Predicting Protein-Protein Interactions from Protein Domains Using a Set Cover Approach

Chengbang Huang[1], Simon P. Kanaan[1], Stefan Wuchty[2], Danny Z. Chen[1], Jesús A. Izaguirre[1,*]

[1]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, United States

[1]Department of Physics, University of Notre Dame, Notre Dame, IN 46556, United States

Abstract

The goal of contemporary proteome research is the elucidation of 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. Our approach features two stages: First, we select high quality protein interactions based on a clustering measure. Second, we use MSSC to assign probabilities to domain pairs. This approach allows us to predict previously unknown protein-protein interactions with a degree of sensitivity and specificity that clearly outscores other approaches. We achieve 86% sensitivity and 62% specificity using 80% of the high quality interactions in the DIP database. We find that the predicted interaction network preserves the characteristics of the initial web of known protein interactions. We also observe high levels of co-expression among putative interactions.

Index Terms

F.2.2.b *Computations on discrete structures*, G.2.2.a *Graph algorithms*, H.2.8.a *Bioinformatics (genome or protein) databases*, J.3.a *Biology and genetics*

I. Introduction

A goal of contemporary proteome research is the elucidation of the structure, interactions and functions of the proteins that constitute cells and organisms. Genomics has already produced an incredible quantity of molecular interaction data, contributing to maps of specific cellular networks. Indeed, large-scale attempts have unraveled the complex web of protein interactions in organisms as diverse as *H. pylori* [42] and *S. cerevisiae* [14], [23], [26]–[28], [45], [50]. Most recently, attention focused on the first protein interaction maps of complex multicellular organisms such as *C. elegans* [52] and *D. melanogaster* [15].

Although large-scale experimental attempts to uncover the complex webs of protein interaction in various organisms are still in progress, theoretical considerations focus on the prediction of potential protein interactions. Pioneering

* Corresponding author

JAI, CH, and SPK were partially funded by NSF grants IBN-0083653, IBN-0313730, and ACI-0135195. The simulations were run in a cluster funded by Notre Dame's high performance cluster grant to JAI.

methods drew on the observation that interacting protein domains tend to combine into a fusion protein [12], [34]. Another approach focused on the observation that functionally linked proteins tend to be either preserved or eliminated in evolution. Proteins having matching phylogenetic profiles strongly tend to be functionally linked [33], [41]. The domain architectures of the interacting proteins account for the basic structure of a protein and offer a framework for prediction models. Interaction domain pair profiles [54] assess 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 [9], [25]. Further ideas include overrepresented domain signatures [48], domain combination [21], graph-theoretical methods [17] and other probabilistic approaches [18], [49].

Assuming that protein domains facilitate the interactions among proteins, we introduce a novel method for the inference of protein interactions, which we test in *S. cerevisiae*. Utilizing a maximum-specificity set cover procedure (MSSC), we calculate the probabilities of putative protein interactions on an interaction network of yeast proteins. Our algorithm clearly outscores previous methods in terms of sensitivity and specificity. The predicted web of protein interactions that keeps the modular scale-free topology of the initial network. Our predictions correlate significantly with elevated levels of co-expression of micro-array data. We refine our predictions by utilizing a set of highly clustered interactions for our analysis. We observe that the proteins which constitute the predicted interactions are strongly co-expressed. Since interactions which are embedded in a highly clustered neighborhood tend to have an elevated degree of quality we conclude that this clustering preprocessing is a crucial step to significantly enhance the specificity of our predictions. Furthermore, we observe that such a set of preprocessed interactions improves MSSC's ability to deal with significantly flawed data. Thus, we conclude that the combination of our algorithm with clustered interaction data helps to eliminate false positive and negative interaction signals, allowing for high quality predictions.

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 [10] while sequences of multicellular eukaryotes appear as multi-domain proteins of up to 130 domains [32].

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 Interactions

The first comprehensive, albeit weakly overlapping protein interaction maps of *S. cerevisiae* have been provided with the yeast-two-hybrid method [27], [45]. Currently, there exists a variety of yeast specific protein interaction databases.

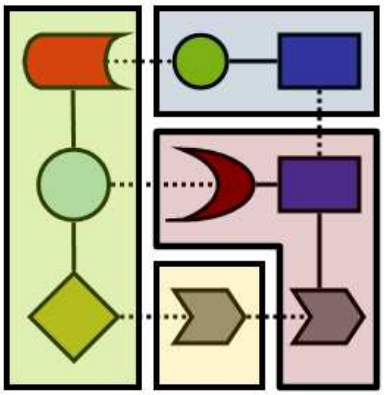


Fig. 1. The fundamental units of proteins (shaded areas) are the domains (geometrical figures), mediating a distinct structure and biological functionality. We assume that the underlying protein domain architectures facilitate the interactions among proteins, allowing us to design a novel method for the inference of protein interactions in *S. cerevisiae*.

Most of them, such as MINT [59], MIPS [39] and BIND [2], collect experimentally determined protein interactions. PREDICTOME [36] and STRING [37] collect functional links between proteins, derived from genome scale two hybrid sets, domain fusion events, phylogenetic history and gene proximity. These databases lack an assessment of the data's quality. In contrast, the GRID database, a compilation of BIND, MIPS and other datasets, as well as the DIP database [58] 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 analyses. Here, we use the DIP database (http://dip.doe-mbi.ucla.edu) which is the qualitatively best compilation of yeast 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 Domains

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 [5] (http://pfam.wustl.edu). The current version 10.0 contains $6,190$ fully annotated PFAM-A families. PFAM-B provides additional PRODOM-generated [8] alignments of sequence clusters in SWISSPROT and TrEMBL [6] that are not modeled in PFAM-A. In order to elucidate the PFAM domain architecture, we browsed swisspfam, a compilation of the domain structure of SWISSPROT and TrEMBL proteins according to PFAM.

C. Microarray Data

Genes with similar expression profiles are likely to encode interacting proteins [20]. We assess MSSC's ability to predict pairs of potentially interacting yeast proteins, by utilizing gene expression data of Eisen *et al.* [11]. This com-

pilation of co-expression patterns consists of $2,467$ yeast genes whose co-expression 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. Conserved Network Features

Almost all biological networks are characterized by a series of organizing principles [3]. The most dramatic is their scale-free nature, indicating that the probability that a node has degree k follows a power law, $P(k) \sim k^{-\gamma}$ [1], [4]. Indeed, we find this inhomogeneity in protein-protein interaction networks of numerous organisms [15], [28], [51]: While most nodes have a small degree, a few highly connected hubs hold the network together [4]. The hubs' crucial role for the protein network's integrity is further indicated by the observation that highly interacting proteins exhibit a significantly elevated propensity to be simultaneously lethal and conserved in evolution [28], [55], [56].

