Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 37349, Pages 1–14
DOI 10.1155/ASP/2006/37349

Towards Inferring Protein Interactions: Challenges and Solutions

Ya Zhang,[1,2] Hongyuan Zha,[3] Chao-Hsien Chu,[4] and Xiang Ji[5]

[1] *Information and Telecommunication Technology Center, The University of Kansas, Lawrence, KS 66045, USA*
[2] *Department of Electrical Engineering and Computer Science, The University of Kansas, Lawrence, KS 66045, USA*
[3] *Department of Computer Science and Engineering, School of Engineering, Pennsylvania State University, University Park, PA 16802, USA*
[4] *College of Information Sciences and Technology, Pennsylvania State University, University Park, PA 16802-6823, USA*
[5] *NEC Laboratories America, Inc., Cupertino, CA 95014, USA*

Received 1 May 2005; Revised 13 October 2005; Accepted 15 December 2005

Discovering interacting proteins has been an essential part of functional genomics. However, existing experimental techniques only uncover a small portion of any interactome. Furthermore, these data often have a very high false rate. By conceptualizing the interactions at domain level, we provide a more abstract representation of interactome, which also facilitates the discovery of unobserved protein-protein interactions. Although several domain-based approaches have been proposed to predict protein-protein interactions, they usually assume that domain interactions are independent on each other for the convenience of computational modeling. A new framework to predict protein interactions is proposed in this paper, where no assumption is made about domain interactions. Protein interactions may be the result of multiple domain interactions which are dependent on each other. A conjunctive norm form representation is used to capture the relationships between protein interactions and domain interactions. The problem of interaction inference is then modeled as a constraint satisfiability problem and solved via linear programing. Experimental results on a combined yeast data set have demonstrated the robustness and the accuracy of the proposed algorithm. Moreover, we also map some predicted interacting domains to three-dimensional structures of protein complexes to show the validity of our predictions.

Copyright © 2006 Hindawi Publishing Corporation. All rights reserved.

1. INTRODUCTION

Proteins usually perform their functions in a collaborative fashion by interacting with each other. Uncovering the complex structures of protein interaction network is essential for understanding how proteins in a cell function together. Many computational efforts have been made to predict interacting proteins. The gene fusion/Rosetta method [1, 2] predicts a pair of proteins to interact if they are encoded separately as two distinct genes in one organism and are encoded by one single gene (fused) in another organism. Several other algorithms explore the use of protein sequences [3], protein structure [4], phylogenetic profiles [5], protein homology [6], gene neighborhood [7], and gene expression correlation [8] for inferring protein-protein interactions. Those methods are mostly based on protein sequence homology or structure homology. For example, Goffard et al. [6] infer two proteins to interact if they are considered to be, respectively, homologous to a pair of interacting proteins according to BLAST search [9]. However, similarity in sequence or structure does not necessarily guarantee similarity in function. Hence the predictions are generally associated with high error rates.

Recent advances in proteomics have opened up new opportunities for studying protein interactions. A large volume of protein interaction data has been generated with high-throughput experimental approaches including yeast two-hybrid genetic screens [10, 11] and mass spectrometric analysis [12], making possible genome-wide analysis of protein interactions. However, these high-throughout experiments inevitably contain many false positives and false negatives [13]. For example, two genome-wide yeast interaction data sets obtained via independent experiments [10, 11, 14] have less than 4% overlap of the identified interactions. This fact implies that these high-throughput interactions 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 experimentally verify individual

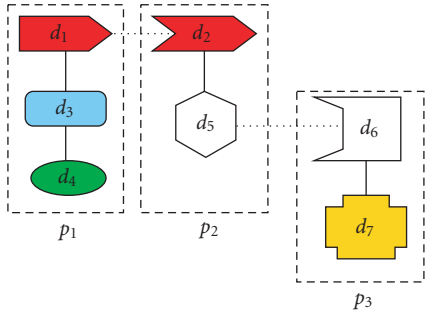


Figure 1: A sketch illustration of how domain interaction contributes to protein interaction. Protein $p1$ and protein $p2$ interact through the binding of domain $d1$ and domain $d2$, while the interaction between domain $d5$ and domain $d6$ is responsible for the interaction of protein $p2$ and protein $p3$.

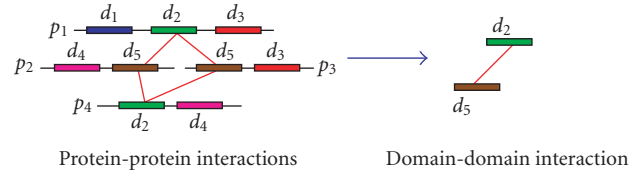


Figure 2: Domain-domain interaction provides an abstract representation of protein-protein interaction. Binding of domain $d2$ to $d5$ mediates the interaction between four pairs of proteins: proteins $p1$ and $p2$, proteins $p1$ and $p3$, proteins $p2$ and $p4$, and proteins $p3$ and $p4$.

interactions. The question—can we infer useful protein-protein interaction information from those high-throughput data—arises.

An important factor contributing to protein interactions is the domain composition of the proteins. Domains are believed to be responsible for protein interactions—proteins interact through their interacting domains (Figure 1). Because domains are deemed as the building blocks of proteins, an abstract representation of interactome is achieved at the domain level (Figure 2). Moreover, this representation facilitates the discovery of unobserved protein-protein interactions. Several computational approaches were motivated by this representation and predict protein interactions based on domain composition of proteins [15–20]: first domain-domain interactions are inferred from high-throughput protein interactions and then the putative domain interactions are used to predict interacting proteins.

As one of the pioneering studies, an association method was proposed for inferring over-represented sequence-signature (domain) pairs [19]. Association methods generally assume that co-occurrence of a domain pair in many interacting proteins indicates association—in this case, interaction among the pair of domains. This simple association method may assign high scores to some domain pairs with low frequency and the score does not correspond well to the possibility of interaction. Later Kim et al. [17] improved this association method by taking into consideration the number of domains in each protein, and Hayashida et al. [16] extended this method to numerical interaction data. The above association methods are limited in the sense that domain-domain interactions are computed locally, which ignores the contextual information for each domain, such as the neighbors of the domains.

A graph-theoretical approach, which combines sequence similarity search with clustering based on interaction patterns and interaction domain information, was proposed in [20]. The use of domain profile pairs were showed to provide better predictions than those solely using protein sequences. However, this method requires a high-quality protein interaction map, which is very expensive to obtain in the first place, to infer protein interactions in another organism.

More recently, several other studies adopted an optimization framework. Deng et al. [15] proposed a probabilistic model for protein interactions and developed a global method to inferring interacting domains by maximizing the likelihood of the observed data. Experimental errors were integrated into the likelihood function as two additional parameters (false positive and false negative). The expectation and maximization (EM) algorithm was used to optimize the parameters. Hayashida et al. [21] added 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. The authors tried to minimize the sum of differences between the computed strength and the predicted probabilities in training data with linear programing. One advantage of the method is that constraints can be easily integrated and thus this method can be easily combined with other existing methods. However, for the ease of computational modeling, the above probabilistic models assume that the domain interactions are independent of each other. This conjecture might be the major source of errors for these domain-based predictions because protein-protein interaction could be mediated by multiple domain interactions and these domain interactions may not be independent.

