

Algorithms to identify protein complexes from
high-throughput data

Wasinee Rungarityotin

In partial fulfillment towards a doctoral degree in Computer Science

submitted to

Department of Mathematics and Computer Science

Freie Universität Berlin

Berlin, Germany

June 2007

Reviewers:

Prof. Dr. Martin Vingron

Prof. Dr. Knut Reinert

Reviewers:

Prof. Dr. Martin Vingron

Prof. Dr. Knut Reinert

Date of the defense: 9. November 2007

To my mother

Abstract

Recent advances in proteomic technologies such as two-hybrid and biochemical purification allow large-scale investigations of protein interactions. The goal of this thesis is to investigate model-based approaches to predict protein complexes from tandem affinity purification experiments. We compare a simple overlapping model to a partitioning model. In addition, we propose 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 a model of protein complexes from experimental data. 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 Field theory to calculate a maximum-likelihood assignment of proteins to clusters.

We compare the effectiveness of the two models by evaluating them against two benchmarks. In our evaluation, the partitioning model performs much better than the overlapping model. This indicates that protein clustering in nature is likely to be a pairwise phenomenon, despite individual examples to the contrary. The performance of the second model is as good as previous techniques based on heuristics, and in contrast to them it has no adjustable parameters, making us confident that it will perform well on a wide range of datasets.

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.

Acknowledgements

I would like to especially thank Florian Markowetz for discussion on mathematical formulation in Chapter 6 and Roland Krause in introducing me to problems commonly found when analyzing datasets from protein purification experiment. I thank Alexander Schliep for discussions on graph-based clustering algorithms. I thank both Alexander Schliep and Martin Vingron for their support.

Furthermore, many ideas in this thesis were developed during my visit at the Institute for Pure And Applied Mathematics (IPAM), University of California, Los Angeles Spring 2004 program “Proteomics: Sequence, Structure, Function” as well as the reunion program in 2005.

Most importantly, I thank Arno Schödl for all his support during the last year of my Ph.D., especially his suggestion on trying out Markov Random Field.

Notation

$G = (V, E)$	A graph G with a vertex set V and an edge set E
$ V $	Number of vertices
$ E $	Number of edges
C_i	The clustering coefficient of a vertex i
D_G	The density of a graph G
$M(G)$	A transition matrix derived from a graph G corresponding to random walks of length at most one
t	The number of trials for testing an interaction between two particular proteins
s	The number of experimental observations (successes) that two proteins interact ($0 \leq s \leq t$)
H_{ij}	A binary random variable representing interaction between proteins i and j
O_{ij}	An observation value for a pair of proteins i and j , defined as (t, s)
ν	The false negative error rate for any given protein interaction
$\hat{\nu}$	The <i>estimated</i> false negative error rate for any given protein interaction
ϕ	The false positive error rate for any given protein interaction
$\hat{\phi}$	The <i>estimated</i> false positive error rate for any given protein interaction
ρ	the prior probability for any given protein interaction
$\hat{\rho}$	the <i>estimated</i> prior probability for any given protein interaction
$\mathcal{L}(\nu, \phi, \rho t, s)$	The likelihood of observing a set of parameter values ν , ϕ and ρ , given values of t and s
\mathbf{Z}	A matrix whose entry $\mathbf{Z}[t, s]$ is the number of times a particular pair of (t, s) occurred in multiple experiments.
$\mathcal{L}(\nu, \phi, \rho \mathbf{Z})$	The likelihood of observing a set of parameter values ν , ϕ and ρ , given the count matrix \mathbf{Z}
N	Number of items or number of proteins in the universe \mathbb{P}

\mathbb{P}	A universe of items or a universe of proteins
T	A <i>transaction</i> which is a set of items, $T \subseteq \mathbb{P}$
D	A database of transactions
M	Number of transactions in the database D
I, X, Y	An exact <i>itemset</i> which is always a subset of T , $\exists T \in D$
$\text{support}(X)$	The fraction of the transactions in D that support the itemset X
σ	A user-defined <i>minimum support threshold</i> , $\sigma \in [0, 1]$
\hat{D}	An input database with false negative and false positive rate
$E[I]$	The expectation of an itemset I given the input \hat{D}
γ	Annealing factor
Q_i	a discrete random variable associated with a protein i indicating a cluster assignment ranging from $\{1, \dots, K\}$
Λ	The negative log-likelihood of Markov Random Fields
$U(Q)$	A potential function of Markov Random Fields
C_{ik}	a cost for a protein i assigned to a cluster k .
q_{ik}	a probability of a protein i in a cluster k . It is a function of C_{ik} .
$B(i)$	An identifier for a bait protein of the i th purification, $B(i) \in \{1, \dots, N\}$
H_i	An N -vector of binary random variable indicating the <i>true</i> i th purification with the bait $B(i)$
\hat{O}_{ik}	An observation value (t, s) for a protein interaction between the bait protein $B(i)$ and a prey protein k

Table of Contents

1	Introduction	11
1.1	Large-scale experimental methods	12
1.1.1	Yeast two-hybrid arrays	12
1.1.2	Tandem Affinity Purifications-Mass Spectrometry (TAP-MS)	13
1.1.3	Available data sets	14
1.1.4	Clustering solutions	15
1.1.5	Protein complex annotation	15
1.2	Overview of the thesis	16
2	Previous work	17
2.1	Protein-protein interaction graph	18
2.2	Molecular complex detection algorithm (MCODE)	19
2.3	Markov clustering	20
2.4	A statistical model of protein interaction	21
2.4.1	A hypothetical dataset	22
2.4.2	A Bayesian model for interaction probability	23
2.4.3	Estimating model parameters	26
2.5	Prediction of protein function using protein-protein interaction graph	27
3	Exact frequent itemset model	29
3.1	Problem setting	30
3.1.1	Maximal frequent itemsets (MFI)	31
3.2	Monotonicity property	32
3.3	The Apriori algorithm	33
3.4	Integer linear programming formulation	34
3.5	Significance of exact frequent itemsets	35
3.6	Result and discussion	36
4	Frequent itemsets with errors	43
4.1	Problem setting	44
4.2	Combinatorial error	44

4.2.1	Algorithm for finding FTIs	45
4.2.2	Algorithm for finding ETIs	47
4.3	Probabilistic error	51
4.3.1	Fragmentation of patterns by noise	52
4.3.2	Algorithm for probabilistic frequent itemsets	53
4.4	Merging maximal frequent itemsets	56
4.5	Result and discussion	57
5	Markov Random Fields	65
5.1	Method	66
5.1.1	The likelihood	68
5.1.2	Mean Field Annealing	70
5.1.3	Estimation of false negative and false positive rate	72
5.1.4	Minimization strategy	73
5.2	Performance measures	74
5.3	Result	77
5.3.1	Quality of clusters	78
5.3.2	Comparison with clustering algorithms for protein-protein interaction networks	78
5.3.3	Comparison with maximal frequent itemsets (MFI)	84
5.4	Discussion	86
6	Visualization of purifications	93
6.1	Systems and Methods	95
6.1.1	Directed-graph of purifications	95
6.1.2	Method outline	96
6.1.3	A model of subset inclusion for purifications	96
6.1.4	Decomposition of the SCC graph	98
6.2	Implementation	99
6.3	Results	100
6.3.1	Examples	103
6.4	Conclusion	103
7	Conclusion	107
	Bibliography	109
A	Zusammenfassung	115
B	Software for MRF	117
B.1	Usage	117
B.2	Input file format	118

B.3 Output file format 118

