Chapter 2

Previous work

Predicting molecular complexes from experimental data is important because it provides a global view on how proteins work together in the cell. Furthermore, for many proteins whose functions are unknown, the knowledge of their protein complex formation can be used to help predict their functions.

Yet, prediction of protein complexes from individual, unprocessed protein purifications remains a challenge. The methods being used can be roughly classified into pairwise and more-than-pairwise techniques. Pairwise techniques model the affinity between proteins on a pairwise basis, and allow clusters of more than two proteins only as a consequence of pairwise interactions. For example, if proteins A, B and C form a cluster, this is caused by pairwise affinity between A and B, B and C or C and A. The simple assumption of pairwise interaction greatly reduces the number of variables in the model and makes estimating them from a realistic number of experiments computationally tractable.

However, Hollunder et al. [21] pointed out that in nature, protein interactions often involve more than a pair of proteins. For example, a protein cluster $\{A, B, C\}$ could require the presence of all three proteins A, B and C for the complex to form, but none of the pairs of proteins would lead to a cluster. These phenomena lead to interesting cluster relationships, such as overlapping clusters with shared modules and hierarchies between clusters, but cannot be captured by pairwise models. We were thus motivated to try frequent itemsets as a non-pairwise technique (see Chapter 3).

In this chapter, we examine previous works that try to model the interaction graph from experimental data and subsequently predict protein complexes using graph-based clustering methods. The

methods described in this chapter all assume pairwise interactions, leading to a partitioning model of protein complexes where one protein is assigned to one group.

2.1 Protein-protein interaction graph

A protein-protein interaction network is an undirected graph $G = (V, E)$ where V is a set of nodes representing proteins and E is a set of edges. Edges are interpreted, depending on the particular model, as physical interaction or protein complex co-membership and may be weighted to designate interaction probability. Some methods take a protein-protein interaction graph curated by experts [4] as input. Others produce an interaction graph as an intermediate result [24, 26].

For estimating protein-protein interaction graphs, several protein-protein interaction databases are available, in particular for the yeast proteome. They include data based on the yeast two-hybrid system [22, 45] and the TAP-MS analysis of protein complexes [14, 15, 26]. However, creating a protein interaction network from high-throughput experiments is difficult due to high error rates. Therefore, with present techniques, the resulting networks are often inaccurate [10]. Current approaches merge the results of different types of experiments such as two-hybrid systems, co-immunoprecipitation and TAP-MS. Two-hybrid results are inherently pairwise, whereas results from other experiments are sets of one or more proteins.

Biochemical purifications can be modeled as observations of protein complexes caused by some underlying pairwise protein interaction topology that is not directly observable. In the general case of the purifications used by Gavin et al. [14, 15] and Krogan et al. [26], one affinity tagged protein is used as a bait to pull associated proteins out of a yeast cell lysate. The two extreme cases for the topology underlying the population of complexes from a single purification experiment are a minimally connected *spoke* model, where the data is modeled as pairwise interactions only between bait and preys, and a maximally connected *matrix* model, where the data is modeled as all proteins connected to all others in the set [4]. The real topology of the set of proteins must lie somewhere between these two extremes. Both have been previously used, for example, Gilchrist et al. [17] uses the spoke model.

For method validation, the most widely accepted protein-protein interaction graph can be obtained from physical interaction data provided by the Munich Information Center for Protein Se-

quences (MIPS) [32]. The MIPS data includes interactions collected from small-scale experiments and the core data of Ito et al. [22]. The data is regarded as highly reliable [11]. Other high-quality databases of protein-protein interactions are the Biomolecular Interaction Network Database (BIND) [2]. MIPS is hosted at `http://mips.gsf.de` and BIND at `http://www.bind.ca`.

In the following, we will describe two specific computational methods from Bader and Hogue [4] and Krogan et al. [26] designed to obtain a set of protein-protein interactions from experimental data and to predict protein complexes given such pairwise protein-protein interactions.

2.2 Molecular complex detection algorithm (MCODE)

The MCODE algorithm [4] transforms each individual purification into an interaction graph using the spoke model of interactions. The MCODE algorithm uses a vertex-weighting scheme based on a measure called the graph density which measures the total connectivity of a given graph [4]. The graph density D_G of a graph, $G = (V, E)$, with number of vertices $|V|$ and number of edges $|E|$ is defined as

$$D_G = \frac{|E|}{|E|_{max}},$$

where $|E|_{max} = \frac{|V|(|V|-1)}{2}$.

