

Protein Interaction Inference as a MAX-SAT Problem

Ya Zhang¹ Hongyuan Zha² Chao-Hisen Chu¹

Keywords: Protein interactions, domain interactions, inference, Satisfiability problem

1 Introduction

Discovering interacting proteins is essential for understanding protein functions in a systematic fashion. However, high throughput interaction data are inherently noisy and only cover a small portion of the interactome. The question - can we infer useful protein-protein interaction information from those high throughput data - arises.

Proteins are considered to interact through their interaction domains. Domain-domain interactions not only represent protein-protein interactions at a more abstract level but also enable the discovery of unobserved protein-protein interactions. Many domain-based approaches have been proposed to predict protein-protein interactions, including association-based methods[4] and Maximum Likelihood Estimation method[1]. Although promising results have been achieved by these methods, the incompleteness and uncertainty of high throughput interaction data form big obstacles in accurate interaction inference.

Existing methods tend to oversimplify the problem by introducing the assumption that the domain interactions are independent from each other. We proposed a new framework of learning with no assumption about the independence between domain interactions. The problem of interaction inference is modeled as a Maximum Satisfiability (MAX-SAT) problem and is solved via linear programming. Experimental results on a combined yeast data set have demonstrated the robustness of and accuracy of the proposed algorithm.

2 Method

Our method for inferring interacting domains is based on the widely accepted fact that two proteins interact if and only if at least one pair of domains from the two proteins interact. Let us denote the set of domains as $D = \{d_1, d_2, \dots, d_N\}$. The set of proteins under investigation are represented as $P = \{p_1, p_2, \dots, p_M\}$. Denote Ω_{ij} as the set of domain pairs in the protein pair (p_i, p_j) . Logically, the relationship between domain-domain interaction and protein-protein interaction can be expressed as:

$$P_{ij} = \vee_{d_{nm} \in \Omega_{ij}} D_{nm}, \quad (1)$$

where \vee means *or*, P_{ij} is the indicator of whether proteins p_i and p_j interact, and D_{nm} is the indicator of whether domains d_n and d_m interact. Both P_{ij} and D_{nm} take binary values with

$$P_{ij} = \begin{cases} 1 & \text{if proteins } p_i \text{ and } p_j \text{ interact,} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$D_{nm} = \begin{cases} 1 & \text{if domains } d_n \text{ and } d_m \text{ interact.} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

With this representation, the problem of inferring interacting domains is naturally formulated as a Maximum Satisfiability (MAX-SAT) problem. Given a set of m clauses in conjunctive normal form over n variables, the MAX-SAT problem is to find a truth assignment for the n variables that satisfies a maximum number of clauses. Here the variables

¹School of Information Sciences and Technology, ²Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802

are the indicators for domain interactions and protein interactions. The clauses are the expressions to show the relationships between protein interactions and domain interactions (Equation 1). Considering the interaction data is associated with high false negatives, we set the objective to be maximizing the number of positive interactions. For each protein pair (p_i, p_j) , we associate an indicator variable $P'_{ij} \in \{0, 1\}$ to indicate whether or not the proteins (p_i, p_j) interact, based on the assignment of domain interaction indicators $D_{nm} \in \{0, 1\}$. The problem can be formed as:

$$\begin{aligned} & \text{maximize } \sum_{P_{ij}=1} P'_{ij} \\ & \text{subject to: } \sum_{d_{nm} \in \Omega_{ij}} D_{nm} \geq P'_{ij} \\ & D_{nm} \in \{0, 1\} \forall m, n \text{ and } P'_{ij} \in \{0, 1\} \forall i, j \end{aligned} \quad (4)$$

where P_{ij} indicates whether proteins p_i and p_j interact based on experimental interaction data. The inequality constraints in Eq. 4 ensure that a protein pair is deemed to be interacting only if at least one of the domain pair in the protein pair is considered interacting.

MAX-SAT problems has been known to be NP-hard and their search space is large. Linear programming is used to solve the satisfiability problem [3]. We solve the relaxation linear program in which we relax the integer constraints on D_{mn} and P'_{ij} by allowing them to assume real values in the interval $[0, 1]$. These real number values obtained represent the probability of picking the integer value 1.

3 Result

A combined yeast data set [1] is used to predict domain interactions. The interaction data set provides positive interaction data for our training. The domain definitions of the yeast proteins are according to Pfam. The objective is to find a set of domain interactions so that a maximum number of protein pairs in the training data are predicted to be interacting according to the domain interactions.

The putative domain interactions are used to predict interacting proteins in the MIPS data [2], which contain physical interaction pairs. The performance of the algorithm is evaluated in terms of sensitivity and specificity. The method achieved a sensitivity of 62.8% and a specificity of 91.2% at the cutoff of 0.4 based on the experiments (see Table 1), which outperforms the previous methods.

Table 1: Number of matched protein pairs between predictions at two training settings

Threshold	Prediction	MIPS	MIPS1*	SN	SP
≥ 0.20	1541	1400	242	62.8%	90.9%
≥ 0.40	1516	1383	225	62.0%	91.2%
≥ 0.60	1442	1352	194	60.6%	93.7%
≥ 0.80	1376	1316	158	59.0%	95.6%

* MIPS1 is the MIPS excluding the training data.

References

- [1] M. Deng, S. Mehta, F. Sun, and T. Chen. Inferring domain-domain interactions from protein-protein interactions. In *Proceedings of RECOMB*, 2002.
- [2] H. W. M. et al. Mips: A database for genomes and protein sequences. *Nucleic Acids Res.*, 30(1):31–34, 2000.
- [3] J. Hooker. Resolution and the integrality of satisfiability problems. *Mathematical Programming*, 74:1–10, 1996.
- [4] E. Sprinzak and H. Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–692, 2001.