To overcome the above limitation, we propose here a new framework of learning without enforcing the independence assumption between domain interactions. The protein-protein interactions are interpreted as the result of domain interactions, either dependent or independent. Hence, our approach is more inclusive than the previous ones. We express the relationships between protein interactions and domain interactions in conjunctive norm forms. This representation naturally leads to the formulation of the interaction inference problem as a satisfiability (SAT) problem. This problem is then solved with linear programing. The prediction framework is characterized in the following two aspects. First, the proposed framework makes no assumption on the dependency/independency 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 as additional constraints. The validity of the prediction method

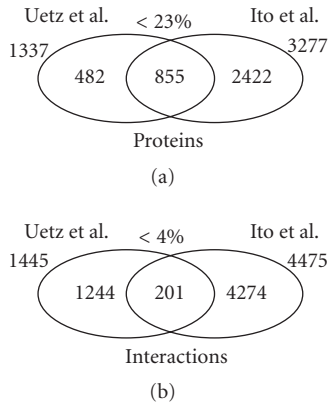


FIGURE 3: Overlap among the results of two independent large-scale yeast two-hybrid screens. The Venn diagram indicates the overlap among the interaction data obtained in two independent experiments [10, 11, 14]. (a) The overlap in terms of proteins. (b) The overlap in terms of interactions.

is evaluated with yeast protein interactions. Experimental results have demonstrated the robustness and accuracy of the proposed algorithm.

2. CHARACTERISTICS OF THE DATA

Although high-throughput experiments have greatly facilitated the study of protein interactions, the high-throughput data generally contain a large number of false negatives, creating big challenges in deciphering the interactome. For example, the genome-wide interaction data for yeast obtained in two independent experiments [10, 11, 14] only have less than four percentage of overlap for protein interactions (Figure 3). This lack of overlap between the data sets indicates that the screens to date are far from exhaustive and the yeast interactome may be much larger than previously estimated. Moreover, the observed protein-protein interaction matrix is quite sparse as shown in Figure 4. Most of the proteins are discovered to interact with only one protein. However, Hazbun and Fields [22] estimated that each protein interact with about 5 to 50 proteins. This fact again suggests that two-hybrid screens reveal a very small portion of the interactome. It is thus necessary to computationally predict potential interactions from experimentally identified interacting proteins.

Another significant feature of the data set is that the distribution of domain frequencies is highly skewed. Most domains occur in one or a few proteins and a few domains are observed frequently in the data set (Figure 5), which leads to substantially different frequencies among some domains. The difference in the frequencies could be problematic for association-based methods for interaction prediction; for example, if domain d_1 occurs only once in protein p_1, and domain d_2 occurs in all proteins. Although we only observed the domain pair d_{12} once, it could still be significant because domain d_1 only occurs once. Most association-based methods do not perform well when the pair of domains have very different frequencies.

3. INFERRING INTERACTING DOMAIN PAIRS

Our framework of inferring interacting domain pairs is built upon a widely accepted hypothesis 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 proteins under investigation as $P = \{p_1, p_2, \ldots, p_M\}$ and their corresponding domains as $D = \{d_1, d_2, \ldots, d_N\}$, where M and N are the number of proteins and domains. The set of domain pairs contained in the protein pair $\langle p_i, p_j \rangle$ is then denoted with Ω_{ij}:

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

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

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

where $\vee$ means *logical* "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

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

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 the set of domain pairs $\Omega_{12} = \{d_{11}, d_{13}, d_{15}, d_{21}, d_{23}, d_{25}\}$. P_{12}, the interaction indicator of the protein pair $\langle p_1, p_2 \rangle$, is expressed in terms of the set of related domain indicators $P_{12} = D_{11} \vee D_{13} \vee D_{15} \vee D_{21} \vee D_{23} \vee D_{25}$.

The problem of inferring interacting domains from protein interactions is essentially to discover the set of domain interactions that best fit the protein interaction data. With the conjunctive norm form of representation, the inference task essentially is to assign values to domain interaction indicators D_{nm} $(n, m = \{1, \ldots, N\})$ and protein interaction indicators P_{ij} $(i, j = \{1, \ldots, M\})$ so that all the protein-domain interaction relationships expressed in (2) are satisfied. This objective naturally leads the formulation of the interaction inference problem 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 obtain a set of assignment for domain interaction indicators that could simultaneously fit into the whole interaction data. Therefore, rather than requiring the assignment to accommodate all the protein interactions, we set the

