

# Chapter 3

## Inferring Protein–Protein Interactions from Multiple Protein Domain Combinations

Simon P. Kanaan, Chengbang Huang, Stefan Wuchty, Danny Z. Chen, and Jesús A. Izaguirre

### Abstract

The ever accumulating wealth of knowledge about protein interactions and the domain architecture of involved proteins in different organisms offers ways to understand the intricate interplay between interactome and proteome. Ultimately, the combination of these sources of information will allow the prediction of interactions among proteins where only domain composition is known. Based on the currently available protein–protein interaction and domain data of *Saccharomyces cerevisiae* and *Drosophila melanogaster* we introduce a novel method, Maximum Specificity Set Cover (MSSC), to predict potential protein–protein interactions. Utilizing interactions and domain architectures of domains as training sets, this algorithm employs a set cover approach to partition domain pairs, which allows the explanation of the underlying protein interaction to the largest degree of specificity. While MSSC in its basic version only considers domain pairs as the driving force between interactions, we also modified the algorithm to account for combinations of more than two domains that govern a protein–protein interaction. This approach allows us to predict the previously unknown protein–protein interactions in *S. cerevisiae* and *D. melanogaster*, with a degree of sensitivity and specificity that clearly outscores other approaches. As a proof of concept we also observe high levels of co-expression and decreasing GO distances between interacting proteins. Although our results are very encouraging, we observe that the quality of predictions significantly depends on the quality of interactions, which were utilized as the training set of the algorithm. The algorithm is part of a Web portal available at <http://ppi.cse.nd.edu>.

**Key words:** Domain combinations, set cover, protein interaction prediction.

---

### 1. Introduction

Contemporary proteome research attempts to elucidate the structure, interactions, and functions of the proteins that constitute cells and organisms. Large-scale methods determine the molecular

interactions and unravel the complex web of protein–protein interactions in single-cellular organisms such as *Helicobacter pylori* (1) and *Saccharomyces cerevisiae* (2–7). Most recently, attention focused on the first protein–protein interaction maps of complex multicellular organisms such as *Caenorhabditis elegans* (8, 9) and *Drosophila melanogaster* (10).

Such experimental results provide the basis for theoretical considerations that focus on the prediction of potential protein–protein interactions. Pioneering methods drew on the observation that interacting protein domains tend to combine into a fusion protein (11, 12) in higher organisms. Another method utilizes the observation that proteins having matching phylogenetic profiles strongly tend to be functionally linked (13, 12). The domain architecture of interacting proteins offers a framework (14) for assessing the potential presence of a particular interaction by clustering protein domains, depending on sequence and connectivity similarities. Another approach estimates the maximum likelihood of domain interaction (15, 16). Further ideas include overrepresented domain signatures (17), graph-theoretical methods (18), and other probabilistic approaches (19). Support vector machines also were employed to predict potential interactions by modeling network motifs that exhibit higher reliability of the underlying protein–protein interactions (20).

---

## 2. Materials

### 2.1. Protein–Protein Interactions

The first comprehensive, albeit weakly overlapping protein–protein interaction maps of *S. cerevisiae* have been provided with the yeast-two-hybrid method (2, 4). Currently, there exists a variety of yeast-specific protein–protein interaction databases. Most of them, such as MINT (21), MIPS (22), and BIND (23), collect experimentally determined protein–protein interactions. These databases lack an assessment of the data’s quality. In contrast, the GRID database, a compilation of BIND, MIPS, and other data sets, as well as the DIP database (24), provides sets of manually curated protein–protein interactions in *S. cerevisiae*. The majority of DIP entries are obtained from combined, non-overlapping data mostly obtained by systematic two-hybrid analyses. Here, we used a compilation of yeast interactions that have been evaluated by a logistic regression method, allowing the assessment of 47,773 interactions among 4,627 proteins (25). Similarly, experimentally determined interactions in *D. melanogaster* were evaluated, allowing for 20,047 interactions among 6,996 proteins (10).

## 2.2. Protein Domain Data

The advent of fully sequenced genomes of various organisms has facilitated the investigation of proteomes. The *Integr8* database (<http://www.ebi.ac.uk/integr8>) has been set up to provide comprehensive statistical and comparative analyses of complete proteomes of fully sequenced organisms. The initial version of the application contains data for the genomes and proteomes of 182 sequenced organisms (including 19 archae, 150 bacteria, and 13 eukaryotes) and proteome analysis derived through the integration of UniProt (26), InterPro (27), CluSTR (28), GO/GOA (29), EMSD, Genome Reviews, and IPI (30). In particular, we utilized IPI (International Protein Index) files of Yeast, which provide full annotations of each protein with its corresponding domains. In particular, we elucidated the domain architecture of the corresponding proteins by focusing on PFAM domain information as of the corresponding IPI files (31).

## 2.3. Microarray Data and Co-expression Correlation Coefficients

Genes with similar expression profiles are likely to encode interacting proteins (32, 33). We assess MSSC's ability to predict pairs of potentially interacting yeast proteins (Section 3.5), by utilizing the gene expression data of *S. cerevisiae* and *D. melanogaster*. By downloading 1,051 expression profiles of Yeast and 157 of fly from the Stanford Microarray Database (SMD, <http://genome-www5.stanford.edu>), we calculated the Pearson's correlation coefficient  $r_P$  for each pair of interacting proteins. Provided there exist data for both proteins over  $m$  time points, the Pearson correlation coefficient is calculated by

$$r_P = \frac{\frac{1}{m} \sum_{i=1}^m x_i y_i - \bar{x} \bar{y}}{\sigma_i \sigma_j}, \quad [1]$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means, and  $\sigma_i$  and  $\sigma_j$  are the standard deviations of  $i$  and  $j$ .

