Methods 110 (2016) 44–53

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Identifying protein complexes based on brainstorming strategy

Xianjun Shen [a,b], Jin Zhou [a], Li Yi [a], Xiaohua Hu [a], Tingting He [a], Jincai Yang [a,*]

[a] School of Computer, Central China Normal University, Wuhan 430079, China
[b] Collaborative & Innovative Center for Educational Technology, Central China Normal University, Wuhan 430079, China

ARTICLE INFO

Article history:
Received 15 April 2016
Received in revised form 17 June 2016
Accepted 9 July 2016
Available online 9 July 2016

Keywords:
Protein-protein interaction network
Protein complex
Brainstorming strategy
Gene ontology

ABSTRACT

Protein complexes comprising of interacting proteins in protein-protein interaction network (PPI network) play a central role in driving biological processes within cells. Recently, more and more swarm intelligence based algorithms to detect protein complexes have been emerging, which have become the research hotspot in proteomics field. In this paper, we propose a novel algorithm for identifying protein complexes based on brainstorming strategy (IPC-BSS), which is integrated into the main idea of swarm intelligence optimization and the improved K-means algorithm. Distance between the nodes in PPI network is defined by combining the network topology and gene ontology (GO) information. Inspired by human brainstorming process, IPC-BSS algorithm firstly selects the clustering center nodes, and then they are separately consolidated with the other nodes with short distance to form initial clusters. Finally, we put forward two ways of updating the initial clusters to search optimal results. Experimental results show that our IPC-BSS algorithm outperforms the other classic algorithms on yeast and human PPI networks, and it obtains many predicted protein complexes with biological significance.

1. Introduction

The occurrence of complicate life phenomenon is influenced by multiple factors, which suggests the proteins are not isolated in the cell during life activities [1–3]. That is, biological functions are carried out by protein complexes consisting of closely associated proteins [4–6]. Thus the research of identifying protein complexes in PPI networks is meaningful. It provides strong support to further study the biological behavior, predict protein functions and design drugs [7]. In the meanwhile, it promotes researches and developments in various disciplines, such as biology, medicine, bioinformatics, etc. Although it provides strong support to plumb the mysteries of human life, it is challenging to devise effective algorithm for discovering complexes from human PPI network [8,9].

So far, the traditional clustering methods have been successfully developed to identify protein complexes in PPI networks but each has defects, such as partition-based method, level-based method, and density-based method. The partition-based method is a practical way to cluster, but its efficiency strongly depends on the pre-known knowledge, especially it is necessary to give the clusters' number in advance [10]. This directly affects the forecast results, because of the actual number is unknown. The level-based algorithm can identify diverse modules and present clearly hierarchy but is very sensitive to the noise of data [10]. Besides, the density-based and local search algorithms typically focus on the relatively dense sub-graphs, while they neglect many peripheral proteins that have few connections with the core protein clusters [10]. Thus, biologically meaningful protein complexes without highly connected topology are ignored. Moreover, one common major drawback of these approaches is that the nodes which have been erroneously placed into complexes in the previous steps have no chance to be corrected in the subsequent steps, which results in "module barrier". Such influences will be accumulated and amplified on the follow-up processes.

Recently, swarm intelligence algorithms have been widely applied and shown excellent performance in the field of network routing, path planning, image processing and clustering analyzing to solve many complex problems. Thus a new kind of clustering mechanisms based on the swarm intelligence to detect protein complexes has become a new research hotspot in this field. In 2008, Sallim et al. [11] firstly propose algorithm ACOPIN that is based on the process of finding an optimal path by ACO to cluster proteins in the PPI network. Using a global optimization model based on PPI networks, ant colony optimization algorithm for predicting protein functions is proposed by Wu et al. [12] in 2011. Next year, Ji et al. [13] introduce the PPI network's topology and protein functional information into the ant colony algorithm, which is named NACO-FAMD algorithm. In addition, Lei [14] defines joint strength to improve ant colony optimization, and then

* Corresponding author.
E-mail addresses: xjshen@mail.ccnu.edu.cn (X. Shen), dreamxiaojin@163.com (J. Zhou), yili@mails.ccnu.edu.cn (L. Yi), xh29@drexel.edu (X. Hu), tthe@mail.ccnu.edu.cn (T. He), jcyang@mail.ccnu.edu.cn (J. Yang).

http://dx.doi.org/10.1016/j.ymeth.2016.07.006
1046-2023/

applies it to cluster proteins in PPI networks. These researches have shown that the mechanism of intelligence optimization algorithms can effectively identify protein complexes in PPI network, while the complicated algorithm designing process leads the time complexity relatively high.

In classic swarm intelligence algorithms, each individual is represented by a simple object such as birds in PSO (Particle Swarm Optimization) [15], ants in ACO (Ant Colony Optimization) [16], bacteria in BFO (Bacteria Foraging Optimization) [17], etc. In PSO algorithm, its searching speed and efficiency are suitable for the real deal, however, for the discrete optimization problem, each individual is greatly influenced by the best one in each iteration so that it's easy to lose population diversity and then fall into premature convergence. The ACO algorithm is derived from the foraging behavior of ants, so it's easy to form "simple obedience" because of the influences of pheromone between individuals within a group. Critical spirit and creativity are weaken by group thinking, thereby the quality of decision is influenced. In BFO algorithm, it is not sensitive to the initial value and parameter. It is robust, simple and easy to achieve, as well as the advantages of parallel processing and global search, but there is not enough high precision and fast convergence speed during the application process. Nevertheless, human beings as social animals with innovative thinking are the most intelligent animals in the world. Sharing of creative thinking between groups generates the sensitivity of nature and strong learning ability which are useful for clustering and partition in IPC-BSS algorithm.

