

USING INDIRECT PROTEIN-PROTEIN INTERACTIONS FOR PROTEIN COMPLEX PREDICTION

Hon Nian Chua¹, Kang Ning², Wing-Kin Sung², Hon Wai Leong² and Limsoon Wong²

¹*Graduate School of Integrated Sciences and* ²*Department of Computer Science, National University of Singapore*
g0306417@nus.edu.sg, {ningkang, ksung, leonghw, wongls}@comp.nus.edu.sg

Protein complexes are fundamental for understanding principles of cellular organizations. Accurate and fast protein complex prediction from the PPI networks of increasing sizes can serve as a guide for biological experiments to discover novel protein complexes. However, protein complex prediction from PPI networks is a hard problem, especially in situations where the PPI network is noisy. We know from previous work that proteins that do not interact, but share interaction partners (level-2 neighbors) often share biological functions. The strength of functional association can be estimated using a topological weight, FS-Weight. Here we study the use of indirect interactions between level-2 neighbors (level-2 interactions) for protein complex prediction. All direct and indirect interactions are first weighted using topological weight (FS-Weight). Interactions with low weight are removed from the network, while level-2 interactions with high weight are introduced into the interaction network. Existing clustering algorithms can then be applied on this modified network. We also propose a novel algorithm that searches for cliques in the modified network, and merge cliques to form clusters using a “partial clique merging” method. In this paper, we show that 1) the use of indirect interactions and topological weight to augment protein-protein interactions can be used to improve the precision of clusters predicted by various existing clustering algorithms; 2) our complex finding algorithm performs very well on interaction networks modified in this way. Since no any other information except the original PPI network is used, our approach would be very useful for protein complex prediction, especially for prediction of novel protein complexes.

Keywords: protein-protein interaction, protein complex prediction, level-2 interaction, partial clique merging

1 INTRODUCTION

Identification of functional modules in protein interactions network is a first step in understanding the organization and dynamics of cell functions. Protein-protein interaction networks (PPIs) are rapidly becoming larger and more complete as research on proteomics and systems biology proliferates [1]. As a result, more protein complexes are being identified [2]. A protein complex is a group of two or more associated proteins. Protein complex is a form of quaternary structure. Similar to phosphorylation, complex formation often serves to activate or inhibit one or more of the associated proteins. Many protein complexes are established, particularly in the model organism *Saccharomyces cerevisiae* (Bakers’ yeast). With a wealth of and constantly increasing size of PPI datasets, efficient and accurate intelligent tools for identification of protein complexes are of great importance. In this paper, we have focused on predicting protein complexes from PPI data.

Currently, there are several approaches to the protein complex prediction problem [3-8]. Spirin et. al. [3] proposed using clique finding and super-paramagnetic clustering with Monte Carlo optimization to find clusters of proteins. They found a significant number of protein complexes that overlap with experimentally derived ones.

While clique finding [3] imposes stringent search criterion, and generally results in greater precision, recall is limited because: 1) protein interaction networks are incomplete; and 2) protein complexes may not necessary be complete subgraphs. Another approach, such as MCODE [5], are clustering based. MCODE makes use of local graph density to find protein complex. PPI networks are transformed to weighted graphs in which vertices are proteins and edges represent protein interactions. The algorithm operates in three stages: vertex weighting, complex prediction and optimal post-processing. Each stage involves several parameters that can be fine-tuned to get better predictions. However, clustering approaches [5, 8] yield good recall but sacrifice precision. To make clustering based approaches more viable, [4, 7] show that it is possible to identify high precision subsets of clusters from clustering results by post-processing based on functional homogeneity, cluster size and interaction density. While post processing significantly improves precision, recall is drastically reduced. Moreover, the approach makes use of functional information, which limits its applicability in less studied genomes such as *Homo sapiens*, *Mus mucus* and *Arabidopsis thaliana*. Recently, a popular clustering algorithm, Markov

