Inferring Protein-Protein Interactions from Protein Domain Combinations

Simon P. Kanaan¹, Chengbang Huang¹, Stefan Wuchty², Danny Z. Chen¹, Jesús A. Izaguirre¹,*

Abstract—A goal of contemporary proteome research is the elucidation of the protein-protein interactions in the cell. Based on currently available protein-protein interaction and domain data of S. cerevisiae, we introduce a novel method, Maximum Specificity Set Cover (MSSC), to predict protein-protein interactions. This algorithm features two stages: First, we select high quality protein-protein interactions based on a clustering measure. Second, we use MSSC to assign probabilities to domain pairs. MSSC is also modified to include the possibility of having more than one domain from each protein causing the protein-protein interaction. This approach allows us to predict previously unknown protein-protein interactions with a degree of sensitivity and specificity that clearly out-scores other approaches. We find that the predicted interaction network preserves the characteristics of the initial web of protein-protein interactions. We also observe high levels of coexpression among putative interactions. We extend our method to infer protein-protein interactions in multicellular organisms where interaction data currently does not exist. Starting from predictions in yeast, we find a set of orthologous interactions in A. thaliana, C. elegans, D. melanogaster, M. musculus, and H. sapiens.

Index Terms—protein interaction prediction, domain combination, orthologous protein interactions.

I. 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 H. pylori [49] and S. cerevisiae [18], [27], [31]–[33], [58]. Most recently, attention focused on the first protein-protein interaction maps of complex multicellular organisms such as C. elegans [39], [62] and D. melanogaster [19].

Such experimental results set 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 [15], [42] in higher organisms. Another method utilizes the observation that proteins having matching phylogenetic profiles strongly tend to be functionally linked [42], [48]. The domain architecture of interacting proteins offers a framework [64] 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 that domains interact [12], [30]. Further ideas include overrepresented domain signatures [56], graph-theoretical methods [21] and other probabilistic approaches [57]. Recently, the observation that highly clustered regions exhibit higher reliability of the underlying protein-protein interactions was used to conceive support vector models to predict potential interactions [1].

Assuming that protein domains facilitate the interactions among proteins, we introduce a novel method for the inference of protein-protein interactions. Utilizing a maximum-specificity set cover procedure (MSSC), which also accounts for multi domain interactions, we calculate the probabilities of putative protein-protein interactions. Based on protein interactions in S. cerevisiae, we observe that our algorithm clearly out-scores previous methods in terms of sensitivity and specificity. We also observe that our predictions correlate significantly with elevated levels of coexpression of current micro-array data. This paper extends results of [29] to the case of using domain combinations and orthologous data.

As a refinement of our prediction method, we utilize protein-protein interaction data of Yeast that is pre-assessed with a clustering procedure. Interactions which are embedded in a highly clustered neighborhood not only tend to have an elevated degree of quality [21], but also exhibit a strong tendency to be coexpressed. Since this step helps to eliminated false signals drastically, we observe that such a set of preprocessed interactions significantly improves MSSC’s ability to produce high quality predictions. High local clustering of interacting proteins coincides with evolutionary conservation [66] and coexpression. It has been shown that highly con-

¹ Dept. of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556
² Dept. of Physics, University of Notre Dame, Notre Dame, IN 46556
* Corresponding author
nected proteins display a lower evolutionary distance to their orthologous counterparts than sparsely connected ones [17]. These proteins might act as an evolutionary core of interactions in different organisms. Such a basic set of interactions combined with the domain architectures of different organisms potentially could serve as the basis for our prediction method to infer protein-protein interactions in organisms where interaction data currently does not exist.

II. MATERIALS AND METHODS

Investigations of the spatial 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. The majority of proteins contains only one domain [13] while sequences of multicellular eukaryotes appear as multi-domain proteins of up to 130 domains [40].

