Chapter 7

Conclusion

In this thesis, we have investigated model-based approaches to predict protein complexes from tandem affinity purification experiments. We compared simple overlapping models to a partitioning model. In addition, we proposed a visualization framework to delineate overlapping complexes from experimental data.

Previous techniques for protein interaction analysis rely on heuristic algorithms. They yield useful results, but make no attempt to provide an explanatory model of the obtained observations. In addition, heuristic algorithms often have a plethora of adjustable parameters, with very little guidance on how to adjust them for a particular dataset. We believe that model-based techniques provide a more rigorous framework for protein interaction analysis. A probabilistic model explicitly and quantitatively states the assumptions about how protein interactions are exposed by the experimental technique. The actual algorithm then uses the model to compute an estimate of the clustering.

We propose two models to predict protein complexes from experimental data. Our first model is in some sense the simplest possible one. It is based on frequent itemset mining, which merely counts the incidence of certain sets of proteins within the experimental results. The affinity of two sets of proteins to form clusters is modeled to be independent, regardless of any overlapping members between these sets. Our second model assumes that formation of protein complexes can be reduced to pairwise interactions between proteins. Interactions between proteins are more likely for pairs of proteins if they come from the same cluster. Based on this model, we use Markov Random Fields theory to calculate a maximum-likelihood assignment of proteins to clusters.

We compare the effectiveness of the two models by evaluating them against the MIPS and Reguly benchmarks. In our evaluation, the partitioning model based on Markov Random Fields performs much better than the overlapping model. This indicates that protein clustering in nature is mainly a pairwise phenomenon, despite individual examples to the contrary. Unlike previous work, our model incorporates observational error, which enables us to directly use the experimental data without requiring an intermediate interaction graph. The assignment to protein complexes are estimated with the Mean Field Annealing algorithm. We can find complexes in the data obtained from high-throughput experiments without prior elimination of proteins from the sets. Because there are proteins which cannot be clustered well, we also provide a model-based quality score to each predicted complex. Our method does not rely on heuristics, which is particular important for applications on protein complex studies in organisms that do not have an established reference frame, for instance many prokaryotes. The model has few parameters, all of which are elegantly estimated from the experimental data using maximum likelihood. The results compare favorably to reference data sets, notably for the larger unfiltered data sets, making us confident that it will perform well on a wide range of datasets. For future work, the hard assignment imposed by our model can be relaxed to capture overlapping complexes, but the model and minimization algorithm must be changed.

Finally, we developed a useful visualization method for tandem affinity experimental data. Purification results are modeled as a directed graph. Edge weights are defined by the inclusion probability between two purifications. This measure captures the asymmetric nature of the bait-prey experiment. We demonstrate the effectiveness of the method by presenting a visualization of the most recent large-scale experiments.