Protein Complex Discovery from Protein Interaction Network with High False-Positive Rate*

Yunku Yeu[1], Jaegyoon Ahn[1], Youngmi Yoon[2], and Sanghyun Park[1]

[1] Dept. of Computer Science, Yonsei University,
3rd Enginnering Bldg. 533-1, Shinchon-dong, Seodaemun-gu Seoul, Korea
[2] Division of Information Technology, Gachon university of Medicine & Science,
1108 Gachon-Kwan, Yonsu-dong, Yonsu-gu, Incheon, Korea
{yyk,ajk}@cs.yonsei.ac.kr, ymyoon@gachon.ac.kr,
sanghyun@cs.yonsei.ac.kr

Abstract. Finding protein complexes and their functions is essential work for understanding biological process. However, one of the difficulties in inferring protein complexes from protein-protein interaction(PPI) network originates from the fact that protein interactions suffer from high false positive rate. We propose a complex finding algorithm which is not strongly dependent on topological traits of the protein interaction network. Our method exploits a new measure, GECSS(Gene Expression Condition Set Similarity) which considers mRNA expression data for a set of PPI. The complexes we found exhibit a higher match with reference complexes than the existing methods. Also we found several novel protein complexes, which are significantly enriched on Gene Ontology database.

Keywords: data mining; machine learning; protein interaction; protein complex.

1 Introduction

Most proteins are known to function within complicated cellular pathways, interacting with other proteins either in pairs or as components of larger complexes [1]. Therefore finding protein complexes is fundamental process for understanding biological functions and cellular organization. Protein complex can be modeled as an undirected graph whose node is a protein, and edge is a physical interaction between two protein nodes. Large scale of PPI information can be abstracted into a PPI network, so finding protein complexes is same as finding subgraph in this global map of protein interactions. Because protein complexes are groups of proteins that interact with others, they are shown as dense subgraphs in PPI network. Several algorithms based on clustering dense region or cliques were proposed to discover protein complexes from PPI networks.

MCODE [2] gives weights to nodes that are densely connected, predicts complexes, and then does postprocess for optimal result. Markov clustering algorithm (MCL) [3] partitions the graph by discriminating strong and weak flows in the graph. DPClus

* This work was supported by National Research Foundation of Korea funded by the Korean Government under Grant (No. 2010-0003965).

C. Pizzuti, M.D. Ritchie, and M. Giacobini (Eds.): EvoBIO 2011, LNCS 6623, pp. 177–182, 2011.