Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 5313050, 7 pages
http://dx.doi.org/10.1155/2016/5313050

Research Article

Statistical Approaches for the Construction and Interpretation of Human Protein-Protein Interaction Network

Yang Hu,[1] Ying Zhang,[2] Jun Ren,[1] Yadong Wang,[3] Zhenzhen Wang,[4] and Jun Zhang[2]

[1] *School of Life Science and Technology, Harbin Institute of Technology, Harbin 150001, China*
[2] *Department of Pharmacy, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin 150088, China*
[3] *School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China*
[4] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China*

Correspondence should be addressed to Yadong Wang; ydwang@hit.edu.cn, Zhenzhen Wang; wangzz@ems.hrbmu.edu.cn, and Jun Zhang; zhangjun13902003@163.com

Received 3 June 2016; Revised 23 July 2016; Accepted 1 August 2016

Academic Editor: Qin Ma

The overall goal is to establish a reliable human protein-protein interaction network and develop computational tools to characterize a protein-protein interaction (PPI) network and the role of individual proteins in the context of the network topology and their expression status. A novel and unique feature of our approach is that we assigned confidence measure to each derived interacting pair and account for the confidence in our network analysis. We integrated experimental data to infer human PPI network. Our model treated the true interacting status (yes versus no) for any given pair of human proteins as a latent variable whose value was not observed. The experimental data were the manifestation of interacting status, which provided evidence as to the likelihood of the interaction. The confidence of interactions would depend on the strength and consistency of the evidence.

1. Introduction

Individual proteins cannot perform their biological functions by themselves, and actually they need to perform their functions in the biological process through interacting with other proteins [1]. Usually the interaction between two proteins means either they perform a biological function corporately or there is physical direct contact between them [2]. Most of the important molecular processes in cell, such as DNA replication, need to be performed by a large number of protein complexes. And these complexes are made up by the interactions between proteins. The study of PPIs is also considered to be a central problem in proteomics for living cells. Due to the dynamic interaction between proteins, the impact of surrounding environment should also be taken into account. The study of human PPI network can help to enhance the understanding of the disease but also provide a theoretical foundation for finding new treatment.

With the continuous progress and development of high-throughput experimental technology, more and more large quantities of interactions between human proteins had been confirmed by a variety of experimental methods. And many kinds of biological interaction networks have been investigated [3–7]. However, current high-throughput experimental techniques also indicated the shortcomings of high error; not only might the different experimental methods induce different experimental results, but also even different research groups using the same experimental method could not guarantee the exact same result. Therefore, it was urgent to integrate the data from different biological experiments, and even different species, to construct a highly credible network of PPIs. So in this paper, a Bayesian hierarchical model of human PPI network was constructed with a variety of sources of protein interaction data. Meanwhile, a Monte Carlo expectation maximization algorithm was used to estimate the parameters of the model. Then the confidence of protein interaction relationship was calculated based on Bayesian model, and human PPI network with high-confidence level could be obtained.

Thereafter, the role of intrinsic disordered proteins (IDPs) was investigated in the high-confidence PPI network. First of all, different functional modules were obtained through

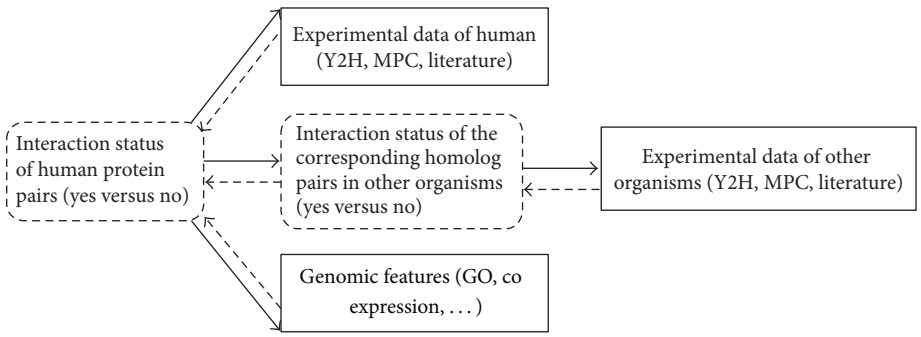

Figure 1: Overall scheme to construct the human protein-protein interaction network. The interaction status of a given pair of human proteins and their homolog in other organisms are unobserved (dashed box) and the experimental data and genomic features are observed evidence (solid boxes). Solid arrows represent model hierarchy and dashed arrows represent inference steps.

Table 1: Data sets or databases used to construct the human protein-protein interaction network.