The first stage of MCODE, called vertex weighting, weights all vertices based on their local network density using the highest k-core subgraphs. A k-core is a subgraph G of minimal degree k, for all v in G, degree(v) $\geq k$. The highest k-core of a graph is the most densely connected k-core. The weight of a vertex v is defined to be the density of the highest k-core of the immediate neighbors of v including v. We summarize the vertex weighting procedure in Algorithm 1. The density of the k-core subgraph is used because it amplifies the weighting of highly interconnected graph regions while removing the many less connected vertices that are normally part of a protein interaction network. A given highly connected vertex v in a dense region of a graph may be connected to many vertices of degree one (single linked vertices). These low degree vertices do not interconnect within the neighborhood of v and thus would reduce the clustering coefficient of v, but not the core-clustering coefficient.

The second stage, called complex prediction, takes as input the vertex weighted graph and a parameter for a vertex weight threshold, seeds a complex with the highest weighted vertex v and

computes the connected components of v where all vertices have weights above a given threshold. A vertex is not visited more than once and therefore complexes cannot overlap. The algorithm is repeated for the next highest unseen weighted vertex remaining in the network. In this way, the highly connected subgraphs of the network are identified. The vertex weight threshold parameter controls the density of the resulting complexes. Resulting complexes from the algorithm are scored and ranked. Let $C = (V, E)$ be a complex subgraph. The complex score is defined as $D_C|V|$. This ranks larger and denser complexes higher in the results. Other scoring schemes are possible, but are not evaluated by MCODE.

The algorithm is slower than the fastest min-cut graph clustering algorithm, a popular method to find highly connected subgraphs, at $O(N^2 \log N)$ [18], but MCODE is easy to implement and visualize because it is based on local density. The view of local density is useful for examining protein interaction graph. The implementation of MCODE is available from `http://cbio.mskcc.org/~bader/software/mcode/index.html`.

Algorithm 1: MCODE-VERTEX-WEIGHTING

Input : A protein-protein interaction graph $G = (V, E)$
Output : Vertex weights W for all vertices in G
forall $v \in V$ **do**
  $N = \{w : (v, w) \in E\}$;
  K = the highest k-core graph among vertices in N;
  k = the minimal degree of the highest k-core K;
  d = the density of K;
  $W(v) = k \times d$;

2.3 Markov clustering

Krogan et al. [26] created a high quality data set of protein-protein interactions from experiments using tandem affinity purification. They processed $4,562$ different tagged proteins of yeast *Saccharomyces cerevisiae* and used mass spectrometry to identify purified proteins. The main difference from MCODE is that it learns a probabilistic model of protein interactions from the data. Machine learning is used to integrate the mass spectrometry scores and assign probabilities to the protein-protein interactions. The final result is the core data set comprising of $7,123$ protein-protein interactions involving $2,708$ proteins. A protein interaction graph is created from this data set and

Markov clustering is used to cluster these proteins into 547 disjoint complexes.

The key intuition behind Markov clustering (MCL) is that a random walk that visits a dense cluster will likely not leave the cluster until many of its vertices have been visited. Rather than simulating random walks, MCL iteratively modifies a matrix of transition probabilities. Given a weighted undirected graph $G = (V, E)$, starting from $M = M(G)$, a transition matrix derived from the graph corresponding to random walks of length at most one, the following two operations are iteratively applied:

1. *expansion*, in which $M = M^e$, where $e \in \mathbb{N}_{>1}$ thus simulating e steps of a random walk with the current transition matrix

2. *inflation*, in which M is re-normalized after taking every entry to its rth power, $r \in \mathbb{R}^+$.

Note that for $r > 1$, inflation emphasizes the heterogeneity of probabilities within a row, while for $r < 1$, homogeneity is emphasized. r also controls the number of clusters at each step and increasing heterogeneity results in a higher number of clusters. The algorithm converges with probability ~ 1 and the calculation stops upon reaching a recurrent state of a fix-point [46]. A recurrent state of period $k \in \mathbb{N}$ is a matrix that is invariant under k expansions and inflations, and a fix-point is a recurrent state of period 1. The clustering is induced by connected components of the graph underlying the final transition matrix. Pseudo-code for MCL is given in Algorithm 2. The implementation of MCL is available at `http://mican.org/mcl`.

2.4 A statistical model of protein interaction

Both techniques described above are heuristic techniques that, while yielding useful results, make no explicit statements regarding their optimality with respect to any criteria. Gilchrist et al. [17] have developed the only approach so far that introduces a probabilistic method based on maximum likelihood estimation. It assumes an indirectly observable model of protein interaction with random experimental errors. We will describe it in details here because the observation model motivated our approach based on Markov Random Fields described in Chapter 5.