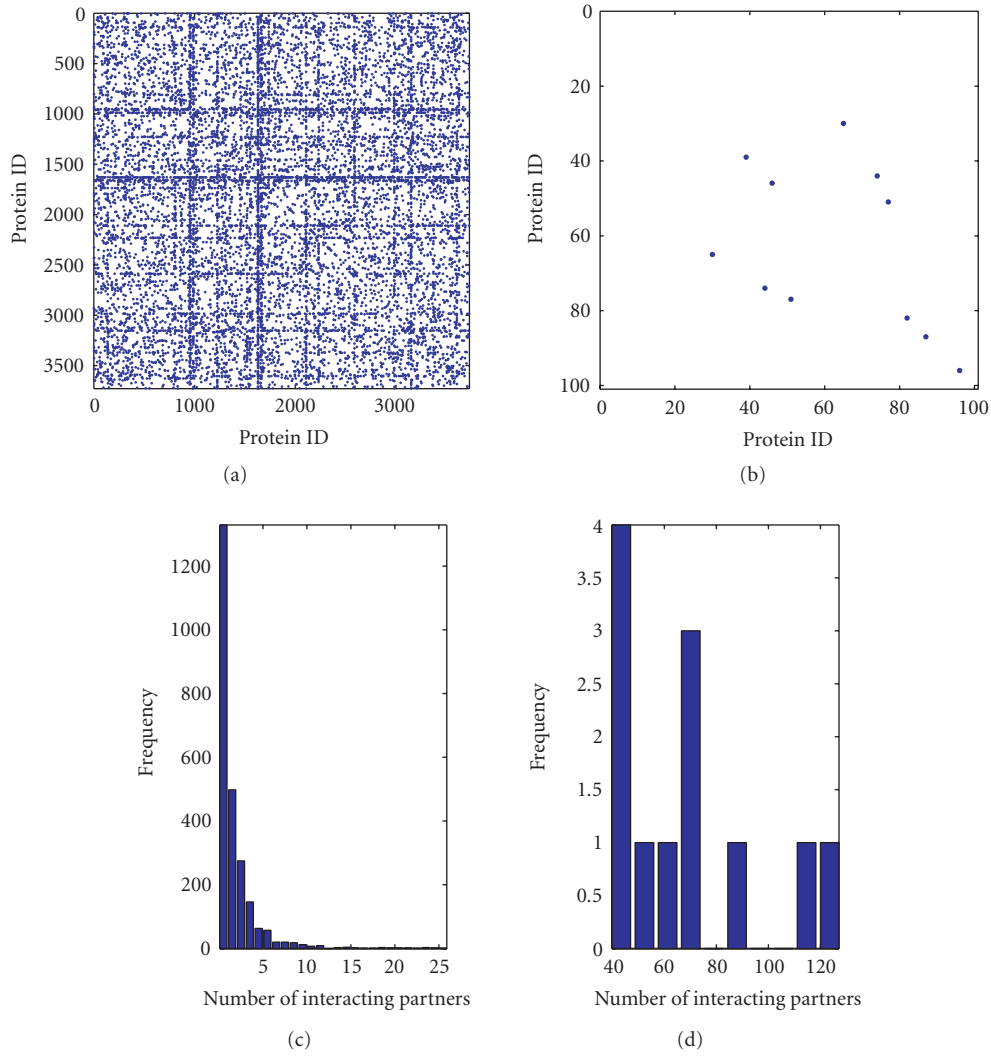


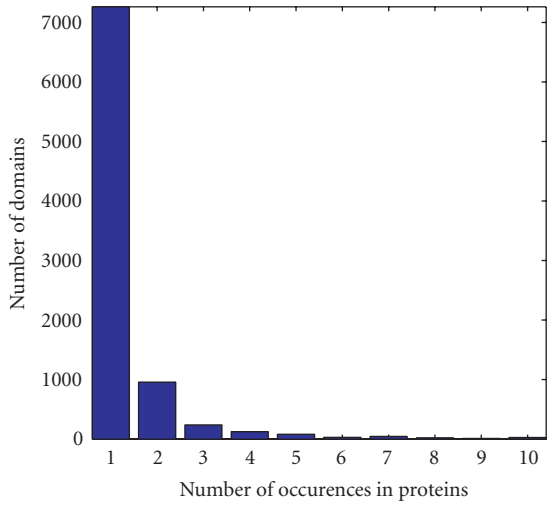
Figure 4: The interaction matrix is very sparse. Most proteins interact with one or a few proteins. (a) The interaction matrix of a combined yeast interaction data set obtained by [10, 11, 14]. (b) A submatrix of the interaction matrix in (a). (c), (d) Histograms for the number of interacting partners of a protein.

objective as to maximize the number of relationships (as expressed in (2)) that are satisfied based on the domain-protein interaction indicators assigned. This objective coincides with those of maximum satisfiability (MAX-SAT) problems.

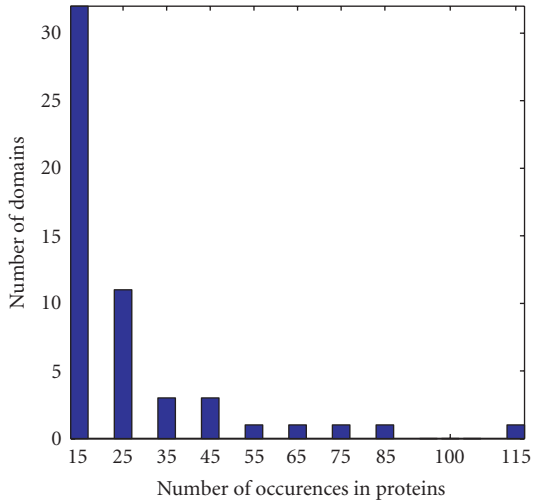
Definition 2. Given a set of p clauses in conjunctive normal form over q variables, the *maximum satisfiability* (MAX-SAT) problem is to obtain a truth assignment for the q variables so that a maximum number of the clauses are satisfied.

SAT and MAX-SAT problems are difficult to solve because of their large search space, and they have been known to be NP-hard [23]. Although a number of techniques have been developed to solve SAT and MAX-SAT problems [24, 25], finding optimal solutions for SAT and MAX-SAT problems is still an active research topic in artificial intelligence, logic, theory of computation, and many other 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 programing [26], a widely used techniques for MAX-SAT problems, is used to solve the inference problem. We employed linear programing for the solution of the MAX-SAT problem for several appealing reasons. First, the running time of linear programing is usually polynomial, while a pure combinatorial algorithm to solve the same problem usually requires exponential time complexity. Considering the unique variable in the MAX-SAT problem is usually quite large, the polynomial solution of linear programing is preferred. Later in this section, we will show two additional advantages of linear programing solution: ability to model the strength of the interaction and to easily incorporate prior knowledge.

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



(a)



(b)

Figure 5: Histogram for the number of proteins in which each domain occurs. If a domain occurs in a protein multiple times, only one is counted.

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, that is,

$$\begin{aligned} \max f &= \sum_{ij} \left(1 - \left|P_{ij} - P'_{ij}\right|\right) \\ \text{subject to } P'_{ij} &= \vee_{d_{nm}\in\Omega_{ij}} D_{nm} \quad (\forall i, j), \end{aligned} \tag{4}$$

where $D_{nm} \in \{0,1\}$ and $P_{ij} \in \{0,1\}$ (for all m, n, and i, j). P_{ij} is the interaction indicator for proteins p_i and p_j according to experimental 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, $|P_{ij} - P'_{ij}| = 1$. Thus, the above objective function counts the number of protein-domain interaction relationships satisfied. This objective is equivalent to minimizing 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:

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

The inequality constraints in (5) are from the constraints in (4) and they ensure that a protein pair is deemed to be interacting only if at least one of the domain pairs in the protein pair is considered interacting, as P_{ij} is either 1 or 0. Equation (6) may be reformulated as

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

The linear programing problem is NP-hard when the variables are restricted to integers. A suitable approximation is to use probabilistic methods. We solve the relaxed linear program by loosing the integer constraints on the matrixes D and P' in (6). D_{nm} and P'_{ij} are allowed to assume any real value in the interval of $[0, 1]$:

$$\begin{aligned} \text{minimize } & \sum_{P_{ij}=0} P'_{ij} - \sum_{P_{ij}=1} P'_{ij} \\ \text{subject to } & \sum_{d_{nm}\in\Omega_{ij}} D_{nm} \geq P_{ij} \quad (\forall i, j), \\ & 0 \leq P'_{ij} \leq 1 \quad (\forall i, j), \\ & 0 \leq D_{nm} \leq 1 \quad (\forall n, m). \end{aligned} \tag{7}$$

Let $\widehat{D_{nm}}$ be the value obtained for variable D_{nm} and $\widehat{P_{ij}}$ for P'_{ij} after solving the linear program. These real number values obtained for D_{nm} and P'_{ij} represent the probability of picking the integer value 1 for them. The real-number solutions have advantages over Boolean solutions for their ability to capture the probabilities of protein interactions and domain interactions. To convert the interactions into Boolean format, we only need to select a threshold and quantize the values to 0 or 1 based on the threshold. Another advantage of using linear programing to solve the MAX-SAT problem is that the formulation as an optimization problem subject to constraints naturally facilitates the integration of prior knowledge about interaction as additional constraints.

4. EXPERIMENTAL RESULTS

To infer the interacting proteins, we use the yeast interaction data set as prepared in [15], which is a combination of interactions obtained from large-scale yeast two-hybrid screens on *Saccharomyces cerevisiae* genome [11, 14]. The data set includes 5719 interactions. The domain definitions of the yeast proteins are according to Pfam [27]. In total, 2918 Pfam domains are defined on the set of proteins. Proteins without defined domains are treated as superdomains.

For validation, the MIPS (Munich Information Center for Protein Sequences) physical interaction pairs [28] are used to evaluate the predictions. The MIPS data set contains 2575 pairs of interacting proteins but does not include any pair of noninteracting proteins. We randomly generate a set of noninteracting protein pairs of size comparable to the number of the interacting protein pairs. Protein pairs which do not contain any domain pair in the training set are deleted because no information about their interaction may be obtained from the training set. This deletion results in a test set of 2099 interactions.

The GNU Linear Programing Kit[1] (version 4.7) is used for solving linear programs on Unix. In particular, a polynomial time linear programing algorithm using an interior point method is used to solve the linear programs. Interior point method is known to be more efficient than the simplex method. This former method achieves optimization by going through the middle of the solid defined by the problem rather than around its surface. The prediction algorithm is mainly implemented in Perl, and the experiments are performed on a SUN Ultra 60 server (450 MHz) with 1 GB RAM.