Another important feature of complex networks is their tendency to cluster. The clustering coefficient [53] of a node i is defined as

$$C_i = \frac{2n_i}{k_i(k_i-1)}, \tag{1}$$

where n_i denotes the number of links connecting the k_i neighbors of node i to each other. The network's inherent modularity is reflected by the distribution of C as a function of the nodes' degree k. If $C(k)$ follows $C(k) \sim k^{-1}$, the network has a hierarchical architecture, indicating that sparsely connected nodes are part of highly clustered areas [43]. This topology allows communities and the scale-free topology to seamlessly coexist [43], suggesting that complex networks are best described as the accumulation of discernible, yet topologically overlapping, functional modules. Apparently, networks featuring such functional modules are observed in almost all types of biological systems [24], [43], [44], [47] where a small subset of hubs play the important role of linking the networks modules [16], [22], [31], [43], [46]. Utilizing available yeast protein interactions as a training set of the MSSC, we expect that the web emerging from the predicted interactions will preserve these network characteristics.

E. Assessment of Protein Interactions

Although the current results concerning the structure of protein interaction networks are impressive, the error-proneness of experimental methods for the determination of protein interactions jeopardizes the strength of the obtained results. A recent estimation of the rate of inaccurately determined yeast protein interaction data uncovered a startling false negative rate of $90\,\%$ while false positives show a $50\,\%$ error rate [38]. Despite these data inconsistencies, a network topology based approach [17] uncovered a remarkable correlation between enhanced quality and network clustering around a certain protein interaction. Considering an interaction network of N nodes, the hypergeometric clustering coefficient, defined as

$$C_{vw} = -\log \sum_{i=|N(v)\cap N(w)|}^{\min(|N(v)|,|N(w)|)} \frac{\binom{|N(v)|}{i}\binom{N-|N(v)|}{|N(w)|-i}}{\binom{N}{|N(w)|}}, \tag{2}$$

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 [17]. 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 pair of nodes. By applying different cut-off values, we elucidated the corresponding interaction network, serving as the basis for further protein interaction predictions. We expect that the interaction webs that exhibit an elevated degree of clustering raise the quality of our predictions.

F. Quality Measures

The accuracies of the our predictions are measured by specificity and sensitivity. The specificity is defined as the ratio of the number of matched interactions between the predicted set, P, and the observed testing set, T, over the total number of predicted interactions, $Sp = \frac{|P \cap T|}{|P|}$. The sensitivity is defined as the ratio of the number of matched interactions, P, over the total number of observed interactions, T, in the testing set, $Sn = \frac{|P \cap T|}{|T|}$. Thus, it is obvious these metrics are testing set dependent.

III. Previous Prediction Methods

In order to have an estimate of the quality of our predictions, we compare the performance to previous methods. In the following, we will give a brief description of the most relevant algorithms that utilize protein interactions and their corresponding domain profiles to predict otherwise unknown protein interactions in *S. cerevisiae*.

A. Association Method (AM)

The association method [48] assigns an interaction probability

$$P(d_m, d_n) = \frac{I_{mn}}{N_{mn}} \tag{3}$$

to each domain pair (d_m, d_n). I_{mn} is the number of interacting protein pairs that contain (d_m, d_n), and N_{mn} is the total number of protein pairs that contain (d_m, d_n).

B. Maximum Likelihood Estimation (MLE)

The maximum likelihood estimation method [9] 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), the probability of a potential interaction is

$$E(P_i, P_j) = 1 - \prod_{(d_m, d_n) \in (P_i, P_j)} (1 - P(d_m, d_n)). \tag{4}$$

where $P(d_m, d_m)$ denotes the probability that domain d_m interacts with domain d_n. So, the maximum likelihood is

$$L = \prod (E(O_{ij} = 1)^{O_{ij}} (1 - E(O_{ij} = 1))^{1-O_{ij}}, \tag{5}$$

where

$$O_{ij} = \begin{cases} 1 & \text{if interaction between proteins } \mathrm{P_i} \text{ and } \mathrm{P_j} \text{ exists,} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The likelihood L is a function of $\theta(E(d_i, d_j), f_p, f_n)$, where $E(d_i, d_j)$ 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 [9] achieved 42.5% specificity and 77.6% sensitivity on a combined yeast protein interaction set compiled from [27], [45].

IV. 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 say that a domain pair covers a protein-protein interaction if the two interacting proteins contain the two domains respectively.

We define the protein interaction problem as the problem of finding a set of domain pairs to represent the given protein-protein interactions. Ideally, the set of domain pairs should give as few false positives as possible. False positives are the predicted protein-protein interactions not included in the input interaction network.

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 $\mathcal{C}$ of $\mathcal{F}$ to cover X,

$$X = \bigcup_{S \in \mathcal{C}} S, \tag{7}$$

and $\mathcal{C}$ is also required to satisfy certain conditions according to different specific problems. For example, the minimum exact set-cover (MESC) problem requires that $\sum_{S \in \mathcal{C}} |S|$ is minimized, and the minimum set-cover (MSC) problem is to find a $\mathcal{C}$ with the minimum cardinality $|\mathcal{C}|$ [7], [29].

We generalize the set-cover problem by enclosing X into a bigger set Y (Figure 2). 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 $\mathcal{C}$ of $\mathcal{F}$ to cover X,

$$X \subseteq \bigcup_{S \in \mathcal{C}} S \tag{8}$$

and $\mathcal{C}$ is also required to satisfy certain conditions according to different specific problems, as before.

With respect to the generalized set-cover setting, the MESC problem requires that $\sum_{S \in \mathcal{C}} |S|$ be minimized. This criterion implies that both the overlap of $\mathcal{C}$ with X and the overlap of $\mathcal{C}$ with $Y - X$ are minimized.

