Biometrics 63, 824–833
September 2007

DOI: 10.1111/j.1541-0420.2007.00755.x

Bayesian Methods for Predicting Interacting Protein Pairs Using Domain Information

Inyoung Kim,[1] Yin Liu,[2] and Hongyu Zhao[1,3,*]

[1]Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut 06520, U.S.A.

[2]Program of Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, U.S.A.

[3]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, U.S.A.

[*]*email:* hongyu.zhao@yale.edu

Summary. Protein–protein interactions (PPIs) play important roles in most fundamental cellular processes including cell cycle, metabolism, and cell proliferation. Therefore, the development of effective statistical approaches to predicting protein interactions based on recently available large-scale experimental data is very important. Because protein domains are the functional units of proteins and PPIs are mostly achieved through domain–domain interactions (DDIs), the modeling and analysis of protein interactions at the domain level may be more informative and insightful. However, due to the large number of domains, the number of parameters to be estimated is very large, yet the amount of information for statistical inference is quite limited. In this article we propose a full Bayesian method and a semi-Bayesian method for simultaneously estimating DDI probabilities, the false positive rate, and the false negative rate of high-throughput data through integrating data from several organisms. We also propose a model to associate protein interaction probabilities with domain interaction probabilities that reflects the number of domains in each protein. Our Bayesian methods are compared with the likelihood-based approach (Deng et al., 2002, *Genome Research* **12,** 1504–1508; Liu, Liu, and Zhao, 2005, *Bioinformatics* **21,** 3279–3285) developed using the expectation maximization algorithm. We show that the full Bayesian method has the smallest mean square error through both simulations and theoretical justification under a special scenario. The large-scale PPI data obtained from high-throughput yeast two-hybrid experiments are used to demonstrate the advantages of the Bayesian approaches.

Key words: Bayesian method; Domain–domain interaction; Expectation maximization algorithm; Protein–protein interaction.

1. Introduction

Because identifying protein–protein interactions (PPIs) is critical for understanding cellular processes, various high-throughput experimental approaches have been developed and enormous amounts of data have been generated to identify interacting proteins. The protein interaction data used in our study are generated by genome-wide yeast two-hybrid assays. In this method, one protein is fused to a DNA-binding domain, and the other is fused to a transcription activation domain. The interaction between the protein pair can be detected by the formation of a transcription activator that activates a reporter construct (Uetz et al., 2000). However, this experimental approach suffers from high false negative and false positive rates due to the limitations of these techniques (Mrowka, Patzak, and Herzel, 2001; von Mering et al., 2002). For example, a self-activating protein being tested in the experiment can lead to a false positive result, and a protein that cannot be targeted to the yeast nucleus may not yield positive results though it may potentially interact with other proteins, which leads to false negative results. It is reported that the false negative rate of the yeast two-hybrid assay used to construct *S. cerevisiae* interaction maps to be larger than 70% (Deng et al., 2002).

A number of computational approaches have been proposed to predict PPIs (Enright et al., 1999; Tsoka and Ouzounis, 2000; Marcotte, Xenarios, and Eisenberg, 2001; Pazos and Valencia, 2001; Goh and Cohen, 2002; Jansen et al., 2003; Lu et al., 2003; Ramani and Marcotte, 2003; Aloy et al., 2004). However, most methods do not consider the fact that domains are the functional units of proteins and PPIs are achieved mostly through domain–domain interactions (DDIs). Protein domains are defined as the basic modules of the overall protein structure and are conserved during evolution. Some proteins consist of only a single domain, but many proteins contain more than one domain to perform multiple functions. For example, the protein DNA-directed RNA polymerase II subunit 9 is a multidomain protein that contains two domains, the TFIIS domain for DNA binding and the RNA polymerase M

domain for RNA synthesis. Protein domains serve as the units for PPIs and the specificity of PPIs is achieved from the binding of a modular domain to another in proteins (Pawson and Nash, 2003). Therefore, the modeling and analysis of protein interactions at the domain level may be more informative and insightful.

Several methods have been proposed for PPI predictions based on protein domains (Gomez, Lo, and Rzhetsky, 2001; Sprinzak and Margalit, 2001; Deng et al., 2002; Gomez, Noble, and Rzhetsky, 2003). The likelihood-based approach (Deng et al., 2002) has been compared with three other methods (Gomez et al., 2001, 2003; Sprinzak and Margalit, 2001) and was shown to be among the best performing methods (Liu, Liu, and Zhao, 2005). Liu et al. (2005) further extended the likelihood-based approach to improve the PPI predictions by pooling information from three organisms: *S. cerevisiae*, *C. elegans* and *D. melanogaster*. In the likelihood-based approach, all DDIs and PPIs are treated as missing data. For a given specified set of false negative and false positive rates, DDI probabilities are estimated using the expectation maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) and then the estimated DDI probabilities are used to infer PPI probabilities.

However, in general, the false negative rate (f_n) and the false positive rate (f_p) of PPI data are both unknown and they may also depend on many factors, for example, data set specific. Therefore, it is more appropriate to estimate the DDI probabilities, the PPI probabilities, f_n, and f_p, simultaneously. In addition, the number of domains in many proteins is more than one. When we infer PPIs using DDIs, it is important to take into account the varying number of domains across different proteins. Furthermore, because the number of parameters is very large, the amount of information for statistical inference may be limited.

Hence our goals in this article are to develop statistical methods to estimate DDI probabilities, PPI probabilities, f_n, and f_p simultaneously, and also propose a model (described in Section 2) to relate PPI probabilities with DDI probabilities that reflects the number of domains in each protein. Our first approach is called "semi-Bayesian" because we use the EM algorithm to estimate DDI and PPI probabilities, coupled with a Bayesian method to estimate f_n and f_p. We specify a uniform prior distribution for f_n and f_p within a reasonable range that can be established from prior biological knowledge. Our second approach is called "full Bayesian" because we estimate DDI probabilities, PPI probabilities, f_n, and f_p using a Bayesian approach, exclusively. We note that the number of parameters to be estimated is very large. For example, if there are more than 2000 annotated domains within an organism, the number of domain pairs is in the order of millions. In this case, some prior information can be easily incorporated in the analysis through a Bayesian method. The setup of our full Bayesian method lends itself easily to information sharing among all domain pairs. In our full Bayesian approach, we consider both Beta and uniform distributions as priors for DDI probabilities, and use uniform distributions for f_n and f_p within a reasonable range that can be established from biological knowledge. Because we consider the possibility of different organisms having different false negative and false positive rates, the ranges are varied based on the organisms so that f_n and f_p can be organism specific.

