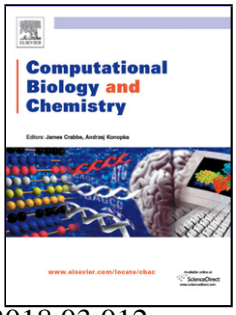

Title: Protein complex prediction by date hub removal

Author: Iana Pyrogova Limsoon Wong

PII: S1476-9271(18)30144-0
DOI: https://doi.org/doi:10.1016/j.compbiolchem.2018.03.012
Reference: CBAC 6817

To appear in: *Computational Biology and Chemistry*

Received date: 7-3-2018
Accepted date: 13-3-2018

Please cite this article as: Iana Pyrogova, Limsoon Wong, Protein complex prediction by date hub removal, <*![CDATA[Computational Biology and Chemistry]]*> (2018), https://doi.org/10.1016/j.compbiolchem.2018.03.012

Protein complex prediction by date hub removal

Iana Pyrogova[a], Limsoon Wong[a]

[a]*Department of Computer Science, National University of Singapore, Singapore*

Abstract

Proteins physically interact with each other and form protein complexes to perform their biological functions. The prediction of protein complexes from protein–protein interaction (PPI) network is usually difficult when the complexes are overlapping with each other in a dense region of the network. To address the problem of predicting overlapping complexes, a previously proposed network–decomposition approach is promising. It decomposes a PPI network by e.g. removing proteins with high degree (hubs) which may participate in different complexes. This motivates us to examine a list of proteins, which bind their different partners at different time or at different location (viz. date hubs), manually collected from literature, for network decomposition. Results show that the CMC complex discovery algorithm after removing date hubs recalls more overlapping complexes that were missed earlier. Further improvement in performance is achieved when we predict date hub proteins based on simple network features and remove them from PPI networks.

Keywords: Protein complex prediction, PPI networks, PPI network decomposition, date hub proteins

1. Introduction

Proteins play a vital role in cellular processes. Generally, they do not act alone but form complexes with other proteins to carry out their biological functions. Protein complexes are formed by physical interaction among proteins at specific time and space. Detecting protein complexes is important for understanding the dynamics of biological processes within an organism. Using advances from high-throughput proteomic techniques, such as yeast two-hybrid system [8] and tandem affinity purification with mass spectrometry [20], it has become possible to compile a large network of protein interactions. However, extracting protein complexes from such networks is a non-trivial task. Therefore, a wide variety of sophisticated PPI network analysis algorithms to detect protein complexes have been proposed in the last two decades. They are often designed to detect sub-graphs with specific topological structures in a PPI network, such as cliques [17, 13], dense sub-graphs [10, 15], and core-attachment structures [24]. Some algorithms incorporate topological features to assign weights to vertices or edges, such as the number of common neighbors [10, 11] and density [2]. A comprehensive review of complex discovery algorithms is given by Srihari et al. [22].

For a better understanding of protein interaction networks, Han et al. [9] studied hub proteins with gene expression data, and introduced the distinction between date and party hub proteins. First, they defined hubs as proteins with degree greater than 5. Then for each hub, they calculated the average of Pearson correlation coefficients (PCC) between the hub protein and each of its neighbors for mRNA expression. Their results suggest that party hubs are co-expressed with their interacting partners (have higher average PCC), while date hubs have significantly more diverse localization of partners. Therefore, party hubs are hubs that interact with their partners at the same time, whereas date hubs bind their different partners, which belong to multiple complexes, at different times or at different locations. Using an arbitrary average PCC threshold, the hub proteins with higher values of average PCC than the threshold were defined as party hubs, and all other proteins with the degree higher than 5 were indicated as date hubs.

Some concerns have been raised [1, 3] regarding the date/party hub classification and the analysis methodology. The major criticism of the study of Han et al. [9], made by Batada et al. [3] and Agarwal et al. [1], is that the distinction between date and party hubs is an artifact of prior data. Batada et al. [3] noted that the bimodality

Email addresses: ianapyrogova@gmail.com (Iana Pyrogova), WongLS@comp.nus.edu.sg (Limsoon Wong)