We believe that the protein-protein interaction problem is NP-hard, although we have not proved it yet. We solve the protein-protein interaction problem by transforming it into a set-cover problem. The experimentally known protein-protein interaction network can be modeled by a graph $G = (P, E)$, where P is the set of proteins and E is the set of

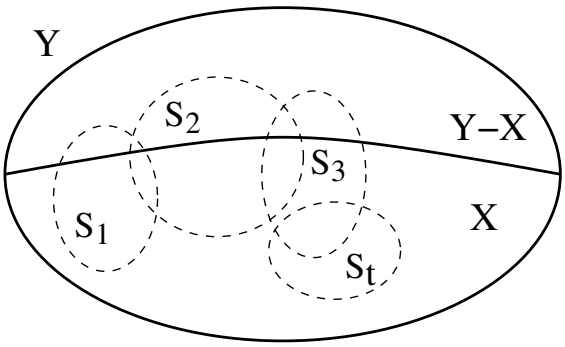


Fig. 2. The generalized set cover problem: X is a subset of Y, and $\mathcal{F} = \{S_i, 1 \leq i \leq t\}$ is a family of subsets of Y.

edges. The proteins are the vertices of G. 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 interaction network G by taking

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

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

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 $\mathcal{C}$ of $\mathcal{F}$ to cover every element (P_i, P_j) in X. An element in $\mathcal{C}$ 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 represent the interaction between P_i and P_j. Therefore, we also have a set of domain pairs to represent the protein network G.

Suppose there is a set D of domain pairs to represent the network G. For every element (P_i, P_j) in X, there is a domain pair (d_m, d_n) from D to represent 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 G 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 represent the protein interaction network. AM simply uses all the possible domain pairs to explain the protein-protein interactions, i.e., it uses $\mathcal{F}$ to cover X, so the resulting specificity is very low [9]. We are interested in using a subset of domain pairs to represent the protein-protein interaction network, and we choose the subset so that both the specificity and sensitivity are maximized, assuming that the training and testing set are the same. This is a reasonable assumption whenever the training set is sufficiently representative of the real testing sets, as is confirmed in our simulations.

The maximum specificity set cover (MSSC) problem is to find a subset $\mathcal{C}$ of $\mathcal{F}$ to cover X such that

$$m(\mathcal{C}) := \sum_{S \in \mathcal{C}} |S - X| \tag{9}$$

is minimized.

Comparing MSSC with MESC, we can see that MSSC allows the subcover $\mathcal{C}$ to cover the overlap with 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* $Y - X$.

Algorithm 1 is our greedy algorithm for MSSC. U represents the uncovered part of X. $\mathcal{E}$ is the subset of $\mathcal{F}$ that has not been chosen by the algorithm.

Algorithm 1 Greedy algorithm for MSSC.

GREEDY_MSSC($Y, X, \mathcal{F}$)
  $U \leftarrow X$
  $\mathcal{E} \leftarrow \mathcal{F}$
  $\mathcal{C} \leftarrow \emptyset$
  while $U \neq \emptyset$
    do select an $S \in \mathcal{E}$ 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 $\mathcal{C}$

In this algorithm, at each step when a subset needs to be chosen, we choose the one whose ratio between the part outside X and the part inside U is minimized. Note that the difference between MSSC and MESC is that MESC chooses the subset minimizing $\frac{|S-U|}{|S \cap U|}$, instead of $\frac{|S-X|}{|S \cap U|}$; MSSC allows the overlapping with X (Figure 3).

The number of iterations of the *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 $\mathcal{O}(|X||\mathcal{F}|\min(|X|, |\mathcal{F}|))$. If we apply proper data structures, it can be realized in $\mathcal{O}(\log |\mathcal{F}| \sum_{S \in \mathcal{F}} |S|)$ time. 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. Second, store all elements in $\mathcal{F}$ 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 $\sum_{S \in \mathcal{F}} |S|$ total. For a single S, the updating of the heap takes $|S| \log |\mathcal{F}|$. Therefore, the total time is $\mathcal{O}(\sum_{S \in \mathcal{F}} |S| + \sum_{S \in \mathcal{F}} |S| \log |\mathcal{F}|)$, which is $\mathcal{O}(\log |\mathcal{F}| \sum_{S \in \mathcal{F}} |S|)$. 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 $\mathcal{O}(|X|^2 + \sum_{S \in \mathcal{F}} |S|)$.

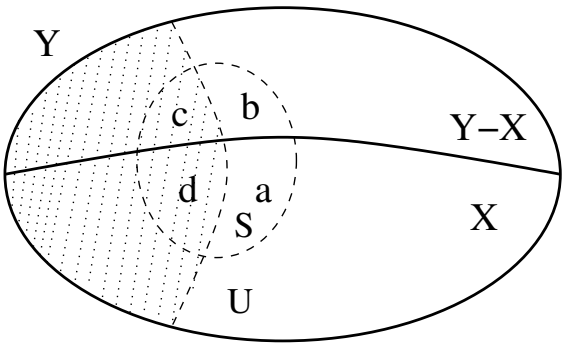


Fig. 3. The shaded area is already covered by $\mathcal{C}$. U is the unshaded area in X. The candidate set S is divided into 4 parts a, b, c and d. MSSC chooses a set S with the minimum $\frac{b+c}{a}$ while MESC chooses one with the minimum $\frac{b+c+d}{a}$. The greedy algorithm for MSSC allows overlapping of subcover inside X. The overlap actually increases the interaction probability for a protein pair.

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

Theorem 4.1: Suppose $\mathcal{C}_a$ is the approximation of MSSC found by the above greedy algorithm, and $\mathcal{C}_o$ is an optimal subcover for MSSC. Let $k = \max_{S \in \mathcal{F}} |S|$. If $m(\mathcal{C}_o) = 0$, then $m(\mathcal{C}_a) = 0$; otherwise, we have

$$\frac{m(\mathcal{C}_a)}{m(\mathcal{C}_o)} \leq [\ln(k-1) + 1]. \tag{10}$$

Proof: If $m(\mathcal{C}_o) = 0$, it means that all elements in $\mathcal{C}_o$ are subsets of X. *GREEDY_MSSC* cannot choose a set that is not completely in X, because at any given time when $U \neq \emptyset$ there exists a set $S \in \mathcal{C}_o \cap \mathcal{E}$ such that $\frac{|S-X|}{|S \cap U|} = 0$, i.e., $S \subseteq X$. Hence $m(\mathcal{C}_a) = 0$. We assume $m(\mathcal{C}_o) \neq 0$ from this point. Without loss of generality, we can also assume that $|\mathcal{C}_o| \leq |X|$. If $|\mathcal{C}_o| > |X|$, it means that at least one element S in $\mathcal{C}_o$ is redundant to cover X, so we can remove S from $\mathcal{C}_o$, and the remaining set is still an optimal solution.

