Extending the MCL-CA Algorithm for Protein Complex Detection from Weighted PPI Networks

Sriganesh Srihari and Hon Wai Leong
Department of Computer Science, National University of Singapore, Singapore 117590
{srigsri | leonghw}@comp.nus.edu.sg

Background: Protein complexes are responsible for most of vital biological processes within the cell. Understanding the machinery behind these biological processes requires detection and analysis of complexes and their constituent proteins. Recent biological experiments on yeast by Gavin et al. (2006) revealed that proteins within a complex are organized into two parts: core and attachment. Based on these insights, we recently [2] proposed a new algorithm called MCL-CA, based on Markov clustering coupled with core-attachment based refinement, to detect protein complexes from yeast protein-protein interaction networks (PPI). We showed that the core-attachment based refinement significantly improved the accuracies of predicted complexes when matched with manually-curated yeast complexes. However, MCL-CA was still prone to noise (false positives) in the interaction datasets.

MCL-CAw: Recent research showed that complex detection algorithms performed significantly better when protein interaction networks were weighted (scored) using measures like Iterative CD distance [1] that reflect the reliability of physical interactions. In this work, we extend our algorithm to MCL-CAw that detects complexes from weighted PPI networks. Our experiments on large weighted datasets show remarkable improvement in accuracies of predicted complexes. A direct result of this is that MCL-CAw is able to cover significantly larger number of manually-curated complexes than MCL-CA. This reflects the sensitivity of our algorithm towards biologically-meaningful weighting measures, and therefore strongly adheres to the findings in [1] about better coverage on weighted datasets.

Results: We considered unweighted yeast protein interaction dataset (from Biogrid) consisting of 4041 proteins and 26748 interactions for our experiments. The corresponding weighted dataset consisted of 2209 proteins and 20399 interactions. A total of 463 manually-curated yeast complexes from three sources (MIPS, Aloy et al. 2004, Wodaklab) were accumulated as the benchmark. We used Jaccard score with a threshold 0.5 as the condition for valid match between predicted and known complexes. MCL-CAw covered 121 out of 246 known complexes (Recall: 0.4917) actually present in the dataset as against only 108 out of 371 known complexes (Recall: 0.2911) covered by MCL-CA. About 73 out of 136 predicted complexes of MCL-CAw correctly matched some known complex (Precision: 0.5367) as against 69 out of 488 (Precision: 0.1413) predicted by MCL-CA. Consequently, the overall performance of MCL-CAw (F-measure: 0.5134) was significantly higher than that of MCL-CA (F-measure: 0.1987).

References: