

Protein Interaction Inference as a MAX-SAT Problem

Abstract

Discovering interacting proteins is essential in understanding protein functions. However, high throughput interaction data are inherently noisy and only cover a small portion of the whole interactome. Domains, the building block of proteins, are believed to be responsible to the protein-protein interactions. An abstract representation of interactome is achieved at domain level and this representation also facilitates the discovery of unobserved protein-protein interactions. Many domain-based approaches have been proposed to predict protein-protein interactions and achieved promising results. They usually assume that domain interactions are independent of each other for the convenience of computational modelling. In this paper, a new framework of learning is proposed. The framework makes no assumption about domain interactions and consider protein interactions that resulted from multiple domain interactions which may be dependent of each other. With a conjunctive norm form representation for the relationship between protein interactions and domain interactions, the problem of interaction inference is modeled as a constraint satisfiability problem and solved via linear programming. Experimental results on a combined yeast data set have demonstrated the robustness of and accuracy of the proposed algorithm.

1 Introduction

Understanding the functions of proteins is a major task of contemporary proteome research in the post-genomic era. Proteins usually perform their functions in a collaborative fashion by interacting with other proteins either in a pair or as components of larger complexes. Uncovering the complex structures of protein interaction network is hence an essential part of protein function assignment. The spectacular advances in proteomics during the last few years have opened up new opportunities for studying protein interaction networks. A tremendous amount of protein interaction data have been generated by high throughput experimental approaches such as the yeast two-hybrid genetic screen [12, 18] and mass spectrometric analysis [10], making possible genome-wide analysis of protein interactions. However, these high-throughput experiments are inevitably associated with high false-positives as well as false-negatives[15]. For example, the genome-wide interaction

data obtained in two independent experiments [13, 12] and [18] only overlap in less than 4% of the identified interactions. Moreover, the high false-negatives indicate that the interaction data only represent a small portion of the whole interactome. However, The large size of such high throughput data makes it impractical, if not impossible, to verify individual interactions by the experimental methods. The question - can we infer useful protein-protein interaction information from those high throughput data - arises.

A number of computational methods have been proposed for the prediction of protein interaction networks. Information about gene fusion events [5], phylogenetic profile [16], protein homology [7], and gene neighborhood [2] are often informative in inferring protein linkage and are explored in several interaction inference algorithms. Another important factor that determining protein interactions is the domain composition of the proteins. Domains are considered responsible for protein interactions – proteins interact through their interacting domains (Figure 1). They are used as the building blocks to form different proteins. As a result, domain-domain interactions provide a more abstract level of representation for the protein-protein interactions and the interactome.

The relationship between domain and protein has motivated domain-based approaches to predict protein interactions[3, 9, 14, 17, 19], which first infer domain-domain interactions from protein interaction data and then use the putative domain interactions to predict interacting proteins. However, most of existing domain-based interaction prediction methods assume that the domain interactions are independent of each other for the convenience of computational modelling. The conjecture might be the major reason for the relatively low precision of the conventional domain based prediction approaches because protein-protein interaction could be mediated by multiple domain interactions in synergy. To overcome this limitations, we propose a new framework of learning without enforcing the independence assumption between domain interactions. The protein-protein interactions are interpreted as the result of one or more domain interactions, which may be either dependent or independent. Our approach is more inclusive than the previous ones. The relationships between protein interactions and domain interactions are expressed in conjunctive norm forms. This representation enables us to naturally formulate the problem of interaction inference as a

satisfiability (SAT) problem. The inference problem is then solved with linear programming. The prediction framework is characterized in the following two aspects. First, the proposed framework makes no assumption on the dependency of domain interactions. Second, when formulating the inference problem as a SAT problem, prior knowledge about domain interaction or protein interaction may be easily input into the framework. The validity of the prediction method is evaluated with yeast protein interactions and protein interactions from multiple species. Experimental results have demonstrated the robustness of and accuracy of the proposed algorithm.