$|\mathcal{C}_o| \leq |X|$ implies that $|m(\mathcal{C}_o) \leq k|X|$. Suppose

$$m(\mathcal{C}_o) = a|X|, \text{ for some value } a, 0 < a \leq k.$$

At a given time, assume that the minimum $\frac{|S-X|}{|S \cap U|}$ is r, where U is defined as in *GREEDY_MSSC*. For any $Z \in \mathcal{C}_o \cap \mathcal{E}$,

$$\frac{|Z - X|}{|Z \cap U|} \geq r,$$

so

$$\frac{|Z \cap U|}{|Z - X|} \leq \frac{1}{r}. \tag{11}$$

We have

$$\begin{aligned} |U| &= |\bigcup_{Z\in\mathcal{C}_o} Z\cap U| \\ &\leq \sum_{Z\in\mathcal{C}_o}\frac{|Z-X|}{r}, \text{ by Equation (11)} \\ &= \frac{m(\mathcal{C}_o)}{r} \\ &= \frac{a|X|}{r}. \end{aligned}$$

Therefore, there are $|X|-|U|\geq(1-\frac{a}{r})|X|$ points of X that are already covered when S is the next set to be chosen, i.e., *GREEDY_MSSC* cannot choose a set S with

$$\frac{|S-X|}{|S\cap U|}\geq r$$

until a fraction $(1-\frac{a}{r})$ of X has been covered. Conversely, if $x=\frac{|X-U|}{|X|}=1-\frac{a}{r}$ (the covered part of X), then $r=\frac{a}{1-x}$, and for the set S chosen by *GREEDY_MSSC*,

$$f(x):=\frac{|S-X|}{|S\cap U|}=\frac{a}{1-x}.$$

x is increasing from 0 to 1. Every time a new subset S is chosen, x "jumps" to a new value, so $f(x)$ is a step function of x. Since $|S\cap U|\geq 1$ and $|S-X|\leq k-1$, $f(x)\leq k-1$. Note that $\frac{a}{1-x}=k-1$ if and only if $x=1-\frac{a}{k-1}$.

When *GREEDY_MSSC* chooses a set S, S covers $|S\cap U|=|X|\Delta x$ more points of X, where $\Delta x=\frac{|S\cap U|}{|X|}$. The contribution of S to $m(\mathcal{C}_a)$ is

$$|S-X|=f(x)|S\cap U|=f(x)|X|\Delta x.$$

Therefore,

$$\begin{aligned} m(\mathcal{C}_a) &= \sum_{S\in\mathcal{C}_a}|S-X| \\ &= \sum_{S\in\mathcal{C}_a} f(x)|X|\Delta x \\ &= |X|\int_0^1 f(x)dx,\ f(x) \text{ is a step function} \\ &\leq |X|[\int_0^{1-\frac{a}{k-1}}\frac{a}{1-x}dx+\int_{1-\frac{a}{k-1}}^1(k-1)dx] \\ &= |X|(a\ln(k-1)-a\ln a+a) \\ &\leq a|X|[\ln(k-1)+1] \\ &= m(\mathcal{C}_o)[\ln(k-1)+1]. \end{aligned}$$

■

The theorem shows the relationship between the approximation by *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 contain a domain pair in the protein network.

When $X = Y$, MSSC is reduced to MSC, which is well known to be NP-hard. In the case of MSC, a logarithmic approximation is the best known approximation.

C. Prediction

Once the domain pairs are chosen by MSSC, each pair is assigned the same interaction probability (Equation (3)) as in AM. The unchosen domain pairs are given an interaction probability 0. Equation (4) is used to calculate the interaction probability for each putative protein pair.

V. Results

We use two sources of protein-protein interactions: one is the combined data set of Uetz *et al.* [50] and Ito *et al.* [27], which we call *CombUI*; the other is a complete protein-protein interactions set retrieved from the *DIP* database [58], which we simply name *DIP*. The combined Uetz and Ito is also used in [9]. We also study the interactions with $C_{vw} \geq 5$, they are embedded in highly clustered neighborhoods in *DIP*. This subset of *DIP* we call *DIP-5*

A. Performance Against Other Methods

We compare the ability of MSSC to predict protein-protein interactions against AM and MLE using *CombUI*. The training set is equal to the testing set in order to compare against published results. Figure 4a shows that MSSC clearly outscores AM [48] as well as MLE [9] in both specificity and sensitivity.

MSSC uses a different criterion than MSC. MSC chooses fewer domain pairs to cover the protein interaction network, but it actually covers more false positives. We observe that MSSC clearly outscores MSC (Figure 4b, using *DIP* as both the training and the testing set), so we can conclude that the design of the MSSC is much more suitable for the appropriate detection of potential protein interactions than is MSC.

Algorithm *GREEDY_MSSC* selects a set $S \in \mathcal{E}$ with the minimum $\frac{|S-X|}{|S \cap U|}$ (a tie is broken by $|S \cap U|$). If two sets have the same $\frac{|S-X|}{|S \cap U|}$ and $|S \cap U|$, *GREEDY_MSSC* chooses one randomly. The inset in Figure 4b shows the error bar for 15 different runs, suggesting that the performance of MSSC is consistent.

MESC chooses a subcover $\mathcal{C}$ with minimum $\sum_{S \in \mathcal{C}} |S|$. The overlap inside X is reduced as much as possible, while MSSC allows the overlapping inside X. For the data files we are using, the prediction difference between MSSC and MESC is very slight. We expect this difference to be more significant for networks of higher eukaryotes, where more redundancy in the protein interaction network should be present.

B. Robustness of MSSC

We take different percentages of the protein network as the training set and the network itself as the testing set and compare MSSC against AM and MSC. In Figure 5, six different training sets are used, 10%, 20%, 40%, 60%, 80% and 100% respectively. The result shows that MSSC consistently outscores AM and MSC when the specificity is high enough, regardless of the size of training set.

The same test is also performed on *DIP-5*. Figure 6 shows that with a highly clustered interaction network, the corresponding percentages for specificity and sensitivity are higher.

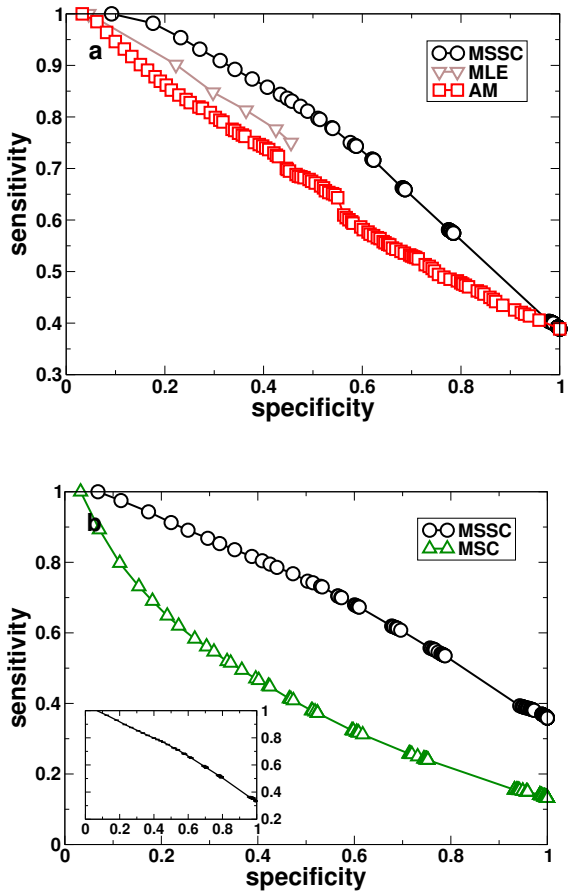


Fig. 4. **(a)** Using *CombUI* as both the training set and the testing set, we compare the performance of MSSC, AM [48] and MLE [9]. MSSC shows significantly higher specificity and sensitivity. **(b)** Using *DIP* as both training set and testing set, we observe that the MSSC algorithm clearly outscores MSC, allowing us to conclude that the design of the MSSC is much more suitable for the appropriate detection of potential protein interactions. We carried out an error analysis by running each algorithm 15 times (inset), allowing us to conclude that the performance of MSSC is consistent.

C. Conservation of Network Characteristics

The underlying protein-protein interaction network has some unique statistical characteristics. The scale-free nature is exemplified by the presence of a power-law in the networks degree distribution. The presence of modularity is indicated by a power-law dependence of the clustering coefficient $C(k)$ from degree k. In Figure 7a, we focus on interactions that score above a certain probability cutoff, allowing us to observe that the power-law dependence of the degree distribution of the networks thus emerging remains untouched (inset). In order to support this qualitative observation quantitatively, we applied a two dimensional Kolmogorov Smirnov test. Comparing the degree distribution of the original yeast protein interaction network with the predicted networks emerging from the application of different probability thresholds, we find small differences ranging from 0.17 (threshold $t = 0.4$) to 0.28 (threshold $t = 1.0$).

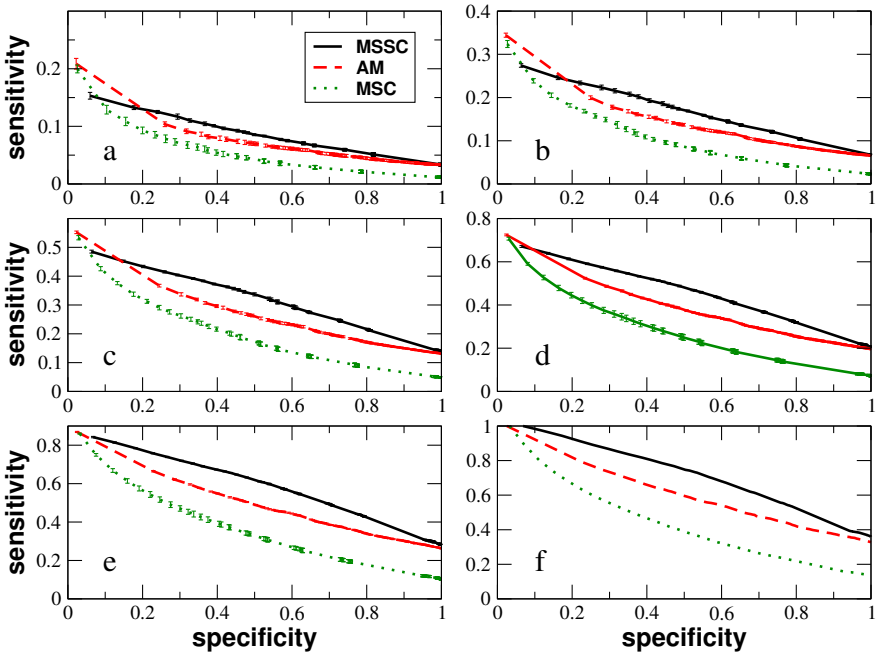


Fig. 5. The testing set is the whole *DIP* set, and different training sets are tried: **(a)** 10%, **(b)** 20%, **(c)** 40%, **(d)** 60%, **(e)** 80% and **(f)** 100% of *DIP*. The result shows that MSSC is consistent and outscores AM and MSC. The error bars are obtained by performing 10 runs for randomly chosen training sets.

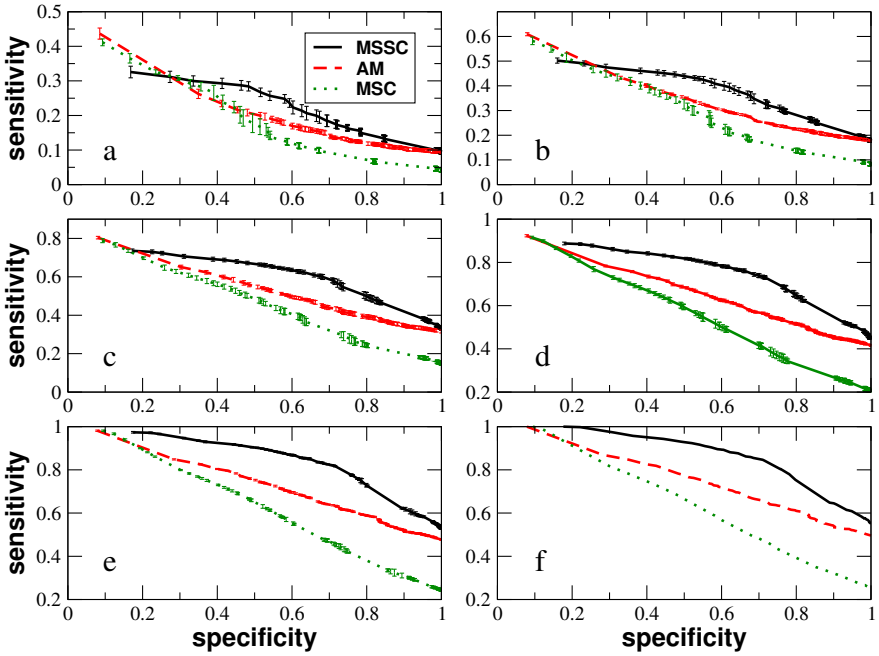


Fig. 6. The testing set is *DIP-5*, and different training sets are tried: **(a)** 10%, **(b)** 20%, **(c)** 40%, **(d)** 60%, **(e)** 80% and **(f)** 100% of *DIP-5*. The result shows that MSSC consistently outscores AM and MSC. The error bars are obtained by performing 10 runs for randomly chosen training sets.

