Chapter 5

Markov Random Fields

The frequent itemset techniques described in the previous two chapters use the most general model of clusters. Different clusters of proteins are considered to be entirely independent of each other. The algorithm makes no assumption that two (or more) proteins appearing together in one cluster are any more or less likely to appear together again in any other cluster. This weak, general model of frequent itemsets has the potential of uncovering complex relationships between proteins. But since every protein can be assigned to not only one, but any number of different clusters, the number of variables that must be estimated is very large, requiring a correspondingly large set of data for estimation. Thus, the disappointing performance of the frequent itemset methods described in the previous two chapters is likely a consequence of insufficient input data, in particular considering the relatively high error rate of purification experiments.

Protein clustering by partitioning such as described in Chapter 2 represents the other extreme of possible models. There, any pair of proteins is assumed to either interact or not, entirely independent of the context of other proteins in which it appears. As a consequence, clusters never overlap, proteins are assigned only to a single cluster and the number variables to estimate is much smaller.

Of course, reality lies somewhere in between. Pairs of proteins that interact with each other in one context are likely to do so again in another context, but there are situations where other proteins interfere, and cause the behavior of proteins to change. Still, before setting out to design such a model, we should explore how far the simpler model of non-overlapping clusters can take us. This is the purpose of this chapter.

Our work was inspired by the probabilistic approach of Gilchrist et al. [17] that makes maximum-likelihood estimates of false negative error rate, false positive error rate and prior probability of interaction (see Chapter 2 for details). It may seem natural to use these estimated parameters to first construct the most likely interaction graph, and then derive protein clusters from this graph. Unfortunately, in this process we lose the maximum likelihood optimality criterion. The uncertainty contained in the observation is no longer represented in the interaction graph, and cannot be properly accounted for when computing the clustering. All approaches that use an unweighted (e.g., thresholded) interaction graph as intermediate step suffer from this problem. We can reasonably expect better results by deriving the clustering directly from the observation, taking into account observational error. We will use Markov Random Fields to model protein complexes while accounting for observational error.

Markov Random Fields have been successfully applied as a probabilistic model in many research areas. In image processing, they were applied as a model for image segmentation [29]. In bioinformatics, MRFs were used to model protein-protein interaction networks to predict protein functions [11] and to discover molecular pathways [40] by combining the MRF model of protein-interaction graphs with gene expression data. Our model differs from these previous works because we use MRF to model protein complexes without assuming an intermediate interaction graph and we model the observation error that previous work did not account for.

Following Gilchrist et al. [17], we consider each purification experiment to be a statistically independent set of observations about the interaction or non-interaction of proteins. We model the observational error as a false positive and false negative error rate and assume it to be the same for all purifications. We try both the spoke model by considering only the interactions between bait and each prey protein, and the matrix model, considering the interactions between bait and preys as well as between pairs of preys as interactions (Figure 5.1).

5.1 Method

We assume that clusters do not overlap and each protein only belongs to one cluster. Each protein i is assigned to a single cluster $Q_i \in \{1, \ldots, K\}$, where K is the number of clusters. We expect proteins in the same cluster to interact, and proteins belonging to different clusters not to interact. Our observation contains errors, with a false negative error rate ν that proteins of the same cluster

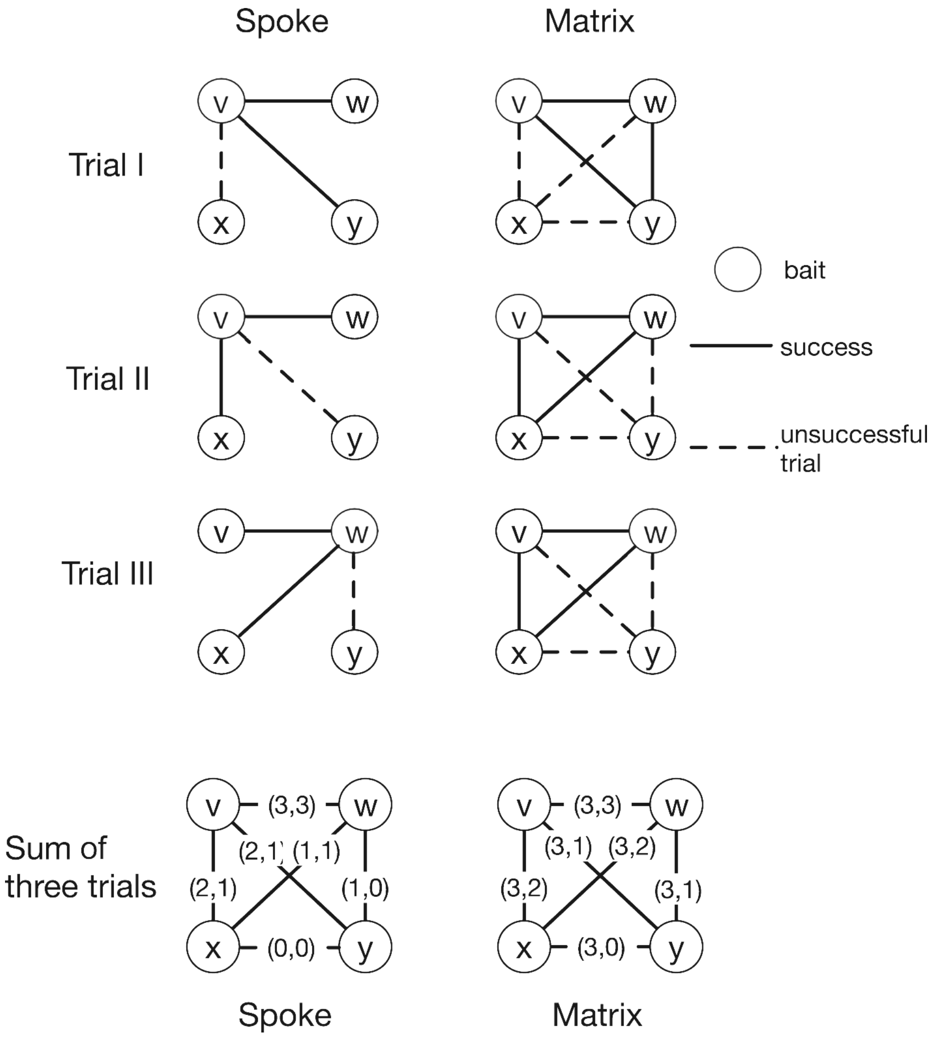


Figure 5.1: Observation model of protein interaction.

are observed to not interact, and a false positive error rate ϕ, that proteins belonging to different clusters are observed to interact. These error rates are assumed to be the same for all interactions. We estimate them along with the cluster assignments of proteins.

If we consider Q_i to be a random variable for the cluster assignment of protein i, the entire cluster assignment is a Markov Random Field because (1) $\mathbf{P}[Q_i = k] > 0$ and (2) its conditional distribution satisfies the Markov property,

$$\mathbf{P}[Q_i|Q_1, \ldots, Q_{i-1}, Q_{i+1}, \ldots, Q_N] = \mathbf{P}[Q_i|Q_j, j \in Neighbor(i)].$$

In other words, the joint probability $\mathbf{P}[Q]$ and the likelihood function only depend on the values of pairs of random variables Q_i and Q_j. In the terminology of Markov Random Fields as a statistical model [11, 29], each protein i is a site that is labeled with the identity of its cluster Q_i. The neighborhood of each site i is all those proteins j for which we have any observation for the protein pair (i, j), either interaction or non-interaction. To compute the cluster assignment Q using a Markov Random Field, we must define the potential function $U(Q)$ which in this setting will be derived from the negative logarithm of the likelihood.

