Inferring Strengths of Protein-Protein Interactions Using Linear Programming

Morihiro Hayashida morihiro@kuicr.kyoto-u.ac.jp
Nobuhisa Ueda ueda@kuicr.kyoto-u.ac.jp
Tatsuya Akutsu takutsu@kuicr.kyoto-u.ac.jp
Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

Keywords: strengths of protein-protein interaction, linear programming

1 Introduction

Several computational methods have been proposed for inference of protein-protein interactions. Most of the existing methods assume that protein-protein interaction data are given as binary data (i.e., whether or not each protein pair interacts). However, multiple biological experiments are performed for the same protein pairs and thus the ratio (strength) of the number of observed interactions to the number of experiments is available for each protein pair.

We propose a new method [2] for inference of protein-protein interactions from such experimental data. This method tries to minimize the errors between the ratios of observed interactions and the predicted probabilities in training data, where this problem is formalized as a linear program based on a probabilistic model. We compared the proposed method with the association method [4], the EM method [1] and ASNM method (association method for numerical data). It is shown that a variant of the method is comparable to existing methods for binary data. It is also shown that the LPNM method outperforms existing methods for numerical data.

2 Probabilistic Model

We use the probabilistic model proposed in [1]. We treat protein-protein interactions and domain-domain interactions as random variable: $P_{ij} = 1$ if P_i and P_j interact with each other, and $D_{mn} = 1$ if D_m and D_n interact with each other. We assume that domain-domain interactions are independent and two proteins interact if and only if at least one domain pairs from the two proteins interact (see Figure 1). Under this assumption, the probability that P_i and P_j interact with each other is given by $\Pr(P_{ij} = 1) = 1 - \prod_{D_{mn} \in P_{ij}} (1 - \lambda_{mn})$, where λ_{mn} means the probability that D_m and D_n interact with each other (i.e., $\lambda_{mn} = \Pr(D_{mn} = 1)$).

3 LPNM: LP-Based Method for Numerical Data

We describe an LP-based method for numerical interaction data, which is the most important variant of the LP-based method. we set ρ_{ij} to be the ratio of interactions between proteins P_i and P_j in a series of experiments, that is, $\rho_{ij} = \frac{K_{ij}}{M}$, where K_{ij} is the number of observed interactions between proteins P_i and P_j, M is the total number of experiments. Since ρ_{ij} is the ratio of interactions, we consider here to minimize the difference between $\Pr(P_{ij} = 1)$ and ρ_{ij}.

$$\Pr(P_{ij} = 1) = 1 - \prod_{D_{mn} \in P_{ij}} (1 - \lambda_{mn}) \approx \rho_{ij} \Leftrightarrow \sum_{D_{mn} \in P_{ij}} \ln(1 - \lambda_{mn}) \approx \ln(1 - \rho_{ij}) \Leftrightarrow \sum_{D_{mn} \in P_{ij}} \gamma_{mn} \approx \beta_{ij},$$

where $\gamma_{mn} = \ln(1 - \lambda_{mn})$, $\beta_{ij} = \ln(1 - \rho_{ij})$. Thus, we obtain the linear programming as Figure 2.

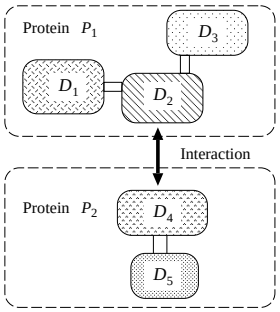

Figure 1: Inference of protein-protein interactions through domain-domain interactions. In this case, we infer that proteins P_1 and P_2 interact with each other since domains D_2 and D_4 interact with each other.

$$
\begin{aligned}
&\text{minimize} \quad \sum_{P_{ij}} \alpha_{ij} \,, \\
&\text{subject to} \\
&\qquad \sum_{D_{mn} \in P_{ij}} \gamma_{mn} - \beta_{ij} \leq \alpha_{ij}, \\
&\qquad \beta_{ij} - \sum_{D_{mn} \in P_{ij}} \gamma_{mn} \leq \alpha_{ij}, \\
&\qquad \gamma_{mn} \leq 0 \ \text{ for all } \gamma_{mn}, \\
&\qquad \alpha_{ij} \geq 0 \ \text{ for all } \alpha_{ij}, \\
&\qquad \beta_{ij} < 0.
\end{aligned}
$$

Figure 2: Linear programming on LPNM method.

Table 1: Root mean squared errors and average elapsed time for numerical interaction data.

	LPNM	ASNM	EM	ASSOC
Average of Errors	0.0307893	0.0404935	0.295138	0.276771
Time (sec)	1.203068	0.0077122	1.620078	0.0088252

4 Results

We used the full data of Ito's Yeast Interacting Proteins (YIP) database [3]. We evaluated the method by 5-fold cross validation. We used 1,586 interaction pairs of proteins and the numbers of their IST hits. Table 1 shows the average of the errors. LPNM outperformed the other methods. it to avoid overfitting.

References

[1] Deng, M., Mehta, S., Sun, F., and Chen, T., Inferring domain-domain interactions from protein-protein interactions, *Genome Research*, 12:1540–1548, 2002.

[2] Hayashida, M., Ueda, N., and Akutsu, T., Inferring strengths of protein-protein interactions from experimental data using linear programming, *Bioinformatics*, 19:58–65, 2003.

[3] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y., A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc. Natl. Acad. Sci.*, 98:4569–4574, 2001.

[4] Sprinzak, E. and Margalit, H., Correlated sequence-signatures as markets of protein-protein interaction, *J. Mol. Biol.*, 311:681–692, 2001.