Furthermore, we find that the corresponding P-values gradually decrease from $P_{t=0.4} = 0.27$ to $P_{t=1.0} = 9.6 \times 10^{-2}$, allowing us to conclude that the observed distributions are basically drawn from the same statistical sample. In the same way, the choice of the cutoff value does not seem to impair the emergence of modules as well. We still find the power-law dependence of the clustering coefficients of the networks emerging from the application of different thresholds t, observations that are strongly supported by KS-scores (KSS) ranging from $KS_{t=0.4} = 0.17$ ($P_{t=0.4} = 0.99$) to $KS_{t=1.0} = 0.24$ ($P_{t=1.0} = 0.88$).

A different assessment of the predictions quality is the tendency of interactions toward co-expression. Utilizing the initial protein interaction data of *S. cerevisiae* and a set of co-expression data [11], we observe a bell-shaped curve peaking around a zero co-expression coefficient. If there exists a correlation between the presence of an interaction between a pair of proteins and their co-expression, we expect a shift to higher expression coefficients. Figure 7b shows that higher probability cutoffs let the resulting networks exhibit an enrichment of co-expressed interacting proteins. Assuming that the observed distributions roughly have the same variance, we apply a Student's t-test to uncover possibly different means of the predicted co-expression profiles. Applying different thresholds t, we find statistically significant t-scores (TTS) $TTS_{t=0.4} = 18.38$ ($P_{t=0.4} = 2.9 \times 10^{-72}$) to $TTS_{t=1.0} = 15.86$ ($P_{t=1.0} = 6.1 \times 10^{-53}$), confirming our observation that an elevated amount of interacting proteins scores higher expression coefficients.

D. Results with High Quality Interactions

Currently available sets of protein interactions contain startling rates of false positives ($\sim 50\%$) and negatives ($\sim 90\%$) [37]. Recently, the quality of a protein interaction was observed to correlate well with the degree of clustering of its immediate networks neighborhood [17]. 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/sensitivity curves. Assessing the ability of the reduced networks to provide a high quality sample of interactions we compared their corresponding sensitivity/specificity curves, allowing us to observe best results with a protein interaction network emerging from links that score above $C_{vw} \geq 5$. We obtained best results with a rather small network constituted by 354 nodes and $2,660$ interactions.

Figure 8a shows that the specificity/sensitivity curve is significantly shifted to higher values than the corresponding one obtained from the full protein interaction network in the worst case scenario, which is to use disjoint training and testing sets. Encouraged by these results, we assume that the proteins which participate in the corresponding interactions will significantly be co-expressed. Indeed, we find that the co-expression coefficient emerging from the predicted interacting proteins peaks around 0.2 (Figure 8b). Allowing different thresholds t, we find that the corresponding t-test scores (TTS) gradually decrease from $TTS_{t=0.4} = 28.2$ ($P_{t=0.4} = 8.1 \times 10^{-171}$) to $TTS_{t=1.0} = 19.3$ ($P_{t=1.0} = 8.4 \times 10^{-83}$), indicating that the limitation to clustered proteins which participate in clustered interactions indeed significantly elevates the quality of our predictions.

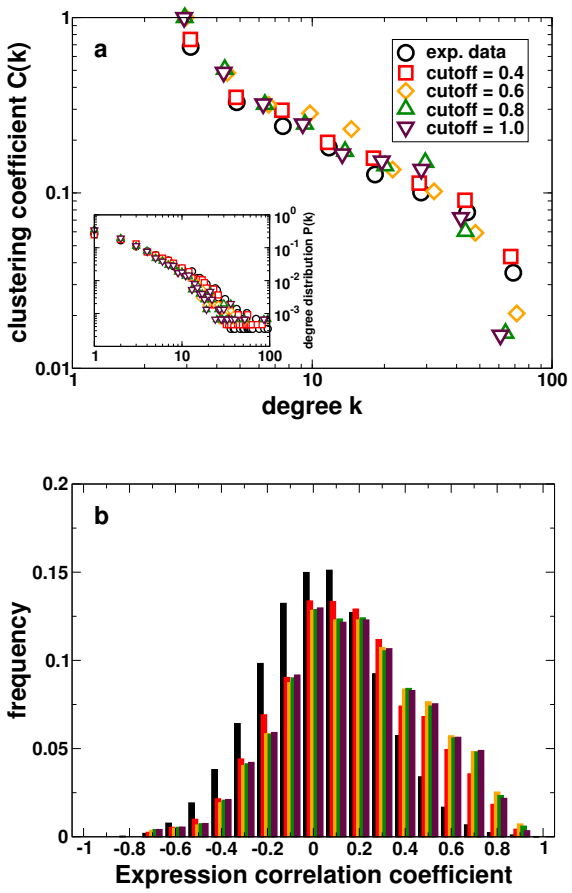


Fig. 7. To assess the quality of protein interaction predictions of MSSC, we determine the statistical characteristics of the networks which emerges from our predictions. **(a)** Utilizing *DIP*, we observe that the degree distribution follows a power-law (inset). Accounting for interactions which score above a certain probability, we find that the power-law dependence of the degree remains largely unchanged. Similarly, we observe that the power-law dependencies of the clustering coefficient $C(k)$ from the degree as exemplified by the experimental data does largely not depend on the choice of the cutoff value. **(b)** Utilizing a set of co-expression data, we observe a bell-shaped distribution curve. Accounting only for interactions that score above certain cutoffs we observe that proteins which participate in interactions scoring higher probability strongly tend to be co-expressed.