The article is organized as follows. In Section 2, we describe the modeling of PPI using DDI information. We give a detailed discussion of our model by explaining the difference among Deng et al. (2002)'s model, Liu et al. (2005)'s model, and our model. In Section 3, we discuss three methods for PPI predictions: the likelihood-based approach using the EM algorithm, our semi-Bayesian method, and our full Bayesian method. Under a simplified scenario described in Web Appendix C, we show that the mean square error (MSE) of the Bayes estimator is smaller than that of the maximum likelihood estimator (MLE). In Section 4, we report the results of simulations comparing these methods. We calculate the MSE of the likelihood-based approach based on the EM algorithm in the ideal case when true f_n and true f_p are known. We compare this MSE with our methods' MSEs. Our simulations suggest that a full Bayesian method appears most efficient in terms of MSE and a semi-Bayesian method is as good as the likelihood-based approach in this ideal case. In Section 5, we apply our methods to large-scale PPI data obtained from high-throughput yeast two-hybrid experiments analyzed by Liu et al. (2005). Section 6 contains concluding remarks.

2. Model

This section describes our model to estimate DDI probabilities and relate them with PPI probabilities. Before we explain our model in detail, we define some notations and briefly describe the model studied by Deng et al. (2002) and Liu et al. (2005). The model proposed by Deng et al. (2002) has two assumptions: (A1) two proteins P_i and P_j interact if and only if at least one domain pair from the two proteins interact; (A2) DDIs are independent, that is, whether two domains interact or not does not depends on the interactions of other domain pairs. To pool information across different organisms, Liu et al. (2005) made another assumption: (A3) the probability that two domains interact is the same among all organisms based on the fact that domains are evolutionally conserved across different organisms. This assumption allows the integration of large-scale PPI data from different organisms to estimate DDI probabilities. By considering each protein as a collection of domains, we can then estimate PPI probabilities in any organism based on the inferred DDI probabilities. More specifically, let λ_{mn} represent the probability that domain D_m interacts with domain D_n. Define $D_{mn}^{(ij)} = 1$ if D_m and D_n interact in protein pair P_i and P_j and $D_{mn}^{(ij)} = 0$ otherwise. Let $(D_{mn}^{(ij)} \in P_{ijk})$ denote all pairs of domains from protein pair P_i and P_j in organism k, where $k = 1, \ldots, K$. Let P_{ijk} represent the interaction event between P_i and P_j in organism k, with $P_{ijk} = 1$ if they interact in organism k and $P_{ijk} = 0$ otherwise. Further let $O_{ijk} = 1$ if P_i and P_j are observed to interact in organism k, and $O_{ijk} = 0$ otherwise. In our example, we focus on $K = 3$ organisms, where $k = 1, 2, 3$ represents *S. cerevisiae* (yeast), *C. elegans* (worm), and *D. melanogaster* (fruit fly), respectively.

The definitions of false negative rate (f_n) and false positive rate (f_p) of protein interaction data are

$$f_p = \Pr(O_{ijk} = 1 \,|\, P_{ijk} = 0),$$

$$f_n = \Pr(O_{ijk} = 0 \,|\, P_{ijk} = 1).$$

We further define $O = \{O_{ijk} = o_{ijk}; \forall\, i \le j\}$, $\Lambda = \{\lambda_{mn}; D_{mn}^{(ij)} \in P_{ijk},\ \forall\, m \le n,\ \forall\, i \le j\}$. With the above assumptions and notations, we have

$$\begin{aligned}\Pr(P_{ijk} = 1) &= 1 - \prod_{\left(D_{mn}^{(ij)} \in P_{ijk}\right)} (1 - \lambda_{mn}) \\ &=_{\text{def.}} h_{ij}^{1}(\Lambda),\end{aligned} \tag{1}$$

and

$$\Pr(O_{ijk} = 1) = \Pr(P_{ijk} = 1)(1 - f_n) + \{1 - \Pr(P_{ijk} = 1)\} f_p. \tag{2}$$

The likelihood for the observed PPI data across all K organisms is then

$$L(f_n, f_p, \Lambda \mid O) = \prod_{ijk} \Pr(O_{ijk} = 1)^{O_{ijk}} \{1 - \Pr(O_{ijk} = 1)\}^{1 - O_{ijk}},$$

which is a function of (λ_{mn}, f_n, f_p). Deng et al. (2002) and Liu et al. (2005) specified values for f_n and f_p, and then estimated λ_{mn} using the EM algorithm by treating all DDIs and PPIs as missing data.

In contrast to these previous methods, we assume that f_n and f_p are unknown, and they are also organism dependent to allow data from different organisms to have different f_n and f_p. That is, we have organism specific rates f_{n_k} and $f_{p_k}, k = 1, \ldots, K$. In this case, equation (2) is replaced by

$$\begin{aligned}\Pr(O_{ijk} = 1) = \Pr(P_{ijk} = 1)(1 - f_{n_k}) \\ + \{1 - \Pr(P_{ijk} = 1)\} f_{p_k}.\end{aligned}$$

We also extend equation (1) to incorporate varying numbers of domains across different proteins. This extension is motivated from observing that the value of equation (1) increases as the number of domains increases. For example, if all domain pairs have 1/2 chance to interact, the PPI probability approaches 1 when the number of domain pairs associated with the protein pair is large. Therefore, we formulate function $h_{ij}(\Lambda)$, where $h_{ij}(\Lambda) = \Pr(P_{ijk} = 1)$, which satisfies the following 4 conditions:

- **C1:** If $\lambda_{mn} = 1$ for at least one domain pair, $h_{ij}(\Lambda) = 1$;
- **C2:** If $\lambda_{mn} = 0$ for all domain pairs, $h_{ij}(\Lambda) = 0$;
- **C3:** If $\lambda_{mn} = 1/2$ for all domain pairs, $h_{ij}(\Lambda) = 1/2$;
- **C4:** (Strictly increasing condition) If $\lambda_{mn} < \lambda_{m'n'}$ and all other λs are the same, $h_{ij}(\Lambda) < h_{ij}(\Lambda')$.

We note that $h_{ij}^{1}(\Lambda)$ in equation (1) does not satisfy **C3**. In this article, we consider one possible function for $h_{ij}(\Lambda)$ as the following:

$$h_{ij}^{a}(\Lambda) = 1 - \prod_{\left(D_{mn}^{(ij)} \in P_{ijk}\right)} \left(1 - \lambda_{mn}^{a}\right),$$

where a can be derived from condition **C3** as,

$$a = \frac{\log\left\{1 - (1/2)^{\frac{1}{M_{ij}}}\right\}}{\log(1/2)}$$

and M_{ij} represents the total number of domain pairs between P_i and P_j. If $M_{ij} = 1$, $a = 1$. Therefore, $h_{ij}^{a}(\Lambda)$ is the same as $h_{ij}^{1}(\Lambda)$ in this special case.

3. Methods

3.1 *Introduction*

This section describes three methods for PPI predictions. Section 3.2 reviews the likelihood-based approach realized through the EM algorithm. Section 3.3 describes our semi-Bayesian approach, while Section 3.4 develops the full Bayesian approach.

3.2 *Likelihood-Based Approach*

In this approach, PPIs and DDIs are treated as random variables. The DDI probability can be estimated using Liu et al. (2005)'s approach, which is an extension of the likelihood-based approach proposed by Deng et al. (2002), so that it can incorporate information from all K organisms.