## 2.4. GO Annotation Data and GO Distance

For any two interacting proteins, we calculate an annotation-based distance between proteins, taking into account all Gene Ontology terms (29) (GO, <http://www.geneontology.org>) that are common to the pair and terms which are specific to each protein. Any two proteins can have several shared GO terms (common terms) and a variable number of terms specific for each protein (specific terms). This distance between interacting proteins  $i$  and  $j$  is based on the Czekanowski-Dice formula (34):

$$d_{i,j} = \frac{|T_{GO}(i) \Delta T_{GO}(j)|}{|T_{GO}(i) \cup T_{GO}(j)| + |T_{GO}(i) \cap T_{GO}(j)|}. \quad [2]$$

In this formula,  $T_{GO}$  are the sets of the proteins with associated GO terms while  $|T_{GO}|$  stands for their number of elements, and  $\Delta$  is the symmetrical difference between two sets. This distance

formula emphasizes the importance of the shared GO terms by giving more weight to similarities than to differences. Consequently, for two genes that do not share any GO terms the distance value is 1, while for two proteins sharing exactly the same set of GO terms the distance value is 0.

---

## 3. Methods

### 3.1. General Outline of the Protein Cover Problem

Investigations of the three-dimensional protein structure suggest that the fundamental unit of protein structure is a domain. Independent of neighboring sequences, this region of a polypeptide chain folds into a distinct structure and mediates the proteins' biological functionality. A domain can also be defined as an amino acid sequence motif with an associated function. Largely, proteins contain only one domain (35) while the majority of sequences from multicellular eukaryotes appear as multidomain proteins of up to 130 domains (36).

We identify proteins in the proteome that give rise to protein interactions through the selection of domain combinations that explain the known protein interaction network. In the simplest case, we use only a selected set of domain pairs in a training set of protein interactions  $R = (P_R, E_R)$ , where  $P_R$  is the set of proteins, and  $E_R$  defines a set of edges between proteins if and only if they interact with each other. The protein interactions  $R$  induce a set of domain pairs  $D_R = \{(d_i, d_j)\}$  where the domains  $d_i$  and  $d_j$  belong to the proteins involved in the interactions  $E_R$ . Schematically, we show these relations in **Fig. 3.1a**. The protein–protein cover problem thus arising is to choose an “optimal” subset of domain pairs  $D_R, D \subseteq D_R$ , such that  $D$  covers all the interactions in  $R$  (**Section 3.5**).

### 3.2. Domain Combinations

We conceptualize domains as the driving force behind the formation of protein interactions. Since the vast majority of proteins in single cellular organisms carry a single domain, domain pairs are sufficient to explain the presence of a protein interaction. However, in more complex organisms the number of multidomain proteins increases. Indicating that protein interactions might be also facilitated by multidomain interactions, we also allow  $D_R$  to include combinations of interacting domains that are potentially involved in a given protein interaction. As such, we handle domain combinations in our framework as new “domains.” Assuming that  $P_1$  has domains  $d_1, d_2$ , and  $d_3$ , we label  $d_1d_2, d_1d_3, d_2d_3$ , and  $d_1d_2d_3$  as “new” domains (**Fig. 3.1b**). Depending on the complexity of the proteins, we might only want to look at combinations up to a certain number of domains.

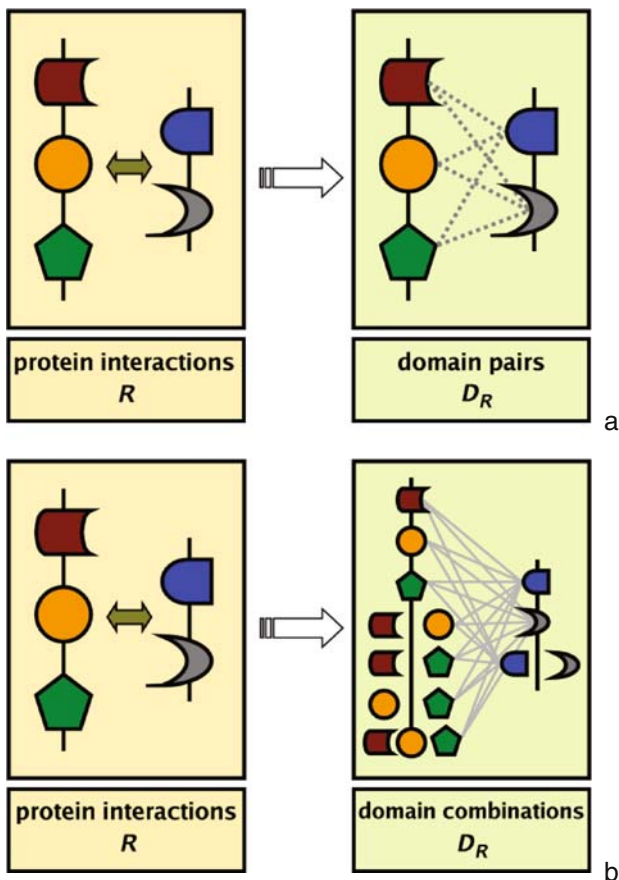


Fig. 3.1. **Combinatorics of protein interactions.** (a) In the simplest case, we consider protein domain pairs as the driving force behind the formation of protein interactions. (b) In a more sophisticated way, we also consider combination of domain pairs that potentially can give rise to observed protein interactions.