VI. Discussion

Our results suggest that the quality of predicted protein interactions depends basically on two different aspects. On the one hand, the quality of our predictions is strongly enhanced if we pre-assess the quality of the underlying protein interactions by determining the degree of clustering of the interaction's immediate neighborhood in the network. The observation that highly clustered links exhibit an elevated reliability is an important step toward the reliable prediction of potential interactions, since the considerable error-proneness of protein interaction data clearly influences the quality of results. The correlation between well clustered neighborhoods around the considered links and their interaction quality is further supported by the significantly elevated degree of co-expressed proteins that participate in present and predicted interactions. This observation is not only a proof of concept, it also suggests that potential strategies for

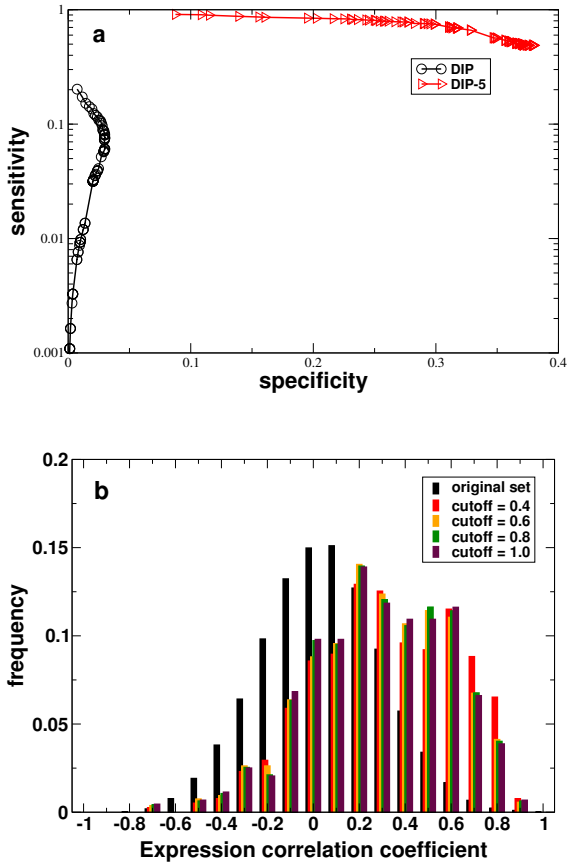


Fig. 8. **(a)** Performance of MSSC in tests of disjoint training and testing sets. The training sets are 80% of DIP and $DIP-5$, and the testing sets are the remainder. The predicted interactions in *DIP-5* exhibit a significantly higher degree of quality. **(b)** Compared to a random set of protein pairs, the quality of *DIP-5* increases with an elevated probability cutoff.

the determination of potential protein interactions have to focus on co-expressed areas of the underlying interaction network.

On the other hand, by design, our MSSC approach selects a set of domain pairs that both cover the experimental observations and that maximize the specificity in the training set. Our results indicate that there is a strong correlation between high specificity in high quality training sets and high specificity in realistic testing sets.

In this paper, we showed a new way of integrating protein interaction and domain data with a quality assessment of the underlying web of interactions. A further improvement of the algorithm will focus on the systematic integration of such pre-assessed interaction data in order to ensure highly reliable predictions. Once large-scale protein interaction sets of organisms other than *S. cerevisiae* are available, we expect that our algorithm will contribute significantly to the elucidation of complete organism-specific interactomes.

REFERENCES

[1] Albert, R. & Barabási, A.-L. (2002) Statistical Mechanics of Complex Networks. *Rev. Mod. Phys.,* **74**, 67–97.

[2] Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F., Pawson, T. & Hogue, C.W. (2001) BIND - The biomolecular interaction network database *Nucl. Acids. Res.* (29), 242–245.

[3] Barabaśi, A. & Oltvai, Z. (2004) Network biology: understanding the cell's functional organization. *Nature Rev. Gen.,* (5), 101–113.

[4] Barabási, A.-L. & Albert, R. (1999) Emergence of scaling in random networks. *Science,* **286**, 509.

[5] 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. (2004) The PFAM protein families database. *Nucl. Acids Res.,* **32**, D138–D141.

[6] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. & Schneider, M. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEmbl in 2003. *Nucl. Acids. Res.,* **31**, 365–390.

[7] Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. (2001) *Introduction to Algorithms, Second Edition*. McGraw Hill.

[8] Corpet, F., Servant, F., Gouzy, J. & Kahn, D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucl. Acids. Res.,* **28** (1), 267–269.

[9] Deng, M., Mehta, S., Sun, F. & Cheng, T. (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Res.,* **12**, 1540–1548.

[10] Doolittle, R. (1995) The Multiplicity of Domains in Proteins. *Ann. Rev. Biochem.,* **64**, 287–314.

[11] Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA,* **95**, 1483–14868.

[12] Enright, A., Iliopoulos, I., Kyrpides, N. & Ouzounis, C. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature,* **402**, 86–90.

[13] Flajolet, M., Rotondo, G., Daviet, L., Bergametti, F., Inchauspe, G., Tiollais, P., Transy, C. & Legrain, P. (2000) A genomic approach to the Hepatitis C virus. *Gene,* **242**, 369–379.

[14] Gavin, A., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J., Michon, A.-M., Cruciat, C.-M., 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, M.-A., 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. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature,* **415**, 141–147.

[15] Giot, L., Bader, J., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y., Ooi, C., Godwin, B., Vitols, E., Vijayadamodar, 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., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C., Finley Jr., R., White, K., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R., McKenna, M., Chant, J., & Rothberg, J. (2004) A protein interaction map of *Drosophila melanogaster*. *Science,* **302**, 1727–1736.

[16] Girvan, M. & Newman, M. (2002) Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA,* **99**, 7821–7826.

[17] Goldberg, D. & Roth, F. (2003) Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. USA,* **100**, 4372–4376.

[18] Gomez, S., Lo, S. & Rhetsky, A. (2001) Probabilistic prediction of unknown metabolic and signal transduction networks. *Genetics,* **159**, 1291–1298.

[19] Gomez, S. M., Noble, W. S. & Rzhetsky, A. (2003) Learning to predict protein-protein interactions from protein sequences. *Bioinformatics,* **19**, 1875–1881.

[20] Grigoriev, A.,(2001) 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.,* **29**, 3513–3519.

[21] Han, D., Kim, H.-S., Seo, J., & Jang, W. (2003) A domain combination based probabilistic framework for protein-protein interaction prediction. *Genome Informatics,* **14**, 250–259.

[22] Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. (1999) From molecular to modular cell biology. *Nature,* **402**, C47–C52.

[23] Ho, Y., Gruhler, A., Heilbut, A., Bader, G., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutillier, K. & coauthors, . (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature,* **415**, 180 – 183.

[24] Holme, P., Huss, M. & Jeong, H. (2003) Subnetwork hierarchies in biochemical pathways. *Bioinformatics,* **19** (4), 532–538.