Let A_m be the set of proteins containing domain D_m and N_{mn} be the total number of protein pairs between A_m and A_n across all K organisms. The observed data are the experimentally observed interactions $O = \{O_{ijk} = o_{ijk};\ i \le j\}$. The complete data include all DDIs for each protein pair. Define the complete data as (O, D), in which O is given above and $D = \{D_{mn}^{(ij)};\ P_i \in A_m,\ P_j \in A_n,\ \forall m, n\}$.

If we specify the values for f_n and f_p, the likelihood function is a function of $\theta = (\lambda_{mn})$ only. The λ_{mn} can be estimated using the EM algorithm as follows. For a given $\theta^{(t-1)} = \lambda_{mn}^{(t-1)}$ obtained from the $(t-1)$th EM iteration, the next E-step can be computed as

$$\begin{aligned}&E\left(D_{mn}^{(ij)} \mid O_{l_1 l_2 k} = o_{l_1 l_2 k}, \forall l_1, l_2, \theta^{(t-1)}\right) \\ &= \frac{\lambda_{mn}^{(t-1)}(1 - f_n)^{O_{ijk}} f_n^{1 - O_{ijk}}}{\Pr\left(O_{ijk} = o_{ijk} \mid \theta^{(t-1)}\right)} = \tau_{mn}^{(ij)}(\theta^{(t-1)}),\end{aligned}$$

where the denominator can be calculated using equation (2).

Because the MLE of λ_{mn} is the fraction of $\{D_{mn}^{(ij)};\ P_i \in A_m,\ P_j \in A_n,\ \forall k\}$ such that $D_{mn}^{(ij)} = 1$, we thus obtain a recursive formula for the M-step:

$$\begin{aligned}\lambda_{mn}^{(t)} &= \frac{\sum_{i \in A_m, j \in A_m, \forall k} \tau_{mn}^{(ij)}(\theta^{(t-1)})}{N_{mn}} \\ &= \frac{\lambda_{mn}^{(t-1)}}{N_{mn}} \sum_{i \in A_m, j \in A_m, \forall k} \frac{(1 - f_n)^{o_{ijk}} f_n^{1 - o_{ijk}}}{\Pr\left(O_{ijk} = o_{ijk} \mid \theta^{(t-1)}\right)},\end{aligned} \tag{3}$$

where N_{mn} was defined above and the summation is over all these protein pairs.

We update the parameter estimates of the λ_{mn} by iterating between the E-step and the M-step until the value of the likelihood function and the MLEs of the λ_{mn} for all the domain pairs converge. The estimated values of the λ_{mn} allow us to compute $\Pr(P_{ijk} = 1)$ and $\Pr(O_{ijk} = 1)$ by equations (1) and (2), respectively.

3.3 *Semi-Bayesian Approach*

Unlike the likelihood-based approach reviewed above, in our semi-Bayesian approach, we treat f_n and f_p as unknown but within a reasonable range that can be established from biological knowledge. We estimate λ_{mn}, f_n, and f_p, simultaneously.

For a given $\theta^{(t-1)} = (\lambda_{mn}^{(t-1)},\ f_n^{(t-1)},\ f_p^{(t-1)})$ obtained from the $(t-1)$th EM iteration, we calculate $\lambda_{mn}^{(t)}$ by replacing

$f_n^{(t-1)}$ and $f_p^{(t-1)}$ in equation (3) and then compute $h_{ij}(\Lambda^{(t)})$. We obtain the expectation of the log likelihood for complete data and define it as $Q(\theta^{(t)})$, which is given in Web Appendix A. Finally, we find solutions, f_n and f_p, that satisfy the following equations:

$$\frac{\partial Q(\theta^{(t)})}{\partial f_n} = \sum_{ijk} \left[\frac{-O_{ijk} h_{ij}(\Lambda^{(t)})}{h_{ij}(\Lambda^{(t)})(1-f_n) + \{1-h_{ij}(\Lambda^{(t)})\} f_p} + \frac{(1-O_{ijk}) h_{ij}(\Lambda^{(t)})}{1-h_{ij}(\Lambda^{(t)})(1-f_n) - \{1-h_{ij}(\Lambda^{(t)})\} f_p} \right] = 0;$$

$$\frac{\partial Q(\theta^{(t)})}{\partial f_p} = \sum \left[\frac{O_{ijk}\{1-h_{ij}(\Lambda^{(t)})\}}{h_{ij}(\Lambda^{(t)})(1-f_n) + \{1-h_{ij}(\Lambda^{(t)})\} f_p} + \frac{-(1-O_{ijk})\{1-h_{ij}(\Lambda^{(t)})\}}{1-h_{ij}(\Lambda^{(t)})(1-f_n) - \{1-h_{ij}(\Lambda^{(t)})\} f_p} \right] = 0.$$

Because these are nonlinear equations of f_n and f_p, we can use the Newton–Raphson method which requires the inverse of the second derivatives of $Q(\theta^{(t)})$ with respect to f_n and f_p. However, if λ_{mn} goes to 0 or 1 for all m, n, the second derivatives are zero (see Web Appendix A). Hence, the inverse of second derivative does not exist.

To avoid this problem, we use a Bayesian approach and assume that $f_{n_k} \sim \text{Unif}[u_{n_k}, v_{n_k}]$ and $f_{p_k} \sim \text{Unif}[u_{p_k}, v_{p_k}]$. Define (P_{ij}^k) to be all pairs of proteins in organism k, where $k = 1, \ldots, K$. The posterior distributions of f_{n_k} and f_{p_k} are proportional to

$$\begin{aligned}[f_{n_k} \mid rest] &\propto L(O \mid f_{n_k}, f_{p_k}, \Lambda) f(f_{n_k} \mid \Lambda, f_{p_k}) f(f_{p_k} \mid \Lambda) \\ &\propto \prod_{(P_{ij}^k)} \left[h_{ij}(\Lambda)(1-f_{n_k}) + (1-h_{ij}(\Lambda)) f_{p_k} \right]^{o_{ijk}} \\ &\quad \times \left[1 - \left\{ h_{ij}(\Lambda)(1-f_{n_k}) + (1-h_{ij}(\Lambda)) f_{p_k} \right\} \right]^{1-o_{ijk}} \\ &\quad \times f(f_{n_k} \mid \Lambda, f_{p_k}); \end{aligned} \tag{4}$$

$$\begin{aligned}[f_{p_k} \mid rest] &\propto L(O \mid f_{n_k}, f_{p_k}, \Lambda) f\{f_{p_k} \mid \Lambda^{(ij)}, f_{n_k}\} f(f_{n_k} \mid \Lambda) \\ &\propto \prod_{(P_{ij}^k)} \left[h_{ij}(\Lambda)(1-f_{n_k}) + (1-h_{ij}(\Lambda)) f_{p_k} \right]^{o_{ijk}} \\ &\quad \times \left[1 - \left\{ h_{ij}(\Lambda)(1-f_{n_k}) + (1-h_{ij}(\Lambda)) f_{p_k} \right\} \right]^{1-o_{ijk}} \\ &\quad \times f(f_{p_k} \mid \Lambda, f_{n_k}). \end{aligned} \tag{5}$$

The posterior distributions of f_{n_k} and f_{p_k} are log-concave functions regardless of the choice of the prior distribution. The proof is given in Web Appendix B. We sample $[f_{n_k} \mid \lambda_{mn}^{(t-1)}, f_{p_k}]$ and $[f_{p_k} \mid \lambda_{mn}^{(t-1)}, f_{n_k}]$ using the adaptive rejection method (Gilks and Wild, 1992). We note that f_{n_k} and f_{p_k} are estimated iteratively through the Gibbs sampling instead of using the iterative formula from the Newton–Raphson method for the M-step in the EM algorithm.