Figure 1 illustrates these assumptions. Our objective is to select domain pairs (pairs of geometrical shapes in the figure) that explain the known protein interaction network. This network is the training set of the algorithm. Using the selected domain-pairs, we predict protein-protein interactions in a testing set of proteins. In order to assess the quality of our predicted interactome, typically the interactions among the proteins in the testing set are known. Thus, we can count how many real interactions we predict, and how many false positives. For these assumptions to hold, it is important to start with a curated network where the false positives have been reduced.

A. 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 [32], [58]. Currently, there exists a variety of yeast specific protein-protein interaction databases. Most of them, such as MINT [70], MIPS [45] and BIND [4], 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 [69] provide sets of manually curated protein-protein interactions in *S. cerevisiae*. The majority of DIP entries is obtained from combined, non-overlapping data mostly obtained by systematic two-hybrid analyzes. Here, we use the DIP database (http://dip.doe-mbi.ucla.edu) which is the qualitatively best compilation of yeast protein-protein interaction data. The current version contains 3,677 proteins involved in 11,249 interactions for which there is domain information. DIP also provides a high quality core set of 2,609 yeast proteins that are involved in 6,355 interactions which have been found with more than one different experimental method.

B. 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 analyzes 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 [3], InterPro [46], CluSTr [36], GO/GOA [9], EMSD, Genome Reviews and IPI [35]. In particular, we utilized IPI (International Protein Index) files of Yeast to elucidate the domain architecture of the corresponding proteins. For our analysis, we focused on domain data retrieved from the PFAM database, a reliable collection of multiple sequence alignments of protein families and profile hidden Markov models [7].

C. Microarray Data

Genes with similar expression profiles are likely to encode interacting proteins [24]. We assess MSSC’s ability to predict pairs of potentially interacting yeast proteins, by utilizing gene expression data of Eisen et al. [14]. This compilation of coexpression patterns consists of 2,467 yeast genes whose coexpression patterns have
been investigated for 79 data points. Considering the strength of our predictions, we expect that potentially interacting proteins show an elevated degree of coexpression.

D. Assessment of Protein-Protein Interactions

Although the current results concerning the structure of protein-protein interaction networks are impressive, the error-proneness of experimental methods for the determination of protein-protein interactions jeopardizes the strength of the obtained results. A recent estimation of the rate of inaccurately determined yeast protein-protein interaction data uncovered a startling false negative rate of 90% while false positives show a 50% error rate [60]. False positives are the predicted protein-protein interactions not included in the input interaction network, while false negatives are the protein-protein interactions in the protein-protein interaction network which are not predicted. Despite these data inconsistencies, a network topology based approach [21] uncovered a remarkable correlation between enhanced quality and network clustering around a certain protein-protein interaction. Considering an interaction network of \( N \) nodes, the hypergeometric clustering coefficient, defined as

\[
C_{vw} = -\log \sum_{i=|N(v)\cap N(w)|}^{\min(|N(v)||N(w))} \left( \frac{|N(v)|}{|N(v)|-i} \right) \left( \frac{|N(w)|-i}{|N(w)|} \right),
\]

where \( N(x) \) represents the neighborhood of a vertex \( x \), reflects the probability that an interaction between proteins \( v \) and \( w \) indeed exists. Given the number of immediate neighbors around the considered proteins, \( N(v) \) and \( N(w) \), the hypergeometric clustering coefficient increases with elevated overlap between the protein’s neighborhoods. Provided that the neighborhoods are independent, the summation can be interpreted as a \( p \) value reflecting the probability of obtaining a number of mutual neighbors between proteins \( v \) and \( w \) at or above the observed number by chance [21]. We excluded the interaction between \( v \) and \( w \) from the calculation, rendering \( C_{vw} \) independent from direct experimental evidence of the considered edge.

In our study, we calculated the link specific clustering coefficients \( C_{vw} \) for each edge. By applying different cut-off values, we elucidated the corresponding interaction network, serving as the basis for further protein-protein interaction predictions. We expect that the interaction webs that exhibit an elevated degree of clustering raise the quality of our predictions.

E. Quality Measures

Predictions accuracy is measured by specificity and sensitivity. The specificity is defined as 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, \( S_P = \frac{|I \cap T|}{|P|} \). The sensitivity is defined 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, \( S_N = \frac{|I \cap T|}{|T|} \).