The performance of the algorithm is evaluated in terms of sensitivity (Sen) and specificity (Spe). 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 ($tp + fp$):

$$\begin{aligned} \text{Sen} &= \frac{tp}{tp + fn}, \\ \text{Spe} &= \frac{tp}{tp + fp}. \end{aligned} \tag{8}$$

4.1. Training

The yeast interaction data set only contains pairs of interacting proteins, which are so-called positive training examples. We are lack of negative training examples because the yeast data set provides no information about the noninteracting proteins. A common approach to obtain negative examples is to use the set of all pairs of proteins excluding the interacting proteins as negative training examples. However, several major issues are raised regarding this solution. First, considering high false negatives (≥ 0.64, according to [15]) of the yeast interaction data set, many interacting protein pairs remain undiscovered. Using all pairs of proteins excluding the interacting proteins as negative training examples will guarantee to include all those false negatives. Secondly, the number of all pairs of proteins is $n(n+1)/2$, where n is the number of proteins in the data set. In the case of the yeast data set, we have 6359 yeast proteins and 5719 interactions. The number of all pairs of proteins is in the order of 2×10^7, four magnitude larger than that of the positive examples. Therefore, the training examples would be very imbalanced if all pairs of proteins are used for training. Moreover, using all pairs of proteins for training demands considerable computational costs.

Considering the above limitations, we generate a subset of noninteracting protein pairs by randomly coupling the proteins which are not observed to interact in the experiments. Now what we need decide is the number of "negative" examples selected. We express the training data in a parametric form as

$$\text{Train}(t) = |\text{Positive}| + |\text{AllPair} - \text{Positive}| \times t, \tag{9}$$

where t is a real number ($0 < t < 1$), $|\cdot|$ represents the size of the set, and $\text{Train}(t)$ is the size of the training data with parameter t. In the actual experiments, we use the parameter

$$\text{NegRatio} = \frac{|\text{Negative}|}{|\text{Positive}|} \tag{10}$$

to indicate the number of "negative" examples selected. As |Positive| is fixed, this ratio is clearly in proportion to the parameter t. We perform experiments with different values of NegRatio and report the results in Figure 6. We start with a training setting of positive examples only, and gradually include more and more negative examples. Intuitively, including a proper number of negative examples increases the specificity of the prediction with minimal loss of sensitivity. Seen from the plots, initially, adding more negative examples for training results in an increased specificity and a reduced sensitivity. However, for NegRatio > 10, the specificities tend to be stable and only slightly fluctuate by random. In the mean while, the sensitivity still keeps decreasing. This phenomenon may be related to the fact that the number of interacting protein pairs treated as negative examples increases with the growing number of negative examples. A reasonable value for NegRatio is 10.

4.2. Results

As the EM method is considered the best among existing methods [21], we here compare the performance of our method with that of the EM method. Our method is referred to as the SAT method thereafter. Setting NegRatio $= \{0, 1, \ldots, 20\}$, we test the SAT method and the EM method on the same sets of interaction data and report their results in Table 1. For all predictions, the threshold is set to 0.6. The experimental results show that the EM method generally predicts at relative high sensitivities while the SAT method

[1] http://www.gnu.org/software/glpk/glpk.html (accessed on April 8th, 2005)

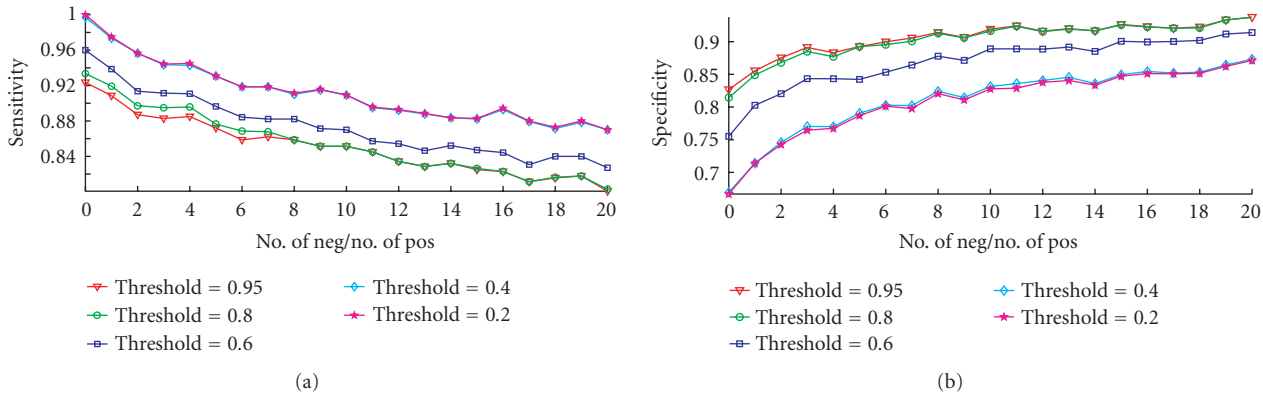


Figure 6: The impact of negative training examples on specificity and sensitivity. The x axis indicates the ratio of the number of randomly selected negative examples to the number of positive examples. The y axis is the sensitivity (a) and specificity (b). The circles, squares, diamonds, triangles, and pentagrams represent the sensitivity/specificity at different interaction thresholds (0.95, 0.8, 0.6, 0.4, and 0.2, resp.).

Table 1: Performance comparison of the SAT method and the EM method at different NegRatio. The threshold for the predictions is set at 0.6. The metrics reported here are sensitivity, specificity, and F-score.

NegRatio	SAT			EM		
	Sen	Spe	F-Score	Sen	Spe	F-Score
0	0.96	0.755	0.845	0.965	0.733	0.833
1	0.939	0.803	0.865	0.967	0.731	0.833
2	0.914	0.820	0.865	0.967	0.729	0.831
3	0.911	0.843	0.876	0.968	0.743	0.840
4	0.911	0.843	0.876	0.974	0.745	0.844
5	0.896	0.842	0.869	0.958	0.738	0.834
6	0.884	0.853	0.869	0.967	0.740	0.838
7	0.882	0.864	0.873	0.970	0.735	0.836
8	0.882	0.878	0.880	0.973	0.743	0.843
9	0.871	0.871	0.871	0.967	0.745	0.842
10	0.87	0.889	0.879	0.970	0.736	0.837
11	0.857	0.889	0.873	0.962	0.741	0.837
12	0.854	0.889	0.871	0.960	0.751	0.843
13	0.846	0.895	0.868	0.967	0.738	0.837
14	0.852	0.885	0.868	0.959	0.751	0.842
15	0.847	0.901	0.873	0.968	0.748	0.844
16	0.844	0.900	0.871	0.967	0.743	0.840
17	0.831	0.900	0.864	0.967	0.742	0.840
18	0.84	0.902	0.870	0.964	0.743	0.839
19	0.84	0.912	0.874	0.971	0.743	0.842
20	0.827	0.914	0.868	0.959	0.744	0.838

predicts at relative high specificity. Moreover, the sensitivity and specificity of the EM method seem to be uncorrelated to the number of negative examples included in the training set (see Table 1 and Figure 7). On the other hand, the number of negative examples included has a clear impact on the performance of SAT approach. Including more negative examples increases the specificity of SAT method at the cost of a lower sensitivity. To compare the two methods, in addition to sensitivity and specificity, we introduce F-score which combines the two former metrics to score the prediction,

$$F\text{-score} = \frac{2\,\text{Spe} \times \text{Sen}}{(\text{Spe} + \text{Sen})}. \tag{11}$$

We calculate F-score for each training run and the results are also listed in Table 1. The F-scores of the SAT methods are higher than those of the EM method (P-value less than 0.0001).

