Contents lists available at ScienceDirect

Information Sciences

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

Protein complex identification through Markov clustering with firefly algorithm on dynamic protein–protein interaction networks

Q1 Q2 Xiuujuan Lei[a,*], Fei Wang[a], Fang-Xiang Wu[b], Aidong Zhang[c], Witold Pedrycz[d,e,f]

Q3
[a] School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi 710062, China
[b] Division of Biomedical Engineering, University of Saskatchewan, Saskatoon SK S7N 5A9, Canada
Q4 [c] Department of Computer Science and Engineering, State University of New York at Buffalo, NY 14260-2000, USA
[d] Department of Electrical & Computer Engineering, University of Alberta, Edmonton T6R 2V4 AB, Canada
[e] Department of Electrical and Computer Engineering Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[f] Systems Research Institute, Polish Academy of Sciences Warsaw, Poland

ARTICLE INFO

Article history:
Received 31 December 2014
Revised 8 September 2015
Accepted 12 September 2015
Available online xxx

Keywords:
Dynamic protein–protein interaction network (DPIN)
Markov clustering (MCL) algorithm
Firefly algorithm (FA)
Protein complex

ABSTRACT

Markov clustering (MCL) is a commonly used algorithm for clustering networks in bioinformatics. It shows good performance in clustering dynamic protein–protein interaction networks (DPINs). However, a limitation of MCL and its variants (e.g, regularized MCL and soft regularized MCL) is that the clustering results are mostly dependent on the parameters whose values are user-specified. In this study, we propose a new MCL method based on the firefly algorithm (FA) to identify protein complexes from DPIN. Based on three-sigma principle, we construct the DPIN and discuss an overall modeling process. In order to optimize parameters, we exploit a number of population-based optimization methods. A thorough comparison completed for different swarm optimization algorithms such as particle swarm optimization (PSO) and firefly algorithm (FA) has been carried out. The identified protein complexes on the DIP dataset show that the new algorithm outperforms the state-of-the-art approaches in terms of accuracy of protein complex identification.

1. Introduction

A protein–protein interaction (PPI) network is a biomolecule relationship network that plays an important role in biological activities. The identification of protein complexes each of which contains a number of proteins that play a similar role in a living cell becomes essential to understand the organization, evolution and interaction of cellular systems. The study of all proteins called "Proteomics" is the systematic approach to diverse properties of proteins that provides detailed descriptions of the structure, function and control of biological systems in health and disease [6]. Usually, proteins seldom act as independent or isolated entities. However, proteins involved in the same cellular processes often interact with each other to incorporate into a large molecule to accomplish the biological functions [49]. This is the protein–protein Interaction (PPI) network which involves many proteins and their interactions. As biological functions are time-sensitive, proteins and interactions do not always exist. In response to a stimulus or a new condition occurring in a living cell, the amounts and locations of proteins change from time to time [12], thus the structures of PPI networks change too [44]. Moreover, there are a large number of false positives or false

Q5
[*] Corresponding author. Tel.: +86 29 85310161.
E-mail addresses: xjlei@snnu.edu.cn (X. Lei), wpedrycz@ualberta.ca (W. Pedrycz).

http://dx.doi.org/10.1016/j.ins.2015.09.028
0020-0255/© 2015 Published by Elsevier Inc.

negatives existing in the current available PPI data [30]. To express the dynamic traits and reduce the effect of the false positives, many dynamic data, including gene expression profiles, have been used to construct the dynamic protein–protein interaction network (DPIN) [26]. In DPIN, gene expression data are used to determine which gene expresses and to demarcate some timestamps. It is well known that not does a gene express at all timestamps. Some genes express at one timestamp and other genes express at another timestamp. If a gene expressed at one timestamp, it is reasonable to assume that the gene product exists at that timestamp. Then the entire set of proteins are partitioned, and at each timestamp there is a static PPI network which exhibits high accuracy in simulating the real-world system.