To summarize, the semi-Bayesian approach works as follows: (1) Choose initial values for λ_{mn}, f_{n_k}, and f_{p_k}; (2) For given f_{n_k} and f_{p_k}, estimate λ_{mn} using the EM algorithm and repeat until the value of the likelihood function converges; (3) For given λ_{mn}, sample f_{n_k} and f_{p_k} from their posterior distributions, respectively and then estimate f_{n_k} and f_{p_k} using the posterior means; and (4) Repeat steps 2–3 until the values of the likelihood function, f_{n_k}, and f_{p_k} converge, simultaneously.

3.4 *Full Bayesian Approach*

In this subsection, we describe our full Bayesian approach. As for the semi-Bayesian method, we also treat f_{n_k} and f_{p_k} as unknown but are within a reasonable range that can be inferred using prior biological knowledge. We assume uniform distributions of f_{n_k} and f_{p_k}: $f_{n_k} \sim \text{Unif}[u_{n_k}, v_{n_k}]$ and $f_{p_k} \sim \text{unif}[u_{p_k}, v_{p_k}]$. Unlike the semi-Bayesian method, however, we also assume that λ_{mn} has a Beta prior distribution: $\lambda_{mn} \sim \text{Beta}(\alpha, \beta)$. The prior parameters are chosen to be proper but vague. We have varied (α, β) in our analysis without appreciable change in results. Recall that (P_{ij}^k) represents all protein pairs in organism k, where $k = 1, \ldots, K$.

The posterior distribution of λ_{mn} is proportional to

$$\begin{aligned}[\Lambda \mid rest] &\propto L(O \mid f_{n_k}, f_{p_k}, \Lambda) f(\Lambda \mid f_{n_k}, f_{p_k}) \\ &\propto \prod_{ijk} \left[h_{ij}(\Lambda)(1-f_{n_k}) + \{1-h_{ij}(\Lambda)\} f_{p_k} \right]^{o_{ijk}} \\ &\quad \times \left[1 - \left\{ h_{ij}(\Lambda)(1-f_{n_k}) + \{1-h_{ij}(\Lambda)\} f_{p_k} \right\} \right]^{1-o_{ijk}} \\ &\quad \times f(\Lambda \mid f_{n_k}, f_{p_k}). \end{aligned}$$

The posterior distributions of f_{n_k} and f_{p_k} are proportional to equations (4) and (5), respectively. The posterior distributions of λ_{mn}, f_{n_k}, and f_{p_k} have the following properties:

- (*P.1*) Under $h_{ij}^1(\Lambda)$, the posterior distribution of λ_{mn} is a log-concave function;
- (*P.2*) Under $h_{ij}^a(\Lambda)$, the posterior distribution of λ_{mn} is not a log-concave function but the behavior of the mean and tails is the same as that of $h_{ij}^1(\Lambda)$;
- (*P.3*) Under both $h_{ij}^1(\Lambda)$ and $h_{ij}^a(\Lambda)$, the posterior distributions of f_{n_k} and f_{p_k} are log-concave functions.

The proofs of these properties are given in Web Appendix B. In order to generate the posterior samples, we use the adaptive rejection sampling method (Gilks and Wild, 1992) and the adaptive rejection Metropolis sampling method (Gilks, Best, and Tan, 1995).

4. Comparisons of the Likelihood-Based Approach and the Bayesian Approach

Compared to the likelihood-based approach, we can prove that, for a wide range of interacting probability, DDI probability estimates from the Bayesian approach have smaller MSEs under the following special scenario: (S.1) f_n and f_p are zeros; and (S.2) for the domain pair of interest, say, D_m and D_n, a protein either has one copy of domain D_m, or one copy of domain D_n, or neither. Under (S.1)–(S.2), DDI probability between two domains can be estimated directly both for the likelihood-based approach and the Bayesian approach. The proof is given in Web Appendix C. In other cases when each

protein pair have a different set of domain pairs and are not independent of other protein pairs, the MLE and the Bayes estimates of PPI probabilities depend on other domain pairs. There is in general no closed form expression for these estimators. In addition, there is also no closed form expression using nonzero constant values of f_n and f_p.

In the more general case, we have conducted a simulation study to empirically compare the performance of the following different approaches:

- EM: a likelihood-based approach using (1) true f_n, f_p, and (2) $h^1_{ij}(\Lambda)$ for protein interaction;
- SemiBay: a semi-Bayesian approach using (1) the estimated $\hat{f}_n, \hat{f}_p$, and (2) $h^1_{ij}(\Lambda)$ for protein interaction;
- Bay: a full Bayesian approach using (1) the estimated $\hat{f}_n, \hat{f}_p$, and (2) $h^1_{ij}(\Lambda)$ for protein interaction;
- BayHa: a full Bayesian approach using (1) the estimated $\hat{f}_n, \hat{f}_p$, and (2) $h^a_{ij}(\Lambda)$ for protein interaction.

In this simulation study, we considered only one organism, that is, $K = 1$, as we expect similar results for multiple organisms. As for the number of domains, we considered two cases: one with 50 domains and the other with 100 domains. For each case, we considered PPIs involving 200 or 400 proteins. Because each protein has at least one domain, the number of domains in each protein was assumed to be $[1, 2, \ldots, 10]$ with probabilities $(\pi_1, \pi_2, \ldots, \pi_9, \pi_{10}) = (0.585, 0.276, 0.069, 0.046, 0.009, 0.009, 0.004, 0.001, 0.0003, 0.0005)$. These probabilities are based on the distribution of the number of domains in proteins in *S. cerevisiae*. We let the true f_n to be 0.8 and the true f_p to be 0.0003. The true DDI probabilities (λ_{mn}) were generated from two distributions: (a) $r_1\text{Beta}(2,\ 2 \times 10^7) + (1 - r_1)\text{Beta}(6,\ 2)$ with $r_1 = 0.9$ and (b) Beta(6, 2), which has mean 0.75. We note that Beta(2, 2×10^7) has mean 10^{-6}. The true PPI probabilities were generated from the model $h^1_{ij}(\Lambda) = 1 - \prod_{D^{(ij)}_{mn} \in P_{ij}} (1 - \lambda_{mn})$.