F. Orthologous Data

The InParanoid database [51] provides orthologous sequence information for \( S. cerevisiae \) and the complete protein sets of \( H. sapiens \), \( D. melanogaster \), \( M. musculus \), \( C. elegans \) and \( A. thaliana \). Utilizing all-versus-all BLASTP searches in protein sets of two species, sequence pairs with mutually best scores were selected as central ortholog pairs. Proteins of both species showing an elevated degree of homology were clustered around these central pairs, a procedure that forms orthologous groups. The quality of the clustering was then assessed by a standard bootstrap procedure. The central ortholog sequence pair that provides a confidence level of 100% was considered as the real orthologous relationship while proteins with a lower level of confidence were considered as their in-paralogs. In our study, we selected only the central ortholog sequence pairs of each group, resulting in 1,847 yeast proteins with orthologs in \( H. sapiens \), 1,975 in \( A. thaliana \), 1,795 in \( C. elegans \), 2,350 in \( M. musculus \), and 1,565 in \( D. melanogaster \). We also compiled a list of 976 proteins that have an ortholog in all organisms under consideration.

III. PROTEIN-PROTEIN COVER PROBLEM

We now describe our modeling framework in detail. The next section describes how we choose domain pairs from \( R \). We are given a training set of protein interactions \( R = (P_R, E_R) \), where \( P_R \) is the set of proteins and \( E_R \) is the set of edges. There is an edge between two proteins if and only if they interact with each other. \( R \) induces a set of domain pairs \( D_R = \{(d_i, d_j)\} \) where \( d_i \) and \( d_j \) belong to the vertices in an edge in \( E_R \). Figure 2 illustrates an interaction between two proteins.

The protein-protein cover problem is to choose a subset of \( D_R, \text{ } D \subseteq D_R \), such that \( D \) covers all the interactions in \( R \). We say that a domain pair covers a protein-protein interaction if the two interacting proteins contain the two domains. We augment each element \( D \)
with the probability of domain-domain interaction,

\[ P(d_i, d_j) = \frac{I_{ij}}{N_{ij}}, \quad (2) \]

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 \((d_i, d_j)\). Given \( D_m \) the associated \( P(d_i, d_j) \), and a protein testing set \( T = (P_T, E_T) \), we predict interactions among proteins in \( P_T \). The output is the prediction interactome \( I = (P_T, E_I) \), where \( E_I \) is the set of predicted interactions. We also associate an interaction probability to each protein pair:

\[ P(P_i, P_j) = 1 - \prod_{(d_m, d_n) \in (P_i, P_j)} (1 - P(d_m, d_n)). \quad (3) \]

A. Including Domain Combinations

We also allow \( D_R \) to include domain combinations, that is the fusion of two or more domains. In effect, the power set of all domains in a protein can appear in one of the elements of \( D_R \), except for the empty set. Domain combinations are handled in our framework as new “domains”. Assume \( P_1 \) has domains \( d_1, d_2 \) and \( d_3 \). We create domains \( d_1d_2, d_1d_3, d_2d_3, \) and \( d_1d_2d_3 \). See Figure 3 for an illustration. Depending on the complexity of the proteins we might only want to look at combinations up to a certain number of domains. MSSC.m stands for MSSC with up to \( m \) domain combinations. For the data sets we looked at we did not find a significant improvement when going to greater then two domain combinations. This might not be the case for more complex data sets.

B. Choosing the Best Domain-Pair Set

Consider the following example. Proteins \( P_1, P_2, P_3, \) and \( P_4 \) contain the domains \( \{d_1\}, \{d_1, d_2\}, \{d_1, d_3\}, \{d_1, d_4\} \) respectively. \( (P_1, P_2), (P_1, P_3), \) and \( (P_1, P_4) \) are the given protein-protein interactions. These protein-protein interactions can be explained by \( \{(d_1, d_2), (d_1, d_3), (d_1, d_4)\} \) or by \( \{(d_1, d_1)\} \). Choosing the larger set gives a lower specificity than choosing the smaller set.