5.1.1 The likelihood

Define S_{ij} to be the event that proteins i and j are observed to interact, and, likewise, F_{ij} the event that they are observed not to interact. The probabilities of these two events, given ν, ϕ and Q, are

$$\mathbf{P}[S_{ij}|\nu, \phi, Q] = \begin{cases} (1-\nu) & : \quad Q_i = Q_j \\ \phi & : \quad Q_i \neq Q_j, \end{cases}$$

and

$$\mathbf{P}[F_{ij}|\nu, \phi, Q] = \begin{cases} \nu & : \quad Q_i = Q_j \\ (1-\phi) & : \quad Q_i \neq Q_j. \end{cases}$$

A single purification experiment generates a whole set of such observations. Under the spoke model, an observation of interaction is made for each pair of bait and all its preys. An observation of non-interaction is made for each pair of bait and all other proteins not being prey. Similarly, under the matrix model, there is an observed interaction between all pairs of proteins purified together, both pairs of bait and preys, and pairs of preys. Non-interaction is observed between the set of bait and preys, paired with all non-purified proteins. For illustration, see Figure 5.1.

Over the course of multiple purification experiments, each pair of proteins may be observed multiple times. To summarize all these observations, using the same definition of trials and successes as introduced in Section 2.4, we define t_{ij} to be the total number of observations made for the protein pair (i, j), and s_{ij} to be the number of these observations where an interaction was observed.

Then, given ν, ϕ and a configuration Q, the likelihood of observing a particular sequence of experimental outcomes (t_{ij}, s_{ij}) for all pairs (i, j) is,

$$\begin{aligned} \mathbf{P}[\{(t_{ij}, s_{ij})\}|\nu, \phi, Q] &= \prod_{(i,j)} \mathbf{P}[S_{ij}|\nu, \phi]^{s_{ij}} \mathbf{P}[F_{ij}|\nu, \phi]^{t_{ij}-s_{ij}} \\ &= \prod_{(i,j):Q_i=Q_j} (1-\nu)^{s_{ij}} \nu^{(t_{ij}-s_{ij})} \\ &\quad \times \prod_{(i,j):Q_i \neq Q_j} \phi^{s_{ij}} (1-\phi)^{(t_{ij}-s_{ij})}. \end{aligned} \tag{5.1}$$

The negative logarithm Λ of the above term is,

$$\begin{aligned} \Lambda &= \sum_{(i,j):Q_i=Q_j} [s_{ij}(-\ln(1-\nu)) + (t_{ij}-s_{ij})(-\ln(\nu))] \\ &\quad + \sum_{(i,j):Q_i \neq Q_j} [s_{ij}(-\ln(\phi)) + (t_{ij}-s_{ij})(-\ln(1-\phi))] \end{aligned} \tag{5.2}$$

We then separate Λ into terms that depend on Q and terms that do not depend on Q. Λ can then be written as

$$\Lambda = \sum_{(i,j):Q_i \neq Q_j} s_{ij}\beta + \sum_{(i,j):Q_i=Q_j} (t_{ij}-s_{ij})\alpha + C, \tag{5.3}$$

where

$$\alpha = -\ln(\nu) + \ln(1-\phi),$$

$$\beta = -\ln(\phi) + \ln(1-\nu),$$

and

$$C = \sum_{(i,j)} (-s_{ij} \ln(1-\nu)) + (-(t_{ij} - s_{ij}) \ln(1-\phi)).$$

C does not depend on Q and is thus irrelevant for minimization with respect to Q. The minimum is also unaffected by changes in α and β as long as the *ratio* between α and β stays unchanged. Incorporating these observations leads to

$$U(Q) = \sum_{(i,j):Q_i = Q_j} (t_{ij} - s_{ij}) + \sum_{(i,j):Q_i \neq Q_j} \psi s_{ij}, \tag{5.4}$$

where

$$\psi = \frac{-\ln(\phi) + \ln(1-\nu)}{-\ln(\nu) + \ln(1-\phi)}.$$

It is noteworthy that this cost function is the same for certain pairs of ϕ and ν that are related by a common ψ. Minimization with respect to Q_i, ν and ϕ yields our desired solution.

5.1.2 Mean Field Annealing

Mean Field Annealing is a popular technique to compute a maximum-likelihood label assignment for Markov Random Fields. We will replace the random variables Q_i with explicit probabilities

$$q_{ik} = \mathbf{P}[Q_i = k].$$

As stated before, the configuration of cluster assignments Q is an MRF because (1) $\mathbf{P}[Q_i = k] > 0$ and (2) its conditional distribution satisfies the Markov property. Following this assumption, it is well known (e.g., see [29]) that the joint probability distribution of Q is a Gibbs distribution, given by

$$\mathbf{P}[Q] = Z^{-1} \exp[-\gamma U(Q)]$$

where $U(Q)$ is the energy function (Eq. 5.4) and γ is the annealing factor. Z is the normalization factor, also called the partition function, with

$$Z = \sum_{Q} \exp[-\gamma U(Q)].$$

Mean Field Theory provides a framework to compute **P**$[Q]$. For our clustering problem, we will apply it to estimate the probability $\hat{q}_{ik}$ of assigning protein i to a cluster k, defined by

$$\begin{aligned}\hat{q}_{ik} &= \frac{\mathbf{P}[Q_i = k|Q_j, j \neq i]}{\sum_{l=1}^{K} \mathbf{P}[Q_i = l|Q_j, j \neq i]} \\ &= \frac{\exp[-\gamma U(Q_i = k|Q_j, j \neq i)]}{\sum_{l=1}^{K} \exp[-\gamma U[Q_i = l|Q_j, j \neq i)]}. \end{aligned} \tag{5.5}$$

Computing the actual energy function is infeasible because it requires us to evaluate the clustering assignment of the whole MRF which is not known. By assuming the Markovian property and replacing the random variables Q_i and Q_j with the expected values of cluster assignments within each protein's neighborhood, we can estimate $U(Q_i = k|Q_j, j \neq i)$ by

$$\begin{aligned}U(Q_i = k|Q_j, j \neq i) &= U(Q_i = k|Q_j, j \in Neighbor(i)) \\ &= \sum_{j \in Neighbor(i)} (\sum_{l=1}^{K} q_{il} q_{jl})(t_{ij} - s_{ij}) + (1 - \sum_{l=1}^{K} q_{il} q_{jl}) \psi s_{ij}.\end{aligned}$$

We evaluate the conditional energy function at a fixed point by assuming that $q_{ik} = 1$ and $q_{il} = 0$ for $l \neq k$. We then approximate $U(Q_i = k|Q_j, j \neq i)$ by

$$C_{ik} = \sum_{j \in Neighbor(i)} q_{jk}(t_{ij} - s_{ij}) + (1 - q_{jk}) \psi s_{ij}. \tag{5.6}$$

Thus, the assignment probability q_{ik} can be computed by

$$\hat{q}_{ik} = \frac{\exp[-\gamma C_{ik}]}{\sum_{l=1}^{K} \exp[-\gamma C_{il}]}. \tag{5.7}$$

In terms of computation, notice that in order to find the mean field at i, we needs to know the mean field at the neighbors of i. Therefore, the mean field is usually computed by iterative procedures shown in details in Algorithm 11.