Method	Organism	Reference
Y2H	Human	Stelzl et al. [8]
Y2H	Human	Rual et al. [9]
MPC	Human	Ewing et al. [10]
Literature	Human	HPRD [11], http://www.hprd.org/
Y2H	Yeast	Ito et al. [12]
Y2H	Yeast	Uetz et al. [13]
MPC	Yeast	Gavin et al. [14]
MPC	Yeast	Ho et al. [15]
MPC	Yeast	Gavin et al. [16]
MPC	Yeast	Krogan et al. [17]
Literature	Multiple	IntAct [18], http://www.ebi.ac.uk/intact/
Literature	Multiple	MIPS [19], http://mips.gsf.de/proj/ppi/
Multiple	Multiple	DIP [20], http://dip.doe-mbi.ucla.edu

clustering of high-confidence PPI network based on the network topology structure. Then we found the functional modules which were significantly correlated with intrinsically disordered proteins and analysed the effect of IDPs in these functional modules, while searching for the associations between these functional modules and diseases.

2. Materials and Methods

2.1. Data Collection. In Table 1, we show the experimental data that will be used for the construction of the human PPI network [8–20]. Note that the literature or text mining approach represents most of the low-throughput experimental studies of individual protein-protein interaction. It is possible that the result from the same experiment will be recorded in multiple databases. We will eliminate this type of redundancy. It should be emphasized that the MPC experiments provide result in the format of protein complexes instead of pair-wise protein-protein interactions. Since proteins located in the same complex might not interact with one another directly, we will account for this factor in our model.

2.2. Statistical Modeling of Various Data Sources. The overall scheme of our approach is illustrated in Figure 1. We consider an empirical Bayes approach to integrate various sources of evidence. Let Z_{ij} be the binary indicator such that $Z_{ij} = 1$ means that human proteins i and j have a direct physical interaction and it is 0 otherwise. Hence, Z_{ij} is the true interacting status that is not observed. To infer Z_{ij}, we consider individual model for each type of observed data and integrate the evidence to compute the probability of $Z_{ij} = 1$.

2.2.1. Human Y2H Data. It has been found that there are a number of mechanisms that can lead to the expression of the reporter gene in a Y2H experiment, which means that an observed interaction might not necessarily mean a true interaction. In our model, we consider the following mechanisms: (a) true interaction; (b) self-activation; and (c) unknown process. Let Y_{ij} be the binary indicator such that $Y_{ij} = 1$ if proteins i and j are observed to interact in a Y2H experiment and it is 0 otherwise. Then $Y_{ij} = 1$ only if at least one of the three above mechanisms is functional. Let $X_i = 1$ if protein i is a self-activation protein and let it be 0 otherwise. We define

$$\alpha_I = \Pr\left[a \text{ is functional} \mid Z_{ij} = 1\right], \tag{1}$$

$$\alpha_S = \Pr\left[b \text{ is functional} \mid X_i + X_j > 0\right], \tag{2}$$

$$\alpha_U = \Pr\left[c \text{ is functional}\right]. \tag{3}$$

Then we have

$$\begin{aligned}&\Pr\left[Y_{ij} = 1 \mid Z, X\right]\\ &\quad = 1 - \left(1 - \alpha_I\right)^{Z_{ij}} \left(1 - \alpha_S\right)^{X_i + X_j} \left(1 - \alpha_U\right).\end{aligned} \tag{4}$$

2.2.2. Human MPC Data. MPC experiment reveals protein complexes instead of individual pairwise PPI. We say protein B is an n-step neighbour of protein A if the shortest path between A and B in the PPI network is of length n. We conjecture that the bait will mostly fish out its 1-step neighbours, and 2-step neighbours and distant proteins (at least three

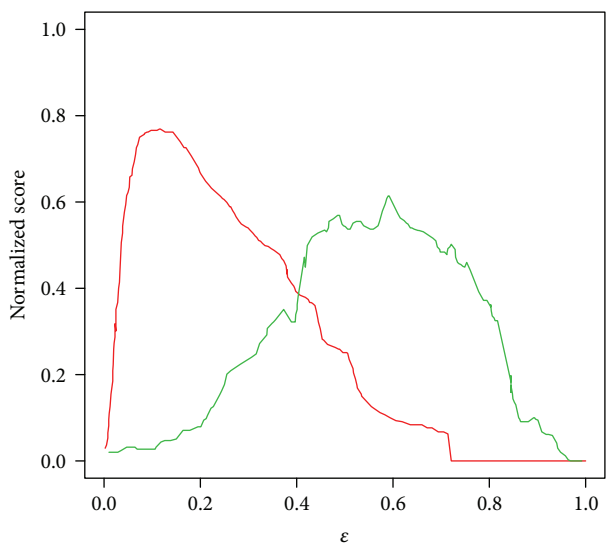

FIGURE 2: The optimization of Q_N and Q_S for different ε. Red line and green line correspond to Q_N and Q_S separately.