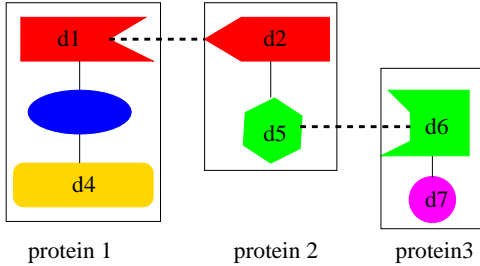


Figure 1: A sketch illustration of how domain interaction contributes to protein interaction. Proteins interact through their interacting domains.

This paper is organized as follows. In section 2, we introduce related researches on the domain-based approaches to predict protein-protein interactions. In section 3, the prediction framework is described in detail. In section 4, the validation result of the framework is illustrated. Finally, we draw conclusion in section 5.

2 Related Work

Two major classes of domain-based interaction prediction approaches are association-based methods and optimization-based methods. Association methods [14, 17] generally assume that co-occurrence of domain pairs in a pair of interacting proteins indicates association – in this case, interaction. The methods may assign high scores to some domain pairs with low frequency and the score does not correspond well to the possibility of interaction.

An optimization framework is adopted by several studies. [3] proposed a probabilistic model for protein interactions and developed a global method to inferring interacting domains by maximizing the likelihood of the observed data. The Expectation and Maximization (EM) algorithm is used to optimize the parameters. The experimental errors are integrated into the likelihood function. [8] add a notion of interaction ‘strength’, to the probabilistic model, in which the strength is computed as the ratio of the number of observed interactions to the number of experiments. They try

to minimize the sum of differences between the computed strength and the predicted probabilities in training data with linear programming. For the ease of computational modelling, the above probabilistic model assumes that the domain interactions are independent of each other.

3 Inferring interacting domain pairs

A common assumption for domain-based protein interaction prediction is that two proteins interact if and only if at least one pair of domains from the two proteins interact. Our framework of inferring interacting domain pairs is also built upon this hypothesis. Let us denote the set of proteins and domains under investigation $P = \{p_1, p_2, \dots, p_M\}$ and $D = \{d_1, d_2, \dots, d_N\}$, where M and N are the number of proteins and domains, respectively. Denote Ω_{ij} to be the set of domain pairs contained in the protein pair $\langle p_i, p_j \rangle$.

$$\Omega_{ij} = \{ \langle d_1, d_2 \rangle \mid \langle d_1, d_2 \rangle \in p_i \times p_j \text{ or } p_j \times p_i \} \quad (1)$$

For any pair of proteins, whether the two proteins interact or not is determined by the interaction of the set of domain pairs contained in the pair of proteins. This relationship may be expressed in conjunctive normal form as:

$$P_{ij} = \bigvee_{d_{nm} \in \Omega_{ij}} D_{nm}, \quad (2)$$

where \bigvee 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

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

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

Example 1: Suppose that protein p_1 contains domains $\{d_1, d_2\}$ and protein p_2 contains domains $\{d_1, d_3, d_5\}$. We then have $\Omega_{12} = \{d_{11}, d_{13}, d_{15}, d_{21}, d_{23}, d_{25}\}$. P_{12} , the interaction indicator of proteins p_1 and p_2 , is expressed in term of the set of related domain indicators: $P_{12} = D_{11} \bigvee D_{13} \bigvee D_{15} \bigvee D_{21} \bigvee D_{23}$.

The problem of inferring potential interacting domains from protein interactions is essentially to discover the set of domain interactions that best represent the protein interaction data. With the conjunctive norm form of representation, the inference task essentially is to assign values to the domain interaction indicators D_{nm} ($n, m = \{1, \dots, N\}$) so that all the protein-domain interaction relationships expressed as in Equation 2 are satisfied. This objective naturally leads the inference problem to be formulated as a satisfiability problem.

Definition 1: Given a set of p clauses in conjunctive normal form over q variables, the *satisfiability* (SAT) problem is to decide whether there is a truth assignment for the q variables that satisfies all the clauses.

Due to the high error rates in the interaction data, it is unlikely to have the set of assignment for domain interaction indicators that can simultaneously fit into the whole interaction data. Therefore, rather than requiring the assignment to be able to accommodate all the protein interactions, we set the objective as maximizing the number of protein interactions that are satisfied based on the domain interaction indicators assigned. This objective coincides with those of maximum satisfiability (MAX-SAT) problems.