Tandem affinity purifications use *bait* proteins to purify protein complexes. Other proteins detected in a purified complex are called *prey* proteins. Because prey proteins are thought to interact

Algorithm 2: Markov Clustering

Input : A weighted undirected graph $G = (V, E)$, expansion parameter e, inflation parameter r

Output : A partitioning of V into disjoint components

$M \leftarrow M(G)$

while M *is not fixpoint* **do**

  $M \leftarrow M^e$

  forall $i \in V$ **do**

    forall $j \in V$ **do**

      $M[i][j] \leftarrow M[i][j]^r$

    forall $j \in V$ **do**

      $M[i][j] \leftarrow \frac{M[i][j]}{\sum_{k \in V} M[i][k]}$

$H \leftarrow$ graph induced by non-zero entries of M

$C \leftarrow$ clustering induced by connected components of H

with the bait within a protein complex, we define a *true* interaction as occurring between proteins which are members of the same complex. The systematic and random errors generated by any experimental method fall into two categories, false negative and false positive errors. False negative errors occur when an experiment fails to identify members of a protein complex. False positive errors occur when an experiment identifies additional proteins that are not part of a complex.

The statistical model proposed by Gilchrist et al. [17] takes random errors into account based on a mechanistic description of how the data in a single experiment is generated. The model allows us to estimate the false negative and false positive error rates of a data set without the use of a manual reference set of protein complexes, and produces a complete undirected graph of pairwise protein interactions, weighted with the respective likelihoods of interaction. In contrast to the techniques described in sections 2.2 and 2.3, it does not calculate a clustering.

2.4.1 A hypothetical dataset

From a sampling perspective, each experiment given a certain bait protein provides a "trial" to gather information on which proteins interact with the bait protein. For illustration, we use the example given in Gilchrist et al. [17] for a scenario involving four proteins v, w, x, y (Fig. 2.1). When we use v as a bait protein, we can view this experiment as a trial to observe three interactions between v and the proteins w, x, y. In repeating this experiment, we would have a second trial to observe

these three interactions. A third experiment, now using protein w as a bait, provides a third trial to observe an interaction between v and w, as well as the first trial to observe an interaction between w and proteins x or y. At the end of these three experiments, we have had three trials for observing an interaction between v and w, two trials for observing an interaction between v and x and no trials for observing an interaction between x and y. We define t as the number of trials for observing an interaction between two particular proteins. For example, from these three experiments, t is equal to 3, 2, 1 and 0 for the protein pairs (v, w), (v, x), (w, x) and (x, y), respectively.

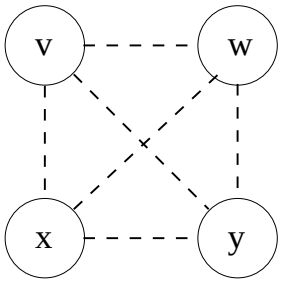
However, in each trial we may or may not observe an interaction. Consequently, we define s (success) as the number of experimental observations that two proteins interact $(0 \leq s \leq t)$. In Figure 2.1(c), we illustrate how the experimental results from the three experiments can be summarized as a set of observation values (t, s) for each possible pair of proteins. The transformation of experimental data into observations (t, s) forms the basis of the statistical model.

2.4.2 A Bayesian model for interaction probability

As we have just shown, we can represent experimental information on any protein-protein interaction by the number of experimental trials (t) and successes (s) in a data set. We will next define the statistical model to interpret this information in a quantitative manner.

We begin by defining a binary random variable H_{ij} for each interaction of a pair of proteins i and j. $H_{ij} = 1$ if two proteins i and j interact by occurring within the same protein complex. $H_{ij} = 0$ if i and j do not occur within the same complex. Because the two events are complementary, we can compute $\mathbf{P}[H_{ij} = 0] = 1 - \mathbf{P}[H_{ij} = 1]$. The statistical model will calculate the posterior probability based on the likelihood of observing s successes and the prior of interaction.

