Journal of Fiber Bioengineering and Informatics 8:3 (2015) 565–574
doi:10.3993/jfbim00123

A Parallel Algorithm for Detecting Complexes in Protein-protein Interaction Networks with MapReduce ⋆

Zhenmei Yu [a,*], Zhangxu Li [b]

[a] *School of Information Technology, Shandong Women's University, Ji'nan 250002, China*
[b] *College of Computer Science and Technology, Jilin University, Changchun 130012, China*

Abstract

Detecting protein complexes from Protein-protein Interaction (PPI) networks has been the focus of many recent efforts on protein. With the appearance of big data and large scale PPI networks, traditional sequential methods, which analyze interaction networks and detect protein complexes, do not utilize high performance computing. In this paper, we propose a parallel algorithm using cloud computing method to improve the computational efficiency and detect protein complexes. Because MapReduce programming model simplifies the implementation of many data parallel applications, firstly we use it to calculate the value of each edge and the value of each node from PPI networks, then expand complexes. At last, we perform the algorithm on different data to test the speedup of the algorithm. Moreover, through the parallel algorithm is compared with sequential method, experimental results show that the running time of parallel algorithm is short. We get a conclusion that parallel algorithm can also accurately assign proteins with similar functions to a complex.

Keywords: PPI Network; Protein Complexes; Parallel Algorithm; MapReduce

1 Introduction

Proteins are the essential materials to constitute the cells and tissues and the interactions between proteins are important for numerous biological functions [1]. Moreover, large amounts of the protein-protein interaction data have been generated by various experimental technologies, e.g., yeast-two-hybrid [2], affinity purification [3-5], and so on. How to mine biological significance from massive protein interaction data, almost become a common problem that every researchers must deal with. The traditional approach usually improved the complexity of algorithm, but it is impractical and expensive overhead that put large-scale mass data into memory to process at the same time, and the scale of network has been limited [6]. Cloud computing, distributed computing, flow computing methods have been proposed to deal with this kind of challenge [7].

⋆This work was supported in part by the Natural Science Foundation of China (Nos. 61272478, 61472416) and the Found of Scientific Research and Innovation Team of Shangdong Women's University on Data Mining and Intelligent Application.

*Corresponding author.
Email address: yuzhenmei@gmail.com (Zhenmei Yu).

The open source Hadoop cloud computing project has aroused widespread concern. The main advantage of Hadoop is that its ability can scale to hundreds or thousands of nodes in a cluster and further handle vast amount of data efficiently over a set of computers [8]. Moreover, MapReduce, as a distributed computing paradigm, is inspired from the map and reduce functions available in functional languages and MapReduce programming model simplifies the implementation of many data parallel applications [7].

In this paper, we present a new parallel algorithm for network clustering, particularly, the discovery of protein complexes in PPI networks. Our discussion focuses on MapReduce model. Firstly, we describe the crucial concepts for finding protein complex in PPI networks. Then we describe in detail the implementation of MapReduce model. Through several important experiments, we show the effectiveness of our methodology using the different PPI networks. We give an analysis of speedup of parallel algorithm. Besides, in order to evaluate the feasibility and validity of our algorithm, we compare parallel algorithm with sequential algorithm using F-measure and Coverage Rate as two important evaluation criterions. As a result, we found that our proposed Parallel algorithms is almost as effective as the sequential algorithm. But the running time of parallel algorithm is more efficient than sequential algorithm. Moreover, to validate the biological significance, we evaluate the predicted results by using Gene Ontology annotations [9] which includes Biological Process, Molecular Function and Cellular Component. We get a conclusion that parallel algorithm can also accurately assign proteins with similar functions to a GO-enriched complex.

The rest of this paper is organized as follows. In Section 2, we detailed introduce the basic knowledge of protein network and MapReduce and how to parallel detect and expand the protein complex with an example. Through several important experiments, we show the effectiveness of our methodology using the different PPI networks in Section 3. In Section 4, we conclude our parallel algorithm in discovering the protein complex in PPI networks.

2 Methodology

In this paper, we discuss the value of edges. We propose the definitions to describe the value of each edge and node in a network. In addition, we introduce the basic knowledge of the MapReduce paradigm and the process of expanding a protein complex from a PPI network as an simple example.

A protein network can be modeled as a graph $G = (N, E)$. N represents the set of nodes and E represents edges. A Graph $G(N, E)$ is composed of many nodes connected by edges, and an edge consists of two nodes x and y. The protein complex in protein networks is a group of nodes and edges, whose biggest feature is that it is densely connected in the group and the connection with the other groups is relatively sparse. Detecting complexes from the protein network is to reveal the true community structure in the different types of protein network.

Definition 1. For an edge $e(x, y)$, it is connecting node x and node y, the value of edge $E(x, y)$ in the network can be defined as:

$$E(x, y) = N(x) \bigcap N(y) \tag{1}$$

where, $N(x)$ is defined as the inclusive neighbors of node x. The more two nodes share the same neighbors, the larger value of the edge connecting two nodes are.