Knowledge-Based Systems 105 (2016) 270–277

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Identification of dynamic protein complexes based on fruit fly optimization algorithm

Xiujuan Lei[a,*], Yulian Ding[a], Hamido Fujita[b], Aidong Zhang[c]

[a] School of Computer Science, Shaanxi Normal University, Xi'an, 710062, China
[b] Faculty of Software and Information Science, Iwate Prefectural University, 152-52 Sugo, Takizawa, 020-0693 Iwate, Japan
[c] Department of Computer Science and Engineering, State University of New York at Buffalo, NY 14260-2000, USA

ARTICLE INFO

Article history:
Received 26 February 2016
Revised 4 April 2016
Accepted 9 May 2016
Available online 10 May 2016

Keywords:
Dynamic protein-protein interaction network (DPIN)
Fruit fly optimization algorithm
Gene expression
Protein complex

ABSTRACT

Protein complexes play a significant role in understanding cellular life in postgenomic era. Yet up to now, the existing protein complex detection algorithms are mostly applied to static PPI networks and their performance is not very ideal for the deficiency of low efficiency and sensitive to noisy data. In this paper, a novel algorithm named Fruit fly Optimization Clustering Algorithm (FOCA), is proposed to identify dynamic protein complexes by combining Fruit fly Optimization Algorithm (FOA) and gene expression profiles. Particularly, we first find the always active proteins by the stable interactions of the dynamic PPI network and detect protein complex cores from those always active proteins. Then, FOA is used to merge of the rest proteins in every dynamic sub-network to their corresponding protein complex cores. The experimental results on DIP dataset demonstrate that FOCA is very effective in detecting protein complexes than the state-of-the-art complex detection techniques.

1. Introduction

Proteins seldom act alone, and they often bind together to form complexes to carry out their biological functions [1,2]. Thus correctly detecting protein complexes play a significant role in understanding the underlying mechanism of most cellular functions and predicting the functions of un-annotated proteins [3–5]. The real protein-protein interaction (PPI) network in cell keeps changing over different stages of the cell cycle [6] and they can be classified into stable or transient PPIs [7], which are usually described as dynamic protein-protein interaction networks (DPIN). Thus it is important to construct dynamic PPI networks to investigate the temporal properties of individual proteins and protein interactions.

Dynamic protein complexes are typically constructed by the dynamic assembly or disassembly to perform various biological functions. To detect dynamic protein complexes, we need to leverage the dynamic information from gene expression data to construct time-evolving dynamic protein interaction networks [8,9]. In [10], the authors incorporated the "time" factor for proteins in the form of cell-cycle phases into the analysis of complexes and studied the dynamic phenomena of complexes assembly and disassembly across various cell cycles. Then we identified dynamic protein complexes from the dynamic PPI networks mainly by applying static complex detection methods for each time point [9].

In recent years, most of the computational clustering methods are based on the assumption that a protein complex corresponds to a dense subgraph or cluster. Many heuristic graph clustering methods consist of a set of nodes that are highly connected to rest nodes of the networks to find clusters. Such as Liu et al. proposed maximal clique algorithm (CMC) [11] which detect protein complexes by finding all the maximal cliques. Bader and Hogue proposed molecular complex detection (MCODE) algorithm [12] which first weight every node based on its local neighborhood densities, then selects nodes with high weights as seeds, and eventually detects protein complexes by extending the seeds. Altaf-Ul-Amin et al. [13] proposed the DPClus algorithm which is different from MCODE for that it assigns weights to nodes by their weighted edges' degrees. There have been emerged many other algorithms, such as Markov Clustering (MCL), HC-PIN and ClusterONE [14–16] and so on. Although the above methods have been shown to identify protein complex effectively, the result data is highly false positive and false negative due to that those methods ignore the inherent architecture of protein complexes. With respect to the core-attachment structure of protein complex from a topological view, Leung et al. [17] design CORE algorithm which calculates the *p-value* for all pairs of proteins to detect cores. Wu et al. [18] proposed COACH algorithm which detected dense subgraphs as protein-complex cores. Although COACH achieves better predi-

* Corresponding author. Fax: +8602985310161.
E-mail address: xjlei168@163.com (X. Lei).

http://dx.doi.org/10.1016/j.knosys.2016.05.019
0950-7051/

cation performance than other methods which fail to consider intrinsic structure [19], there are no uniform definitions for the core and attachments and how to detect the attachments for each core are still the challenge we have to face.

Bio-inspired algorithms provide a new perspective for solving complexes protein complexes detection problem, with the characteristics of high robust, low complexities, excellent optimization ability. And there are a lot of successful clustering models which based on the intelligent methods to tackle the problems of PPI data clustering and perform well. In 2013, Lei proposed Bacteria Foraging Optimization (BFO) clustering model [20] based on BFO mechanism and intuitionistic fuzzy set. And at the same year, she proposed PMABC-ACE [21] clustering model based on the propagating mechanism of artificial bee colony. After that, in 2016, Lei proposed F-MCL [22] clustering model based on Markov clustering and firefly algorithm which automatically adjusts the parameters by introducing firefly algorithm.