### 3.3. Previous Prediction Methods

#### 3.3.1. Association Method

The set of domain pairs  $D$  obtained from a training set of interactions among proteins for which the domain architectures are known can be utilized in different ways to predict protein–protein interactions. The association method (17) assigns an interaction probability  $P(d_m, d_n)$  to each domain pair  $(d_m, d_n)$  in  $D_R$  (such that  $D_R = D$ ) by

$$P(d_i, d_j) = \frac{I_{ij}}{N_{ij}}, \quad [3]$$

where  $I_{ij}$  is the number of interacting protein pairs that contain  $(d_i, d_j)$ , and  $N_{ij}$  is the total number of protein pairs that contain  $(d_i, d_j)$ . The interaction probability for each putative interaction between pairs of proteins is calculated using

$$P(P_i, P_j) = 1 - \prod_{(d_m, d_n) \in (P_i, P_j)} (1 - P(d_m, d_n)). \quad [4]$$

### 3.3.2. Maximum Likelihood Estimation (MLE)

The maximum likelihood estimation method (15) assumes that two proteins interact if at least one pair of domains of the two proteins interacts. Under the above assumption, any protein pair  $(P_i, P_j)$  is the same as the one used in our protein–protein cover problem, Eq. [4]. So, the maximum likelihood is

$$L = \prod P(O_{ij} = 1)^{O_{ij}}(1 - P(O_{ij} = 1))^{1-O_{ij}} \quad [5]$$

where

$$O_{ij} = \begin{cases} 1 & \text{if } (P_i, P_j) \in E_R, \\ 0 & \text{otherwise.} \end{cases} \quad [6]$$

The likelihood  $L$  is a function of  $\theta(P(d_i, d_j), f_p, f_n)$ , where  $P(d_i, d_j)$  represents the probability that domains  $d_i$  and  $d_j$  interact while  $f_p$  and  $f_n$  indicate fixed rates of false positive and false negative interactions in the underlying network. Note that in both the Association Method (AM) and the Maximum-Likelihood-Estimation (MLE), domain pairs were utilized to predict potential protein interactions.

### 3.4. Transformation of Protein Network to Set Cover Problem

Suppose  $X$  is a finite set and  $\mathcal{F}$  is a family of subsets of  $X$  that can cover  $X$ , i.e.,  $X \subseteq \bigcup_{S \in \mathcal{F}} S$ . The set-cover problem is to find a subset  $C$  of  $\mathcal{F}$  to cover  $X$ ,

$$X \subseteq \bigcup_{S \in C} S, \quad [7]$$

where  $C$  is also required to satisfy certain conditions according to different specific problems. For example, the minimum exact set-cover problem requires that  $\sum_{S \in C} |S|$  is minimized, allowing for a  $C$  with minimum cardinality  $|C|$  (37, 38). The minimum set-cover problem is NP-complete. The set-cover problem can be generalized for our purposes by putting  $X$  into a bigger set  $\mathcal{Y}$  (Fig. 3.2a). Suppose  $\mathcal{Y}$  is a finite set,  $X \subseteq \mathcal{Y}$  and  $F$  is a family of subsets of  $\mathcal{Y}$  that can cover  $X$ , i.e.,  $X \subseteq \bigcup_{S \in F} S$ . Thus, the generalized set-cover problem is to find a subset  $C$  of  $F$  to cover  $X$ ,

$$X \subseteq \bigcup_{S \in C} S, \quad [8]$$

where  $C$  is again constrained to certain problem-specific conditions. This theoretical framework allows us to conceive protein interactions as a set-cover problem. As already mentioned, protein–protein interactions can be modeled as a graph  $P_R = (P, E)$ , where  $P$  is the set of proteins and  $E$  is the set of edges between two proteins if and only if they interact with each other. A set-cover problem is set up from the protein–protein interaction network  $P_R$  by taking

$$\mathcal{Y} = \{\text{all protein pairs } (P_i, P_j) | P_i, P_j \in P_R\},$$

$$X = \{\text{protein pairs } (P_i, P_j) | (P_i, P_j) \in E_R\},$$

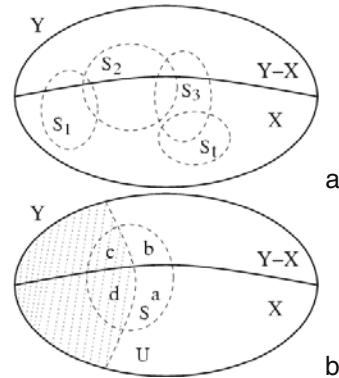


Fig. 3.2. **Schematic representations of set cover problems.** (a) Here, we show the generalized set cover problem:  $X$  is a subset of  $Y$ , and  $F = \{S_i, 1 \leq i \leq t\}$  is a family of subsets of  $Y$ . (b) MSSC chooses a set  $S$  with the minimum  $\frac{b+c}{a}$ . The greedy algorithm for MSSC allows overlapping of subset inside  $X$ . The overlap actually increases the interaction probability for a protein pair.

and  $F$  to be the set of all domain pairs  $(d_m, d_n)$ , where  $(d_m, d_n)$  is contained by at least one element of  $X$ . A domain pair  $(d_m, d_n)$  is considered as a subset of  $\mathcal{Y}$ . Specifically, if a protein pair  $(P_i, P_j)$  (an element in  $X$ ) contains  $(d_m, d_n)$ , then  $(P_i, P_j)$  belongs to the subset  $(d_m, d_n)$ .

Suppose we find a subset  $C$  of  $F$  to cover every element  $(P_i, P_j)$  in  $X$  where an element in  $C$  corresponds to a domain pair  $(d_m, d_n)$ . If  $(d_m, d_n)$  covers  $(P_i, P_j)$ , then the two proteins  $P_i$  and  $P_j$  contain  $d_m$  and  $d_n$ , respectively. Since  $(d_m, d_n)$  can be used to cover the interaction between  $P_i$  and  $P_j$ , we also have a set of domain pairs to cover the protein network  $P_R$ . Suppose there is a set  $D$  of domain pairs to cover the network  $P_R$ . For every element  $(P_i, P_j)$  in  $X$ , there is a domain pair  $(d_m, d_n)$  from  $D$  to cover the interaction between  $P_i$  and  $P_j$ . Since  $(d_m, d_n)$  can be viewed as an element in  $F$ , the collection  $C$  of all the domain pairs from  $D$  is a subset of  $F$ , and  $C$  covers  $X$ .