Algorithm 11: Mean Field Annealing

```
Input  : A graph $G = (V, E)$, with an observation $(t_{ij}, s_{ij})$ for each edge, $\psi$, a number
         of clusters $K$
Output : A probability $q_{ik}$ for a node $i$ belonging to a cluster $k$ for all $i$ and for all $k$
Initialize $q$ to random values;
Initialize annealing factor $\gamma$;
while $\gamma < \gamma_{\max}$ do
    repeat
        forall $i \in V$ do
            forall $k \in K$ do
                $C_{ik} = \sum_{j \in Neighbor(i)} q_{jk}(t_{ij} - s_{ij}) + (1 - q_{jk})\psi s_{ij}$
            forall $k \in K$ do
                $\hat{q}_{ik} = \frac{\exp(-\gamma C_{ik})}{\sum_{l=1}^{K} \exp(-\gamma C_{il})}$
            forall $k \in K$ do
                $q_{ik} = \hat{q}_{ik}$
    until $q$ converges;
    Increase $\gamma$;
```

5.1.3 Estimation of false negative and false positive rate

Given a cluster assignment Q, we can estimate the error rate ν and ϕ by minimizing equation Eq. 5.2 with respect to ν and ϕ. The derivative of Eq. 5.2 with respect to ν is

$$\frac{\partial \Lambda}{\partial \nu} = \frac{a}{1 - \nu} - \frac{b}{\nu}, \tag{5.8}$$

where

$$a = \sum_{(i,j):Q_i = Q_j} s_{ij},$$

and

$$b = \sum_{(i,j):Q_i = Q_j} (t_{ij} - s_{ij}).$$

The derivative of Eq. 5.2 with respect to ϕ is

$$\frac{\partial \Lambda}{\partial \phi} = -\frac{c}{\phi} + \frac{d}{1 - \phi}, \tag{5.9}$$

where

$$c = \sum_{(i,j):Q_i \neq Q_j} s_{ij},$$

and

$$d = \sum_{(i,j):Q_i \neq Q_j} (t_{ij} - s_{ij}).$$

Setting Eq. 5.8 and Eq. 5.9 to zero, the solutions for optimal error rates ν^* and ϕ^*, given the cluster assignments Q, are

$$\nu^* = \frac{\sum\limits_{(i,j):Q_i = Q_j} (t_{ij} - s_{ij})}{\sum\limits_{(i,j):Q_i = Q_j} t_{ij}},$$

and

$$\phi^* = \frac{\sum\limits_{(i,j):Q_i \neq Q_j} s_{ij}}{\sum\limits_{(i,j):Q_i \neq Q_j} t_{ij}}.$$

When evaluating the likelihood of a particular solution Q, we use ν^* and ϕ^* that maximizes the likelihood.

5.1.4 Minimization strategy

Each run of Mean Field Annealing requires two inputs, the number of clusters K and the error rate ratio ψ. We find values for both inputs that maximize the likelihood of solution Q by repeatedly optimizing Q using Mean Field Annealing for different values of K and ψ. Our tests show that on a large scale, the likelihood is roughly convex with respect to these two values, but unfortunately with smaller scale local minima interspersed. To avoid getting stuck in these local minima, we perform iterative line minimization, alternating between minimizing with respect to K and ψ, while holding the other constant. At each step, we computed five to seven values within a progressively smaller range. In our tests, three iterations were sufficient for converging upon the maximum likelihood (minimum negative log-likelihood). As shown in Algorithm 12, we can extend the Algorithm 11 to include the estimation of ν and ϕ by reestimating the error rates after we have reestimated the probability q using the annealing procedure. We implemented the Mean Field Annealing algorithm in C++ running on a single processor machine with a memory of at least 512 MB. The running time of Mean Field Annealing is quadratic in the number of nodes ($O(K|V|^2)$). On a data set of about

3000 proteins, a single minimization for a fixed number of clusters takes an average of 30 hours of CPU time on Athlon 1 Ghz processor.

Algorithm 12: MFA with error rate estimation

```
Input    : A graph $G = (V, E)$, with an observation $(t_{ij}, s_{ij})$ for each edge, a number of
           clusters $K$, an initial value for $\psi$
Output   : A probability $q_{ik}$ for a node $i$ belonging to a cluster $k$ for all $i$ and for all $k$, $\nu$, $\phi$
Initialize $q$ to random values;
Initialize annealing factor $\gamma$;
while $\gamma < \gamma_{\max}$ do
    repeat
        forall $i \in V$ do
            forall $k \in K$ do
                $C_{ik} = \sum_{j \in Neighbor(i)} Q_{jk}(t_{ij} - s_{ij}) + \psi(1 - Q_{jk})s_{ij}$
            forall $k \in K$ do
                $\hat{q}_{ik} = \frac{\exp(-\gamma C_{ik})}{\sum_{l=1}^{K} \exp(-\gamma C_{il})}$
            forall $k \in K$ do
                $q_{ik} = \hat{q}_{ik}$
    until $q$ converges;
    Increase $\gamma$;
$\nu = \frac{\sum_{(i,j)} (\sum_{l=1}^{K} q_{il} q_{jl})(t_{ij} - s_{ij})}{\sum_{(i,j) \in G} (\sum_{l=1}^{K} q_{il} q_{jl}) t_{ij}}$
$\phi = \frac{\sum_{(i,j)} (1 - \sum_{l=1}^{K} q_{il} q_{jl}) s_{ij}}{\sum_{(i,j) \in G} (1 - \sum_{l=1}^{K} q_{il} q_{jl}) t_{ij}}$
if $\phi + \nu = 1$ then
    $\psi = \frac{\nu}{1 - \nu}$
else
    $\psi = \frac{-\ln(\phi) + \ln(1 - \nu)}{-\ln(\nu) + \ln(1 - \phi)}$
Compute the likelihood $\Lambda$;
```

5.2 Performance measures

To extract relevant information from our clusters, we compare the result to the MIPS and Reguly benchmarks. To evaluate prediction accuracy, we consider both all pairs comparison and the accuracy measure [8] derived from a contingency table.

To summarize the evaluation procedure of Brohée, S and van Helden [8], we begin by building a contingency table to compare a clustering result with the annotated complexes. With n complexes and m clusters, the contingency table T is an $n \times m$ matrix whose entry T_{ij} indicates the number of proteins found in common between the ith complex and the jth cluster. Given a contingency table T, overall accuracy and separation value can be computed to measure the correspondence between clustering result and the annotated complexes [8].