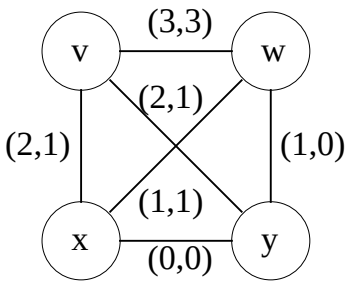
In order to calculate the probability of observing s successes given H_{ij}, we need to define two terms: the false negative error rate ν and the false positive error rate ϕ. Each of these rates is specific to a particular experimental technique and represents the random errors associated with such a technique. The model will ignore systematic errors. The false negative error rate ν is equal to the probability that, for any given trial, we do not observe an interaction between two proteins that occur within the same complex. Conversely, the false positive error rate ϕ is equal to the probability that, for any given trial, we observe an interaction between two proteins that do not occur together within the same complex. Using our example, if proteins v, w and x interact with one another to



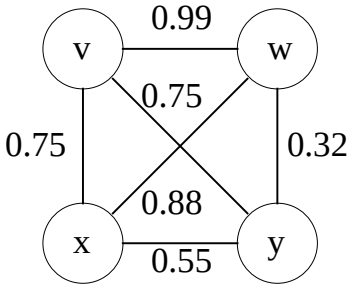
(a) Initial information

Trial	Bait	Prey
I	v	w,y
II	v	w,x
III	w	v,x

(b) Experimental trials



(c) Observation



(d) Data interpretation

Figure 2.1: Illustration of the statistical approach with a hypothetical dataset. (a) Two proteins are connected by an edge if they are part of the same complex. The panel illustrates the lack of confidence in any such association. (b) The results from three experimental trials in which v was used twice as a bait protein and protein w was used once. (c) Observation of the experimental data from the trials in (b) through their (t, s) values. (d) The posterior probability of each protein-protein association based on the data in (b) using a hypothetical false negative error rate, $\nu = 0.35$, false positive error, $\phi = 0.1$, and the prior probability of an association, $\rho = 0.55$.

form a single complex, then the first experiment with v has one false negative interaction because x is not observed with v. This experiment also has one false positive interaction because of protein y. The second and third experiments had no false positive or false negative observations.

If we assume that the random experimental errors are independent of one another, the probability of observing s successes out of t trials follows a binomial distribution. Given $H_{ij} = 1$, the two proteins i and j occur within the same complex. The probability of successfully observing i and j is $1 - \nu$. Define $O_{ij} = (t, s)$ to be an observation value of an interaction between proteins i and j. Thus, given $H_{ij} = 1$, the probability of observing an interaction s times out of t trials for proteins i and j is

$$\begin{aligned} \mathbf{P}[O_{ij} = (t,s)|H_{ij} = 1] = \mathbf{P}[O_{ij} = (t,s)|H_{ij} = 1, \nu] &= \mathbf{P}[s|H_{ij} = 1, t, \nu] \\ &= \binom{t}{s} \nu^{t-s}(1-\nu)^s. \end{aligned} \tag{2.1}$$

In contrast, given the complementary condition $H_{ij} = 0$, the probability that we will observe a false association given the false positive error rate ϕ is

$$\begin{aligned} \mathbf{P}[O_{ij} = (t,s)|H_{ij} = 0] = \mathbf{P}[O_{ij} = (t,s)|H_{ij} = 0, \phi] &= \mathbf{P}[s|H_{ij} = 0, t, \phi] \\ &= \binom{t}{s} (1-\phi)^{t-s}\phi^s. \end{aligned} \tag{2.2}$$

It follows from Bayes' Theorem that

$$\begin{aligned} \mathbf{P}[H_{ij} = 1|O_{ij} = (t,s)] &= \frac{\mathbf{P}[O_{ij} = (t,s)|H_{ij} = 1]\rho}{\mathbf{P}[O_{ij} = (t,s)]} \\ &= \frac{\mathbf{P}[O_{ij} = (t,s)|H_{ij} = 1]\rho}{\mathbf{P}[O_{ij} = (t,s)|H_{ij} = 1]\rho + \mathbf{P}[O_{ij} = (t,s)|H_{ij} = 0](1-\rho)}, \end{aligned} \tag{2.3}$$

where the likelihood term $\mathbf{P}[O_{ij}|H_{ij}]$ is defined by Equations 2.1 and 2.2. The term ρ defines the prior probability for $H_{ij} = 1$; ρ is equal to the probability that two proteins selected at random are found within the same protein complex. In the absence of any experimental data when $t = 0$, the right-hand side of Equation 2.3 simplifies to the prior probability ρ.

Figure 2.1 illustrates the application of the model to the hypothetical example. Given a false positive error rate, false negative error rate and a prior probability, we compute from Equations 2.1

and 2.2 the likelihood of observing s successes out of t trials for each pairwise interaction (Figure 2.1(c)) and from Equation 2.3 the posterior probability. As shown in Figure 2.1(d), we can represent all the possible protein-protein interactions using a complete weighted undirected graph whose nodes correspond to proteins and whose edges have weights that correspond to the posterior probability that two proteins are part of the same complex.