Definition 2: Given a set of p clauses in conjunctive normal form over q variables, the *maximum satisfiability* (MAX-SAT) problem is to obtain a truth assignment for the q variables so that a maximum number of the clauses are satisfied.

SAT and MAX-SAT problems are difficult to solve because of their large search space, and they have been known to be NP-hard [4]. Although a number of techniques have been developed to solve SAT and MAX-SAT problems, finding optimal solutions for SAT and MAX-SAT problems is still active research topics in artificial intelligence, logic, theory of computation, and many related areas. How to optimize the solutions of SAT and MAX-SAT problems, however, is out of the scope of this paper. Therefore, in this study, linear programming [11], a widely used techniques for MAX-SAT problems, is used to solve the inference problem.

For the interaction inference problem, we associate an indicator variable $P'_{ij} \in \{0, 1\}$ with each protein pair $\langle p_i, p_j \rangle$ to indicate whether or not the proteins are predicted to interact, based on the assignment of domain interaction indicator matrix D . The goal is to maximize the number of satisfied protein-domain interaction relationships, i.e.

$$\begin{aligned} \max f &= \sum_{ij} (1 - |P_{ij} - P'_{ij}|) \\ \text{subject to } P'_{ij} &= \vee_{d_{nm} \in \Omega_{ij}} D_{nm} \quad (\forall i, j), \end{aligned} \quad (3)$$

where $D_{nm} \in \{0, 1\}$ and $P_{ij} \in \{0, 1\} (\forall m, n, \text{ and } i, j)$. P_{ij} is interaction indicator for proteins p_i and p_j according to the experimentally-determined interaction data. Here, if the interaction between proteins p_i and p_j is predicted to be identical to that provided in the data, then we have $P_{ij} - P'_{ij} = 0$; otherwise, $|P_{ij} - P'_{ij}| = 1$. The objective of Equation 3 is equivalent to minimize the function $\sum_{ij} |P_{ij} - P'_{ij}|$, which is the total number of protein pairs whose protein-domain interaction relationships are unsatisfied based on the domain interaction assignment. To solve this minimization problem, the following linear program is

formulated:

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

The inequality constraints in Equation 4 are from the constraints in Equation 3 and they 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. As P_{ij} either be 1 or 0. Equation 5 may be reformulated as:

$$\begin{aligned} \text{minimize} & \quad \sum_{P_{ij}=0} P'_{ij} - \sum_{P_{ij}=1} P'_{ij} \\ \text{subject to:} & \quad \sum_{d_{nm} \in \Omega_{ij}} D_{nm} \geq P_{ij} \quad (\forall i, j) \\ & \quad P'_{ij} \in \{0, 1\} \quad (\forall i, j) \\ & \quad D_{nm} \in \{0, 1\} \quad (\forall n, m). \end{aligned} \quad (5)$$

The linear programming problem is NP-hard when the variables are restricted to integers. A suitable approximation is to use probabilistic methods. We solve the relaxation linear program in which we loose the integer constraints on the matrixes D and P' in Eq. 5. D_{nm} and P'_{ij} are allowed to assume any real value in the interval of $[0, 1]$.

$$\begin{aligned} \text{minimize} & \quad \sum_{P_{ij}=0} P'_{ij} - \sum_{P_{ij}=1} P'_{ij} \\ \text{subject to:} & \quad \sum_{d_{nm} \in \Omega_{ij}} D_{nm} \geq P_{ij} \quad (\forall i, j) \\ & \quad 0 \leq P'_{ij} \leq 1 \quad (\forall i, j) \\ & \quad 0 \leq D_{nm} \leq 1 \quad (\forall n, m). \end{aligned} \quad (6)$$

Let \hat{D}_{nm} be the value obtained for variable D_{nm} and \hat{P}'_{ij} for P'_{ij} after solving the linear program. These real number values obtained for D_{nm} and \hat{P}'_{ij} represent the probability of picking the integer value 1 for them.