In this paper, inspired by the concept of swarm intelligence optimization and human collective behavior, we propose a novel algorithm IPC-BSS to identify protein complexes based on brainstorming strategy. In the main idea of our IPC-BSS algorithm, protein is regard as individual while protein complex as group. Individual thinking and discussing in this process is unrestricted. Its aim is to emerge new ideas or inspire innovative ideas. Everyone reflects on a special field with freedom, thus creative brainstorming arises. Once a new viewpoint or solution comes into being, speak it loudly, and then a new one will be put forward on the basis of it. However, these standpoints are only recorded but not assessed until the end of the brainstorming session. Subsequently, a new method with creative thinking begins to take shape. IPC-BSS algorithm aims to simulate the brainstorming process. Firstly, the clustering center nodes in the PPI network are selected and extended to form initial clusters according to the distance between nodes, which is defined by combing network topology with GO (gene ontology) information and thus reflects the close degree between proteins. Then the initial clusters are optimized by dynamically searching the optimal results combining the improved K-means [18] algorithm and the neighbor nodes to update within or between the existing modules, which imitates human idea generation process. Finally we merge and filter the unsatisfied clusters to obtain final protein complexes. We validate our algorithm on yeast and human PPI networks. The experimental results show that IPC-BSS algorithm integrating with the creative thought outperforms the other classic algorithms in practice for detecting protein complexes, which indicates the strong global optimization ability of the new approach.

2. Material and methods

2.1. Terms and definitions

2.1.1. Node clustering coefficient (NCC)

PPI network is typically modeled as an undirected graph $G = (V, E)$, where V denotes the set of nodes (proteins) and $E = \{(u, v)|u, v \in V\}$ denotes the set of edges (protein interactions). For each node i in graph G, all the nodes which connect with i constitute its neighbor set represented as $Nbs(i)$, then the clustering coefficient of node i can be represented as Eq. (1).

$$NCC_i = \frac{2 \times n_i}{|Nbs(i)| \times (|Nbs(i)| - 1)} \tag{1}$$

where n_i denotes the actual number of edges among the nodes in $Nbs(i)$, $|Nbs(i)| \times (|Nbs(i)| - 1)/2$ denotes the theoretical maximal number of edges which can be formed by the neighbor nodes. The node clustering coefficient reflects the local density of a node [19].

2.1.2. Distance between proteins

Protein complex is formed by proteins interacting closely with each other. We define the distance between nodes in PPI network by combining the network topology with GO information [20] to measure how closely do the pairwise proteins interact, as well as reflect the topological similarity and functional consistency between two proteins. The distance $d(i, j)$ between two nodes i and j is defined as Eq. (2):

$$d(i,j) = \frac{IS(i) \Delta IS(j)}{|IS(i) \cap IS(j)| + |IS(i) \cup IS(j)|} + \left(1 - \frac{|f(i) \cap f(j)|}{|f(i) \cup f(j)|}\right) \tag{2}$$

where the part before the plus sign represents the topological distance referenced by Czekanowski-Dice distance [21], $IS(i)$ denotes a set consisted of the neighbors of node i and itself, while Δ, $\cup$ and $\cap$ represent the operation of symmetric difference, union and intersection between two sets, respectively. The distance between two nodes without any common neighbors is equal to 1 otherwise greater than 0. It reflects that the shorter distance between two nodes is, the more neighbors they share and thus they possess the more similar functions. The topological distance suffers certain limitation because of the noise in PPI network, which can be made up by integrating with the functional annotations information. The part after the plus sign denotes the functional similarity of two proteins. $f(i)$ and $f(j)$ denote the set of functional annotations of proteins i and j, respectively. The whole functional annotations set can be obtained from the Uniprot Database [22]. Experiments have showed that the more similar the functions are, the smaller the functional distance is [23]. The distance integrating network topology and GO information measures the probability of two proteins participating into the same biological process.

2.1.3. Module closely associated degree (MCAD)

Taking the relationship among all nodes within a module into account, *MCAD* defined as Eq. (3) quantifies how closely the nodes in a group interact with each other.

$$MCAD_i = \frac{\sum_{x,y \in G_i, \langle x,y \rangle \in E} \frac{|Nbs(x) \cap Nbs(y)|}{min\{|Nbs(x)-1|, |Nbs(y)-1|\}}}{E_i} \tag{3}$$

where E_i and G_i denote the number of edges and the nodes set in group i, respectively. Then we define the average module closely associated degree (*avg_MCAD*) denoted by Eq. (4) to measure the overall module closely associated degree when partition a PPI network into protein complexes.

$$avg\_MCAD = \frac{\sum_{i=1}^{G_{sum}} MCAD_i}{G_{num}} \tag{4}$$

where G_{sum} represents the number of modules in network. When *avg_MCAD* equals to the maximum value, the corresponding partition is viewed as the optimal cluster result in PPI network.