In this transformation, the set of protein–protein interactions  $P_R$  corresponds to the set  $X$  that needs to be covered, and a domain pair corresponds to an element in  $F$  (a subset of  $\mathcal{Y}$ ).

Once a set cover that fulfills these criteria is found, sets of protein domain interactions allow a description and explanation of the underlying protein interactions to the best extent. Such pairs of proteins can be scanned if an interaction among their domains is actually present in the set cover and therefore is a potential candidate for a putative protein interaction.

### 3.5. MSSC Approach

Many ways exist that allow choices of domain pairs which cover the protein–protein interactions in a training set. AM simply uses all the possible domain pairs to explain the protein–protein

interaction network, i.e., it uses  $F$  to cover  $X$ , so the resulting specificity is very low (15). Sometimes, we are only interested in using a subset of domain pairs to cover the protein–protein interaction network, and hopefully the subset is chosen so that both specificity and sensitivity are maximized. So, the MSSC problem is to find a subset  $C$  of  $F$  to cover  $X$  such that

$$m(C) := \sum_{S \in C} |S - X| \quad [9]$$

is minimized.

MSSC allows the subset  $C$  to cover the overlap with  $X$ , but the overlap with  $\mathcal{Y} - X$  (outside  $X$ ) is minimized. MSSC chooses a cover in this way to maximize the specificity because the false positives appear only in  $\mathcal{Y} - X$ . In developing a greedy algorithm for MSSC (**Fig. 3.3**) at each step, it chooses a subset whose ratio between the part outside  $X$  and the part inside  $U$ ,  $(|S - X|)/(|S \cap U|)$ , is minimized (**Fig. 3.2b**).

The number of iterations of the while loop is bounded by  $\min(|X|, |F|)$  where each single iteration takes  $(O(|X||F|))$  time; so the time complexity of this greedy algorithm is  $\mathcal{O}(|X||F| \min(|X|, |F|))$ . If we apply appropriate data structures, it can be realized in  $O(\log |F| \sum_{S \in F} |S|)$  time. In particular, we maintain a bipartite graph between the elements in  $\mathcal{Y}$  and the elements in  $F$ . If the former is contained by the latter, we add an

```

GREEDY_MSSC( $\mathcal{Y}, X, \mathcal{F}$ )
   $U \leftarrow X$ 
   $\mathcal{E} \leftarrow \mathcal{F}$ 
   $\mathcal{C} \leftarrow \emptyset$ 
  while  $U \neq \emptyset$ 
    do pick  $S \in \mathcal{E}$  with the minimum  $\frac{|S-X|}{|S \cap U|}$ 
      (a tie is broken by  $|S \cap U|$ )
       $U \leftarrow U - \{S\}$ 
       $\mathcal{E} \leftarrow \mathcal{E} - \{S\}$ 
       $\mathcal{C} \leftarrow \mathcal{C} \cup \{S\}$ 
  return  $\mathcal{C}$ 

```

**Fig. 3.3. Pseudocode of the greedy algorithm for solving the Maximum Specificity Set Cover (MSSC) problem.** In this representation, we describe the basic steps of the greedy algorithm, which allows us to find a set cover with maximum specificity. In particular, the routine *GREEDY\_MSSC*( $\mathcal{Y}, X, \mathcal{F}$ ) receives  $X$ , the set of actual interactions,  $\mathcal{Y}$  the set of non existing interactions, and  $\mathcal{F}$ , a family of possible families of protein domain interactions that cover  $X$ . In every step of the algorithm a family  $S$  from  $\mathcal{E}$  is chosen, which minimizes the ratio between the cover between  $S$  and  $X$  and the intersection between  $S$  and the already covered set of interactions  $U$ . This optimization procedure allows us to obtain an optimal collection of families of domain interactions that allow the largest possible cover of interactions.

edge between them, so there are  $\sum_{S \in F} |S|$  edges. Furthermore, we store all elements in  $F$  into a heap ordered by  $\frac{|S-X|}{|S \cap U|}$ . When a subset  $S$  is selected, it is excluded from our problem. We update the bipartite graph and the heap accordingly. The bipartite graph will not be updated more than  $\sum_{S \in F} |S|$  times totally. For a single  $S$ , the updating of the heap takes  $|S| \log |F|$ . Therefore, the total time is  $O(\sum_{S \in F} |S| + \sum_{S \in F} |S| \log |F|)$ , which is  $O(\log |F| \sum_{S \in F} |S|)$ . If  $|F|$  is very large, we use an array of  $|X|^2$  instead of a heap to store  $F$ , resulting in a time which is  $O(|X|^2 + \sum_{S \in F} |S|)$ .

The greedy algorithm only allows an approximation. Its solution has the following relationship with the optimal solution of MSSC

**Theorem** Suppose  $C_a$  is the approximation of MSSC found by the above greedy algorithm, and  $C_o$  is an optimal subset for MSSC. Let  $k = \max_{S \in F} |S|$ . If  $m(C_o) = 0$ , then  $m(C_a) = 0$ ; otherwise, we have

$$\frac{m(C_a)}{m(C_o)} \leq \lceil \ln(k-1) + 1 \rceil. \quad [10]$$

The proof for this theorem can be found in (39). Since  $k$  is the maximum number of elements a subset can have, it corresponds to the maximum number of protein pairs that contain a domain pair in the protein network. Therefore, this theorem indicates that the difference between the approximated and the exact solution remains small, if the maximal number of protein interactions covered by a domain pair is kept small too (i.e., we want to get rid of “promiscuous” domain pairs).

The MSSC procedure, which also accounts for multidomain interactions, allows us to predict putative protein–protein interactions in *S. cerevisiae* and *D. melanogaster*. We observe that our algorithm clearly outscores previous methods such as the Association method (AM) and Maximum Likelihood Estimation (MLE) in terms of sensitivity and specificity. We also observe that our algorithm increases the quality of predictions by using a carefully selected training set of protein interactions. As such we observe that a curated set of yeast and fly protein interactions, which aims to evaluate each interaction with a confidence score, can increase the quality of predictions drastically. Indeed, we observe that our predictions correlate significantly with elevated levels of co-expression and low GO distances, a strong indication for the quality of our predictions.