However, we find that the *separation measure* produces undesirable effects when the reference data set contains overlapping complexes. By its definition [8], a good matching of a cluster to more than one complexes will result in a low separation value, contrary to expectation. In our case, this situation occurs because the MIPS and Reguly benchmark are overlapping, while the results of MCL and MRF are not. Furthermore, when we match the reference data set to itself, we find that its separation value can be less than some clustering solutions. For these reasons, we do not apply the separation measure. Thus, we only use the accuracy measure, which we now define.

Accuracy

A *complex-coverage* (denoted CO) is a quantity that characterizes the average coverage of complexes by a clustering result,

$$\mathrm{CO} = \frac{\sum\limits_{i=1}^{n} N_i (\max\limits_{j}\ \mathrm{CO}_{ij})}{\sum\limits_{i=1}^{n} N_i},$$

where $\mathrm{CO}_{ij} = T_{ij}/N_i$, N_i is the number of proteins in the complex i.

A *positive-predictive value* (denoted PPV) is the proportion of proteins in cluster j that belongs to complex i, relative to the total number of members of this cluster assigned to all complexes.

$$\mathrm{PPV}_{ij} = \frac{T_{ij}}{\sum\limits_{i=1}^{n} T_{ij}} = \frac{T_{ij}}{T_{\cdot j}}.$$

Note that the normalization is not the size of cluster j, but the marginal sum of a column j which can be different from the cluster size because some proteins belong to more than one cluster. To characterize the general positive-predictive value of a clustering result as a whole, we use the following

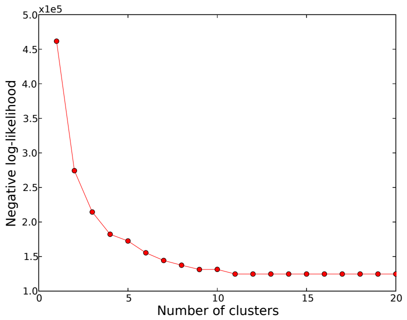


Figure 5.2: MRF on simulated data: given a graph of 500 nodes with 11 clusters randomly assigned, our MRF clustering can recover the true clustering with the minimum negative log-likelihood at 11 clusters. Notice that any more clusters do not affect the cost any further; they simply remain empty.

weighted average quantity,

$$\mathrm{PPV} = \frac{\sum_{j=1}^{m} T_{.j}(\max_i \ \mathrm{PPV}_{ij})}{\sum_{j=1}^{m} T_{.j}}.$$

The *accuracy value* is a geometric average between complex-coverage and positive-predictive value, $\mathrm{Acc} = \sqrt{\mathrm{CO} \times \mathrm{PPV}}$.

All-pair comparison: sensitivity and specificity

As a second metric, we use the standard all pair sensitivity (sens) and specificity (spec). We will refer to an (unordered) pair of proteins from the same complex as a *true* pair, and to a pair of proteins from the same cluster as a *predicted* pair. We will call a true predicted pair *true positive* (TP), a true pair which has not been predicted *false negative* (FN), a false pair predicted to be from the same complex *false positive* (FP) and a correctly predicted false pair *true negative* (TN). The following quantities summarize the performance of all-pair comparison: *Sensitivity*, $\mathrm{SN} = \frac{\#TP}{\#TP+\#FN}$ and *Specificity*, $\mathrm{SP} = \frac{\#TP}{\#TP+\#FP}$. A perfect clustering method would have $\mathrm{SN} = \mathrm{SP} = 1$, which implies that neither FP nor FN errors are made.

Dataset		K	$\hat{\nu}$	$\hat{\phi}$
Gavin02	Spoke model	393	0.423	1.3×10^{-3}
	Matrix model	310	0.752	1.7×10^{-3}
Gavin06	Spoke model	698	0.547	2.4×10^{-3}
	Matrix model	550	0.807	2.7×10^{-3}

Table 5.1: Maximum likelihood solution for the spoke model ($\psi = 3.5$) and the matrix model ($\psi = 10.0$). We choose the number of clusters that yields maximum likelihood by searching over a range of K. For the spoke model, our estimated error rates are slightly higher than the rates estimated by [17]. K: number of clusters. $\hat{\nu}$:estimated false negative rate. $\hat{\phi}$:estimated false positive rate.

5.3 Result

To test the convergence of our algorithm independent of the random starting point, we ran it on simulated data. We simulated the data by creating a set of nodes and randomly assigning them to a fixed number of clusters. To connect any two nodes, first we fixed the number of trials to be the same for all pairs of nodes and set the success value s according to the specified values of ν and ϕ. We ran the algorithm multiple times for different random starting points and initial values for ψ. We computed the average minimum cost at a given number of clusters, as shown in Figure 5.2. Independent of the initial value for ψ, the result shows that the algorithm converges to a correct solution on simulated data. A correct solution includes a correct clustering solution and an accurate estimation of model parameters, ν and ϕ. As expected, we found that the number of trials affects the accuracy of the results. For the same solution, the negative log-likelihood becomes higher (lower likelihood) as we reduce the number of trials. This corresponds to the fact that the more repeated experiments we have, the better the estimation.

Working with experimental data, we compute clusters for two types of observation models: the spoke model and the matrix model of protein interactions (see Chapter 2). To select a maximum likelihood solution, we performed two line minimizations on the negative log-likelihood as described above, first selecting ψ and then selecting the number of clusters. The maximum likelihood solutions are shown in Table 5.1.

Dataset	Num. Proteins		MCL	MRF	MCODE	Gavin06 (all)	Gavin06 (core)
Gavin02	1390	Proteins clustered	1390	1390	112	–	–
		with MIPS	494	494	53	–	–
		with Reguly	136	136	20	–	–
Gavin06	2760	Proteins clustered	2760	2760	243	1488	1147
		with MIPS	819	819	141	633	492
		with Reguly	520	520	120	429	336

Table 5.2: Number of proteins in clustering experiments. MCL and MRF consider the same number of proteins: all proteins in the experiments. However, their clustering solutions are different; MCL will produce more singletons than MRF.

5.3.1 Quality of clusters

We devised a measure to assess the quality of each cluster with respect to the model. For each cluster k, we define $E(k)$ to be a measure of the observational error with respect to the expected error.