As a novel evolutionary optimization approach, Fruit fly Optimization Algorithm (FOA) mimics the foraging behavior of fruit flies for searching global optimum, which is proposed by Pan in 2012 [23]. The FOA has few parameters to be adjusted, and it is easy to understand and implement. Due to its merits, the FOA has already been successfully applied to solve some academic and engineering optimization problems [23–26], including financial distress, PID controller tuning, web auction logistics service, multidimensional knapsack problem and so on. Those have verified that FOA is applicable for solving many types of scheduling problems and is also competitive to other optimization algorithms. And in protein complexes detecting, the searching for a good protein module can also be easily incorporated into the search framework of the FOA to further enhance the detection ability of protein complexes. Therefore, in this study, we propose an improved clustering model aimed at PPI data, which blend the FOA algorithm in it.

In this paper, a new clustering method named Fruit fly Optimization Clustering Algorithm (FOCA) algorithm which based on FOA and the gene expression profiles is proposed. Firstly, we find the stable proteins and transient proteins according to the gene expression profiles. Then we select the highly connected and high density small clusters on the stable proteins as the food of fruit flies, the transient proteins as the fruit flies, and the "closeness" of a transient protein to a core cluster is the smell concentration of fruit fly. Finally, those transient proteins find the corresponding core cluster and form the final protein complexes when the fruit flies find foods. For testing the performance of our algorithm, we compare our method with those traditional clustering methods, such as CMC, MCODE, DPClus, MCL, HC-PIN, ClusterONE, CORE, COACH, [11–18] and so on.

The outline of this paper is as follows. Section 2 describes some preliminary theories and the details of our algorithms, Section 3 shows the experimental results and analysis, and Section 4 concludes the paper.

2. Method

2.1. Fruit fly optimization algorithm

Fruit fly optimization algorithm is a novel swarm intelligent optimization algorithm which mimics the foraging behavior of fruit flies for searching the global optimum. With the outstanding olfactory, fruit flies can perceive the smell in the air even the food source beyond 40 meters and fly toward it. Then, after it gets close to the food location, it can also use its sensitive vision to find food and the company's flocking location, and also fly towards that direction. Fig. 1 shows the iterative food searching process of fruit fly [23].

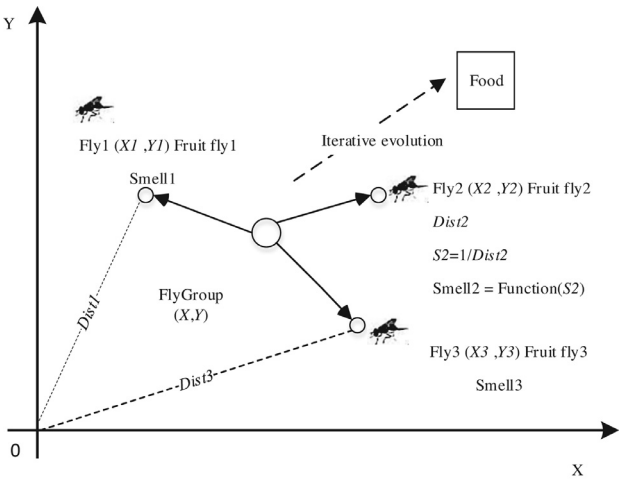

Fig. 1. Illustration of the group iterative food searching of fruit fly.

According to the basic FOA [23], the calculation steps are listed as below:

Step 1. Randomly initialize fruit fly swarm location which is shown in Fig. 1 The initial location is marked as (InitX_axis, InitY_axis).

Step 2. Every individual fruit fly searching for the food by a random direction and distance to the origin, using the osphresis. New location can be calculated using:

$$x(t+1) = x(t) + RV_x$$
$$y(t+1) = y(t) + RV_y \quad (1)$$

where *RV* means *randomvalue which* is the movement value in each coordinate. As shown in Fig. 1, fly group move to the new locations like Fly1, Fly2, Fly3, the new locations compose the new fly group and new locations take place of the former fly group locations for calculation.

Step 3. Due to the food location cannot be known, the distance to the origin is thus estimated first, marked as *Dist* calculated by:

$$Dist_i = \sqrt{x_i^2 + y_i^2} \quad (2)$$

The smell concentration judgment value (S) is calculated, and this value is the reciprocal of *Dist*.

Step 4. Substitute smell concentration judgment value (S) into smell concentration judgment function (or called *Fitness* function) so as to find the smell concentration ($Smell_i$) of the individual location of the fruit fly.

$$Smell_i = Function(s_i) \quad (3)$$

Step 5. Find out the fruit fly with minimal smell concentration (finding the maximal value marked as [*bestSmell bestIndex*]) among the fruit fly swarm.

Step 6. Keep the best smell concentration value (marked as *Smellbest*) and *x, y* coordinates, and at this moment, the fruit fly swarm will use vision to fly towards that location.

Step 7. Enter iterative optimization to repeat the implementation of **Steps 2-5**, then judge if the smell concentration is superior to the previous iterative smell concentration, if so, implement **Step 6**.

2.2. Dynamic PPI network model construction

The dynamic PPI networks are constructed by integrating time-course gene expression data with static PPI networks. The existing