GRAPH-BASED ANALYSIS OF PROTEIN-PROTEIN INTERACTION DATA SETS

Pengjun Pei
ppei@cse.buffalo.edu

The Department of Computer Science & Engineering
State University of New York at Buffalo
Buffalo, New York, 14260

\

Dissertation Committee:	Dr. Aidong Zhang (Chair)
	Dr. Matthew J. Beal
	Dr. Murali Ramanathan
External Reader:	Dr. Yulan Liang

Abstract

High-throughput methods for detecting *protein-protein interactions (PPI)* have recently gained popularity. These rapid advances in technology have given researchers an initial global picture of protein interactions on a genomic scale. The usefulness of this understanding is, however, typically compromised by noisy data and intrinsic complexity of the biological system. In this dissertation, we attempt to solve some problems in effectively analyzing the data.

Firstly, since there are lots of false positives in experimentally detected interactions, we propose a novel topological measurement to select reliable interactions from the noisy data. Our method is based on the small-world network property of the protein interaction network and generalizes purely local measures adopted previously. Based on our observation that the true positive interactions in protein complexes and tightly coupled networks demonstrate dense interactions, we propose to measure the significance of two proteins' co-existence in a dense network as an index of interaction reliability. Our topological measure also integrates the prior confidence of each data set. The experiments demonstrate that our measure can be used to identify reliable interactions and to predict potential interactions with improved performance. Meanwhile, we discovered two additional properties: namely, *short alternative path property* and *local clustering of network property* of the protein interaction network, which are generalizations of previously known protein interaction network properties.

Secondly, we address the problem of effectively incorporating domain knowledge into the protein clustering process. Based on our analysis of the relationship of network topology and biological relevance, we propose a novel semi-supervised clustering algorithm suitable for the noisy protein interaction network. We choose to estimate the pairwise similarity between each protein pair and use this similarity as input to clustering algorithms. Therefore, it is not bounded to any specific clustering methods. We select

topological features in the network and define a model to map these features to pairwise similarities. The known protein annotations are used to train the model. Using this model, we can estimate the pairwise similarity between each pair of proteins. Finally, normal unsupervised clustering algorithms can be applied using the similarity matrix. Since our similarity measure has already incorporated prior protein annotations, our algorithm can detect clusters with improved performance. Also, the unsupervised clustering algorithms we adopt maintain the explorative nature and therefore are capable of detecting new protein functional groups.

Thirdly, we investigate the problem of protein complex detection. Protein complexes can be roughly considered as densely connected subgraphs in the network. The difficulties in this problem are caused by the fact that protein complexes may overlap with each other, i.e. containing shared proteins, and the protein interaction network contains a lot of noise. To overcome these difficulties, we propose a novel subgraph quality measure, and based on the measure, we propose a novel *seed-refine* algorithm. Our subgraph quality measure achieves two goals: 1) it provides a statistically meaningful combination of inside links, outside links and the size of the subgraph and, 2) it provides a statistically meaningful combination of the quality contribution of each vertex in the subgraph. Our *seed-refine* algorithm consists of a two-layer seeding heuristic to find good seeds and a novel subgraph refinement method that controls the overlap between subgraphs. Our algorithm allows to output overlapping subgraphs but methodologically makes it possible only when there is strong evidence to do so. Experiments confirm the effectiveness of our method.