C. Prediction

We use the set \( D \) to predict the protein-protein interactions in \( T \). Each pair in \( D \) is assigned the interaction probability of Eq. (2). The interaction probability for each putative protein pair is calculated using Eq. (3), according to Algorithm 1.

**Algorithm 1 Predicting Novel Protein-Protein Pairs**

```
Prediction(D, T)
For each domain pair \((d_i, d_j) \in D\)
  For each protein \( P_i \in T \) containing \( d_i \)
    For each protein \( P_j \in T \) containing \( d_j \)
      compute \( P(P_i, P_j) \), Eq. (3).
```

IV. PREVIOUS PREDICTION METHODS

In order to have an estimate of the quality of our predictions, we compare performance of our algorithm to the following known methods. These algorithms utilize protein-protein interactions and their corresponding domain profiles to predict otherwise unknown protein-protein interactions in *S. cerevisiae*. 
A. Association Method (AM)

The association method [56] assigns an interaction probability \( P(d_m, d_n) \) Eq. (2), to each domain pair \((d_m, d_n)\) in \(D_R\). AM uses a protein-protein cover approach where \(D_R = D\).

B. Maximum Likelihood Estimation (MLE)

The maximum likelihood estimation method [12] assumes that two proteins interact if at least one pair of domains of the two proteins interacts.

Under the above assumption, for any protein pair \((P_i, P_j)\) is the same as the one used in our protein-protein cover problem, Eq. (3). So, the maximum likelihood is

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

where

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

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 negative interactions in the underlying network. The maximization of \(L\) by an expectation maximization algorithm [12] achieved 42.5% specificity and 77.6% sensitivity on a combined yeast protein-protein interaction set compiled from [32], [58].

V. MAXIMUM SPECIFICITY SET COVER (MSSC)

We present a novel method to predict protein-protein interactions. Our method uses a set-cover approach by choosing some domain pairs to cover the given protein-protein interactions.

We define the protein-protein interaction problem as the problem of finding a set of domain pairs to cover the given protein-protein interactions. Ideally, the set of domain pairs should give as few false positives as possible.

A. 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., \(X = \bigcup_{S \in \mathcal{F}} S\). The set-cover problem is to find a subset \(C\) of \(\mathcal{F}\) to cover \(X\),

\[
X = \bigcup_{S \in C} S
\]

and \(C\) 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, and the minimum set-cover problem is to find a \(C\) with the minimum cardinality \(|C|\) [10], [34]. The minimum set-cover problem is NP-complete. We believe that the protein-protein cover problem NP-complete but have not proved it yet.

The set-cover problem can be generalized for our purposes by putting \(X\) into a bigger set \(Y\) (Figure 4). Suppose \(Y\) is a finite set, \(X \subseteq Y\) and \(\mathcal{F}\) is a family of subsets of \(Y\) that can cover \(X\), i.e., \(X \subseteq \bigcup_{S \in \mathcal{F}} S\). The generalized set-cover problem is to find a subset \(C\) of \(\mathcal{F}\) to cover \(X\),

\[
X = \bigcup_{S \in C} S
\]

and \(C\) is also required to satisfy certain conditions according to different specific problems, as before.

Fig. 4. The generalized set cover problem: \(X\) is a subset of \(Y\), and \(\mathcal{F} = \{S_i, 1 \leq i \leq t\}\) is a family of subsets of \(Y\).

We can transform our protein-protein cover problem into a set-cover problem. The protein-protein interaction problem can be modeled by a graph \(P_R = (P, E)\), where \(P\) is the set of proteins and \(E\) is the set of edges. There is an edge between two proteins if and only if they interact with each other. A set-cover problem is constructed from the protein-protein interaction network \(P_R\) by taking

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

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

and \(\mathcal{F}\) 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 viewed as a subset of \(Y\). 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 \(\mathcal{F}\) to cover every element \((P_i, P_j)\) in \(X\). 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; so \((d_m, d_n)\) can be used to cover the interaction between \(P_i\) and \(P_j\). Therefore, 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 \(\mathcal{F}\), the collection \(\mathcal{C}\) of all the domain pairs from \(D\) is a subset of \(\mathcal{F}\), and \(\mathcal{C}\) 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 \(\mathcal{F}\) (a subset of \(Y\)).