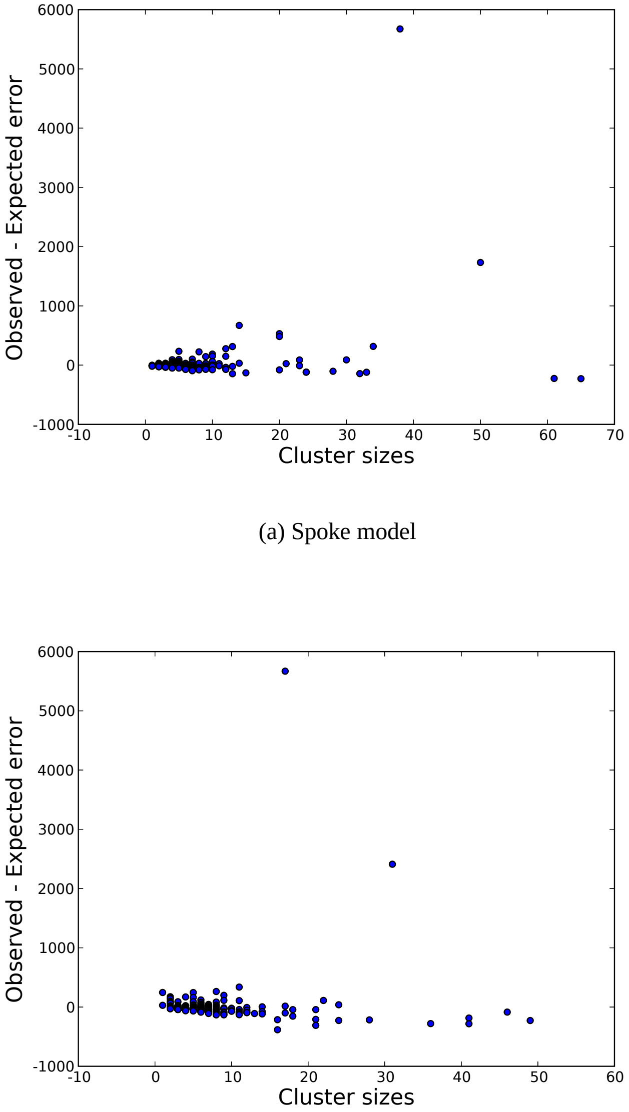
$$E(k) = \sum_{(i,j):Q_i=Q_j=k} (t_{ij} - s_{ij}) + \sum_{(i,j):Q_i\neq Q_j=k} s_{ij} - E_{fn}(k) - E_{fp}(k),$$

where $E_{fn}(k) = \nu^* \sum_{(i,j):Q_i=Q_j=k} t_{ij}$ and $E_{fp}(k) = \phi^* \sum_{(i,j):Q_i\neq Q_j=k} t_{ij}$.

From its definition, this score is always positive for some clusters, indicating that the observed error is higher than the expected error, and less than zero for other clusters, indicating that the observed error is less than the expected error. So rather than giving an absolute measure of quality for the whole solution, the measure indicates, within a given solution, clusters with high confidence and those with low confidence. The scatter plots in Figure 5.3 shows that there is no correlation between the score $E(k)$ and cluster sizes and that the mode of the measure is around zero, as expected. Interestingly, they also show that we have discovered quite reliable observations for some large clusters.

5.3.2 Comparison with clustering algorithms for protein-protein interaction networks

We evaluated the performance of MRF using two data sets : **Gavin02** and **Gavin06**. We compared the performance of the MRF algorithm to both the algorithms MCL and MCODE and the hand-curated results accompanying publications of the data sets. Because MCL and MCODE require an



(a) Spoke model

(b) Matrix model

Figure 5.3: **Gavin06**: a scatter plot of the quality of clusters VS cluster sizes.

Dataset	MCL	MCL with interaction prob. [17]	MRF	MCODE
Gavin02	From [8]	Inflation = 1.8	$\psi = 3.5$	From [8]
Spoke model	Inflation = 1.8	$\nu = 0.346$	Maximum likelihood	Node score percentage = 0.0
		$\phi = 1.07 \times 10^{-3}$		Complex fluff = 0.2
		$\rho = 1.88 \times 10^{-3}$		Depth = 100
Gavin06	Inflation = 3.0	Inflation = 3.0	$\psi = 3.5$	From [8]
Spoke model		$\nu = 0.407$	Maximum likelihood	Node score percentage = 0.0
		$\phi = 1.35 \times 10^{-3}$		Complex fluff = 0.2
		$\rho = 3.89 \times 10^{-3}$		Depth = 100
Gavin06	–	–	$\psi = 10.0$	–
Matrix model			Maximum likelihood	

Table 5.3: Parameter setting for MCL, MRF and MCODE.

interaction graph as input data set, for each of these data sets, we built an interaction graph using a spoke model. MCL accepts both weighted and unweighted graphs as an input. For the weighted interaction graph, we computed the interaction probability using the statistical model in [17] (with no threshold).

We set the parameters of MCODE according to [8]. With respect to the inflation parameter for MCL, we found that the optimal setting as published in [8] is suitable for the smaller data set (Gavin02), but yields a biologically incorrect number of clusters for the larger Gavin06 data set. Therefore, to run MCL on the large data set, we have explored several inflation parameters and choose a solution with the inflation parameter = 3.0 that controls the number of clusters larger than 2 to be close to the published number of 487 complexes [14]. The trade-off in sensitivity and specificity from exploring the inflation parameters is shown in Figure 5.4. The range of recommended inflation parameters is between 1.1 and 5.0 (`http://micans.org/mcl`). We summarize the parameter setting for all three algorithms in Table 5.3. For comparison of the clustering algorithms, we apply the above performance measures to evaluate the clustering solutions against the MIPS and Reguly benchmarks. For each of these measures, we contrast the scores reached with the real clustering results with the random expectation estimated with permuted clusters and found the separation to be high in all algorithms. We summarize the measurement in Table 5.6 for the Gavin02 data set and the Gavin06 data set. For the evaluation, we do not consider singletons as valid clusters and exclude them from the distribution of cluster sizes. See Table 5.4 and Table 5.5.

	MCL	MCL with inter. prob.	MRF (spoke)	MRF (matrix)	MCODE
Num. of clusters	351	352	393	310	24
Num. of singletons	177	178	226	79	0
size ≥ 2	174	174	167	231	24
Mean	6.97	6.97	6.97	5.67	4.67
Median	4	5	5	2	4
1st quantile	3	3	2	2	3
3rd quantile	8	8	10	6	6
90%	15	14	14	13	7
99%	42	40	34	36	9
Largest cluster	51	45	36	44	11

Table 5.4: **Gavin02**: distribution of cluster sizes.

There is a precaution on the set of proteins used to evaluate performance measures. For each data set, we use the same set of proteins for evaluation in all measures, which is the set of proteins in the experiment with annotation. This will in general cause lower sensitivity in algorithms such as MCODE due to many unassigned proteins. We do not want to reduce the set of proteins to the common sets among all algorithms (the minimal set of proteins) as the measurement will not be representative. Bearing in mind these limitations, we can interpret the results as shown in Table 5.6 and the ROC curves in Figure 5.4. As expected, we find clustering solutions of MCODE to have low all pairs sensitivity because it simply does not assign the majority of proteins present in the experiment and hence produces very few clusters. If we change the setting of MCODE to include more clusters and assign more proteins, we significantly lose accuracy in all measures (data not shown). Thus, we focus on comparing MCL and MRF.