Under this setup, for each of the four combinations of the number of domains and the number of proteins, we simulated the observed interacting protein pair set $O = \{O_{ij} = o_{ij}; i \leq j\}$ using a Bernoulli(π) distribution, where $\pi = h_{ij}(\Lambda)(1 - f_n) + \{1 - h_{ij}(\Lambda)\}f_p$, $f_n = 0.8$, and $f_p = 0.0003$. We simulated 100 sets of observed interacting protein pairs.

In the likelihood-based approach, we assumed that the true f_n and f_p were known in the estimation of DDI probabilities. In our semi-Bayesian method and full Bayesian method, a burn-in time of 1000 iterations was followed by 100 iterations from the posterior distribution.

We computed the MSE of λ_{mn} for each domain pair and then calculated the average MSE. We also computed the MSE of $h_{ij}(\Lambda)$ at the protein level. The values of the average MSE are given in Tables 1–2 and Web Tables 1–3. For all cases, the full Bayesian method had the best performance. We also note that although we generated true PPI probabilities using $h^1_{ij}(\Lambda)$, not $h^a_{ij}(\Lambda)$, the full Bayesian approach with $h^a_{ij}(\Lambda)$ was as efficient as that based on the $h^1_{ij}(\Lambda)$. This may be because about 80% of the proteins had one or two domains and we only considered at most 400 proteins. Hence a is close to 1 in many cases. Therefore, $h^1_{ij}(\Lambda)$ is close to $h^a_{ij}(\Lambda)$ in this case.

From Table 2 and Web Table 2, we observe that the estimated values of f_n and f_p obtained from the semi-Bayesian method and the full Bayesian method are quite accurate. We also note from Web Table 3 that the full Bayesian method became more efficient than the other methods as the mixture proportion (r_1) approached to 0.5.

Furthermore, we calculated receiver operating characteristic (ROC) curves for the simulated data sets and then computed the average ROC curve. We treat the two proteins as interacting if the interaction probability is greater than 0.6, which is within 1 standard deviation of the mean of the Beta(6, 2) distribution. The average ROC curves based on two distributions of λ_{mn} with 50 domains and 400 proteins are given in Figure 1 and Web Figure 1. The results were

Table 1

The average MSE values of DDI and PPI probabilities for each method. In this simulation, the true DDI probability (λ_{mn}) was generated from $0.9\text{Beta}(2,\ 2 \times 10^7) + 0.1\text{Beta}(6,\ 2)$. *The true PPI probabilities were generated from the model* $h^1_{ij}(\Lambda) = 1 - \prod_{D^{(ij)}_{mn} \in P_{ij}} (1 - \lambda_{mn})$. *EM = the likelihood-based approach using true f_n, f_p, and h^1_{ij} for PPIs; SemiBay = the semi-Bayesian approach using the estimated $\hat{f}_n, \hat{f}_p$, and h^1_{ij} for PPIs; Bay = the full Bayesian approach using the estimated $\hat{f}_n, \hat{f}_p$, and h^1_{ij} for PPIs; BayHa = the full Bayesian approach using the estimated $\hat{f}_n, \hat{f}_p$, and h^a_{ij} for PPIs.*

Number of domains	Number of proteins		Domain Level				Protein Level			
			EM	Semi-Bay	Bay	BayHa	EM	Semi-Bay	Bay	BayHa
50	200	MSE	0.022	0.024	0.014	0.015	0.113	0.118	0.081	0.082
		Var	0.010	0.010	0.007	0.007	0.095	0.100	0.069	0.067
		Bias^2	0.012	0.014	0.007	0.008	0.018	0.018	0.013	0.015
	400	MSE	0.004	0.004	0.003	0.003	0.084	0.085	0.053	0.053
		Var	0.002	0.001	0.001	0.001	0.078	0.079	0.049	0.048
		Bias^2	0.002	0.003	0.002	0.002	0.006	0.006	0.004	0.005
100	200	MSE	0.072	0.069	0.041	0.042	0.165	0.164	0.132	0.133
		Var	0.029	0.027	0.015	0.015	0.143	0.143	0.114	0.114
		Bias^2	0.043	0.042	0.026	0.027	0.022	0.021	0.018	0.019
	400	MSE	0.023	0.024	0.014	0.015	0.113	0.118	0.081	0.082
		Var	0.010	0.010	0.007	0.007	0.095	0.099	0.068	0.067
		Bias^2	0.013	0.014	0.007	0.008	0.018	0.019	0.013	0.015

Table 2

The average MSE values of f_n and f_p for each method. In this simulation, the true DDI probability (λ_{mn}) was generated from $0.9\text{Beta}(2,\ 2 \times 10^7) + 0.1\text{Beta}(6,\ 2)$, *respectively. The true PPI probabilities were generated from the model* $h_{ij}^1(\Lambda) = 1 - \prod_{D_{mn}^{(ij)} \in P_{ij}} (1 - \lambda_{mn})$. *SemiBay = the semi-Bayesian approach using the estimated $\hat{f}_n, \hat{f}_p$, and h_{ij}^1 for PPIs; Bay = the full Bayesian approach using the estimated $\hat{f}_n, \hat{f}_p$, and h_{ij}^1 for PPIs; BayHa = the full Bayesian approach using the estimated $\hat{f}_n, \hat{f}_p$, and h_{ij}^a for PPIs.*

Number of domains	Number of proteins		SemiBay f_n	SemiBay f_p	Bay f_n	Bay f_p	BayHa f_n	BayHa f_p
50	200	Mean	0.756	0.00056	0.766	0.00055	0.768	0.00054
		MSE	2.00e-3	6.76e-8	1.27e-3	6.51e-8	1.13e-3	6.32e-8
		Var	1.36e-4	2.10e-9	1.16e-4	3.86e-9	1.12e-4	3.54e-9
		Bias^2	1.87e-3	6.55e-8	1.15e-3	6.12e-8	1.02e-3	5.97e-8
50	400	Mean	0.786	0.00035	0.793	0.00036	0.801	0.00035
		MSE	2.96e-4	2.53e-9	2.67e-4	2.91e-9	2.66e-4	2.54e-9
		Var	2.78e-4	2.99e-10	2.48e-4	2.80e-10	2.48e-4	2.21e-10
		Bias^2	1.85e-5	2.23e-9	1.83e-5	2.63e-9	1.80e-5	2.32e-9

similar for the other cases. In Figure 1 and Web Figure 1, the average ROC curves obtained from the full Bayesian method are higher than others. The average ROC curve of the semi-Bayesian method was similar to that of the EM approach when true f_n and f_p values were used. However, true f_n and f_p are unknown in practice. Therefore, the semi-Bayesian approach is likely to be more useful than the likelihood-based approach.