### B. MSSC Approach

There are many ways to choose domain pairs to cover the protein-protein interaction network. AM simply uses all the possible domain pairs to explain the protein-protein interaction network, i.e., it uses \(\mathcal{F}\) to cover \(X\), so the resulting specificity is very low [12]. Sometimes we are only interested in using a “subset” of domain pairs to cover the protein-protein interaction network, i.e., it uses \(Y\) in \(\mathcal{F}\) to cover \(X\), but the overlap with \(Y - X\) (outside \(X\)) is minimized. MSSC chooses a cover in this way to maximize the specificity because the false positives appear only in \(Y - X\).

We develop a greedy algorithm for MSSC, as in Algorithm 2.

#### Algorithm 2 Greedy algorithm for MSSC.

```
GREEDY_MSSC(Y, X, \mathcal{F})
  U ← X
  \mathcal{E} ← \emptyset
  \mathcal{C} ← \emptyset
  \textbf{while} U ≠ \emptyset
    \textbf{do} pick \(S \in \mathcal{E}\) with the minimum \(|S - X|/|S ∩ U|\)
    \hspace{1cm} (a tie is broken by \(|S ∩ U|\))
    \hspace{1cm} U ← U − S
    \hspace{1cm} \mathcal{E} ← \mathcal{E} − \{S\}
    \hspace{1cm} \mathcal{C} ← \mathcal{C} ∪ \{S\}
  \textbf{return} \mathcal{C}
```

In this algorithm, at each step MSSC chooses a subset whose ratio between the part outside \(X\) and the part inside \(U\), \(|S - X|/|S ∩ U|\), is minimized (Figure 5).

The number of iterations of the \textbf{while} loop is bounded by \(\min(|X|, |\mathcal{F}|)\), and each single iteration takes \(|X||\mathcal{F}|\) time; so the time complexity of this greedy algorithm is \(O(|X||\mathcal{F}|\min(|X|, |\mathcal{F}|))\). If we apply proper data structures, it can be realized in \(O(|\mathcal{F}| \sum_{S \in \mathcal{F}} |S|)\) time. More specifically, first, maintain a bipartite graph between elements in \(Y\) and elements in \(\mathcal{F}\). If the former is contained by the latter, we add an edge between them, so there are \(|\sum_{S \in \mathcal{F}} |S|\) edges. Secondly, store all elements in \(\mathcal{F}\) into a heap ordered by \(|S - X|/|S ∩ U|\). 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 \(|\sum_{S \in \mathcal{F}} |S|\) totally. For a single \(S\), the updating of the heap takes \(|S| \log |\mathcal{F}|\). Therefore, the total time is \(O(\sum_{S \in \mathcal{F}} |S| + \sum_{S \in \mathcal{F}} |S| \log |\mathcal{F}|)\), which is \(O(|\mathcal{F}| \sum_{S \in \mathcal{F}} |S|)\). If \(|\mathcal{F}|\) is very big, we use an array of \(|X|^2\) instead of a heap to store \(\mathcal{F}\), and the resulting time will be \(O(|X|^2 + \sum_{S \in \mathcal{F}} |S|)\).

The above greedy algorithm is just an approximation, and the solution found by it has the following relationship with the optimal solution of MSSC.

\textbf{Theorem 5.1:} Suppose \(C_a\) is the approximation of MSSC found by the above greedy algorithm, and \(C_o\) is an optimal subcover for MSSC. Let \(k = \max_{S \in \mathcal{F}} |S|\). If \(m(C_o) = 0\), then \(m(C_a) = 0\); otherwise, we have

\[
\frac{m(C_a)}{m(C_o)} \leq \lfloor \ln(k - 1) + 1 \rfloor. \tag{9}
\]

The proof for this theorem can be found in [29].