### 3.6. Comparison of the Performance of MSSC to Other Algorithms

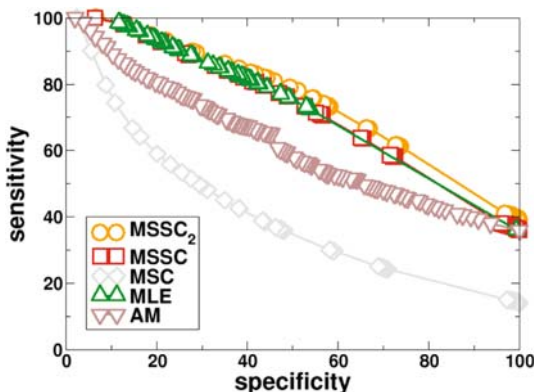
Since it was used in the original paper (15), we use the combined data set of Uetz et al. (4) and Ito et al. (2), which allows a direct performance comparison of AM, MLE, MSC, MSSC, and MSSC<sub>2</sub>. We choose MSSC<sub>2</sub>, the MSSC version that accounts for up to two domain combinations, since we did not find a significant

improvement when going to greater than two domain combinations with the current data. However, this might not be the case for more complex data sets.

We use all interactions in the aforementioned dataset as training and test sets. As for information about the domain architecture of yeast proteins, we utilized PFAM domain data (31) as laid down in the Integr8 database. This induces an overfitting, but results in disjoint training and testing sets are qualitatively similar. We measure the prediction accuracy by specificity, the ratio of the number of matched interactions between the predicted interaction set,  $I$ , and the testing set,  $T$ , over the total number of predicted interactions,  $SP = \frac{|I \cap T|}{|I|}$ . As quality parameters, we define sensitivity as the ratio of the number of matched interactions between the predicted set,  $I$ , and the testing set,  $T$ , over the total number of observed interactions,  $SN = \frac{|I \cap T|}{|T|}$ .

In **Fig. 3.4**, we observe that MSSC – the implementation of our algorithm that accounts for domain pairs only – outperforms AM in terms of both specificity and sensitivity drastically. While MSSC in general allows for results that are very similar to MLE, we observe that MSSC generates significantly more results in areas of high specificity.

Apart from MSSC, we also tried the minimum set cover (MSC), utilizing domain pairs. MSC uses different criteria to choose the subset  $C$  from  $F$  so that  $C$  has minimum cardinality  $|C|$  (38, 39). Compared to MSSC, MSC chooses fewer domain



**Fig. 3.4. Performance comparison of different algorithms.** We compare the performance of the Association Method (AM) (17), Maximum Likelihood Estimation (MLE) (15), Minimum Set Cover (MSC) (38), Maximum Specificity Set cover (MSSC), and its version that uses pairwise domain combinations (MSSC<sub>2</sub>). As training and testing set we utilize a combined set of interactions retrieved from yeast-two-hybrid compilations of Uetz et al. (4) and Ito et al. (2). We observe that MLE and MSSC share the same prediction characteristics while MSSC<sub>2</sub> allows the best predictions. On the other hand, MSSC and MSSC<sub>2</sub> clearly outscore AM and MSC.

pairs to cover the protein–protein interaction network, but actually covers more false positives, as indicated by the comparably low values of sensitivity and specificity. In **Fig. 3.4**, we also observe that an implementation of MSSC that accounts for interactions between up to two domain combinations (MSSC<sub>2</sub>) slightly outperforms MSSC. Note that the solution for MSSC is a subset of MSSC<sub>2</sub> since both use the same algorithm, while MSSC<sub>2</sub> gives the algorithm more choices in order to obtain more accurate predictions.

### 3.7. Results with High-Quality Interactions

Currently available sets of protein–protein interactions contain startling rates of false positives (~50%) and false negatives (~90%) (40). However, there exist a variety of ways to circumvent this problem. One of the most reliable ways to assess the quality of interactions is to integrate different sources interactions to increase their reliability. In particular, training a logistic regression model with parameters such as co-expression, topological and protein interaction related data allows a prediction of an interaction’s reliability quantified by a confidence value (25). Here, we utilize such evaluated interaction data sets of the organisms *S. cerevisiae* (25) and *D. melanogaster* (10), combining 4,627 yeast proteins and 47,783 interactions and 6,996 fly protein, which are involved in 20,047 interactions. All of these interactions are evaluated by a confidence value.

In general, proteomes are composed by a majority of proteins that carry one domain. However, in more complex organisms, proteins carry more than one domain, suggesting that domain combinations are putatively important for protein interactions. In **Fig. 3.5**, we present the sensitivity/specificity curves we obtained by applying MSSC and MSSC<sub>2</sub> to the curated sets of yeast and fly protein interactions. In particular, we utilized sets of increasing reliability (as measured by the threshold of the confidence value  $c$ ) of yeast (25) and fly (10), allowing us to obtain sensitivity/sensitivity curves of predictions by considering these sets as trainings as well as testing set. In general, we observe that both yeast and fly protein interaction sets of increasing reliability outscore the corresponding curves obtained with sets of lower quality. In particular, our results suggest that predicting interactions with MSSC<sub>2</sub> slightly outperforms the results obtained with MSSC, indicating the role of protein domain combinations for the underlying interactome. In the following, we predict protein interactions in yeast and fly by the application of the MSSC<sub>2</sub> algorithm.