4 Experimental Results

4.1 Training settings

We use the yeast interaction data as prepared in [3]. The data set is a combination of interaction data obtained from large scale yeast two-Hybrid screens on *Saccharomyces cerevisiae* genome [13, 18]. It includes 5719 interactions among 3729 proteins. The domain definitions of the yeast proteins are according to Pfam [1], which contains hidden Markov model based profiles (HMM-profiles) of many common protein domains based on multiple sequence alignments. The Pfam database contains two parts: one is the curated part called Pfam-A and the other is automatically generated supplement called Pfam-B which represents small families taken from the PRODOM database that do not overlap with Pfam-A. Both Pfam-A and Pfam-B families

Table 2: Number of matched protein pairs between predictions at two training settings. MIPS is the MIPS data, and MIPS1 is the MIPS excluding the training data.

Threshold	Setting 1: Positive training					Setting 2: Positive and negative training				
	Prediction	MIPS	MIPS1	SN	SP	Prediction	MIPS	MIPS1	SN	SP
≥ 0.20	1541	1400	242	62.8%	90.9%	1223	1043	210	46.8%	85.3%
≥ 0.40	1516	1383	225	62.0%	91.2%	1223	1043	210	46.8%	85.3%
≥ 0.60	1442	1352	194	60.6%	93.7%	1166	1027	196	46.1%	88.1%
≥ 0.80	1376	1316	158	59.0%	95.6%	1130	1016	188	45.6%	89.9%
≥ 0.975	1351	1297	139	58.2%	96.0%	1129	1016	188	45.6%	90.0%

are used here. In total, 2918 Pfam domains are defined on the set of proteins. Proteins without defined domains are treated as superdomains, and domains which always coexist in proteins are merged into one domain. This end up with 4131 domains, which include Pfam domains, superdomains, and merged domains.

The yeast data set only gives information about the interacting protein pairs. No information about the non-interacting protein pairs is provided. A set of non-interacting protein pairs are generated by randomly coupling the proteins who are not observed to interact in the experiments.

Remark 1: Because the yeast interaction data set is associated with high false negatives. When selecting the negative interaction data, it is not guaranteed that all the interacting protein pairs were excluded from the randomly generated set of non-interacting protein pairs.

Considering the potential limitation in selecting negative interaction data, the following two settings are used in our training to infer domain-domain interactions.

Setting 1: During training, only the yeast data set, i.e. positive interaction data, is used as inputs to Equation 6 to infer a set of domain-domain interactions.

Setting 2: The yeast data set, together with the artificially generated negative interaction data, is used to infer a set of domain-domain interactions.

Example 2: Suppose that protein domain composition data are listed in Table 1. An entry of one indicates that the protein contains the corresponding domain. For example, p_1 contains domains $\{d_1, d_2\}$. Suppose the following protein pairs are observed to interact by experiments: $\langle p_1, p_2 \rangle$ and $\langle p_3, p_4 \rangle$. Under the setting 1, the objective function for the linear program is formulated as: $\min f_1 = -P'_{12} - P'_{34}$,

Table 1: A protein domain composition data set.

	d_1	d_2	d_3	d_4	d_5
p_1	1	1	0	0	0
p_2	1	0	1	0	1
p_3	0	0	0	1	1
p_4	0	1	0	1	0

while the objective function under the setting 2 is to minimize $f_2 = P'_{11} - P'_{12} + P'_{13} + P'_{14} + P'_{22} + P'_{23} + P'_{24} + P'_{33} - P'_{34} + P'_{44}$.

The GNU Linear Programming Kit¹ (version 4.7) is used for solving linear programs on Unix. The algorithm is mainly implemented in Perl, and the experiments were performed on SUN Ultra 60 server (450 MHz) with 1 GB RAM.

4.2 Evaluation

For validation, the MIPS (Munich Information center for Protein Sequences) physical interaction pairs [6] are used to evaluate the predictions. The MIPS data set contains 2575 pairs of interacting proteins among 1919 proteins. Proteins which do not contain any domain from the training set are deleted because no information about their interaction is obtained from the training set. This deletion results in a test set of 2230 interactions among 1764 proteins. The MIPS data do not include pairs of non-interacting proteins. We again randomly generate a set of non-interacting protein pairs of size comparable to the number of the interacting protein pairs.