step-away) are occasionally observed. Hence, we define the following parameters for the bait proteins:

$$\begin{aligned} &\Pr\left[\text{1-step neighbour is observed}\right] = \psi_1, \\ &\Pr\left[\text{2-step neighbour is observed}\right] = \psi_2. \end{aligned} \tag{5}$$

Let C_k be the set of proteins in a complex corresponding to bait protein k. Denote by $n_k^{(1)}$, $n_k^{(2)}$ the set of 1-step and 2-step neighbours of the bait protein k under a given value of Z. Then the probability of observing C_k can be written as follows:

$$\begin{aligned} &\Pr\left[C_k \mid Z\right] \\ &\quad = \psi_1^{|n_k^{(1)} \cap C_k|} \left(1-\psi_1\right)^{|n_k^{(1)} \setminus C_k|} \psi_2^{|n_k^{(2)} \cap C_k|} \left(1-\psi_2\right)^{|n_k^{(2)} \setminus C_k|}, \end{aligned} \tag{6}$$

where $|\cdot|$ is the function that maps a set to its size.

2.2.3. Literature Data on Human PPI. Let L_{ij} be the interaction status of proteins i and j reported. We will account for the false positive rate ($\gamma_{0,k}$) and false negative rate ($\gamma_{0,k}$):

$$\Pr\left[H_{ij} = 1 \mid Z_{ij}\right] = \gamma_1^{Z_{ij}} \gamma_0^{1-Z_{ij}}. \tag{7}$$

2.2.4. Data from Other Organisms. We will also collect (Y^*, C^*) from other organisms with corresponding unobserved variables denoted by (Z^*, X^*). Similar models can be used to model (Y, C, L) for inference of (Z^*, X^*). To connect (Z^*, X^*) to (Z, X), we consider the following models:

$$\begin{aligned} &\Pr\left[Z_{i'j'}^* = 1 \mid Z_{ij}\right] \\ &\quad = \left[\Delta_1\left(J_{ii',jj'};\phi_1\right)\right]^{Z_{ij}} \left[\Delta_0\left(J_{ii',jj'};\phi_0\right)\right]^{1-Z_{ij}}, \\ &\Pr\left[X_{i'}^* = 1 \mid X_i\right] = \left[\Omega_1\left(I_{ii'};\lambda_1\right)\right]^{X_i} \left[\Omega_0\left(I_{ii'};\lambda_0\right)\right]^{1-X_i}, \end{aligned} \tag{8}$$

where $J_{ii',jj'}$ is the joint sequence identity between i and i' and between j and j' and $I_{ii'}$ is sequence identity between i and i'; Δ_1, Δ_0, Ω_1, and Ω_0 are functions of the joint or individual sequence identities with parameters ϕ_1, ϕ_0, λ_1, and λ_0, which can be modeled by parametric structure.

2.3. Construction of Hierarchical Bayesian Model. So far we have introduced the distribution models for the experimental data and genomic features that are conditional on the values of Z and X. To finish the model, we also need to specify the distributions of Z and X, which can be modeled with independent Bernoulli distributions:

$$\begin{aligned} \Pr\left(Z_{ij} = 1\right) &= \rho, \\ \Pr\left(X_i = 1\right) &= r. \end{aligned} \tag{9}$$

With the observed data and the unobserved variables, we can infer the posterior probability of Z using the EM algorithm. Note that there are multiple organisms and multiple data sets for some of the organisms. Different parameters will be used to account for difference in the data.

As illustrated in (10), the complete log likelihood function of our model can be expanded below, and the factor of (10) can be substituted by (3)~(9):