For the purpose of interaction prediction, we are more interested in discovering interacting proteins rather than noninteracting proteins. That is, errors in predicted interacting proteins (fp) are less tolerable than those in predicted noninteracting proteins (fn). Thus, specificity is a more important metric than sensitivity. The predictions by the SAT method generally have higher specificities than those by the EM method as seen from Figure 7 (different NegRatio while threshold is set to 0.6) and Figure 8 (different threshold values while NegRatio is set to 10). In this sense, we are more in favor of the SAT method.

We employ a polynomial time linear programing algorithm using an interior point method (provided by the GNU Linear Programing Kit) to solve the linear programs. Table 2 and Figure 9 show the running time of the GNU LP program with different number of variables.

To compare the predictions made by the SAT method and the EM method, we plot the predicted protein-protein interaction matrixes of the two methods as shown in Figure 10(a) (NegRatio = 10 and threshold = 0.6). In these plots, each row and each column represent a protein. A circle means that the proteins at the corresponding row and column interact according to SAT prediction. Similarly, a triangle indicates that the proteins at the corresponding row and column interact according to EM prediction. The protein interactions in the testing set are indicated by dots. The two methods produce about 75.5% overlaps in their predictions about protein interaction (either interacting or noninteracting). When this overlapped portion is compared with the testing interactions

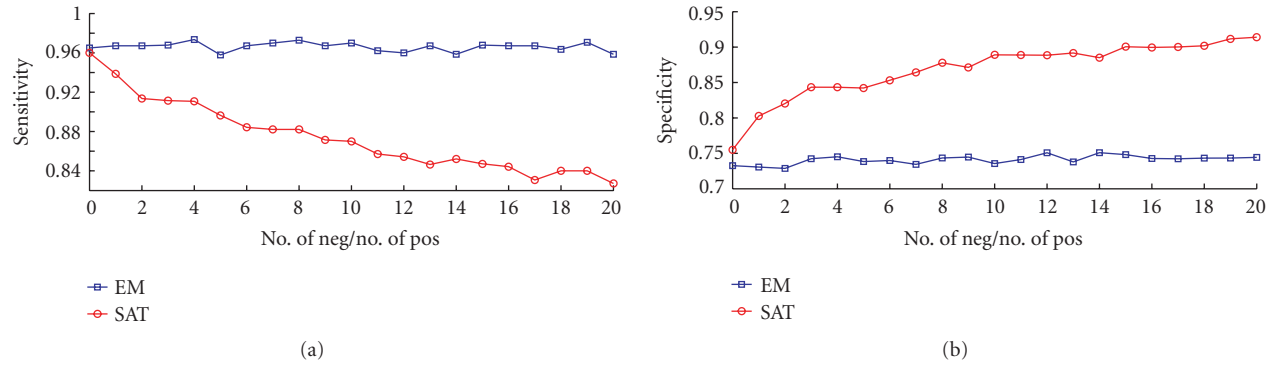


Figure 7: Comparison of how specificity and sensitivity change with different NegRatio for the SAT method and the EM algorithm. The threshold for the predictions is set at 0.6. The lines with circles represent the performance of the SAT method, while the lines with squares represent that of the EM method.

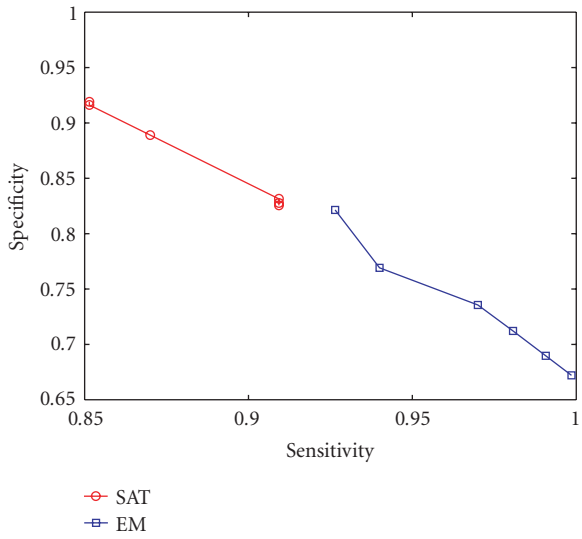


Figure 8: Comparison of specificity and sensitivity of our algorithm to those of the EM algorithm (NegRatio = 10).

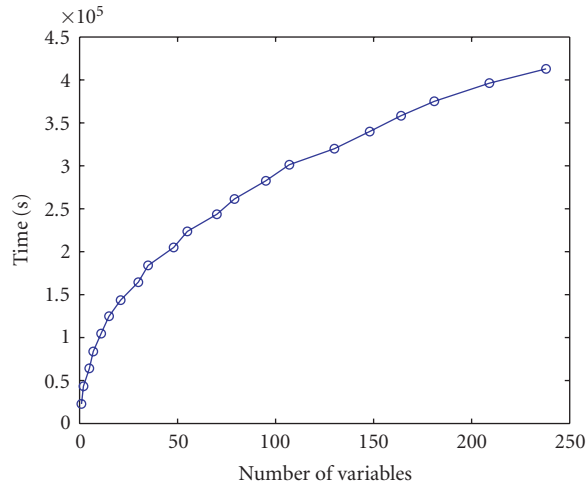


Figure 9: Running time of GNU LP program with different number of variables.

(Figure 10), it results in a slightly higher specificity of 0.899 at a sensitivity of 0.867.

4.3. *Structural evidences for the predicted domain interactions*

Biological validation of the predictions is by no means a trivial task. The lack of a golden test set for domain interactions is the major reason that a statistically significant test is infeasible. Here we use some examples to illustrate some of the predictions.

Recently, iPfam[2] has been built as a resource containing domain-domain interactions observed in protein data bank (PDB) entries. For each entry in PDB, Pfam domains are first projected onto the structure. Then, the distances between each pair of domains are computed to decide whether interactions are formed between these domains. The domain interactions logged in iPfam include inter-protein or intra-protein ones, while our predictions only cover those between proteins. Therefore, it is expected that our prediction only matches to a portion of iPfam interactions. The predicted domain-domain interactions are compared with those contained in iPfam. Table 3 list some of those domain-domain interactions.

As there is very limited information on domain interactions available, here we attempt to draw evidences from structures of interacting proteins or protein complexes to validate our predictions about interacting domains. First let us look at the complex structure of the protein *cyclin a* and the protein *cyclin-dependent kinase 2* (PDB ID 1*fin*). According to Pfam, *cyclin a* contains two copies of PF00069

[2] http://www.sanger.ac.uk/Software/Pfam/iPfam/.

Table 2: The running time of GNU LP with different number of variables.