The performance of the algorithm is evaluated in terms of sensitivity SN and specificity SP . Sensitivity is the ratio of the correctly predicted interacting protein pairs (TP) to the total number of interacting protein pairs ($TP + FN$), while specificity is the ratio of the correctly predicted interacting protein pairs (TP) to the number of protein pairs predicted to be interacting ($TP + FP$).

$$SN = \frac{TP}{TP + FN} \quad (7)$$

$$SP = \frac{TP}{TP + FP} \quad (8)$$

The results are presented in Table 2. According to the results, training with only positive interaction data (setting 1) generates a better result than training with positive interaction data together with the randomly generated negative interactions (setting 2). For example, at the cutoff 0.8, predicting protein-protein interactions at setting 1 achieved a

¹<http://www.gnu.org/software/glpk/glpk.html>

Table 3: Number of matched protein pairs between predictions for our method with setting 1 and MLE-EM method. MIPS is the MIPS data, and MIPS1 is the MIPS excluding the training data.

Threshold	MAX-SAT (Setting 1)				MLE-EM			
	MIPS	MIPS1	SN	SP	MIPS	MIPS1	SN	SP
≥ 0.20	1400	242	62.8%	90.9%	1074	51	48.2%	87.2%
≥ 0.40	1383	225	62.0%	91.2%	993	47	44.5%	88.0%
≥ 0.60	1352	194	60.6%	93.7%	933	43	41.8%	89.3%
≥ 0.80	1316	158	59.0%	95.6%	882	40	39.6%	90.2%
≥ 0.975	1297	139	58.2%	96.0%	845	35	37.9%	90.9%

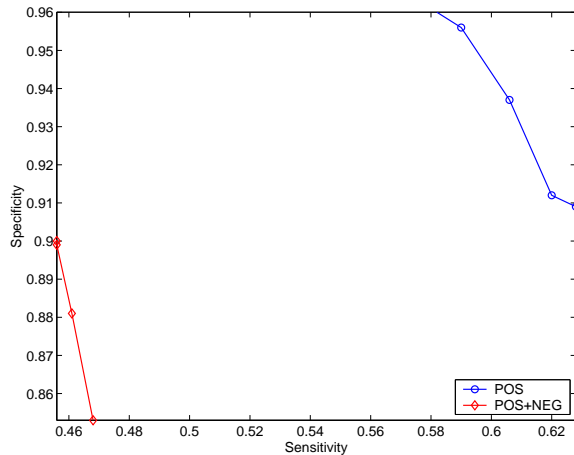


Figure 2: Comparison of specificity and sensitivity of the prediction of protein-protein interactions by the SAT method for the two experimental settings.

sensitively of 59.0% and a specificity of 95.6% while predicting at setting 2 leads to a sensitivity of 45.6% and a specificity of 89.9%, both much lower than those with setting 1. Figure 2 provides a comparison of the results obtained with the two settings at several different thresholds. As we can see from the graph, setting 1 performs consistently better. One possible explanation for the results is that the high throughput data contain a high ratio of false negative. Therefore the randomly selected examples of non-interacting protein pairs include many interacting protein pairs. Using the false non-interacting protein pairs in training prevents many interacting domains being recognized.

We compare the performance of our method (training at setting 1) with that of the EM method presented in [3] and the results of the two methods are presented in Table 3. Our method is able to predict protein-protein interactions at higher sensitivities and specificities. For the MIPS data excluding training data (MIPS1 data), we correctly predicted 194 pairs of interacting proteins when training with only interacting protein pairs with setting 1 at the cut-off of 0.6, while the MLE-EM method is only able to predict 43 pairs correctly at the same cut-off.

4.3 Biological Significance of the Predictions

Table 4 lists the novel interacting protein pairs discovered with our methods. The prediction about the interaction between ADR1 and ZAP1 is very significant because ADR1 and ZAP1 are zinc-responsive transcription factors. It is very likely that the two proteins bind together in response to the presence of zinc and other related stimulates. Another significant prediction we made is the interaction between protein PAP1, a amino acid permease, and protein SEC17, which is a peripheral membrane protein required for vesicular transport. The rationale after their interaction is that when the amino acid permease PAP1 uptakes amino acids, it may need to bind to SEC17 to transport the amino acids to other cellular compartment.