In addition, we conducted further simulation with true PPI probabilities defined by $h_{ij}^a(\Lambda) = 1 - \prod_{D_{mn}^{(ij)} \in P_{ij}} (1 - \lambda_{mn}^a)$, where $a = \log\{1 - (1/2)^{1/M_{ij}}\}/\log(1/2)$. When the number of domains was 50, the number of proteins was 400, and the true DDI probabilities (λ_{mn}) were generated from (a) $r_1\text{Beta}(2,\ 2 \times 10^7) + (1 - r_1)\text{Beta}(6,\ 2)$ with $r_1 = 0.9$, the MSEs of EM, semi-Bayesian, full Bayesian methods with $h_{ij}^1(\Lambda)$, and full Bayesian method with $h_{ij}^a(\Lambda)$ were about 0.009, 0.009, 0.006,

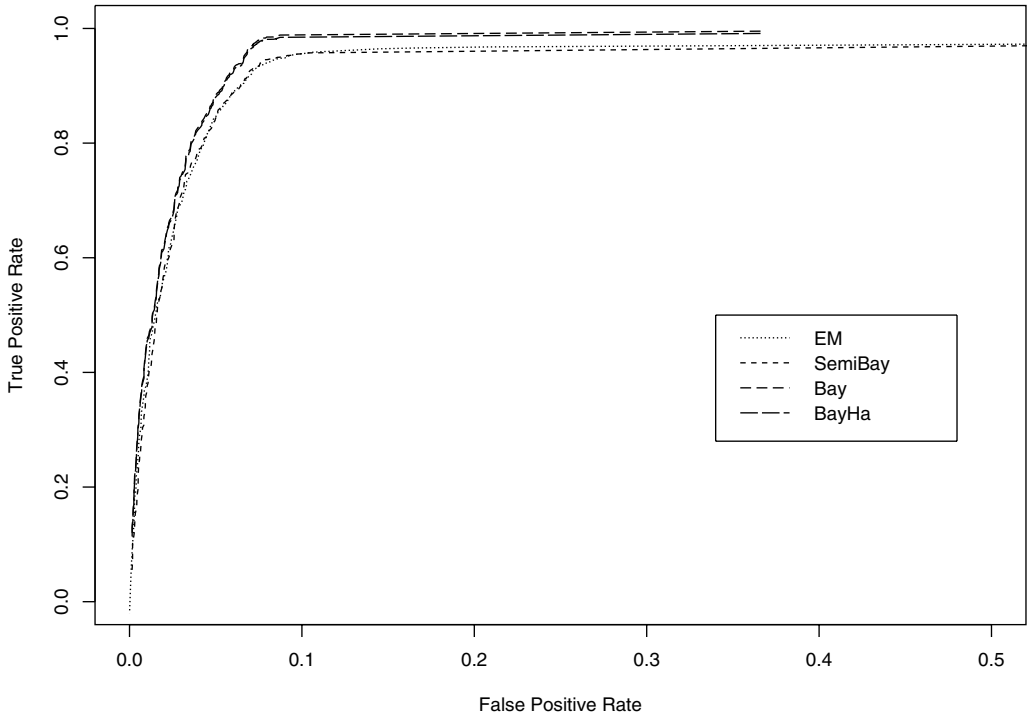

Figure 1. The average ROC curves for the likelihood based approach, the Semi-Bayesian method, and the full Bayesian methods. The number of domains is 50 and the number of proteins is 400. The distribution of DDI probability is $\lambda_{mn} \sim 0.9\text{Beta}(2, 2 \times 10^7) + 0.1\text{Beta}(6, 2)$. The true PPI probabilities were generated from the model $h_{ij}^1(\Lambda) = 1 - \prod_{D_{mn}^{(ij)} \in P_{ij}} (1 - \lambda_{mn})$. EM = the likelihood based approach using true f_n, f_p, and h_{ij}^1 for PPIs; SemiBay = the semi-Bayesian approach using the estimated $\hat{f}_n$, $\hat{f}_p$, and h_{ij}^1 for PPIs; Bay = the full Bayesian approach using the estimated $\hat{f}_n, \hat{f}_p$, and h_{ij}^1 for PPIs; BayHa = the full Bayesian approach using the estimated $\hat{f}_n, \hat{f}_p$, and h_{ij}^a for PPIs.

and 0.003 at the domain level, respectively. The MSEs of EM, semi-Bayesian, and full Bayesian with $h^1_{ij}(\Lambda)$ were about 2.25 times worse than those when the true PPIs were generated using $h^1_{ij}(\Lambda)$. The MSE of the full Bayesian approach with $h^1_{ij}(\Lambda)$ was smaller than EM and semi-Bayesian methods but it was about two times worse than that of the full Bayesian approach with $h^a_{ij}(\Lambda)$. However, the MSE of the full Bayesian approach with $h^a_{ij}(\Lambda)$ was similar to the case when true PPIs were generated using $h^1_{ij}(\Lambda)$. When the true DDI probabilities (λ_{mn}) were generated from (b) Beta(6, 2), the results were similar. The simulation results for both (a) and (b) are given in Web Tables 4–5.

5. Example

We use large-scale PPI data from three organisms, *S. cerevisiae*, *C. elegans*, and *D. melanogaster*, obtained from high-throughput yeast two-hybrid experiments, to infer DDI probabilities. For *S. cerevisiae*, we used 5295 interactions which came from two independent studies (Ito et al., 2000; Uetz et al., 2000). For *C. elegans* and *D. melanogaster*, 4714 and 20,349 interactions were used from yeast two-hybrid experiments (Giot et al., 2003; Li et al., 2004), respectively. The protein-domain relationships for each protein in *S. cerevisiae*, *C. elegans*, and *D. melanogaster* were obtained from Pfam (Bateman et al., 2004) and SMART (Letunic et al., 2004).

5.1 *Estimation Procedure*

Using these data sets, we first estimated DDI probabilities. For the likelihood-based approach, we used $f_n = 0.8$ and $f_p = 0.0003$ that were used by Deng et al. (2002) and Liu et al. (2005) based on prior biological knowledge. The estimated DDI probabilities were then used to infer PPI probabilities in *S. cerevisiae*. For the full Bayesian method, we considered the following four cases of prior and model specification:

- Case 1: $h_{ij}(\Lambda) = h^1_{ij}(\Lambda)$, with prior distributions $f_n \sim$ Unif(0, 1), $f_p \sim$ Unif(0, 1), and $\lambda_{mn} \sim$ Unif(0, 1);
- Case 2: The same setting as case 1 except $\lambda_{mn} \sim$ Beta(2, 2);
- Case 3: $h_{ij}(\Lambda) = h^1_{ij}(\Lambda)$, with prior distributions $f_{n_k} \sim$ Unif$(u_k, 1)$, $f_{p_k} \sim$ Unif$(0, v_k)$, $\lambda_{mn} \sim$ Beta$(2, 2)$, $u_k \sim$ Unif$(0, 0.3)$, and $v_k \sim$ Unif(0.5, 1);
- Case 4: The same setting as case 3 except $h_{ij}(\Lambda) = h^a_{ij}(\Lambda)$.