NegRatio	0	1	2	3	4	5	6	7	8	9	10
n_{negative}	0	5719	11438	17157	22876	28595	34314	40033	45752	51471	57190
n_{positive}	5719	5719	5719	5719	5719	5719	5719	5719	5719	5719	5719
$n_{\text{variables}}$	22738	43417	64030	83801	104718	124775	143744	164518	183948	204905	223661
T_{LP} (seconds)	1.0	2.0	5.0	7.0	11.0	15.0	21.0	30.0	35.0	48.0	55.0

NegRatio	11	12	13	14	15	16	17	18	19	20
n_{negative}	62909	68628	74347	80066	85785	91504	97223	102942	108661	114380
n_{positive}	5719	5719	5719	5719	5719	5719	5719	5719	5719	5719
$n_{\text{variables}}$	243500	261383	282568	301274	319929	339958	358401	375141	396173	412924
T_{LP} (seconds)	70.0	79.0	95.0	107.0	130.0	148.0	164.0	181.0	209.0	238.0

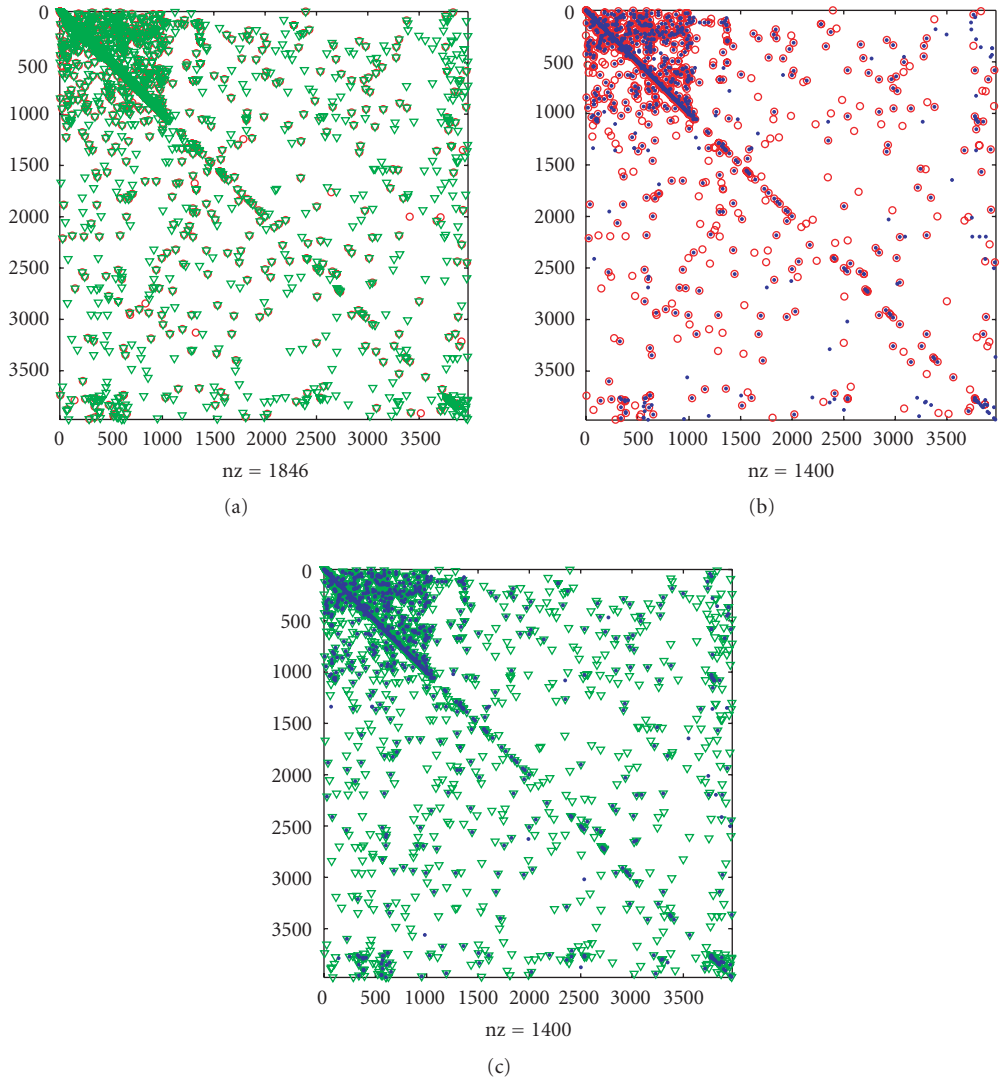


Figure 10: The degree of overlap among testing protein interactions, predicted interactions by SAT approach and EM approach. The NegRatio and threshold of the prediction are set to 10 and 0.6, respectively. (a) Overlap of predicted protein interactions by SAT methods (circles) and those by EM methods (triangles). (b) Overlap of predicted protein interactions by SAT methods (circles) and the testing set (dots). (c) Overlap of predicted protein interactions by EM methods (triangles) and the testing set (dots).

Table 3: Examples of predicted domain-domain interactions that matches the predictions by iPfam.

Domain 1	Domain 2	Domain 1	Domain 2
PF02984	PF00069	PF00134	PF00069
PF00023	PF00069	PF00378	PF00378
PF00786	PF00069	PF00043	PF02798
PF02115	PF00071	PF02826	PF00389
PF02629	PF00389	PF00581	PF00581
PF01842	PF00389	PF00995	PF00804
PF00227	PF00227	PF00227	PF00389
PF00491	PF00491	PF00675	PF00675
PF00631	PF00400	PF00091	PF00389
PF00503	PF00400	PF01111	PF00069
PF00389	PF00137	PF00389	PF00004
PF00291	PF00585	PF00389	PF00400
PF01466	PF00646	PF01466	PF00888

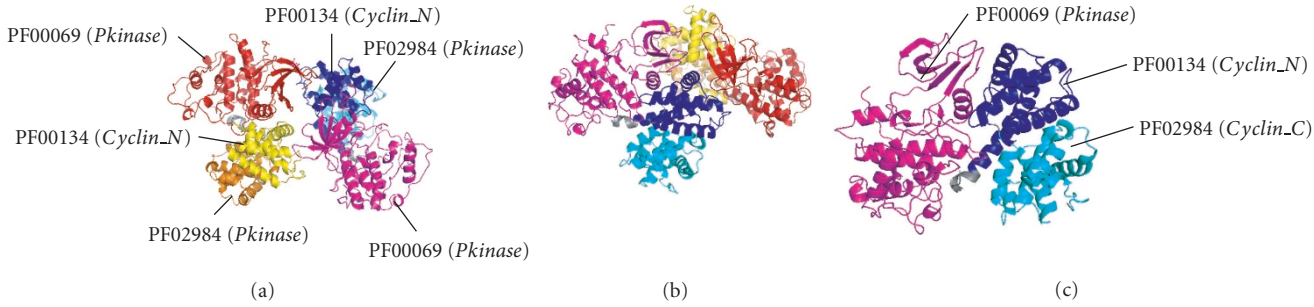


Figure 11: The 3-D structure of *cyclin a—cyclin-dependent kinase 2 complex* (PDB ID 1*fin*). The structure shows how *cyclin-dependent kinase 2* binds to *cyclin a*. The Pfam domains are graphed on the structure and labelled in color. Two PF00069 (*Pkinase*) domains are marked in red and purple, respectively. Two PF00134 (*Cyclin_N*) domains are colored in blue and yellow, respectively. The protein segments in cyan and orange are PF02984 (*Cyclin_C*) domains. (a), (b) The complex structure is captured from different angles to show how the domains contact with each other. (c) Part of the structure is shown to indicate how the three domains contact with each other.

(*Pkinase*) domains, while *cyclin-dependent kinase 2* contains two copies of PF00134 (*Cyclin_N*) domains and two copies of PF02984 (*Cyclin_C*) domains. We graph these domains on the PDB structure (see Figure 11). The complex structure is captured from different angles to show how the domains contact with each other. As shown in the structure, the PF02984 (*Cyclin_C*) domain and the PF00134 (*Cyclin_N*) domain both interact with the PF00069 (*Pkinase*) domain. Moreover, according to our prediction, $D_{\text{PF02984,PF00069}} = 0.58$, and $D_{\text{PF00134,PF00069}} = 1$. From Figure 11(c), we can see that the area of contact between PF00134 and FP00069 is actually larger than that between PF02984 and PF00069. It seems that our algorithm is able to successfully predict not only the domain interactions but also the relative strength of the domain interactions.