Complex-size distribution

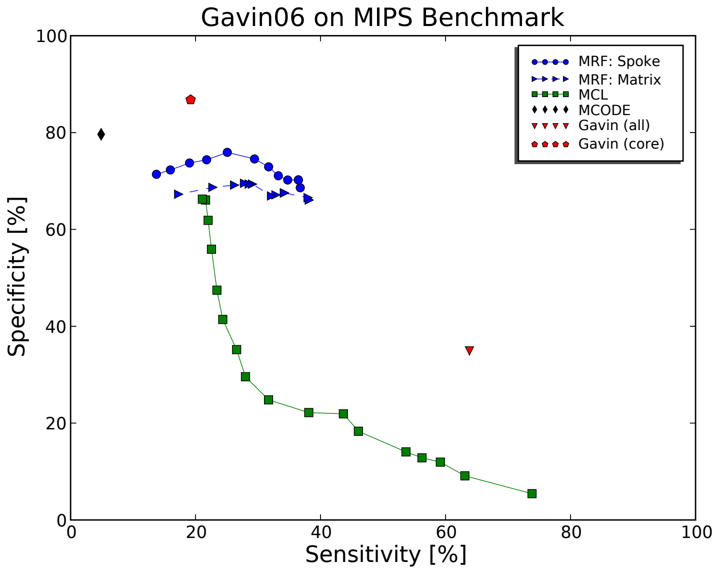
Principle properties and potential artifacts are visible in a simple plot of the population of proteins by cluster size (see Figure 5.5). From the Gavin06 data set, we only consider proteins with MIPS complexes, ignoring singletons; this results in 819 proteins. For each clustering solution, we restrict the size distribution to this set of protein, also ignoring singleton-clusters. It is worth to note that there is an absence of larger MIPS complexes in the range from 30 to 50, which has only one complex of size 47. Obviously, the proteins in the largest complex of size 60 all correspond to a single complex (the ribosome), whereas the 60 proteins in clusters of size 12 correspond to

	MCL	MCL with	MRF	MRF	Gavin06	Gavin06	MCODE
		interaction prob.	(spoke)	(matrix)	(all)	(core)	
Num. of clusters	781	732	698	550	487	477	55
Num. singletons	331	269	4	2	0	55	0
size ≥ 2	450	463	694	548	0	422	0
Mean	5.39	5.38	3.97	5.03	13.46	3.33	4.42
Median	2	3	2	3	9	2	4
1st quantile	2	2	2	3	4	2	3
3rd quantile	4	4	4	5	18	4	5
90%	8	7	7	8	33	6	7
99%	36	29	32	31	66	12	16
Largest cluster	561	607	65	49	96	23	16

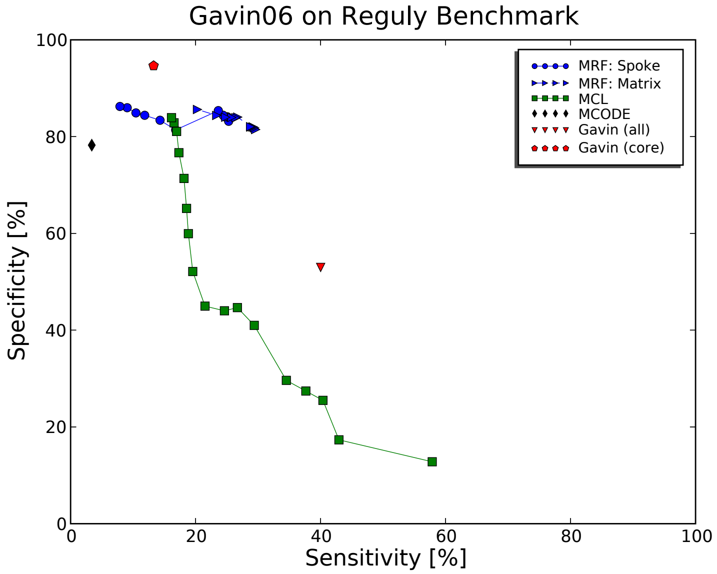
Table 5.5: **Gavin06**: distribution of cluster sizes.

Dataset		MCODE	MCL	MCL with	MRF	MRF
				inter. prob.	(spoke)	(matrix)
Gavin02	CO	25.1	61.0	62.3	60.4	**76.8**
	PPV	**76.1**	71.0	74.9	75.5	70.6
	Acc	43.7	65.8	68.3	67.5	**73.6**
	All pairs					
	SN	2.3	68.6	68.9	**66.7**	62.6
	SP	**92.5**	78.7	82.4	87.9	64.7
	Geo. average	14.7	73.0	75.4	**76.6**	63.6
Gavin06	CO	27.2	63.6	65.3	65.6	**67.4**
	PPV	54.1	59.1	63.2	71.6	**73.8**
	Acc	38.4	61.3	64.3	68.6	**70.5**
	All pairs					
	SN	4.9	44.1	**44.7**	37.2	38.2
	SP	**79.6**	18.0	22.5	70.0	66.1
	Geo. average	19.7	28.2	31.7	**51.0**	50.2

Table 5.6: Clustering performance of MCODE, MCL and MRF: comparison against the MIPS benchmark. We use all proteins in the experiment with annotation.



(a) MIPS



(b) Reguly

Figure 5.4: Gavin06: Comparison of sens. and spec. for all clustering solutions on all proteins with annotation from: (a) MIPS and (b) Reguly. The curve for MRF is generated as we filter out clusters with high observed-errors. The curve for MCL is generated for different inflation parameters: $[1.2 : 0.2 : 5.0]$ which is recommended by the MCL program.

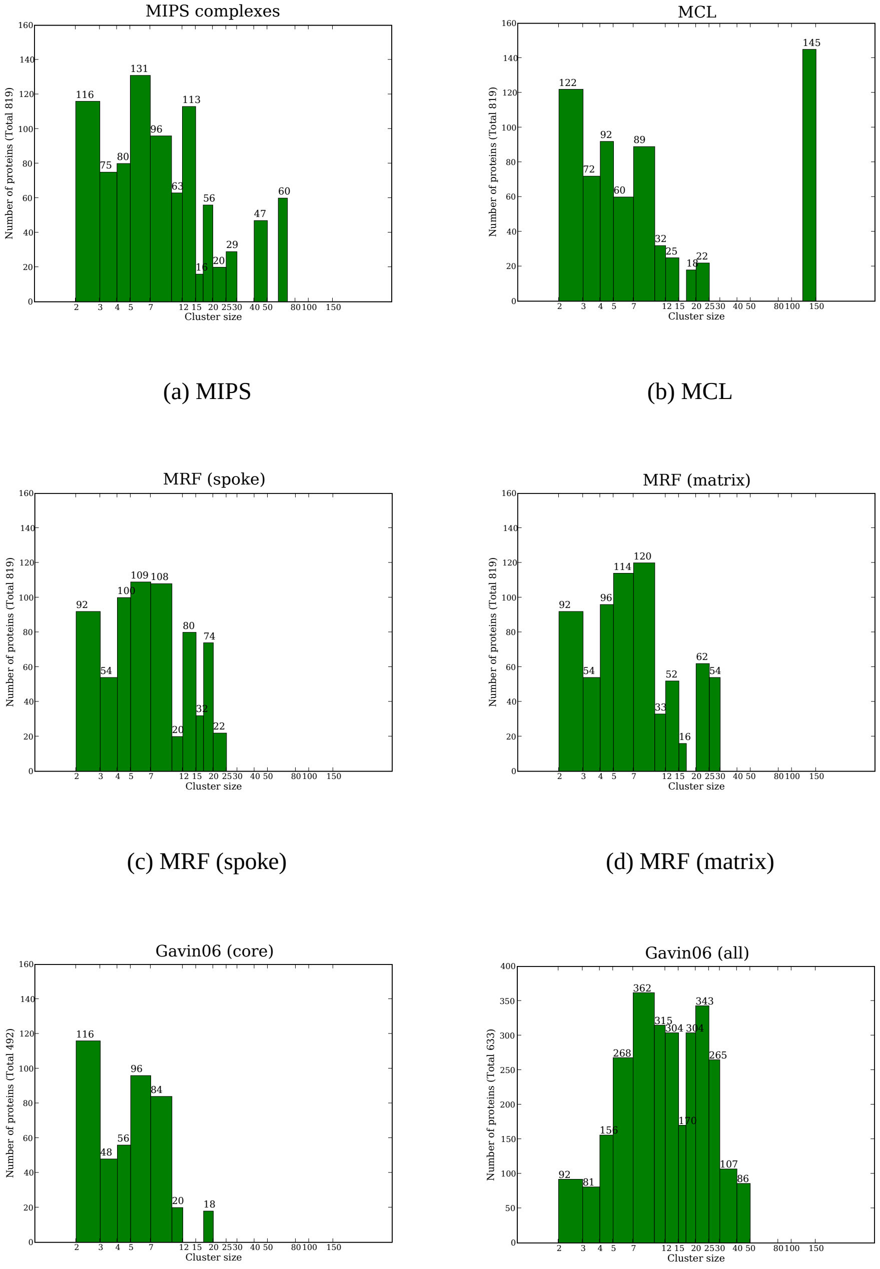
5 different complexes. All clustering solutions substantially deviate from the MIPS distribution. MCL has a large cluster of 145 proteins from different MIPS complexes, a likely artifact. This cluster is part of a giant component when considering all proteins. The overall property can be deducted using our approach better than MCL. The Gavin core set is only a subset and contains a substantial number of small elements and fewer complexes than the MIPS solution, prominently the mitochondrial ribosome and mediator complex. The larger, complete solution (Gavin06 (all)) contains few small clusters; although this solution contains larger clusters (size ≤ 50), they do not accurately map to larger complexes.

