2:42pm
poster
16
spotlight

Identifying Protein Complexes in Saccharomyces cerevisiae

Boyko Kakaradov
Haidong Wang
Sean Collins
Nevan Krogan
Daphne Koller

Motivation:
The identification of protein complexes on a genome-wide scale is an open problem that has recently been illuminated by two high-throughput Tandem Affinity Purification / Mass Spectrometry (TAP/MS) studies on Saccharomyces cerevisiae [1, 2]. While preliminary analytical methods have successfully found biologically relevant clusters in the noisy and incomplete TAP/MS data [3,4], no clearly superior complex identification method exists. High-quality, comprehensive complex predictions are a necessary foundation for the understanding of cellular pathways through computational models. Thus, we propose a method that incorporates multiple sources of evidence to significantly increase the accuracy and coverage of complex predictions.

Materials and Methods:
Our approach consists of two stages. In the first stage, features for each pair of proteins are derived from a heterogeneous set of assays including TAP-MS score, GO annotation, cellular localization, and expression correlation. We then use Boosting with logloss to learn a model that predicts the affinities between two proteins based on the features for the protein pair. The affinity is defined as the likelihood of two proteins being in the same complex. In the second stage, we compute the affinity scores for all pairs of proteins. We construct a graph with proteins as its nodes and the inverse of the affinity scores as the distance between the nodes. We apply hierarchical clustering on the resulting graph to get a set of clusters.

Results:
We compiled a reference complex set of 340 well-known complexes by combining multiple sources: small-scale experiments from SGD and MIPS, and hand-curated lists from collaborating biologists [1, 3]. A recently published method [4] predicts 400 complexes 66 of which match the reference set perfectly. The following ablative analysis describes the relative gains produced by different aspects of our approach. Our baseline - hierarchical clustering over the original TAP/MS scores only, predicts 486 complexes of which 77 are perfect. The use of data integration resulted in 350 predictions including 92 perfect matches.

Conclusion:
Our results show that complementing the TAP/MS score with various indirect evidences improves the complex predictions significantly by mitigating noise and complementing coverage.

References:
[1] Krogan, N.J. et al. "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae" Nature 2006
[2] Gavin, A. C. et al. "Proteome survey reveals modularity of the yeast cell machinery" Nature 2006
[3] Collins, S. et al. "Toward a Comprehensive Atlas of the Physical Interactome of Saccharomyces cerevisiae" Mol Cell Proteomics 2007 6: 439-450
[4] Pu, S. et al. "Identifying functional modules in the physical interactome of Saccharomyces cerevisiae" Proteomics 2007, 7, 944-960