For applications to real data, it is important to estimate the three parameters of the model: ν, ϕ and ρ. We will show next how they can be estimated from the data by solving the maximum likelihood problem.

2.4.3 Estimating model parameters

We begin our estimation by defining the likelihood $\mathcal{L}$ of observing a set of parameter values ν, ϕ and ρ. It follows from Equations 2.1- 2.3 that

$$\mathcal{L}(\nu, \phi, \rho | t, s) = (1-\nu)^s \nu^{t-s} \rho + \phi^s (1-\phi)^{t-s} (1-\rho). \tag{2.4}$$

Any one observed value (t, s) does not contain much information on the parameters, ν, ϕ and ρ. However, because high-throughput datasets contain many experimental trials (t), we have enough information to compute $\mathcal{L}$ over the distribution of values (t, s). We define a matrix $\mathbf{Z}$ whose entry $\mathbf{Z}[t, s]$ is the number of times a particular pair of (t, s) occurred in a high-throughput experiment. Assuming independence between interactions, the total likelihood $\mathcal{L}$ of a set of parameter values ν, ϕ and ρ given the matrix $\mathbf{Z}$ can be written as

$$\mathcal{L}(\nu, \phi, \rho | \mathbf{Z}) = \prod_{t=1}^{t_{max}} \prod_{s=0}^{t} [(1-\nu)^s \nu^{t-s} \rho + \phi^s (1-\phi)^{t-s} (1-\rho)]^{\mathbf{Z}[t,s]}, \tag{2.5}$$

where t_{max} is the maximum number of times any one interaction has been used in the high-throughput experiment. By finding the parameter values that maximize $\mathcal{L}$ of Equation 2.5 for a given $\mathbf{Z}$ matrix, we can obtain maximum likelihood estimates for ν, ϕ and ρ. Gilchrist et al. [17] provide more details on how to estimate these parameters from multiple high-throughput experiments and the accompanying web-site `http://www.tiem.utk.edu/~mikeg/software.html` provides a stand-alone software program to estimate these parameters given a $\mathbf{Z}$ matrix. An example

Dataset	$\hat{\nu}$	$\hat{\phi}$	$\hat{\rho}$
Gavin02	0.346	1.07×10^{-3}	1.88×10^{-3}
Ho02	0.539	1.30×10^{-3}	1.88×10^{-3}
Gavin06	0.407	1.35×10^{-3}	3.89×10^{-3}

Table 2.1: Maximum-likelihood estimates of false negative error rate, ν, false positive error rate, ϕ and global prior ρ for the **Gavin02** (Gavin et al. [15]), **Ho02** (Ho et al. [20]) and **Gavin06** (Gavin et al. [14]) datasets.

of the maximum likelihood estimates for different datasets is shown in Table 2.1.

As discussed in the beginning of this section, Gilchrist et al. [17] unfortunately stop short of calculating a clustering with the same rigorous probabilistic techniques they used to calculate interaction probabilities. Chapter 5 fills this gap and shows that this more rigorous approach leads to better results than previous techniques.

2.5 Prediction of protein function using protein-protein interaction graph

The task of assigning protein functions to novel proteins is closely related to the task of finding protein complexes. Several approaches have been applied to this problem, including the analysis of gene expression patterns, phylogenetic profiles and protein-protein interactions. Assuming that proteins interact with one another to achieve a particular function, we will summarize methods based on protein-protein interaction data.

It should be noted that the interaction partners of a protein may belong to different functional categories. This complex network of within-function and cross-function interactions makes the problem of functional assignment a difficult task. Methods based on chi-square statistics [19] and on frequencies of interaction partners having certain functions of interest [13, 38] have been used to assign functions to unannotated proteins. However, these methods lack a systematic mathematical model. Deng et al. [11] propose a superior probabilistic method that uses Markov Random Fields to model a protein-protein interaction graph and applies Bayesian analysis to assign functions to proteins. It is limited in so far as it only assigns known functional categories to proteins. Although this probabilistic approach also uses Markov Random Fields, our approach differs in that we use

Markov Random Fields to model protein complexes, not a protein-protein interaction graph. Our prediction is an assignment of proteins to complexes using maximum likelihood analysis.

In summary, it should be noted that all of the methods described above assume a curated protein-protein interaction graph constructed from the MIPS database of protein complexes. Their prediction performance depends on the quality of the input interaction graph, because they do not model observational errors. As a result, they cannot be used to predict functions based on the experimental data. Furthermore, they cannot be used to discover novel functional categories.