The theorem shows relationship between the approximation by \textbf{GREEDY_MSSC} and an optimal solution. If \(k\) is small, the difference between them is small too. In this theorem, \(k\) is the maximum number of elements a subset can have, and it corresponds to the maximum number of protein pairs that contains a domain pairs in the protein network.

### VI. Results

We use two sources of protein-protein interactions: one is the combined data set of Uetz \textit{et al.} [58] and
Ito et al. [32], which we call CombUI; the other is a complete protein-protein interactions set retrieved from the DIP database [69], which we simply name DIP. The combined Uetz and Ito is also used in [12]. We also study the yeast interactions that are evolutionary conserved in human, mouse, worm, fly, arabidopsis, and a crossection of proteins that have orthologs in the latter organisms. these interactions were obtained from Integr8 database. Additionally, we tested on DIP interactions that score above a certain threshold.

A. Comparison With Previous Methods

We compare the ability of MSSC to predict protein-protein interactions against AM and MLE. We use sets of protein-protein interactions in [32] and [58]. Figure 6a shows that MSSC algorithm clearly outscores AM [56] as well as MLE [12] in both specificity and sensitivity.

B. Performance of MSC, MSSC and MSSC_2

Besides MSSC we also tried the minimum set cover (MSC). MSC uses different criteria to choose the sub-cover C from F so that C has the minimum cardinality |C| [10]. Compared to MSSC, MSC chooses fewer domain pairs to “cover” the protein-protein interaction network, but it actually “covers” more false positives. This is shown in Figure 6b. Compared to MSSC, MSSC_2 outperforms MSSC by a few percent. Notice that the solution for MSSC is a subset of MSSC_2 since they both use the same algorithm while MSSC_2 gives the algorithm more choices in order to get more accurate predictions.

C. Results with High Quality Interactions

Currently available sets of protein-protein interactions contain startling rates of false positives (≈ 50%) and false negatives (≈ 90%) [60]. Recently, the quality of a protein-protein interaction was observed to correlate well with the degree of clustering of its immediate networks neighborhood [21]. We assume that our prediction results can be significantly improved by focusing on such highly clustered links. Calculating the hypergeometric clustering coefficient for every link in the yeast interaction network, we elucidated only those interactions that score above a certain level of clustering. In order to assess the strength of interactions which are embedded in an increasingly clustered neighborhood to significantly improve the quality of predictions, we calculated the corresponding specificity vs. sensitivity curves. Increasing clustering coefficient reduces the network and gives better results when comparing their corresponding sensitivity vs. specificity curves(Fig. 7a).

Encouraged by these results, we assume that the proteins which participate in the corresponding interactions will significantly be coexpressed. Indeed, we find that the distribution of the coexpression coefficients emerging from the predicted interacting proteins peaks around shifts toward higher values if the threshold is increased (Figure 7b (inset)). Since we predict interactions evaluated by an occurrence probability, we consider predicted pairs that score above p > 0.9. We observe that the initial trends of coexpression are enhanced. These results are further supported by highly significant P-values of a Students t-test which gradually decrease from $P_{t=1} = 3.3 \times 10^{-21}$ to $P_{t=15} = 9.26 \times 10^{-4}$, indicating that the limitation to clustered proteins which participate in clustered interactions indeed significantly elevate the quality of our predictions.
Fig. 7. (a) Observing that interactions which are embedded in a highly clustered neighborhood exhibit a significantly higher degree of quality, we calculated the hypergeometric clustering coefficient $C_{vw}$ of each link in the original protein-protein interaction network of Yeast. A network composed of interactions that score above a certain threshold is the basis for our predictions, which outscore the original predictions in terms of specificity and sensitivity. (b) Compared to a comprehensive set of coexpression values of protein links (background), the quality of the underlying stripped protein interaction network increases with increased threshold (inset). Since we predict interactions evaluated by an occurrence probability, we only consider predicted pairs that score above $p > 0.9$. The quality of predicted pairs thus obtained is well reflected by an elevated tendency of coexpression, measurements that are supported by low P-values ($P < 10^{-3}$) of a Student’s test. For these tests $R = T$.

