Protein Complex Prediction via Bottleneck-Based Graph Partitioning

Jaegyoon Ahn[1]
ajk@cs.yonsei.ac.kr

Dae Hyun Lee[1]
mudjjang@yonsei.ac.kr

Youngmi Yoon[2]
ymyoon@gachon.ac.kr

Yunku Yeu[1]
yyk@cs.yonsei.ac.kr

Sanghyun Park[1,*]
sanghyun@cs.yonsei.ac.kr

[1] Department of Computer Science, Yonsei Univ., 3rd Engineering Bldg. 533-1, Shinchon-dong, Seodaemun-gu, Seoul, Korea, 0082-2-2123-7757

[2] Department of Computer Engineering, Gachon Univ., 1342 Seongnamdaero, Sujeong-gu, Seongnam-si, Gyeonggi-do, Korea, 0082-32-820-4393

ABSTRACT
Detecting protein complexes is one of essential and fundamental tasks in understanding various biological functions or processes. Therefore, precise identification of protein complexes is indispensible. For more precise detection of protein complexes, we propose a novel data structure which employs bottleneck proteins as partitioning points for detecting the protein complexes. The partitioning process allows overlapping between resulting protein complexes. We applied our algorithm to several PPI (Protein-Protein Interaction) networks of Saccharomyces cerevisiae and Homo sapiens, and validated our results using public databases of protein complexes. Our algorithm resulted in overlapping protein complexes with significantly improved F1 score, which comes from higher precision.

Categories and Subject Descriptors
H.2.8 [**Database Management**]: Data mining; J.3 [**Life and Medical Sciences**]: Biology and Genetics

General Terms
Algorithms

Keywords
Network clustering, Protein complex detection, Protein-protein interaction, Bottleneck protein

1. INTRODUCTION
Most proteins are known to be involved in complex biological processes or functions in a cell, forming a protein complex with other proteins [1]. Therefore, detecting protein complexes is one of essential and fundamental tasks in understanding various biological functions or processes. A protein complex can be modeled as an undirected graph of which node is a protein and edge is a physical interaction between two protein nodes. This physical interaction of two proteins is called PPI (Protein-Protein Interaction). Representative methods to find those interactions are two-hybrid system [2] and Mass Spectrometry [3]. Recent development of those high-throughput methods has resulted in abundant PPI network.

A protein complex is a set of proteins that interact with each other, so it is frequently assumed that distances between its member proteins are short, and its members tend to form clique-like structure in the PPI network. Accordingly, a protein complex is often assumed as a dense sub-graph in the PPI network. There have been active researches to develop algorithms for detecting protein complexes, and many of them are based on searching dense sub-graph in the PPI network. MCODE [4] gives high weight to nodes of which degree is high, and searches the network using those nodes as seeds. It enforces local search on the network, and finds sub-network whose nodes are highly interconnected. CMC [5] gives weight to PPIs using an iterative scoring method to assess the reliability of PPI, finds maximal cliques from the weighted PPI network, and then removes or merges overlapping maximal cliques based on their interconnectivity. MCL [6] detects clusters by distinguishing the strong and weak connections in the network and partitioning the network, based on manipulation of transition probabilities or stochastic flows between vertices of the graph. MCL has been reported to have good performance, and many variations of it have been proposed [7, 8, 9]. However, they are known to suffer from imbalance of resulting clusters [9].

These network clustering algorithms commonly do not allow overlapping between identified protein complexes. In other words, a protein can be involved in only one protein complex. Recently, algorithms that allow overlapping have been extensively studied. DPClus [10] detects initial protein complexes starting from the seeds and then including neighbors so as to maintain the edge's density of the sub-network above the threshold. Then it finds overlapped protein complexes extending the initial protein complexes. CFinder [11] is based on Clique Percolation Method

* To whom correspondence should be addressed.

DTMBIO'12, October 29, 2012, Maui, Hawaii, USA.