To evaluate the quality of our predictions, we analyze the distributions of co-expression correlations. In **Fig. 3.6**, we observe that training sets containing high confidence interactions indeed allow a significant shift of distributions of co-expression correlation coefficients. In both cases, yeast and fly predictions show significant different means of their distributions when

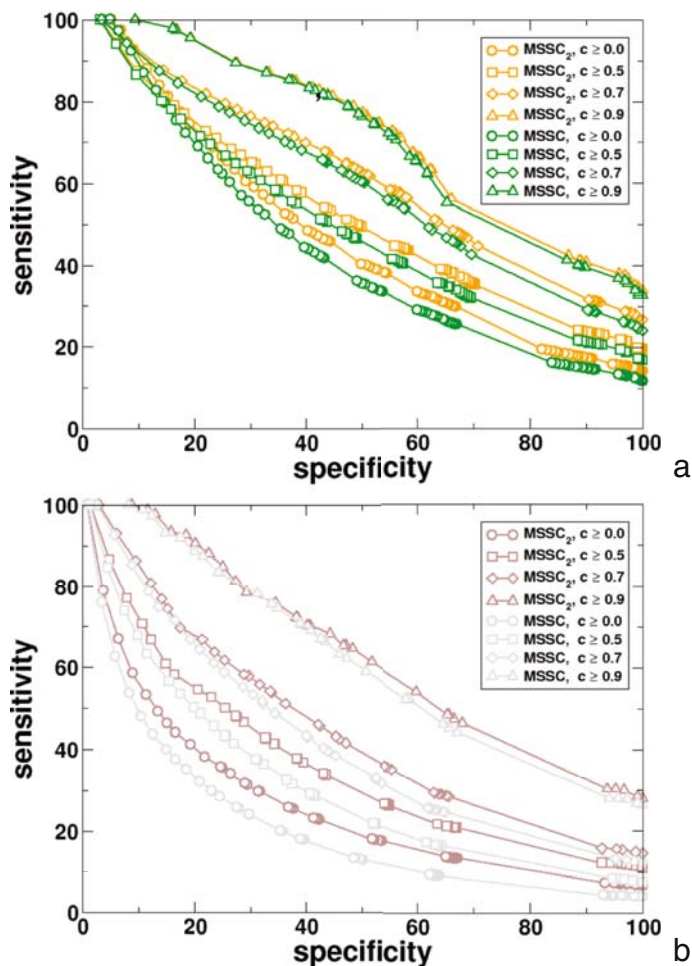


Fig. 3.5. **Sensitivity/specificity curves of predictions in yeast and fly protein interactions sets.** We predicted interactions by feeding the MSSC and MSSC<sub>2</sub> algorithms with protein interactions sets of increasing reliability ( $c$ ). We obtained sensitivity/specificity curves by considering these sets as trainings as well as testing sets. **(a)** In general, we observe that both **(a)** yeast and **(b)** fly protein interaction sets of increasing reliability allow us to obtain sensitivity and specificity values that outscore the corresponding curves obtained with sets of lower quality. In particular, our results suggest that predicting interactions with MSSC<sub>2</sub> slightly outscore the results obtained with MSSC, indicating the dominating role of domain combinations for the underlying interactome.

compared to a background distribution (which is defined as the full set of uncurated yeast (25) and fly interactions (10)). Significant Student's  $t$ -test scores support our conclusion that high-quality interactions allow a higher degree of quality predictions.

As a different measure for the existence of a predicted interaction, we utilize the empirical observation that interacting proteins show a significantly elevated tendency to share similar functions.

