Algorithmic and Statistical Models on Protein Interactions

Tony Chiang

January 30, 2011

Many of the functions of proteins involve their interaction with other molecules, often with other proteins in order to assemble organizational units. The importance of protein interactions is seen from the fact that the presence/absence of certain protein complexes directly corresponds and influences the phenotype exhibited by the cell. Our primary purpose is to investigate the relationship between certain (mainly disease) phenotypes and protein complexes. Lage et al. [2007] have introduced the notion of a phenome-interactome though most of this work still is far from complete. For this reason, a better understanding of protein interactions is fundamental to the study of biological systems in general.

A considerable amount of work has been conducted to map the protein interaction graph (an estimated interactome) for a number of organisms, notably in *S. cerevisiae*. Graphs provide a meaningful way to model relationships between biological subsystems while being flexible enough to represent many types of data. Confusion arises when a experimentally derived data graph (G') is assumed to be identical to the intended, underlying graph (G) of biological interest. In most cases, we make an estimate of G (denoted by $\hat{G}$) based on the data of G'.

I have investigated most of the large-scale publicly available datasets [Chiang et al., 2007a, Scholtens et al., 2007, Chiang et al., 2007b] and have implemented software packages to analyze such data (*ppiStats, ppiData, Rintact, ScISI, and simulatorAPMS*) which are a part of the *Bioconductor* project. Collaborating with the Fred Hutchinson Cancer Research Center and Northwestern University, I have shown, using a binomial error model, that the data are not only affected by stochastic errors but also by systematic bias. Such bias in the data have often confounded researchers who have dismissed protein interaction data as "too noisy" or irreproducible. Knowing how the data might be biased, I will then use a variant maximal clique finding algorithm with a penalized likelihood component to mine for tightly cohesive subgraphs. These subgraphs are a good representation for the modular protein complexes. We will examine how certain variants of complexes can cause disease thus giving clues on how to fight such phenotypes.

While my most recent work focused on the interpretation of interaction data, my future work will be the estimation of the protein complex interactome and the functional annotation for novel protein complexes. In order to test the veracity of protein complex estimates from the data, we will collaborate with Dr. Anne Claude Gavin's group at the European Molecular Biology Laboratory who works with yeast models; in addition, we will also collaborate with Dr. Seth Grant's group at the Wellcome Trust Sanger Institute who works with mouse models. Both groups will use the Tandem Affinity Purification (TAP) methods to assay protein complex co-members for a protein of interest. Working with two different model organisms will help us better understand the role of multi-protein complexes, help us investigate the conservation of modular protein complexes over large evolutionary distances, and help us determine the functional roles that such complexes may have in the organization of the cell.

References

T. Chiang et al. Coverage and error models of protein-protein interaction data by directed graph analysis. *Genome Biology*, 2007a.

T Chiang et al. Rintact: enabling computational analysis of molecular interaction data from the intact repository. *Bioinformatics*, 2007b.

K Lage et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 2007.

D Scholtens et al. Estimating node degree in bait-prey graphs. *Bioinformatics*, 2007.