Traditional clustering methods perform not so well for PPI data considering that the PPI network exhibits the small world and scale free properties [39,45]. Many new methods are proposed for clustering PPI networks. In 2002, Girvan and Newman proposed a new method, named a GN algorithm, to identify protein complexes in networks [15]. Newman also proposed a Newman fast algorithm in 2003, which is based on hierarchical condensations [31]. In 2003, Bader and Hogue proposed MCODE (Molecular complex detection) algorithm [5]. They first weighted every node in PPI networks by the node's local neighbor density, then picked the nodes with high weights as the seed nodes of initial clusters and further augmented these clusters to form the preliminary clusters. In 2006 Adamcsek et al. developed the software CFinder to uncover the overlapping clusters in biological networks [1]. In 2009, Leung developed a core-attachment approach for predicting protein complexes from the PPI network of single species based on a study on the organization of protein complexes [24]. Based on the similar idea, Wu et al. developed the COACH algorithm [46]. Unlike CORE, the main feature of COACH is its ability to deal with the overlapping problem. In 2011, Sun and Gao proposed a novel method based on a fuzzy relation model to detect overlapping and non-overlapping community structures in complex networks [39]. This model uses fuzzy relation to identify community structure instead of involving traversal search of a graph model. In 2012, Ma and Gao also used a graph theory combined with a core-attachment-based algorithm to predict protein complexes in PPI networks [29]. In 2013, Chen and Wu proposed an algorithm based on multiple topological structures to identify protein complexes from PPI networks [8].The prediction results produced by this algorithm show that the multiple topological structure based algorithm cannot only discover a large number of predicted protein complexes, but it can also generate results with high accuracy in terms of *f-measure*, matching with known protein complexes and functional enrichments with GO. There are many other algorithms, such as RNSC [18], DPClus [3], IPCA [25], link community (LinkCom) [2] and Markov clustering (MCL) [13]. Furthermore, Chen et al. developed computational algorithms for identifying protein complexes from PPI networks in terms of used data and detection mechanism [9].

MCL is a graph clustering algorithm based on stochastic flow simulation, which has shown to be effective in clustering biological networks [6, 41]. It offers several advantages. It offers a sound approach based on the probabilities of transition in graphs. It shows to be significantly tolerant to noise and behaves robustly [41]. While not being completely parameter free, varying a single parameter can result in clusters of different granularities [35]. However, in spite of its popularity in the bioinformatics community, MCL has drawn limited attention from the data mining community primarily because it does not scale very well to moderate sized graphs [7]. Additionally, the algorithm tends to break communities, which is not ideal in many cases. To retain the strengths of MCL and alleviate its weakness, Satuluri et al. proposed a Regularized MCL (R-MCL) [35,36]. This improved algorithm runs faster than MCL and improves the accuracy of protein complex identification. Nevertheless, these two algorithms still can only generate non-overlapped clusters and they always assign all proteins into clusters while not all proteins are functionally annotated. Then in 2012 Shih and Parthasarathy proposed a 'Soft' R-MCL (SR-MCL) to construct overlapped clusters [38]. The intuition behind SR-MCL is to produce overlapped clusters by iteratively re-executing R-MCL while ensuring the resulting clusters are not always the same. In order to produce different clusters in each iteration, the stochastic flows are penalized if they flow into a node that was an attractor node in previous iterations. Since, iteratively re-executing R-MCL would produce several redundant and low-quality clusters, a post-processing is applied to remove those clusters. Only a cluster that is not removed by the post-processing is predicted as a protein complex, so not all proteins are assigned into clusters. Although R-MCL and SR-MCL are better than MCL, their parameters are still user-specified which makes it difficult to deal with a variety of PPI data sets effectively. In this paper, we propose a hybridization strategy to automatically adjust the parameters by introducing the firefly algorithm (FA).

Swarm intelligence is a type of population-based meta-heuristic. It seeks near-optimal solutions to the difficult optimization problems by simulating a collective social behavior of individuals, such as birds, bees, ants and fishes (ant colony optimization, ACO [37], artificial fish school algorithm, AFA [28]). The particle swarm optimization (PSO) algorithm, proposed by Kennedy and Eberhart in 1995, which simulates the foraging behavior of birds, is considered as a simple and efficient implementation, solving various optimization problems [17]. The algorithm can also be used in clustering problems. In 2012 Lei et al. proposed an improved functional-flow based approach through the quantum-behaved particle swarm optimization (QPSO) algorithm, which can find the optimum threshold automatically when calculating the lowest similarity between complexes [19]. It showed better performance than functional-flow method in terms of accuracy and number of matched clusters. In 2013, Lei et al. developed an improved bacteria foraging optimization (BFO) based on BFO mechanism and intuitionistic fuzzy sets, with trigonometric membership functions and the indeterminacy degree is introduced to detect the overlapping proteins [23]. Lei et al. proposed a novel PMABC-ACE model based on the propagating mechanism of artificial bee colony (PMABC) and adopted the aggregation coefficient of edge (ACE) in the preprocessing of edges of PPI networks [29]. This algorithm can automatically produce the number of clusters when running the clustering procedure. These two approaches could detect the overlapping proteins while their time complexity was substantially reduced. In addition to the PSO and ABC algorithms and their variants, ACO and AFA algorithm also exhibit a wide range of applications. For instance, Seçkiner introduced ACO to deal with function optimization [37] and Ma used AFA to solve path planning problems [28].