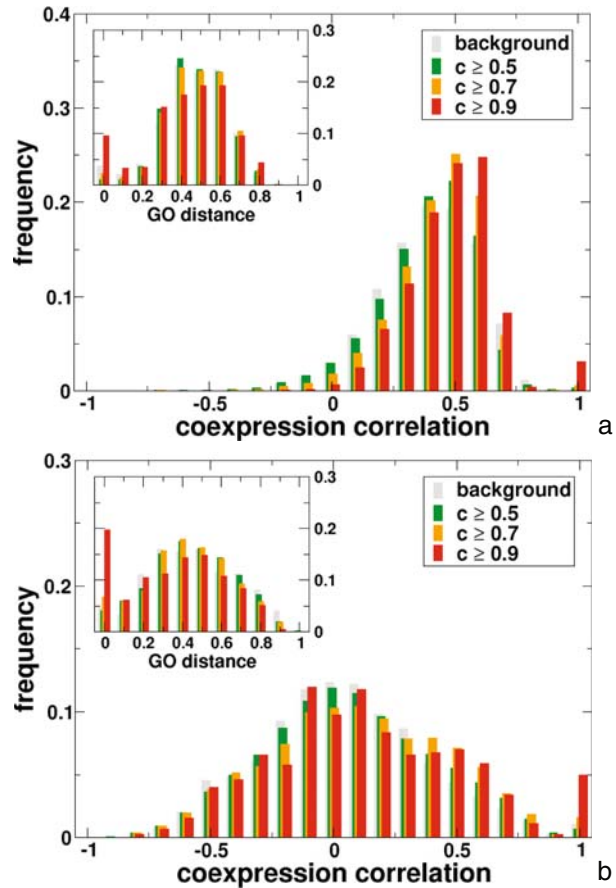


Fig. 3.6. **Co-expression and GO distance analysis of the predictions in yeast and fly.** Compared to a comprehensive set of co-expression correlations values of protein links (background), we observe that in (a) yeast and (b) fly the quality of the underlying protein interaction network increases the quality of our predictions. The shift toward higher coexpression correlation values is further supported by significant Student's *t*-test scores, when testing the curves of the predictions to a background distribution being the full set of uncurated yeast and fly interactions. [ $c \geq 0.5$ : 2.61 ( $P = 9.1 \times 10^{-3}$ ),  $c \geq 0.7$ : 41.66 ( $P < 10^{-10}$ ),  $c \geq 0.9$ : 36.26 ( $P < 10^{-10}$ )] and fly [ $c \geq 0.5$ : 4.20 ( $P = 2.6 \times 10^{-5}$ ),  $c \geq 0.7$ : 10.53 ( $P < 10^{-10}$ ),  $c \geq 0.9$ : 8.81 ( $P < 10^{-10}$ )]. As a different indicator of the existence of a potential interaction we show the GO distance for yeast ((a), inset) and fly ((b), inset). Similarly to the distributions of co-expression coefficients we find that an increasing quality of the training sets allows qualitatively better predictions as exemplified by the shifts toward lower values. These results are further supported by significant Student's *t*-test scores when compared to the background distributions of yeast [ $c \geq 0.5$ : 3.85 ( $P = 1.2 \times 10^{-4}$ ),  $c \geq 0.7$ : 3.10 ( $P = 2.0 \times 10^{-3}$ ),  $c \geq 0.9$ : 5.94 ( $P = 2.9 \times 10^{-9}$ )] and fly [ $c \geq 0.5$ : 1.02 ( $P = 0.31$ ),  $c \geq 0.7$ : 2.88 ( $P = 4.0 \times 10^{-3}$ ),  $c \geq 0.9$ : 6.27 ( $P = 5.9 \times 10^{-10}$ )].

In turn, a measure that represents functional similarity might be used as an indicator of an interaction's existence. Utilizing GO annotations (29), we calculate a GO distance (see Materials) for every predicted interaction. If two proteins do not share any GO

terms, the distance value is 1, while the opposite holds for proteins sharing exactly the same set of GO terms. Indeed, in the insets of **Fig. 3.6**, we observe that both yeast and fly predictions, that were obtained with high-quality training sets show a stronger functional similarity. Again, in both cases distributions have significantly different means as exemplified by significant Student's *t*-test scores, indicating that high-quality interactions indeed influence the quality of the predictions.

---

## 4. Conclusions

In this paper, we present a novel algorithm that allows the selection of a set of domain pairs, which covers the experimental observations and maximizes the specificity in the training set of protein interactions. Compared to previous methods, MSSC is able to improve the specificity for a given sensitivity. As a refinement of this algorithm we also introduced the opportunity to model and predict interactions as the consequence of interactions among many combinations of domains. In particular, we observe that the relatively small amount of multidomain proteins in yeast compared to fly already have a significant impact on the interactions of the underlying interactome. As such, we observe that we obtain better results by applying MSSC<sub>2</sub>, our algorithmic extension that accounts for domain combinations.

---

## 5. Notes



1. *Dependence from training data.* Our results also suggest that the quality of predicted protein–protein interactions strongly depends on the utilized training sets. Although we showed that our algorithm by design reduces the amount of false positives, it allows only high-quality interactions if the training set reflects an elevated degree of quality.

The dependence on high-quality interaction sets also poses a sometimes intricate problem. The increase in quality always is accompanied by a decrease of protein interactions and therefore limits the number of interacting proteins involved. Thus, the number of protein domains that allow these interactions is diminished as well. Since protein interactions and domains are the major data sources of our algorithm, the choice of an

appropriate training set that balances quality and a reasonable number of domains, that still allows predictions on a large scale is a crucial step.

2. *Outlook.* The proposed algorithm can be used to predict protein interactions for every organism for which data of protein interactions and the corresponding protein domain architectures are available. Encouraged by the high quality of our results, a next step is the prediction of potential interactions between proteins in organisms where high-quality interaction data are available to train MSSC<sub>2</sub>. Furthermore, proteins that participate in many interactions are preferentially conserved and change their sequence only to a small extent (41, 42). The observation that high clustering and co-expressed protein–protein interaction sets show preferential evolutionary conservation (43, 44), and increased reliability (18) allows us to not only obtain a high-quality set of interactions potentially serving as the basis for predictions in a reference organism. In fact, such sets of interactions may also indicate evolutionary cores that have been conserved more generally among different organisms. As such they not only allow the evaluation of protein–protein interactions but also could serve as a training set for the predictions of protein interactions in target organisms.

---

## Acknowledgments

Danny Chen was supported in part by the NSF under Grant CCF-0515203. Jesús Izaguirre was supported by partial funding from NSF grants IOB-0313730, CCR-0135195, and DBI- 0450067. Stefan Wuchty was supported by the Northwestern Institute of Complexity (NICO).

## References

1. Rain JC, Selig L, DeReuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schächter V, Chemama Y, Labigne A, Legrain P. The protein–protein interaction map of *Helicobacter pylori*. *Nature* 2001, 409, 211–215.
2. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y. Towards 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 Nat Acad Sci USA* 2000, 97, 1143–1147.
3. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Nat Acad Sci USA* 2001, 98, 4569–4574.
4. Uetz P, Giot L, Cagney G, Mansfield T, Judson R, Knight J, Lockshorn D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M,