$$\begin{aligned} L_C(\Psi) &= f\left(H, Y, W, Z, X, L, Y^*, W^*, Z^*, X^* \mid \theta\right) = f\left(H \mid Z, \theta\right) f\left(Y \mid Z, X, \theta\right) f\left(W \mid Z, \theta\right) f\left(L \mid Z, \theta\right) \\ &\cdot f\left(Y^* \mid Z^*, X^*, \theta\right) f\left(W^* \mid Z^*, \theta\right) f\left(Z^*, X^* \mid Z, X, \theta\right) f\left(Z, X \mid \theta\right) = \prod_{(i,j)\in S_H} f\left(H_{ij} \mid Z_{ij}, \theta\right) \\ &\cdot \prod_{(i,j)\in S_Y} \left[\prod_{t=1}^{r_{ij}} f\left(Y_{ij}^{(t)} \mid Z_{ij}, X_i, X_j, \theta\right)\right] \prod_{(i,j)\in S_M} \left[\prod_{t=1}^{e_{ij}} f\left(W_{ij}^{i(t)} \mid Z_{ij}, \theta\right) \sum_{t=1}^{e_{ji}} f\left(W_{ij}^{j(t)} \mid Z_{ij}, \theta\right)\right] \\ &\cdot \prod_{(i,j)\in S_{Y^*}} \left[\prod_{t=1}^{r_{ij}^*} f\left(Y_{ij}^{(t)*} \mid Z_{ij}^*, X_i^*, X_j^*, \theta\right)\right] \prod_{(i,j)\in S_M^*} \left[\prod_{t=1}^{e_{ij}^*} f\left(W_{ij}^{i(t)*} \mid Z_{ij}^*, \theta\right) \sum_{t=1}^{e_{ji}^*} f\left(W_{ij}^{j(t)*} \mid Z_{ij}, \theta\right)\right] \prod_{(i,j)\in S_L} f\left(L_{ij} \mid Z_{ij}, \theta\right) \\ &\cdot \prod_{(i,j)\in S^*} f\left(Z_{ij}^* \mid Z_{ij}, \theta\right) \prod_{i\in S_Y^*} f\left(X_i^* \mid X_i, \theta\right) \prod_{(i,j)\in S} f\left(Z_{ij} \mid \theta\right) \prod_{i\in S_Y} f\left(X_i \mid \theta\right) \end{aligned}$$

$$
\begin{aligned}
&= \prod_{(i,j)\in S_Y} \left[1-(1-\alpha_I)^{Z_{ij}}(1-\alpha_S)^{X_i+X_j}(1-\alpha_U)\right]^{Y_{ij}^{+}} \left[(1-\alpha_I)^{Z_{ij}}(1-\alpha_S)^{X_i+X_j}(1-\alpha_U)\right]^{Y_{ij}^{\#}} \\
&\cdot \prod_{(i,j)\in S_H} \left(\gamma_1^{Z_{ij}}\gamma_0^{1-Z_{ij}}\right)^{H_{ij}} \left(1-\gamma_1^{Z_{ij}}\gamma_0^{1-Z_{ij}}\right)^{1-H_{ij}} \prod_{(i,j)\in S_M} \left[\psi_1^{Z_{ij}W_{ij}^{+}}(1-\psi_1)^{Z_{ij}W_{ij}^{\#}} \times \psi_2^{(1-Z_{ij})W_{ij}^{+}}(1-\psi_2)^{(1-Z_{ij})W_{ij}^{\#}}\right] \\
&\cdot \prod_{(i,j)\in S_{Y^*}} \left[1-(1-\alpha_I)^{Z_{ij}^*}(1-\alpha_S)^{X_i^*+X_j^*}(1-\alpha_U)\right]^{Y_{ij}^{+*}} \left[(1-\alpha_I)^{Z_{ij}^*}(1-\alpha_S)^{X_i^*+X_j^*}(1-\alpha_U)\right]^{Y_{ij}^{\#*}} \\
&\cdot \prod_{(i,j)\in S_M^*} \left[\psi_1^{Z_{ij}^*W_{ij}^{+*}}(1-\psi_1)^{Z_{ij}^*W_{ij}^{\#*}} \times \psi_2^{(1-Z_{ij}^*)W_{ij}^{+*}}(1-\psi_2)^{(1-Z_{ij}^*)W_{ij}^{\#*}}\right] \prod_{(i,j)\in S_H} \left(\gamma_1^{Z_{ij}}\gamma_0^{1-Z_{ij}}\right)^{H_{ij}} \left(1-\gamma_1^{Z_{ij}}\gamma_0^{1-Z_{ij}}\right)^{1-H_{ij}} \\
&\cdot \prod_{(i,j)\in S_L} \left(\beta_1^{Z_{ij}}\beta_0^{1-Z_{ij}}\right)^{L_{ij}} \left(1-\beta_1^{Z_{ij}}\beta_0^{1-Z_{ij}}\right)^{1-L_{ij}} \prod_{(i,j)\in S} \rho^{Z_{ij}}(1-\rho)^{1-Z_{ij}} \prod_{i\in S_Y} r^{X_i}(1-r)^{1-X_i} \\
&\cdot \prod_{(i,j)\in S^*} \left(\phi_1^{Z_{ij}}\phi_0^{1-Z_{ij}}\right)^{Z_{ij}^*} \left(1-\phi_1^{Z_{ij}}\phi_0^{1-Z_{ij}}\right)^{1-Z_{ij}^*} \prod_{i\in S_Y^*} \left(\lambda_1^{X_i}\lambda_0^{1-X_i}\right)^{X_i^*} \left(1-\lambda_1^{X_i}\lambda_0^{1-X_i}\right)^{1-X_i^*},
\end{aligned}
\tag{10}
$$

where the parameter vector $\theta = \{\rho, r, \alpha_I, \alpha_S, \alpha_U, \psi_1, \psi_2, \gamma_1, \gamma_0, \beta_1, \beta_0, \phi_1, \phi_0, \lambda_1, \lambda_0\}$.