Another evidence supporting our prediction that the PF00023 (*Ank*) domain interacts with the PF00069 (*Pkinase*) domain is obtained from the three-dimensional (3-D) structure of the *P18(Ink4C)-Cdk6-K-Cyclin ternary complex* (PDB ID 1*g*3*n*) (see Figure 12). As indicated by its name, the complex contains three proteins: *cyclin-dependent kinase 6 (cdk6)*, *cyclin-dependent kinase 6 inhibitor (P18(Ink4C))*, and *V-Cyclin (K-Cyclin)* (grey). According to Pfam, cyclin-dependent kinase 6 contains *Pkinase* domains, while cyclin-dependent kinase 6 inhibitor contains *Ank* domains. Two additional examples are shown in Figure 13, where the complexes structure of *rac-rhogdi* shows the interactions between the Pfam domains, PF02115 (*Rho_GDI*) and PF00071 (*Ras*) (Figure 13(a)), and the interaction between the Pfam domains, PF00043 (*GST_C*) and PF02798 (*GST_N*), is illustrated through the structure of the human glutathione s-transferase p1-1 in complex with ethacrynic acid-glutathione conjugate (Figure 13(b)).

4.4. 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

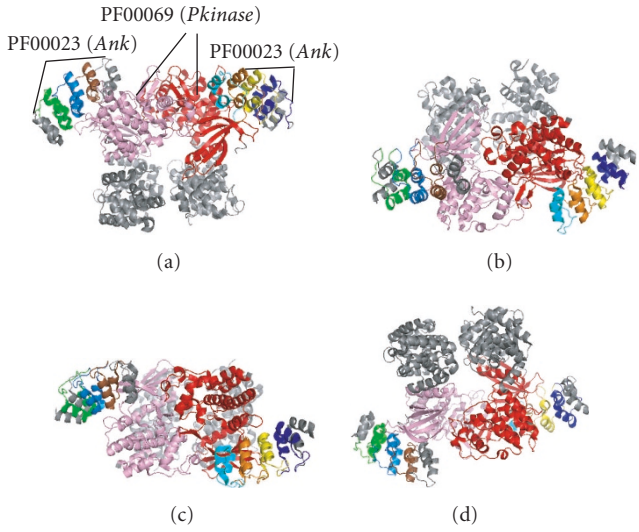


Figure 12: The 3-D structure of a *P18(Ink4C)-Cdk6-K-Cyclin ternary complex* (PDB ID 1*g*3*n*). The complex contains three proteins: *cyclin-dependent kinase 6 (cdk6), cyclin-dependent kinase 6 inhibitor (P18(Ink4C)), and V-Cyclin (K-Cyclin)*. The Pfam domains are graphed on the structure and labelled in color. Two PF00069 (*Pkinase*) domains are marked in red and pink, respectively. Ten copies of PF00023 (*Ank*) domains are marked with other colors except grey. The complex structure is captured from different angles to show how the domains contact with each other.

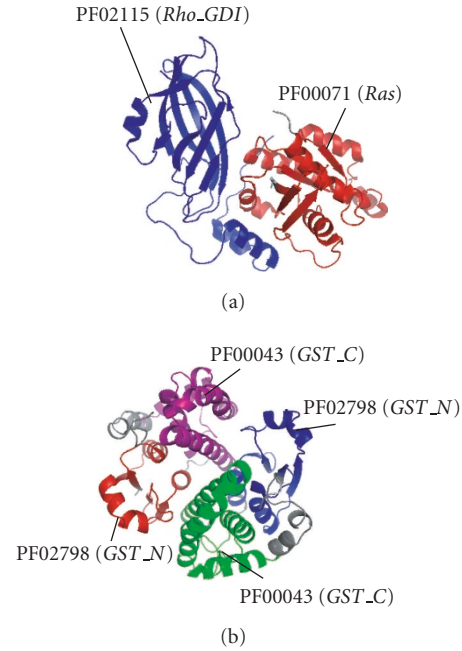


Figure 13: (a) The 3-D structure of a rac-rhogdi complex. The complex contains ras-Related C3 Botulinum Toxin Substrate 2 (P21-Rac2) and rho GDP-Dissociation Inhibitor 2 (rho Gdi 2, rho-Gdi beta, Ly-Gdi). The Pfam domains are graphed on the structure and labelled in color. The PF00071 (*Ras*) domain is marked in red. The PF02115 (*Rho_GDI*) domain is colored in blue. (b) The 3-D structure of the human glutathione s-transferase p1-1 in complex with ethacrynic acid-glutathione conjugate. Two copies of the PF02798 (*GST_N*) domains are marked in red and blue, respectively. Two copies of the PF00043 (*GST_C*) domains are colored in purple and green, respectively.

significant prediction we made is the interaction between protein PAP1, an 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 associated 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 are usually involved in the same cellular processes, we may predict that YMR291W is involved in the response to stresses.

5. DISCUSSIONS AND CONCLUSIONS

Inferring protein interaction is a very challenging problem due to the high level of noise in the interaction data and limited information about the protein interactions. Existing domain-based methods tend to oversimplify the problem by introducing the assumption that the domain interactions are independent from each other. In our study, the protein-protein interactions are interpreted as the result of domain interactions which are not necessarily independent of each other. The relationships between protein interactions and domain interactions are expressed in conjunctive norm forms, which enables us to formulate the problem of interaction inference as a satisfiability (SAT) problem. The inference problem is then relaxed and 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 and is a more natural way of modeling the relationship between protein-protein interactions and domain-domain interactions. Secondly, when formulating the inference problem as a MAX-SAT problem, prior knowledge about domain interaction or protein interaction may be easily input into the framework as additional constraints. The validity of the prediction method is evaluated with yeast protein interactions. Our method achieves a sensitivity of 87.0% and a specificity of 88.9% at the threshold 0.6 (NegRatio = 10) on a combined yeast data set. Compared with the MLE-EM method, our method is able to predict at a higher specificity while maintaining a reasonable sensitivity. Attempts were made to validate our prediction on domain interactions by inspecting the

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
LSM1	Component of small nuclear ribonucleoprotein complexes involved in mRNA decapping and decay	MUD1	U1 snRNP A protein, homolog of human U1-A; involved in nuclear mRNA splicing
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 is responsive to stresses including heat, ethanol, and sodium arsenite	YMR291W	ORF, uncharacterized

positions of the domains in some protein complexes based on their structure information deposited in PDB. Our method correctly predicted the interactions among domains. Further more, the scores assigned to each pair of domains also correspond to the strength of the interaction.

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 a sensitivity around 87%. 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.

The domain-based approaches to infer protein-protein interactions usually do not differentiate interaction domains and catalytic domains. However, the interaction domains are more likely to mediate protein interaction. Interaction domains are believed to be more likely to mediate specific protein-protein interactions. Unique characteristics have been revealed about interaction domains in terms of their lengths, structures, and frequency in genomes [29]. Moreover, proteins containing the same interaction domains are often observed to have very diverse functions. For example, SH2 domain containing proteins perform functions that include regulation of protein/lipid phosphorylation, phospholipid metabolism, transcriptional regulation, cytoskeletal organization, and control of Ras-like GTPases. However, our current understanding of interaction domains is still limited to a few well-studied ones such as SH2 domains. An automatic method may be developed to identify interaction domains in proteins. This result may then be used to help the further identification of interacting domains and proteins and improve the accuracy of protein interaction prediction.