- Fields S, Rothberg J. A comprehensive analysis of protein–protein interactions of *Saccharomyces cerevisiae*. *Nature* 2000, 403, 623–627.
5. Gavin A, Bösch M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick J, Michon AM, Cruciat CM, Remor M, Böfert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley R, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, 415, 141–147.
  6. Ho Y, Gruhler A, Heilbut A, Bader G, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutillier K, coauthors. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 2002, 415, 180–183.
  7. Jeong H, Mason S, Barabási AL, Oltvai Z. Lethality and centrality in protein networks. *Nature* 2001, 411, 41–42.
  8. Walhout A, Sordella R, Lu X, Hartley J, Temple G, Brasch M, Thierry-Mieg N, Vidal M. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 2000, 287, 116–122.
  9. Li S, Armstrong C, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Ha T, et al. A map of the interactome network of the metazoan *C. elegans*. *Science* 2004, 303, 540–543.
  10. Giot L, Bader J, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao Y, Ooi C, Godwin B, Vitols E, Vijayadamar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agce M, Voss E, Furtak K, Renzulli R, Aanensen N, Carroll S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon C, Finley R Jr, White K, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shinkets R, McKenna M, Chant J, Rothberg J. A protein interaction map of *Drosophila melanogaster*. *Science* 2004, 302, 1727–1736.
  11. Enright A, Iliopoulos I, Kyrpides N, Ouzounis C. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 1999, 402, 86–90.
  12. Marcotte E, Pellegrini M, Thompson M, Yeates T, Eisenberg D. A combined algorithm for genomewide prediction of protein function. *Nature* 1999, 402, 83–86.
  13. Pellegrini M, Marcotte E, Thompson M, Eisenberg D, Yeates T. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc Natl Acad Sci USA* 1999, 96, 4285–4288.
  14. Wojcik J, Schächter V. protein–protein interaction map inference using interacting domain profile pairs. *Bioinformatics* 2001, 17, 296S–305S.
  15. Deng M, Mehta S, Sun F, Cheng T. Inferring domain-domain interactions from protein–protein interactions. *Genome Res* 2002, 12, 1540–1548.
  16. Iossifov I, Krauthammer M, Friedman C, Hatzivassiloglou V, Bader J, White K, Rzhetsky A. Probabilistic inference of molecular networks from noisy data sources. *Bioinformatics* 2004, 20, 1205–1213.
  17. Sprinzak E, Margalit H. Correlated sequence-signature as markers of protein–protein interaction. *J Mol Biol* 2001, 311, 681–692.
  18. Goldberg D, Roth F. Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci USA* 2003, 100, 4372–4376.
  19. Tong A, Drees B, Nardelli G, Bader G, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Apoluzzi S, et al. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 2002, 295, 321–324.
  20. Albert I, Albert R. Conserved network motifs allow protein–protein interaction prediction. *Bioinformatics* 2004, 20, 3346–3352.
  21. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesaroni G. Mint – a molecular interaction database. *FEBS Lett.* 513, 2002, 135–140.
  22. Mewes HW, D Frishman UB, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkötter M, Rudd S, Weil B. MIPS: A database for genomes and protein sequences. *Nucl Acids Res* 2002, 30, 31–34.
  23. Bader G, Donaldson I, Wolting C, Ouellette B, Pawson T, Hogue C. BIND – The biomolecular interaction network database. *Nucl Acids Res* 2001, 29, 242–245.
  24. Xenarios I, Salwinski L, Duan X, Higney P, Kim SM, Eisenberg D. Dip, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. *Nucl Acids Res* 2002, 30, 303–305.

25. Bader J, Chaudhuri D, Rothberg J, Chant J. Gaining confidence in high-throughput protein interaction networks. *Nature Biotech* 2004, 22, 78–85.
26. Apweiler R, Bairoch A, Wu C, Barker W, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin M, Natale D, O'Donovan C, Redaschi N, Yeh L. UniProt: The universal protein knowledgebase. *Nucl Acids Res* 2004, 32, D115–D119.
27. Mulder N, Apweiler R, Attwood T, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley R, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard S, Pagni M, Peyruc D, Ponting C, Selengut J, Servant F, Sigrist C, Vaughan R, Zdobnov E. The interpro database, 2003 brings increased coverage and new features. *Nucl Acids Res* 2003, 31, 315–318.
28. Kriventseva E, Fleischmann W, Zdobnov E, Apweiler R. CluSTR: A database of clusters of SWISS-PROT+TrEMBL proteins. *Nucl Acids Res* 2001, 29, 33–36.
29. Consortium G. The gene ontology (go) database and information resource. *Nucl Acids Res* 2004, 32, D258–D261.
30. Kersey P, Duarte J, Williams A, Apweiler R, Karavidopoulou Y, Birney E. The international protein index: An integrated database for proteomics experiments. *Proteomics* 2004, 4, 1985–1988.
31. Bateman A, Coin L, Durbin R, Finn R, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer E, Studholme D, Yeats C, Eddy S. The PFAM protein families database. *Nucl Acids Res* 2004, 32, D138–D141.
32. Grigoriev A. A relationship between gene expression and protein interactions on the proteome scale: Analysis of the bacteriophage t7 and the yeast *Saccharomyces cerevisiae*. *Nucl Acids Res* 2001, 29: 3513–3519.
33. Ge H, Ziu L, Church G, Vidal M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat Genet* 2001, 29, 482–486.
34. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B. GOToolBox: Functional analysis of gene datasets based on gene ontology. *Genome Biol.* 2004, 5, R101.
35. Doolittle R. The multiplicity of domains in proteins. *Ann Rev Biochem* 1995, 64, 287–314.
36. Li WH, Gu Z, Wang H. Evolutionary analyses of the human genome. *Nature* 2001, 409, 847–849.
37. Johnson DS. Approximation algorithms for combinatorial problems. *J Comput System Sci* 1974, 9, 256–278.
38. Cormen TH, Leiserson CE, Rivest RL, Stein C. *Introduction to Algorithms*, Second Edition. McGraw Hill Boston, MA, 2001.
39. Huang C, Morcos F, Kanaan S, Wuchty S, Chen D, Izaguirre J. Predicting protein-protein interactions from protein domains using a set cover approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2007, 4, 78–87.
40. von Mering C, Krause R, Snel B, Cornell M, Oliver S, Fields S, Bork P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2003, 31, 399–403.
41. Wuchty S. Topology and evolution in the yeast protein interaction network. *Genome Res* 2004, 14, 1310–1314.
42. Fraser H, Hirsh A, Steinmetz L, Scharf C, Feldman M. Evolutionary rate in the protein interaction network. *Science* 2002, 296, 750–752.
43. Wuchty S, Oltvai Z, Barabási AL. Evolutionary conservation of motif constituents within the yeast protein interaction network. *Nat Genet* 2003, 35, 176–179.
44. Wuchty S, Barabási AL, Ferdig M. Stable evolutionary signal in a yeast protein interaction network. *BMC Evol Biol.* 2006, 6, pp. 8.