2.4. Monte Carlo Expectation Maximization for Parameter Estimation. In the model, it was not possible to estimate the true value of potential variables and model parameters directly. In order to effectively estimate the potential variables and model parameters, this paper used the Monte Carlo expectation maximization algorithm based on incomplete parameter estimation, as illustrated in Algorithm 1.

In the E-step of Algorithm 1, we use Gibbs sampling to sample (Z, X, Z^*, X^*) from $f(Z, X, Z^*, X^* \mid H, Y, W, L, Y^*, W^*, \widehat{\theta}_0)$ in turn. Repeat the sampling process until the estimations of missing data are obtained. Then in the M-step of Algorithm 1, the parameter vector $\theta = \{\gamma_1, \gamma_0, \alpha_I, \alpha_S, \alpha_U, \beta_1, \beta_0, \phi_1, \phi_0, \lambda_1, \lambda_0\}$ is estimated by Greedy Hill Climbing. Finally the iteration is stopped when diff $>$ 0.01.

3. Results

All the protein names were mapped to the Entrez IDs. Finally we got 32540 proteins, and there were 144603 interactions between these proteins.

3.1. Construction of the Human PPI Network with Reliable Confidence Measure. Four models were established separately using high-throughput Y2H experimental data, high-throughput MPC experimental data, human PPI data, and all the PPI data. The comparisons among these four models were listed in Table 2.

After the estimation of parameter vector θ by Monte Carlo EM, we recalculated the posterior probability of Z, which is $\Pr[Z \mid H, Y, W, L, Y^*, W^*]$, with θ and the observed values H, Y, W, L, Y^*, W^*. And for each pair of PPI, we considered them as reliable confidence interaction if $\Pr[Z_{ij} = 1 \mid H, Y, W, L, Y^*, W^*] > 0.8$. Then we got 48361 PPIs with reliable confidence measure among 23286 proteins.

3.2. Characterization of Network and Roles of IDPs Based on Network Topology. We analysed the role of IDPs in the human PPI networks with reliable confidence measure. A IDP was defined as a protein with continuous intrinsically disorder region whose length was larger than 40 amino acids. And 8735 IDPs were identified from 23286 proteins after predictions.

Firstly, the human PPI network was cut into subnetworks or modules by SCAN. SCAN obtained modules based on the similarity between common neighbors. Then we used modularity and similarity-based modularity as metrics. Modularity is a statistical measure of the quality of network clustering, which is defined as follows:

$$
Q_N = \sum_{s=1}^{N_C} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L}\right)^2\right], \tag{11}
$$

where N_C is the number of clusterings, L is the number of edges, l_s is the number of edges for s_{th} module, and d_s is the degree of all the nodes in s_{th} module. We could obtain the best clustering by optimizing Q_N. And similarity-based modularity is the supplementary for the modularity, which is defined as follows:

$$
Q_S = \sum_{s=1}^{NC} \left[\frac{IS_i}{TS} - \left(\frac{DS_i}{2TS}\right)^2\right]. \tag{12}
$$

As shown in Figure 2, on one hand, the modularity monotonically decreased from the position nearby zero, and it could not be maximized. On the other hand, the similarity-based modularity could be maximized while the threshold ε equals 0.61. Conditional on the $\varepsilon = 0.61$, the reliable human PPI

(1) $i = 0$, initialize the parameters
(2) while (diff > 0.01) {
(3) $j = 0$ // E-Step
(4) while ($j \le T$) {
(5) Sample $X^{(j+1)}$ from $f(Z^{(j)}, X, Z^{*(j)}, X^{*(j)} \mid H, Y, W, L, Y^*, W^*, \widehat{\theta}_i)$
(6) Sample $Z^{(j+1)}$ from $f(Z, X^{(j+1)}, Z^{*(j)}, X^{*(j)} \mid H, Y, W, L, Y^*, W^*, \widehat{\theta}_i)$
(7) Sample $Z^{(j+1)}$ from $f(Z^{(j+1)}, X^{(j+1)}, Z^{*(j)}, X^* \mid H, Y, W, L, Y^*, W^*, \widehat{\theta}_i)$
(8) Sample $Z^{(j+1)}$ from $f(Z^{(j+1)}, X^{(j+1)}, Z^*, X^{*(j+1)} \mid H, Y, W, L, Y^*, W^*, \widehat{\theta}_i)$
(9) $j = j + 1$
(10) }
(11) calculate Q function

$$\widehat{Q}\left(\theta \mid \theta^{(i)}, Y, W, L, Y^*, W^*\right) = \frac{1}{T}\sum_{m=1}^{T} \log L^c\left(\theta \mid Y, W, L, Y^*, W^*, Z^{(m)}, X^{(m)}, Z^{(m)*}, X^{(m)*}\right)$$