The estimated false negative rate and false positive rate from the semi-Bayesian method are 0.898 and 0.000026, respectively. The estimated values from Bayesian method with case 1(2) are 0.920 (0.920) and 0.000025 (0.000024), respectively. We also estimated these rates from Bayesian method with case 3(4). The estimated false negative and false positive rates of the data set for *S. cerevisiae* are 0.912(0.901) and 0.000026(0.000027), respectively. These values for *C. elegans* are 0.950(0.946) and 0.000018(0.000020), respectively. The estimated rates for *D. melanogaster* are 0.931(0.929) and 0.000031(0.000033), respectively. These values are different from the fixed values of Deng et al. (2002) and Liu et al. (2005).

In our semi-Bayesian method, we first chose the initial parameter values as those from the likelihood-based approach. For our full Bayesian method, we chose the initial parameter values as those from the semi-Bayesian method. We used the adaptive rejection sampling and adaptive metropolis rejection sampling by Gilks and Wild (1992) and Gilks et al. (1995) in order to generate samples from the posterior distributions. A burn-in time of 1000 iterations was followed by 100 iterations from the posterior distributions. The Markov chain Monte Carlo trace plots of the sampled false negative rates and false positive rates of three organisms, and the Markov chain Monte Carlo trace plots of samples of six randomly chosen DDI probabilities in full Bayesian method with case 4 are given in Web Figure 3.

The likelihood-based approach, the semi-Bayesian method, and the full Bayesian method took about 1 day, 2 days, and 7 days to run, respectively, using one node in a Linux cluster. Each node has 2 CPUs, 3.2 GHz Xenon, and 2 GB memory. Our code is written in `C++` and is available upon request from the authors.

5.2 *Comparison with MIPS Protein Interaction Database*

We selected top 1000 predicted interacting protein pairs from the likelihood-based approach and top 1000 pairs from the full Bayesian methods and then compared them with 3543 experimentally verified physical interactions in *S. cerevisiae* at the Munich Information Center for Protein Sequences (MIPS; `http://mips.gsf.de/genre/proj/yeast/`). The Venn diagrams on the overlap patterns among three protein pair sets (MIPS, the likelihood-based approach, and the full Bayesian methods) are shown in Web Figure 2. We can observe that the number of common protein pairs between MIPS and the full Bayesian approaches is larger than that between MIPS and the likelihood-based approach. Furthermore, the number of common protein pairs between MIPS and the full Bayesian approach with case 4 is the largest among those from MIPS and the full Bayesian approach with other cases.

We also compared the ROC curves among the likelihood-based approach, the semi-Bayesian method and the full Bayesian methods. For the ROC curves, we considered the 3543 yeast physical interaction pairs in MIPS as positive pairs and the other possible protein pairs, 6,895,215 pairs, as negative pairs. We calculated the true positive rate and the false positive rate for different thresholds. The true positive rate was calculated as the number of predicted protein pairs that were included in the positive pairs divided by 3543, the total number of positives, and the false positive rate was calculated as the number of predicted protein pairs that were included in the negative pairs divided by 6,895,215, the total number of negatives. The ROC curves are shown in Figure 2. We note that the interacting protein pairs we use as the gold standard are incomplete because the number of protein interactions verified by experiments is very limited. As the number of annotated interactions increases, the values of the false positive rate and false negative rate will certainly change.

Based on a thorough comparison of true positive rate and false negative rate, we conclude that the semi-Bayesian and full Bayesian approaches may be more effective for this data set in borrowing information from multiple organisms than the likelihood-based approach by Liu et al. (2005). We also note that the full Bayesian approach based on cases 3 and 4 can further improve PPI predictions. This is likely achieved by allowing different organisms having different false negative and false positive rates and using our model $h^a_{ij}(\Lambda)$ to relate PPI probabilities with DDI probabilities.

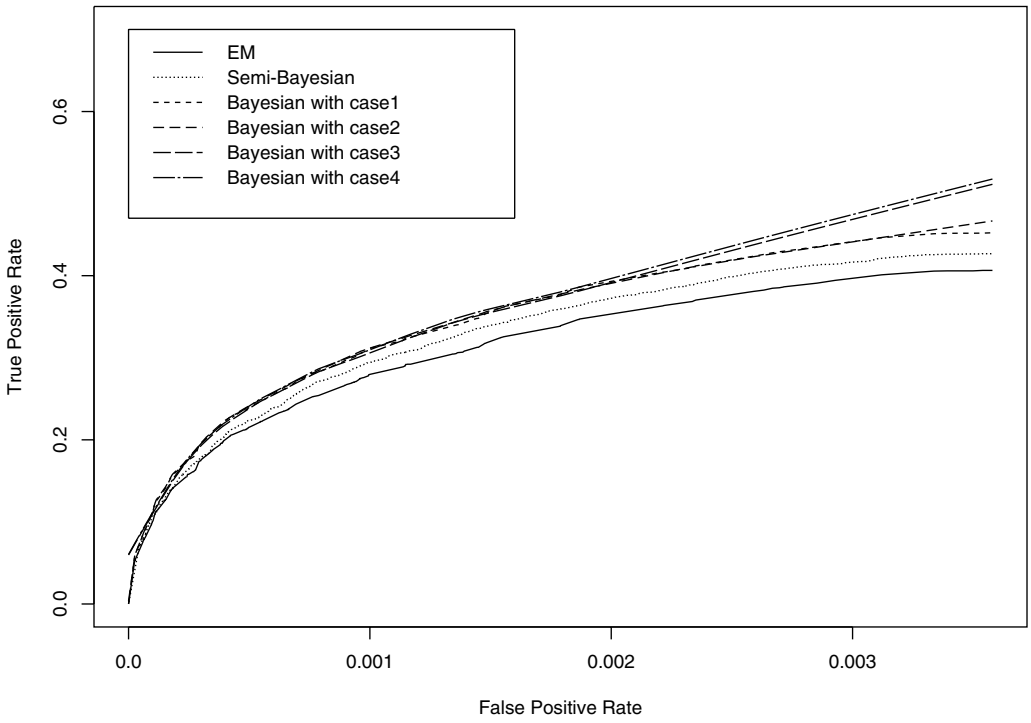

Figure 2. ROC curves for the likelihood based approach, the semi-Bayesian method, and the full Bayesian methods. We use MIPS protein interactions as the gold standard. EM = the likelihood based approach using $f_n = 0.8$ and $f_p = 0.0003$; Semi-Bayesian = the semi-Bayesian method using the estimated f_n and f_p; Bayesian = the full Bayesian method with the estimated organism specific $\hat{f}_{n_k}$ and $\hat{f}_{p_k}$.

