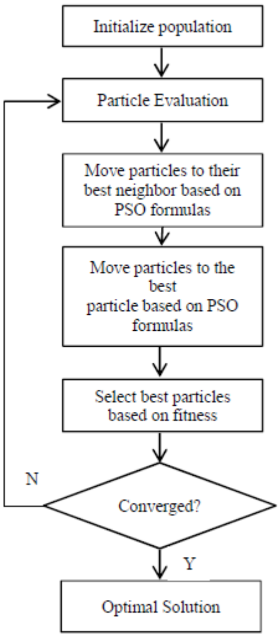
Initialize population
Particle Evaluation
Move particles to their best neighbor based on PSO formulas
Move particles to the best particle based on PSO formulas
Select best particles based on fitness
N
Converged?
Y
Optimal Solution

Fig. 2 Steps of Algorithm

The protein-protein interaction network is derived from the yeast subgroup in the Database of Interacting Proteins (DIP) [10]. It consists of 4963 proteins and 17579 interactions. Near all of these interactions have been achieved by Yeast Two-Hybrid screen (Y2H).

5 Experimental Results

Our clustering is performed on Saccharomyces cerevisiae (yeast). This data is available in [10] and it is free for download. The dataset contains 4963 proteins (nodes) and 17570 interactions (edges) and comprises noise due to experimental errors.

To compare algorithms, network modularity measure proposed by Grivan and Newman [13] was used. A measure of modularity to compare clusters is calculated by (7).

$$Q = \sum_i (e_{ii} - a_i^2) \qquad (7)$$

Where i is clusters index, eii is number of edges in i which both ends are in cluster and ai is number of edges that only one end is in ith cluster. Fig. 3 shows the comparison of our algorithm, MCODE and Cfinder based on formula (7). It presents that network modularity of our proposed method is higher than the other methods.

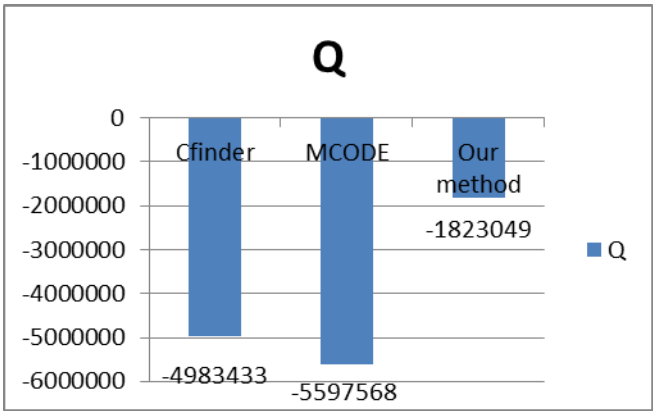

Fig. 3 Network modularity comparison between our method, MCODE and CFinder approaches. X axis is clustering approach and Y axis is network modularity.

The time complexity of the MCODE algorithm is known as polynomial O(nmh^3) where n is number of vertices, m is number of edges and h is the average size of neighborhood for vertexes in input graph G [5]. The complexity of the Cfinder is $O(n^2)$ and our proposed method can cluster protein-protein interaction network in O(knp) where k is the number of iterations and p is the number of particles.

Conclusion

In this paper, we present a protein-protein interaction clustering algorithm based on particle swarm optimization to find dense sub-graphs in protein-protein networks. The topological comparison between other clustering approaches such as Cfinder and MCODE and our algorithm shows that our algorithm is more accurate than these two methods. With explanation of structure of protein interactions network, functions of unidentified proteins could be predicted by functions of other known proteins that are in same clusters.

Acknowledgment

I would like to acknowledge and extend my heartfelt gratitude to Hamid Ravaee who has made the completion of this paper possible, for his understanding and assistance.

References

1. S. Brohée, J. v. Helden., "Evaluation of clustering algorithms for protein-protein interaction networks": BMC Bioinformatics, 2006. 7:488
2. T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, Y.Sakaki., "A comprehensive two-hybrid analysis to explore the yeast protein interactome." : PNAS, 2001, Vol. 98, pp. 4277-4278.
3. P. Uetz, L.Giot, G. Cagney, T.A. Mansfield, R.S. Judson,J. R. Knight,D. Lockshon, V. Narayan, M. Srinivasan, P.Pochart, A. Qureshi-Emili, Y.Li, B. Godwin, D.Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang,M.Johnston, S. Fields, J. M. Rothberg., "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae." : Nature, 2000,Vol. 403, pp. 623-627.
4. L. gao, p. sun, j. song., "Clustering algorithms for detecting functional modules in protein interaction networks." : Journal of Bioinformatics and Computational Biology, 2009, Volume: 7, Issue: 1, pp. 217-242.
5. G. Bader, C. Hogue, "An automated method for finding molecular complexes in large protein interaction networks." : BMC Bioinformatics, 2003, Vol. 4, No. 1, 2
6. B. Adamcsek, "CFinder: locating cliques and overlapping modules in biological networks." Bioinformatics, 2006, Vol. 22, pp. 1021-1023.
7. M. Wu, X. Li, C. Kwoh., "Algorithms for Detecting Protein Complexes in PPI Networks: An Evaluation Study." : PRIB08, 2008. pp. 135-146.
8. H. Ravaee, A. Masoudi-Nejad, S. Omidi, A. Moeini “Improved Immune Genetic Algorithm for Clustering Protein-Protein Interaction Network”, IEEE International Conference on Bioinformatics and Bioengineering, 2010.

9. I. Xenarios, L. Salwínski, X. Joyce Duan, P. Higney, S. Kim and D. Eisenberg, "DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interaction." : Nucleic Acids Res, 2002, Vol. 30, pp. 303-305.
10. D. Maio, D. Maltoni, S. Rizzi., "Topological Clustering Of Maps Using A Genetic Algorithm." : Pattern Recognition Lett, 1995.
11. M. Newman, M. Girvan., "Finding and evaluating community structure in networks." : Physical Review, 2004.
12. C. Eberhart, ‚J. Kennedy. "A new optimizer using particle swarm theory". Proceedings of the Sixth International Symposium on Micro Machine and Human Science", 1995, pp. 39-43.
13. G.D. Bader,. and C.W Hogue, "Analyzing yeast protein-protein interaction data obtained from different sources." : Nat. Biotechnol, 2003, Vol. 20, pp. 991-997.
14. V. Mirny, L. Spirin., "Protein complexes and functional modules in molecular networks." : Proc. Natl Acad. Sci, 2003, Vol. 100(21), pp. 12123–12126.
15. C. Pizzuti, E. Rombo and E. Marchiori, "Complex Detection in Protein-Protein Interaction Networks: A Compact Overview for Researchers and Practitioners", Machine learning and datamining in bioinformatics, 2012, pp. 221-223.
16. X. Cui, T. Potok, "Document Clustering Analysis Based on Hybrid PSO+K-means Algorithm" , Journal of Computer Sciences,2005, pp. 27-33