(12) // M-Step

(13)
$$\widehat{\rho}^{(i+1)} = \frac{1}{T}\sum_{m=1}^{T}\left(\frac{\sum_{(i,j)\in S} Z_{ij}^{(m)}}{|S|}\right)$$
$$\widehat{r}^{(i+1)} = \frac{1}{T}\sum_{m=1}^{T}\left(\frac{\sum_{i\in S_Y} X_i^{(m)}}{\|S_Y\|}\right)$$
$$\widehat{\psi}_1^{(i+1)} = \frac{1}{T}\sum_{m=1}^{T}\left(\frac{\sum_{(i,j)\in S_M} Z_{ij}^{(m)}\left(\sum_{k=1}^{e_{ij}} W_{ij}^{ik} + \sum_{k=1}^{e_{ji}} W_{ij}^{jk}\right) + \sum_{(i,j)\in S_M^*} Z_{ij}^{*(m)}\left(\sum_{k=1}^{e_{ij}^*} W_{ij}^{ik*} + \sum_{k=1}^{e_{ji}^*} W_{ij}^{jk*}\right)}{\sum_{(i,j)\in S_M} Z_{ij}^{(m)}\left(e_{ij} + e_{ji}\right) + \sum_{(i,j)\in S_M^*} Z_{ij}^{*(m)}\left(e_{ij}^* + e_{ji}^*\right)}\right)$$
$$\widehat{\psi}_2^{(i+1)} = \frac{1}{T}\sum_{m=1}^{T}\left(\frac{\sum_{(i,j)\in S_M}\left(1 - Z_{ij}^{(m)}\right)\left(\sum_{k=1}^{e_{ij}} W_{ij}^{ik} + \sum_{k=1}^{e_{ji}} W_{ij}^{jk}\right) + \sum_{(i,j)\in S_M^*}\left(1 - Z_{ij}^{*(m)}\right)\left(\sum_{k=1}^{e_{ij}^*} W_{ij}^{ik*} + \sum_{k=1}^{e_{ji}^*} W_{ij}^{jk*}\right)}{\sum_{(i,j)\in S_M}\left(1 - Z_{ij}^{(m)}\right)\left(e_{ij} + e_{ji}\right) + \sum_{(i,j)\in S_M^*}\left(1 - Z_{ij}^{*(m)}\right)\left(e_{ij}^* + e_{ji}^*\right)}\right)$$

(14) $k = 0$;
(15) $\text{change}^0 = 0.01$
(16) $\theta^k = \theta^{(i)} = \left\{\alpha_I^{(i)}, \alpha_S^{(i)}, \alpha_U^{(i)}, \gamma_1^{(i)}, \gamma_0^{(i)}, \beta_1^{(i)}, \beta_0^{(i)}, \phi_1^{(i)}, \phi_0^{(i)}, \lambda_1^{(i)}, \lambda_0^{(i)}\right\}$
(17) while (1) {
(18) $\alpha_I^{k+1} = \arg\max Q\left(\alpha_I, \alpha_S^k, \alpha_U^k\right)$
(19) $\alpha_S^{k+1} = \arg\max Q\left(\alpha_I^{k+1}, \alpha_S, \alpha_U^k\right)$
$\alpha_U^{k+1} = \arg\max Q\left(\alpha_I^{k+1}, \alpha_S^{k+1}, \alpha_U\right)$
(20) $\text{change}^{k+1} = \widehat{Q}\left(\theta^{k+1}\right) - \widehat{Q}\left(\theta^k\right)$
(21) if (abs(change^{k+1}) < abs($\text{change}^k/20$))
(22) break
(23) $k = k + 1$
(24) }
(25) $\text{diff} = \dfrac{\left|\widehat{Q}\left(\theta^{(i+1)}\right) - \widehat{Q}\left(\theta^{(i)}\right)\right|}{\widehat{Q}\left(\theta^{(i)}\right)}$
(26) $i = i + 1$
(27) $T = T * 1.1$
(28) }

Algorithm 1: Monte Carlo expectation maximization for parameter estimation.

network was cut into 241 modules. Under the significant level $\alpha = 0.05$, the p value of each module was calculated by the formula below:

$$p\text{-value} = \sum_{i=m}^{n} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}, \tag{13}$$

where N is the number of all the proteins and M is the number of all the IDPs. 33 modules among 241 modules were significantly associated with IDPs.