5.3 *Comparison with iPfam Domain Interaction Database*

To verify DDI predictions from the likelihood-based approach, the semi-Bayesian method, and the full Bayesian methods, we compare our predictions with iPfam directly. (`http://www.sanger.ac.uk/Software/Pfam/iPfam/`). We only use a single data source, yeast two-hybrid data. There are 2450 experimentally verified physical domain interactions in *S. cerevisiae* at iPfam. We considered the 2450 yeast physical domain interaction pairs in iPfam as positive pairs and the other possible domain pairs, 3,877,055 pairs, as negative pairs. We compared ROC curves among the likelihood-based approach, the semi-Bayesian method, and the full Bayesian methods (see Web Figure 4). Based on a thorough comparison of true positive rate and false negative rate, the semi-Bayesian and full Bayesian approaches are more effective than the likelihood-based approach for this data set.

6. Discussion

In this article, we have developed a semi-Bayesian method and a full Bayesian method for simultaneously estimating DDI probabilities, the false positive rate, and the false negative rate, from high-throughput yeast two-hybrid data. We have also proposed a model to relate protein interaction probabilities with domain interaction probabilities that more appropriately models the number of domains in each protein. We use this model to infer PPI probabilities using the estimated DDI probabilities. Compared to previous methods by Deng et al. (2002) and Liu et al. (2005), our methods may be more efficient in dealing with a large number of parameters using some prior information, more effective to allow for different false positive and false negative rates across different data sets, and more appropriate when the proportion of proteins having more than three domains increases.

Our simulation study suggests that the semi-Bayesian method is as good as the likelihood-based approach in the best case for the latter approach, that is, when f_n and f_p are assumed to be known, and our full Bayesian method performs better than the likelihood-based approach even in this case. We also have showed this result through theoretical justification under a simplified scenario.

We note that our model $h^a_{ij}(\Lambda)$ is only one possible function to relate protein interaction probabilities with domain interaction probabilities that reflects the number of domains in each protein. This function seems to be working out well in our study. However, other functions may be even more effective and they may be estimated either parametrically or nonparametrically. Future research on identifying other models is warranted.

Our methods are based on the case when the experimentally observed interactions $O = \{O_{ijk} = o_{ijk};\ i \leq j\}$ are binary data. We can extend our methods to the continuous case, for example, a confidence score for each protein pair, similar in spirit to Bader et al. (2004). In this case, it is important to estimate appropriate distributions either parametrically or nonparametrically, which are related to the distribution of confidence scores. One possible distribution for the score values is a mixture of two distributions, with one distribution representing interacting protein pairs and the other corresponding to the other protein pairs.

In our study, we have assumed that the domains are independent of each other. However, domains may be classified into different super families based on structural or functional evidence for a common evolutionary ancestor (Gough and Chothis, 2002). Domains within the same super family would have similar interaction profiles due to the similarity of their structures. We can further develop our Bayesian approach to handle this dependence structure.

In our example considered in this article, we only used yeast two-hybrid data to make predictions. We may further improve our predictions by integrating multiple data sources, for example, gene expression data and gene ontology information (Lee et al., 2004, 2006).

Last but not the least, although many of the predicted pairs are included in the MIPS database in our example, some of them with high interaction probabilities are not in the MIPS database. These predictions need to be validated experimentally as the ultimate test of our proposal methods.

7. Supplementary Materials

Web Appendices, tables, and figures referenced in Sections 3.3–3.4, Section 4, Section 5.2, Section 5.3, and Section 6 are available under the Paper Information link at the Biometrics website `http://www.tibs.org/biometrics`.

Acknowledgements

This study was supported in part by the National Science Foundation grant DMS0241160. We thank the associate editor and referee for their many helpful comments.

References

Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A. C., Bork, P., Superti-Furga, G., Serrano, L., and Russell, R. B. (2004). Structure-based assembly of protein complexes in yeast. *Science* **302,** 2026–2029.

Bader, J. S., Chaudhuri, A., Rothergm, J. M., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology* **22,** 78–85.

Bateman, A., Coin, L., Durbin, R., et al. (2004). The Pfam protein families database. *Nucleic Acids Research* **32,** D138–D141.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likeliood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39,** 1–38.

Deng, M., Mehta, S., Sun, F., and Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Research* **12,** 1504–1508.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Quzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402,** 86–90.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* **41,** 337–348.

Gilks, W. R., Best, N. G., and Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling. *Applied Statistics* **44,** 455–472.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., et al. (2003). A protein interaction map of *Drosophila melanogaster*. *Science* **302,** 1727–1736.

Goh, C. S. and Cohen, F. E. (2002). Co-evolutionary analysis reveals insights into protein-protein interactions. *Journal of Molecular Biology* **324,** 177–179.

Gomez, S. M., Lo, S. H., and Rzhetsky, A. (2001). Probabilistic prediction of unknown metabolic and signal-transduction networks. *Genetics* **159,** 1291–1298.

Gomez, S. M., Noble, W. S., and Rzhetsky, A. (2003). Learning to predict protein-protein interactions from protein sequences. *Bioinformatics* **19,** 1875–1881.

Gough, J. and Chothis, C. (2002). SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Research* **30,** 268–272.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America* **98,** 4569–4574.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302,** 449–453.

Lee, H., Deng, M., Sun, F., and Chen, T. (2006). An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics* **7,** 269.

Lee, I., Date, S., Adai, A., and Marcotte, E. (2004). A probabilistic functional network of yeast genes. *Science* **306,** 1555–1558.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P., and Bork, P. (2004). SMART 4.0: Towards genomic data integration. *Nucleic Acids Research* **32,** D142–D144.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., et al. (2004). A map of the interactome network of the metazoan *C. elegans*. *Science* **303,** 540–543.

Liu, Y., Liu, N., and Zhao, H. (2005). Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics* **21,** 3279–3285.

Lu, L., Arakaki, A. K., Lu, H., and Skolnick, J. (2003). Multimeric threading-based prediction of protein-protein interactions on a genomic scale: Application to the *Saccharomyces cerevisiae* proteome. *Genome Research* **13,** 1146–1154.

Marcotte, E. M., Xenarios, I., and Eisenberg, D. (2001). Mining literature for protein-protein interactions. *Bioinformatics* **17,** 357–363.

Mrowka, R., Patzak, A., and Herzel, H. (2001). Is there a bias in proteome research? *Genome Research* **11,** 1971–1973.

Pawson, T. and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science* **300,** 445–452.

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering* **14,** 609–614.

Ramani, A. K. and Marcotte, E. M. (2003). Exploiting the co-evolution of interacting proteins to discover interaction

specificity. *Journal of Molecular Biology* **327,** 273–284.

Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology* **311,** 681–692.

Tsoka, S. and Ouzounis, C. A. (2000). Prediction of protein interactions: Metabolic enzymes are frequently involved in gene fusion. *Nature Genetics* **26,** 141–142.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., et al. (2000). A comprehensive analysis of protein-protein interaction in *Saccharomyces cerevisiae*. *Nature* **403,** 623–627.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417,** 399–403.

Received April 2006. *Revised August* 2006.
Accepted October 2006.