Our prediction of protein-protein interactions is associate with very low cost and it helps biologists to select important protein pairs out of numerous candidates without experimentation. Based on the prediction, biologists can assign priorities to the proteins or domains to be experimented on. Moreover, the prediction may also be used to assign functions to unknown proteins. For example, the uncharacterized protein, YMR291W, was predicted to interact with HSP104. Since interacting proteins usually involved in the same cellular processes, we may predict that YMR291W is involved the response to stresses.

5 Discussions and Conclusions

We have presented a novel domain-based method the predicting protein-protein interactions. Unlike existing methods, which oversimplify the problem by introducing the assumption that the domain interactions are independent from each other, our methods is based on minimum assumptions about protein-protein interaction – proteins interacts through their interacting domains. We model the problem of interaction inference as a constraint satisfiability problem and solve it as a linear program. Our method (with training setting 1) achieved a sensitivity of 62.0% and a specificity of 91.2% at the threshold 0.4 on a combined yeast data set. The predictions on interacting protein pairs made by our method have more overlaps with MIPS interaction

Table 4: Examples of the discovered novel interacting protein pairs.

Interactor I	Function	Interactor II	Function
ADR1	Zinc-finger transcription factor involved in regulation of ADH2 and peroxisomal genes	ZAP1	Zinc-regulated transcription factor, binds to zinc-responsive promoter elements to induce transcription of certain genes in the presence of zinc
PAP1	Amino acid permease involved in the uptake of cysteine, leucine, isoleucine and valine	SEC17	Peripheral membrane protein required for vesicular transport between ER and Golgi and for the 'priming' step in homotypic vacuole fusion, part of the cis-SNARE complex
CLN1	role in cell cycle START	PKH1	Pkb-activating Kinase Homologue; Ser/Thr protein kinase
SMK1	Mitogen-activated protein kinase required for spore morphogenesis that is expressed as a middle sporulation-specific gene	SWE1	Protein kinase that regulates the G2/M transition by inhibition of Cdc28p kinase activity
DUN1	Cell-cycle checkpoint serine-threonine kinase required for DNA damage-induced transcription of certain target genes, phosphorylation of Rad55p and Sml1p, and transient G2/M arrest after DNA damage; also regulates postreplicative DNA repair	TIF35	Subunit of the core complex of translation initiation factor 3(eIF3), which is essential for translation
BOI1	Protein implicated in polar growth; interacts with bud-emergence protein Bem1p	TIF35	Subunit of the core complex of translation initiation factor 3(eIF3), which is essential for translation
TIF34	Subunit of the core complex of translation initiation factor 3(eIF3), which is essential for translation	WTM2	WD repeat containing transcriptional modulator 2; Transcriptional modulator
GPA1	GTP-binding alpha subunit of the heterotrimeric G protein that couples to pheromone receptors; negatively regulates the mating pathway by sequestering G(beta)gamma and by triggering an adaptive response; activates the pathway via Scp160p	PAC1	Protein involved in nuclear migration, part of the dynein/dynactin pathway; targets dynein to microtubule tips, which is necessary for sliding of microtubules along bud cortex
PRP3	Splicing factor, component of the U4/U6-U5 snRNP complex	TPK3	Involved in nutrient control of cell growth and division; cAMP-dependent protein kinase catalytic subunit
ARO8	Aromatic aminotransferase, expression is regulated by general control of amino acid biosynthesis	SRP1	Cell wall mannoprotein of the Srp1p/Tip1p family of serine-alanine-rich proteins
AHP1	Thiol-specific peroxiredoxin, reduces hydroperoxides to protect against oxidative damage; function in vivo requires covalent conjugation to Urm1p	SRP1	Cell wall mannoprotein of the Srp1p/Tip1p family of serine-alanine-rich proteins; expression is downregulated at acidic pH and induced by cold shock and anaerobiosis; abundance is increased in cells cultured without shaking
CUS2	Protein that binds to U2 snRNA and Prp11p, may be involved in U2 snRNA folding	SAP190	Protein that forms a complex with the Sit4p protein phosphatase and is required for its function
HSP104	Heat shock protein that responsive to stresses including: heat, ethanol, and sodium arsenite	YMR291W	ORF, Uncharacterized

data compared to those by MLE-EM method.