However, due to the fact that acquisition of functional modules is only dependent on the network topology, we analysed the modules with known diseases. And the overlap of PPI in hela cell and a functional module which was highly related with IDPs was shown in Figure 3. The weight of each side is the posterior probability of the real value Z. If a node with more than 5 neighbours was defined as a hub node in this subnetwork, a total of 69% of the hub nodes were IDPs. It is verified that IDPs were easy to become hub nodes of the protein interaction network due to the flexibility

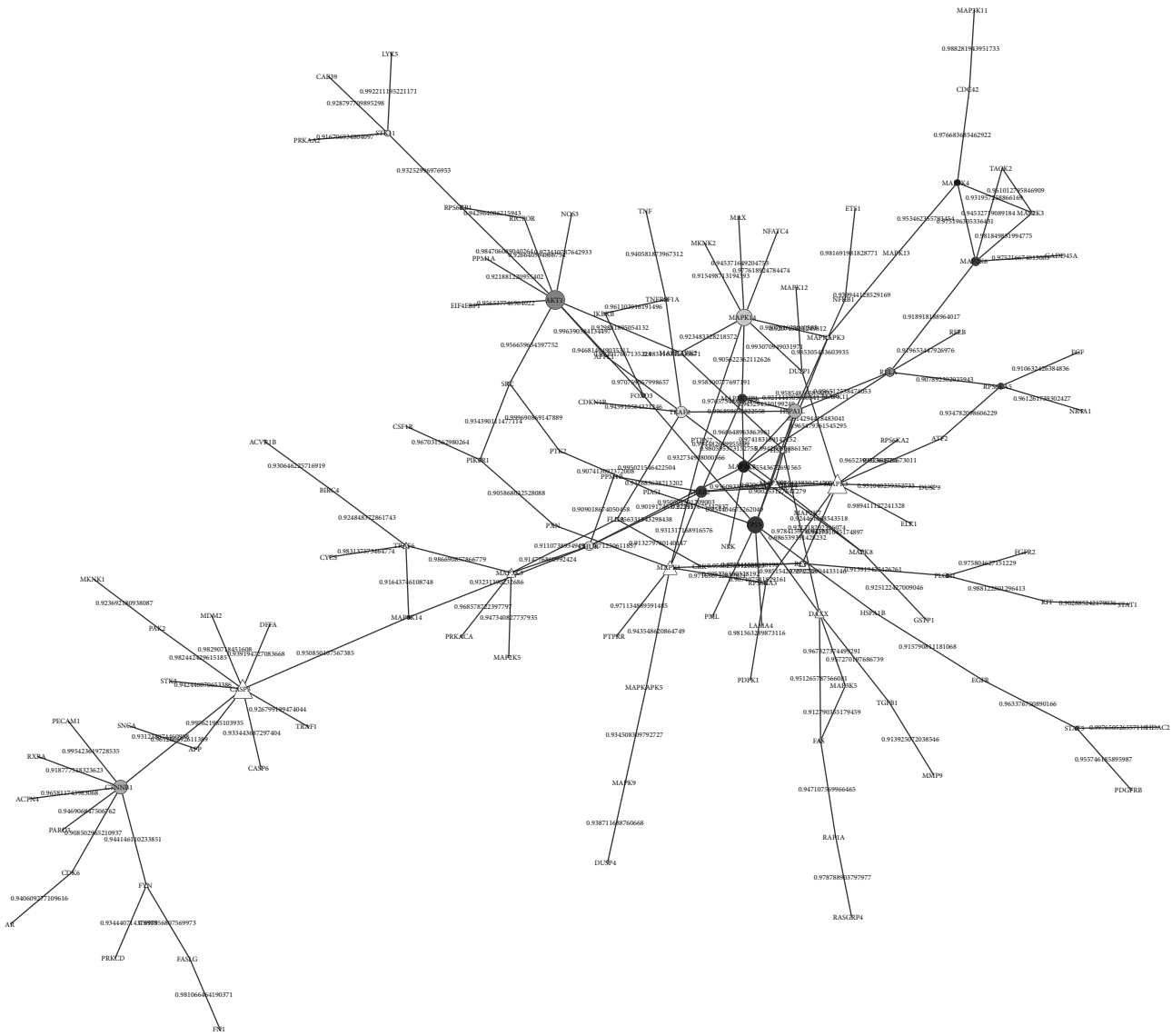

Figure 3: A reliable subnetwork for hela cell. Circles correspond to IDPs. And the degree of grey corresponds to the length of intrinsically disordered region for IDP.

Table 2: Comparison of parameters based on different data.