5.3.3 Comparison with maximal frequent itemsets (MFI)

We obtain a solution of overlapping complexes by computing maximal frequent itemsets (MFI) rather than exact frequent itemsets. It is clear from the two previous chapters that exact frequent itemset mining is not a sensitive method and even with extension to handle errors in the previous chapter, we still cannot improve the sensitivity of frequent itemsets. We can explain the problem of low sensitivity by showing an example on a typical result of frequent itemsets. As illustrated in Figure 5.6, although frequent itemsets can output overlapping clusters, in reality they are overlapping fragments of a true class which should be merged to one cluster. To alleviate the problem of fragmentation, we also merge maximal frequent itemsets when they share more than 3 proteins.

To select a solution from MFI, we have to decide on a suitable minimum support that yields high accuracy when comparing to a benchmark. We search over different values of minimum support and select the one that gives a good balance of all-pair sensitivity and specificity (by computing a geometric average). Figure 5.7 shows the result of exploring different minimum support for MFI and clearly shows that MRF is a better prediction. A minimum support that yields the best geometric average of sens. and spec. is 9 out of 2166 on the Gavin06 data set. After we have selected the best solution for MFI, we can compare this solution against the maximum likelihood solution of MRF (on the spoke model) by computing all the performance measures as shown in Table 5.9.

Although MFI results in more clusters than MRF (Table 5.8), there are only 456 proteins in those clusters indicating that MFI merely constructs fragments of real complexes. This result is also confirmed by low sensitivity and low complex-coverage in Table 5.9. When we evaluate using all annotated proteins, the accuracy measure is much lower than the accuracy value of MRF. The



(a) MIPS

(b) MCL

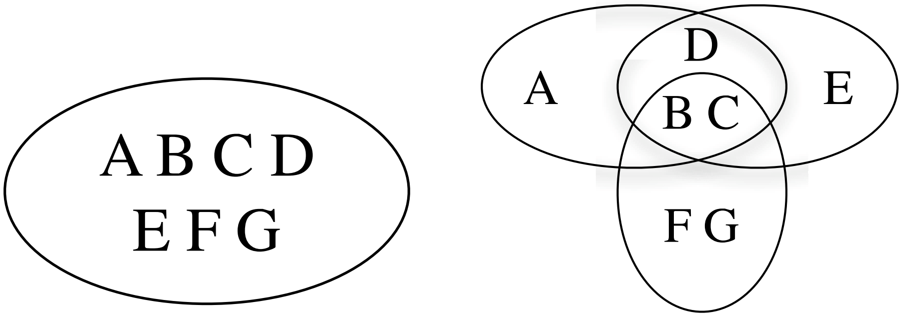
(c) MRF (spoke)

(d) MRF (matrix)

(e) Gavin06 (core)

(f) Gavin06 (all)

Figure 5.5: Comparison of cluster-size distribution with MIPS-size distribution.



(a) True classes

(b) Frequent itemsets

Figure 5.6: Problem with exact frequent itemsets: low sensitivity. Frequent itemsets break the true cluster into three overlapping clusters: $\{A, B, C, D\}$, $\{B, C, D, E\}$ and $\{B, C, F, G\}$ although the correct answer is $\{A, B, C, D, E, F, G\}$.

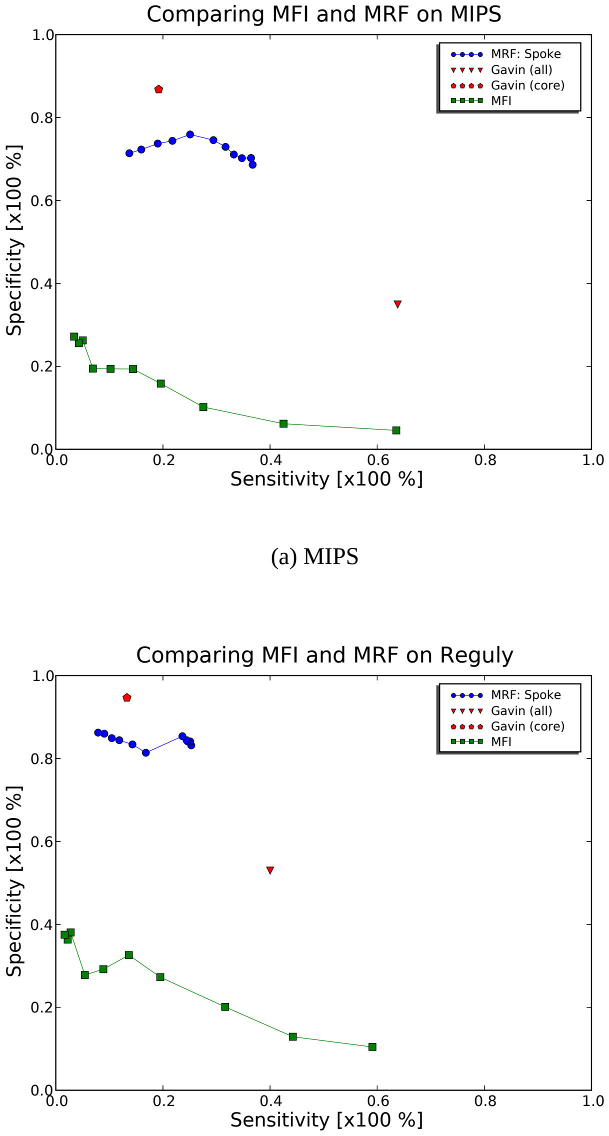
Dataset	Num. Proteins		MRF	MFI
Gavin06	2760	Proteins clustered	2760	456
		Num. purifications	2166	2166
		with MIPS	819	256
		with Reguly	520	195

Table 5.7: Comparing number of proteins in clustering experiments between MRF and MFI. For MFI, we select a solution with the minimum support of 9 (out of 2166) which yields the best average of all-pair sens. and spec..

results in Figure 5.7 and Table 5.9 show that the solution from MRF outperforms MFI and that MFI is not a good model of overlapping complexes. It is interesting to note that Hollunder et al. [21] shows that the results of frequent itemsets are statistically significant and postulate that they correspond to subcomplex structure of multi-protein complexes. The question whether or not the observed subcomplex structure has any biological significance still remains to be investigated and at present there is no experimental method to validate such a hypothesis.