Although our method achieved relatively high sensitivity and specificity. The sensitivity is still low. The reason for the relatively low sensitivity is that the protein-protein interactions provided for the training (the combined data set) only represent a very small fraction of the potential protein-protein interactions due to high false-negative associated with high throughput methods. As proper training instances are necessary for prediction methods to perform well, it is quite reasonable for our method to achieve sensitivities around 60%. With the accumulation of high throughput interaction data, we may be able to include more instance in the training data and improve the sensitivity of the prediction.

One limitation shared by all domain-based interaction inference methods is that domain composition is considered as the solely determining factor for interactions. However, the presence of a pair of interacting domain in a pair of proteins is only a necessary but not sufficient for two proteins to interact. Whether two proteins interact or not may also depends on their expression level, their subcellular location, and many other factors. Proteins are observed to interact with different partners in fulfilling different cellular functions. For example, the 14-3-3 Domain interacts with Cdc25 tyrosine phosphatase during cell cycle regulation, while it interacts c-Raf Ser/Thr Kinase when it functions for signal transduction. Hence, protein interactions cannot be studied in an isolated fashion. A system biology approach, which focuses on the interplay between all components of the cell, may be central to the understanding of protein interactions.

References

- [1] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Stud holme, C. Yeats, and S. R. Eddy. The pfam protein families database. *Nucleic Acids Res. (Database Issue)*, 32:D138–D141, 2004.
- [2] T. Dandekar, B. Snel, M. Huynen, and P. Bork. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, 23:324–328, 1998.
- [3] M. Deng, S. Mehta, F. Sun, and T. Chen. Inferring domain-domain interactions from protein-protein interactions. In *Proceedings of the sixth annual international conference on Computational biology (RECOMB)*, pages 117–126, Washington, DC, USA, April 2002.
- [4] D. Du, J. Gu, and P. Pardalos. *Satisfiability Problem: Theory and Application*, volume 35 of *DIMACS Series in Discrete Mathematics*. American Mathematical Society, 1997.
- [5] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 203(6757):86–90, Nov 1999.
- [6] H. W. Mewes et al. Mips: A database for genomes and protein sequences. *Nucleic Acids Res.*, 30(1):31–34, 2000.
- [7] N. Goffard, V. Garcia, F. Iragne, A. Groppi, and A. de Daruvar. Ippred: server for proteins interactions inference. *Bioinformatics*, 19:903–904, 2003.
- [8] M. Hayashida, N. Ueda, and T. Akutsu. Interring strengths of protein-protein interactions from experimental data using linear programming. *Bioinformatics*, 19(Suppl. 2):ii58–ii65, 2003.
- [9] M. Hayashida, N. Ueda, and T. Akutsu. A simple method for interring strengths of protein-protein interactions. *Genome Informatics*, 15(1):56–68, 2004.
- [10] Y. Ho, A. Gruhler, and A. Heilbut et al. Systematic identification of protein complexes in *saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415:180–183, January 2002.
- [11] J. Hooker. Resolution and the integrality of satisfiability problems. *Mathematical Programming*, 74:1–10, 1996.
- [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. *Proc. Natl. Acad. Sci.*, 98(8):4569–4574, 2001.
- [13] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl. Acad. Sci.*, 97(3):1143–1147, 2000.
- [14] W. K. Kim, J. Park, and J. K. Suh. Large scale statistical prediction of protein-protein interaction by potentially interacting domain (pid) pair. *Genome Informatics*, 13:42–50, 2002.
- [15] R. Mrowka, A. Patzak, and H. Herze. Is there a bias in proteome research? *Genome Res*, 11(12):1971–1973, December 2001.

- [16] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences (PNAS)*, 96:4285–4288, 1999.
- [17] E. Sprinzak and H. Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–692, 2001.
- [18] P. Uetz., L. Cagney, and G. Mansfield et al. A comprehensive analysis of protein-protein interactions in *saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, 2000.
- [19] J. Wojcik and V. Schachter. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 17(Suppl. 1):S296–S305, 2001.