ACKNOWLEDGMENTS

The authors are thankful to Dr. Stephen R. Holbrook, Dr. Chris Ding, and Dr. Xue-Wen Chen for their insightful discussions and comments on the manuscript. The authors would also like to thank the anonymous reviewers and editors for their helpful comments.

REFERENCES

[1] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis, "Protein interaction maps for complete genomes based on gene fusion events," *Nature*, vol. 402, no. 6757, pp. 86–90, 1999.

[2] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg, "Detecting protein function and protein-protein interactions from genome sequences," *Science*, vol. 285, no. 5428, pp. 751–753, 1999.

[3] J. R. Bock and D. A. Gough, "Predicting protein-protein interactions from primary structure," *Bioinformatics*, vol. 17, no. 5, pp. 455–460, 2001.

[4] J. Park, M. Lappe, and S. A. Teichmann, "Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertories in the pdb and yeast," *Journal of Molecular Biology*, vol. 307, pp. 929–938, 2001.

[5] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein fucntions by comparative genome analysis: protein phylogenetic profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 8, pp. 4285–4288, 1999.

[6] N. Goffard, V. Garcia, F. Iragne, A. Groppi, and A. de Daruvar, "Ippred: server for proteins interactions inference," *Bioinformatics*, vol. 19, pp. 903–904, 2003.

[7] T. Dandekar, B. Snel, M. Huynen, and P. Bork, "Conservation of gene order: a fingerprint of proteins that physically interact," *Trends in Biochemical Sciences*, vol. 23, pp. 324–328, 1998.

[8] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, pp. 14863–14868, 1998.

[9] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, pp. 403–410, 1990.

[10] 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.

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

[12] Y. Ho, A. Gruhler, A. Heilbut, et al., "Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, pp. 180–183, 2002.

[13] R. Mrowka, A. Patzak, and H. Herze, "Is there a bias in proteome research?" *Genome Research*, vol. 11, no. 12, pp. 1971–1973, 2001.

[14] T. Ito, K. Tashiro, S. Muta, et al., "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," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 3, pp. 1143–1147, 2000.

[15] M. Deng, S. Mehta, F. Sun, and T. Chen, "Inferring domain-domain interactions from protein-protein interactions," in *Proceedings of the 6th Annual International Conference on Computational Biology (RECOMB '02)*, pp. 117–126, Washington, DC, USA, April 2002.

[16] M. Hayashida, N. Ueda, and T. Akutsu, "A simple method for interring strengths of protein-protein interactions," *Genome Informatics*, vol. 15, no. 1, pp. 56–68, 2004.

[17] 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*, vol. 13, pp. 42–50, 2002.

[18] S. K. Ng, Z. Zhang, and S. H. Tan, "Integrative approach for computationally inferring protein domain interactions," *Bioinformatics*, vol. 19, no. 8, pp. 923–929, 2003.

[19] E. Sprinzak and H. Margalit, "Correlated sequence-signatures as markers of protein-protein interaction," *Journal of Molecular Biology*, vol. 311, no. 4, pp. 681–692, 2001.

[20] J. Wojcik and V. Schächter, "Protein-protein interaction map inference using interacting domain profile pairs," *Bioinformatics*, vol. 17, suppl. 1, pp. S296–S305, 2001.

[21] M. Hayashida, N. Ueda, and T. Akutsu, "Interring strengths of protein-protein interactions from experimental data using linear programming," *Bioinformatics*, vol. 19, suppl. 2, pp. ii58–ii65, 2003.

[22] T. R. Hazbun and S. Fields, "Networking proteins in yeast," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4277–4278, 2001.

[23] D. Du, J. Gu, and P. Pardalos, *Satisfiability Problem: Theory and Application*, vol. 35 of *DIMACS Series in Discrete Mathematics*, American Mathematical Society, Providence, RI, USA, 1997.

[24] J. Gramm, E. A. Hirsch, R. Niedermeier, and P. Rossmanith, "New worst-case upper bounds for max-2-sat with application to maxcut," *Discrete Applied Mathematics*, vol. 130, no. 2, pp. 139–155, 2003.

[25] H. Zhang and H. Shen, "Exact algorithms for maxsat," *Electronic Notes in Theoretical Computer Science*, vol. 86, no. 1, pp. 1–14, 2003.

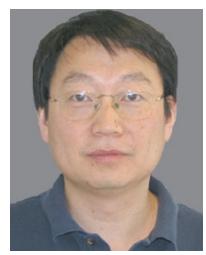
[26] J. Hooker, "Resolution and the integrality of satisfiability problems," *Mathematical Programming*, vol. 74, pp. 1–10, 1996.

[27] A. Bateman, L. Coin, R. Durbin, et al., "The pfam protein families database," *Nucleic Acids Research*, vol. 32, pp. D138–D141, 2004.

[28] H. W. Mewes, D. Frishman, C. Gruber, et al., "MIPS: a database for genomes and protein sequences," *Nucleic Acids Research*, vol. 28, no. 1, pp. 37–40, 2000.

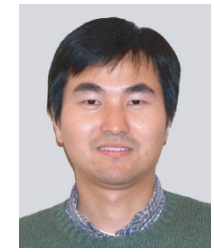
[29] T. Pawsona, M. Rainaa, and P. Nasha, "Interaction domains: from simple binding events to complex cellular behavior," *FEBS Letters*, vol. 513, pp. 2–10, 2002.

Ya Zhang is an Assistant Professor in the Department of Electrical Engineering and Computer Science at the University of Kansas. She received her B.S. degree from Tsinghua University, China, in 2000, and the Ph.D. degree in Information Sciences and Technology from the Pennsylvania State University in 2005. Her research interests include bioinformatics, computational biology, machine learning, data mining, statistical learning, text mining, and system biology.



Hongyuan Zha received the B.S. degree in mathematics from Fudan University, Shanghai, in 1984, and the Ph.D. degree in scientific computing from Stanford University in 1993. He is a Professor in the Department of Computer Science and Engineering at Pennsylvania State University, where he has worked since 1992. His research interests include scientific computing and machine learning, especially statistical and computational methods for nonlinear dimension reduction.

Chao-Hsien Chu is an Associate Professor of information sciences and technology and the Executive Director of the Center for Information Assurance at the Pennsylvania State University, University Park, PA (USA). He was previously on the faculty at Iowa State University (USA) and Baruch College (USA), and a Visiting Professor at the University of Tsukuba (Japan) and Hebei University of Technology (China). He is currently on leave to the Singapore Management University (Singapore) (2005–2006). He received a Ph.D. in business administration from Penn State University. His current research interests are in communication networks design, information assurance and security (especially in wireless security, intrusion detection, and cyber forensics), and intelligent technologies (fuzzy logic, neural network, genetic algorithms, etc.) for data mining (e.g., bioinformatics and privacy preserving) and systems management. His research papers have been published in Decision Sciences, the IEEE Transactions on Evolutionary Computation, IIE Transactions, Decision Support Systems, European Journal of Operational Research, Electronic Commerce Research, Expert Systems with Applications, International Journal of Mobile Communications, Journal of Operations Management, International Journal of Production Research, among others. He is currently on the editorial review board for a number of journals.



Xiang Ji received his B.S. degree from the University of Science and Technology of China in 1999 and his Ph.D. degree in computer science from The Pennsylvania State University in 2004. He has joined the NEC Labs. America as a Research Staff Member on intelligent information system research since 2004. His research interests include data mining, machine learning, and bioinformatics.