5.4 Discussion

Before we discuss the details of the results, we would like to point out that, in contrast to others, MRF is essentially a parameter-free method. Although MFA requires two input – ψ and the number



(a) MIPS

(b) Reguly

Figure 5.7: Comparison of MFI at different minimum support with MRF. Sens. and Spec. for **Gavin06**. Both figures show all pairs specificity and sensitivity on proteins with known complexes from (a) the MIPS benchmark and (b) the Reguly benchmark. Both figures show MRF to be more accurate than frequent itemsets. MFI: Maximal Frequent Itemsets. MRF: Markov Random Field.

	MCL	MRF	MFI
		(spoke)	min. support = 9
Num. of clusters	732	698	898
Num. singletons	269	4	0
size ≥ 2	463	694	898
Mean	5.38	3.97	2.71
Median	3	2	2
1st quantile	2	2	2
3rd quantile	4	4	3
90%	7	7	3
99%	29	32	5
Largest cluster	607	65	241

Table 5.8: **Gavin06**: distribution of cluster sizes. Comparison between MCL, MRF and MFI.

Dataset		MRF	MFI
		(spoke)	min. support = 9
Gavin06	CO	**65.6**	42.7
	PPV	71.6	**74.5**
	Acc	**68.6**	56.4
	All pairs		
	Sens.	**37.2**	19.6
	Spec.	**70.0**	15.8
	Geo. average	**51.0**	17.6

Table 5.9: Clustering performance of MRF and MFI. We evaluate on all proteins in the experiment with MIPS complexes.

of clusters K, we provide a systematic way to estimate them using maximum likelihood. Methods such as MCODE or the Gavin06 solution require more parameters without a systematic way to select them other than trying out several values and comparing the results to a benchmark. If there is no reference data set available, these methods cannot assess the quality of their solution, while MRF at least provides a relative measure of cluster quality. MCL suffers from the same problem of parameter selection and essentially has three parameters, the expansion and inflation values and the number of clusters. So to choose a solution from MCL we must not only compare with the benchmark, but also decide if the number of clusters is biologically plausible. With regard to the number of predicted clusters, it is not surprising that MRF estimates higher number of clusters because it does not eliminate proteins prior to clustering, unlike other methods such as the MCODE or the Gavin06 core solution.

Although we recommend the spoke model over the matrix model due to lower false negative rate, it is worth to note that the solution of the matrix model is also biological meaningful when compared with the MIPS data set, although with a slightly less specific solution than the solution from the spoke model (on the Gavin06 data set comparing to the MIPS benchmark). Our caution is that the reality lies in between and the error rate is only one criterion to select between these two models.

With regard to quality of clusters, we observe that almost all predicted clusters fit the model except some outliers that should not be regarded as complexes due to extremely high observed error (shown as data points on the top of Figure 5.3). Closer inspection reveals that they are clusters consisting mostly of junk proteins which in reality cannot be assigned to any complex. By giving the junk clusters the worst quality score, MRF can separate them from the rest of other complexes. For MCL, there is no such indicator other than eliminating singletons.

The performance of MCL and MRF on the Gavin02 data set is comparable, both achieve high accuracy. The similarity in performance between MCL and MRF also indicates that the Gavin02 data set has lower level of noise and error modeling does not necessarily gain us more accuracy. The low level of noise corresponds to the fact that the Gavin02 has already been filtered. The similarity between MCL and MRF also prevails in their distribution of cluster sizes (see Table 5.4).

The performance gain from error modeling is more noticeable in the larger Gavin06 data set which is not filtered and thus expected to contain more errors. We clearly see in this case that the model of interaction probability only slightly improve MCL in predicting protein complexes and that the error model of MRF is more accurate. Machine learning techniques can be used to filter low accuracy interaction [26] before running MCL, but we cannot compare our model with this model because the raw purification data set, in particular repeated experiments, are not provided with the publication [26].

The Accuracy measure indicates on average how well an individual cluster matches any complex. It penalizes split complexes more than merged complexes. To see if complexes are merged, we have to look at the all pair comparison for high sensitivity with low specificity. When compared with MIPS, due to complexes merging to a giant component, MCL performs quite well on Gavin06 in the accuracy value, but not when considering the all-pair SN and SP. To avoid the giant component, the inflation parameter of MCL must be set to the maximum level recommended (inflation = 5.0) which results in reduced sensitivity. See Figure 5.4. MRF can maintain high specificity without sacrificing the sensitivity. When comparing to highly specific solutions such as MCODE or Gavin06 (core) which assign fewer proteins, MRF loses less than 10% percent in specificity, but gains about 30% in sensitivity and assigns more proteins (Table 5.6).

In general, both MCL and MRF perform better when comparing to the MIPS benchmark than to the Reguly benchmark, with MRF performing better than MCL at matching both benchmarks on the Gavin06 data set. The Reguly benchmark allows many complexes to overlap. Hence, MCL and MRF will never be able to fully reconstruct the Reguly benchmark because they assume a partitioning model of protein complexes. However, we can see that when trying to match the MIPS complexes based on all-pair comparison, MRF outperforms MCL. This indicates that in general the assumption of complex formation based on only pairwise interaction is a reasonable one resulting in few false positive errors. As expected, the high specificity (low false positive errors) and low sensitivity (high false negative errors) of MRF is a result of the predicted error rates by the model; the high false negative rate predicts low sensitivity, while the low false positive rate predicts high specificity.

When comparing all solutions to the MIPS-size distribution in Figure 5.5, we clearly see that MCL is particularly far-off due to the giant component which assign about 140 proteins from different MIPS complexes into the same cluster. The solution from MRF appears to be the closest match in this regard, although it still cannot reconstruct MIPS-complexes larger than 30. Other solutions also have the same problem; the Gavin06 (core) solution only maps to small complexes (size ≤ 20). MRF replaces large complexes by producing more smaller clusters than MIPS (size ≤ 5).

In summary, if the data has already been filtered as in the Gavin02 data set, MRF does not have an advantage over MCL because it is computationally more expensive. When clustering large and noisy data set, the result has shown that MRF is a better method with a rigorous framework to select parameters using maximum likelihood which is in itself an advantage over heuristics-based methods such as MCL or MCODE.