**D. Testing**

For most of the results shown $R \subseteq T$. Figure 8 shows performance for disjoint training and testing sets, $R \cap T = \emptyset$. $R$ is approximately 80% of the whole data set, and $T$ is composed of the remaining protein-protein interactions. We observe that MSSC performs better as the clustering threshold increases. The result suggests that the corresponding interactions are coexpressed. The error analysis is based on 20 different runs. We apply AM to the set with threshold = 1, the result shows that MSSC outscores AM again in this case.

Fig. 8. Performance of MSSC in tests of disjoint training and testing sets, using DIP and Integr8. The training set, $R$, is about 80% of the whole data set, and the testing set, $T$, is the remaining 20%. Error analysis is based on 20 different runs. We also apply AM to the set with threshold = 1, and MSSC outscores AM in this case too.

**E. Results with Orthologous Interactions**

The observation that highly clustered areas provide high reliability of the corresponding links returns in a more biological disguise. In a recent work, densely connected small sub-networks or motifs were found to exhibit a significantly higher degree of conservation than their sparsely connected counterparts [66]. Combining the latter aspect with the already uncovered correlation of high clustering and preferential coexpression, we expect that yeast interaction networks that consist of proteins that have an ortholog to a higher order organism are good candidates to improve our prediction results. Utilizing ortholog data of the higher eukaryotes *H. sapiens, D. melanogaster, M. musculus, C. elegans* and *A. thaliana* we elucidated the corresponding interaction network of yeast proteins. Applying our method to predict potential interactions from these networks we clearly observe that the predictions clearly outscore the respective ones utilizing the whole yeast interaction network. As a further improvement, we span an interaction network by proteins that have an ortholog in all organisms under consideration (crossection).

Indeed, we find that this set of interactions which presents a core of the higher eukaryotes once again exceeds the results of the latter predictions (Fig. 9a). Utilizing yeast coexpression data, we find our initial assumption that clustered interactions show preferential coexpression to be confirmed. In all cases, we observe a clear trend toward coexpression of the predicted interactions (Fig. 9b, inset), tendencies that are enhanced by limiting to interactions that score above occurrence probability $p > 0.9$ (Fig. 9b). Supporting their significance, we observe gradually decreasing P-values of Student’s
MSSC selects a set of domain pairs which covers the experimental observations and maximizes the specificity in the training set. Our results indicate that there is a strong correlation between high specificity in high quality training sets and in predicted protein-protein interaction sets.

The results also suggest that the quality of predicted protein-protein interactions depends basically on two different aspects. First, we clearly find that the quality of predictions is strongly enhanced protein-protein interaction training sets with a high degree of clustering. Second, highly clustered proteins show an elevated degree of coexpression in the training and testing data sets used.

Furthermore, proteins that participate in many interactions are preferentially conserved and change their sequence only to a small extent \[17\], \[68\]. We observe that high clustering and coexpressed protein-protein interaction sets show preferential evolutionary conservation. This allows us to refine our predictions by relying on evolutionary relevant interactions. Such sets of interactions may indicate evolutionary cores from where the proteome and interactome developed further. Combining these two observations, we propose that these interactions will be conserved more generally among different organisms. Thus, this evaluation of protein-protein interactions focusing on the level of orthologous clustering around a certain link can underpin new tools to infer protein-protein interactions.

This framework allows us to get a rough evolutionary core of interactions in those species for which sequenced genomes, high-resolution genome expression data but no protein-protein interaction data currently exist. Such a characterization will shed new perspectives on the evolution of medically or economically important disease organisms by uncovering organism-specific features, ultimately allowing the design of appropriate control strategies.

VII. Discussion

MSSC is able to improve the specificity for a given sensitivity over other protein-protein prediction methods. An open theoretical question is whether the protein-protein cover problem is NP-complete.

**References**