Parameters	High-throughput Y2H	High-throughput MPC	Human PPI data	All PPI data
ρ	6.8×10^{-3}	1.9×10^{-3}	6.1×10^{-3}	1.4×10^{-2}
r	7.7×10^{-5}	—	5.3×10^{-5}	8.9×10^{-5}
α_I	0.658	—	0.543	0.933
α_S	0.426	—	0.496	0.852
α_U	4.5×10^{-3}	—	9.7×10^{-4}	0.007
ψ_1	—	0.738	0.755	0.809
ψ_2	—	0.623	0.764	0.788

of the structure, revealing an important role of IDPs in the regulation of cervical cancer hela cell.

4. Discussion

Our model is unique and novel in the following perspectives. First, it integrates Y2H and MPC data in a cohesive and unified model that connect the two types of data through the unobserved true status of direct physical interaction Z. Second, the model allows a natural calculation of the confidence of each interacting pair via the posterior probability. This is a critical measurement in downstream analysis and will be accounted for. To our knowledge, no previous study has considered uncertainty in the PPI network analysis.

The inference of the interacting probability involves a large number of latent variables. The combinatorial effects make it impractical to compute the expectation of the missing variables analytically during the E-step. It is likely that various data sets carry different amount of information regarding the true interaction status. Hence, the inference can be

made by appropriately weighing data of various types instead of treating them equally. This can be achieved by setting parameter constrain.

Competing Interests

The authors confirm that there is no conflict of interests related to the content of this article.

Authors' Contributions

Yang Hu and Ying Zhang contributed equally to this work.

References

[1] C. Wu, F. Zhang, X. Li et al., "Composite functional module inference: detecting cooperation between transcriptional regulation and protein interaction by mantel test," *BMC Systems Biology*, vol. 4, article 82, 2010.

[2] Y.-A. Huang, Z.-H. You, X. Chen, K. Chan, and X. Luo, "Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding," *BMC Bioinformatics*, vol. 17, article 184, 2016.

[3] X. Chen, "Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA," *Scientific Reports*, vol. 5, Article ID 13186, 2015.

[4] X. Zeng, X. Zhang, and Q. Zou, "Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks," *Briefings in Bioinformatics*, vol. 17, no. 2, pp. 193–203, 2016.

[5] Q. Zou, J. Li, Q. Hong et al., "Prediction of microRNA-disease associations based on social network analysis methods," *BioMed Research International*, vol. 2015, Article ID 810514, 9 pages, 2015.

[6] Q. Zou, J. Li, L. Song, X. Zeng, and G. Wang, "Similarity computation strategies in the microRNA-disease network: a survey," *Briefings in Functional Genomics*, vol. 15, no. 1, pp. 55–64, 2016.

[7] F. Zhang, B. Gao, L. Xu et al., "Allele-specific behavior of molecular networks: understanding small-molecule drug response in yeast," *PLoS ONE*, vol. 8, no. 1, Article ID e53581, 2013.

[8] U. Stelzl, U. Worm, M. Lalowski et al., "A human protein-protein interaction network: a resource for annotating the proteome," *Cell*, vol. 122, no. 6, pp. 957–968, 2005.

[9] J.-F. Rual, K. Venkatesan, T. Hao et al., "Towards a proteome-scale map of the human protein-protein interaction network," *Nature*, vol. 437, no. 7062, pp. 1173–1178, 2005.

[10] R. M. Ewing, P. Chu, F. Elisma et al., "Large-scale mapping of human protein-protein interactions by mass spectrometry," *Molecular Systems Biology*, vol. 3, article 89, 2007.

[11] T. S. Keshava Prasad, R. Goel, K. Kandasamy et al., "Human Protein Reference Database—2009 update," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D767–D772, 2009.

[12] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569–4574, 2001.

[13] P. Uetz, L. Glot, G. Cagney et al., "A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*," *Nature*, vol. 403, no. 6770, pp. 623–627, 2000.

[14] A.-C. Gavin, P. Aloy, P. Grandi et al., "Proteome survey reveals modularity of the yeast cell machinery," *Nature*, vol. 440, no. 7084, pp. 631–636, 2006.

[15] Y. Ho, A. Gruhler, A. Heilbut et al., "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[16] A.-C. Gavin, M. Bösche, R. Krause et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[17] N. J. Krogan, G. Cagney, H. Yu et al., "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," *Nature*, vol. 440, no. 7084, pp. 637–643, 2006.

[18] S. Kerrien, Y. Alam-Faruque, B. Aranda et al., "IntAct—open source resource for molecular interaction data," *Nucleic Acids Research*, vol. 35, no. 1, pp. D561–D565, 2007.

[19] H. W. Mewes, D. Frishman, U. Güldener et al., "MIPS: a database for genomes and protein sequences," *Nucleic Acids Research*, vol. 30, no. 1, pp. 31–34, 2002.

[20] I. Xenarios, Ł. Salwínski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, no. 1, pp. 303–305, 2002.