[25] Iossifov, I., Krauthammer, M., Friedman, C., Hatzivassiloglou, V., Bader, J., White, K. & Rzhetsky, A. (2004). Probabilistic inference of molecular networks from noisy data sources.

[26] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. & Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Nat. Acad. Sci. USA,* **98** (8), 4569–4574.

[27] Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. & Sakaki, Y. (2000) 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,* **97** (3), 1143–1147.

[28] Jeong, H., Mason, S., Barabási, A.-L. & Oltvai, Z. (2001) Lethality and centrality in protein networks. *Nature,* **411**, 41–42.

[29] Johnson, D. S. (1974) Approximation algorithms for combinatorial problems. *J. Comput. System Sci.,* **9**, 256–278.

[30] Lappe, M., Park, J., Niggeman, O. & Holm, L. (2001) Generating protein interaction maps from incomplete data: application to fold assignment. *Bioinformatics,* **17**, S149–S156.

[31] Lauffenburger, D. (2000) Cell signaling pathways as control modules: Complexity for simplicity. *Proc. Natl. Acad. Sci. USA,* **97**, 5031–5033.

[32] Li, W.-H., Gu, Z., Wang, H., & Nekrutenko, A. (2001) Evolutionary analyses of the human genome. *Nature,* **409**, 847–849.

[33] Marcotte, E., Pellegrini, M., Ng, H.-L., Rice, D., Yeates, T. & Eisenberg, D. (1999*a*) Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science,* **285**, 751–753.

[34] Marcotte, E., Pellegrini, M., Thompson, M., Yeates, T. & Eisenberg, D. (1999*b*) A combined algorithm for genomewide prediction of protein function. *Nature,* **402**, 83–86.

[35] McGraith, S., Holtzman, T., Moss, B. & Fields, S. (2000) Genome-wide analysis of Vaccinia virus protein-protein interactions. *Proc. Natl. Acad. Sci. USA,* **97**, 4879–4884.

[36] Mellor, J.C., Yanai, I., Clodfelter, K.H., Mintseris, J. & DeLisi, C. (2002) Predictome: a database of putative functional links between proteins. *Nucl. Acids Res.,* **30**, 306–309.

[37] von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P. & Snel, B. (2003) STRING: a database of predicted functional associations between proteins. *Nucl. Acids Res.,* **31**, 258–261.

[38] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. & Bork, P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature,* **31**, 399–403.

[39] Mewes, H.W., Frishman, D., Büldener, U.G., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S. & Weil, S. (2002) MIPS: a database for genomes and protein sequences. *Nucl. Acids Res.,* **30**, 31–34.

[40] Ng, S.-K., Zhang, Z., Tan, S.-H. & Lin, K. (2003) Interdom: a database of putative interactive protein domains for validating predicted protein interactions and complexes. *Nucl. Acids Res.,* **31**, 251–254.

[41] Pellegrini, M., Marcotte, E., Thompson, M., Eisenberg, D. & Yeates, T. (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA,* **96**, 4285–4288.

[42] Rain, J.-C., Selig, L., DeReuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schächter, V., Chemama, Y., Labigne, A. & Legrain, P. (2001) The protein-protein interaction map of *Helicobacter pylori*. *Nature,* **409**, 211–215.

[43] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. (2002) Hierarchical Organization of Modularity in Metabolic Networks. *Science,* **297**, 1551–1555.

[44] Rives, A. & Galitski, T. (2003) Modular organisation of cellular networks. *Proc. Natl. Acad. Sci. U.S.A.,* **100**, 1128–1133.

[45] Schwikowski, B., Uetz, P. & Fields, S. (2000) A network of protein-protein interactions in yeast. *Nature Biotechn.,* **18**, 1257–1261.

[46] Shen-Orr, S., Milo, R., Mangan, S. & Alon, U. (2002) Network motifs in the transcriptional regulation network of *E. coli*. *Nature Genet.,* **31**, 64 – 68.

[47] Spirin, V. & Mirny, L. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA,* **100**, 12123–12128.

[48] Sprinzak, E. & Margalit, H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.,* **311** (4), 681–692.

[49] Tong, A., Drees, B., Nardelli, G., Bader, G., Branetti, B., Castagnoli, L., Evangelista, M., ferracuti, S., Nelson, B., Apoluzzi, S. & coworkers (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science,* **295**, 321–324.

[50] 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. (2000) A comprehensive analysis of protein-protein interactions of saccharomyces cerevisiae. *Nature,* **403**, 623–627.

[51] Wagner, A. (2001) The Yeast Protein Interaction Network Evolves Rapidly and Contains Few Redundant Duplicate Genes. *Mol. Biol. Evol,* **18** (7), 1283–1292.

[52] Walhout, A., Sordella, R., Lu, X., Hartley, J., Temple, G., Brasch, M., Thierry-Mieg, N. & Vidal, M. (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science,* **287**, 116–122.

[53] Watts, D. & Strogatz, S. (1998) Collective dynamics of 'small-world' networks. *Nature,* **393**, 440–442.

[54] Wojcik, J. & Schächter, V. (2001) Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics,* **17**, 296S–305S.

[55] Wuchty, S. (2002) Interaction and domain networks in yeast. *Proteomics,* **2** (12), 1715–1723.

[56] Wuchty, S. Topology and evolution in the yeast protein interaction network. *Genome Res.,* **14**, 1310–1314.

[57] Wuchty, S. Peeling the yeast interaction network. *Proteomics, in press*.

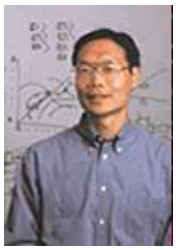
[58] Xenarios, I., Salwinski, L., Duan, X., Higney, P., Kim, S.-M. & Eisenberg, D. (2002) DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucl. Acids Res.,* **30**, 303–305.

[59] Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., & Cesareni, G. (2002) MINT - A Molecular INTeraction database. *FEBS Lett.,* **513**, 135–140.

Chengbang Huang is a Ph.D. student in the department of computer science and engineering at the University of Notre Dame, Notre Dame, Indiana, directed by Dr. Izaguirre. His current research interest is to use a multi-model framework to simulate avian limb growth, and algorithms for predicting protein-protein interactions. Huang has an M.S. in Mathematics from the University of Notre Dame.

Simon P. Kanaan is a graduate student in the Department of Computer Science and Engineering at the University of Notre Dame, Notre Dame, Indiana. He works as a teaching and research assistant under the direction of Dr. Izaguirre. His current research includes predicting interactions between proteins based upon domain configurations.

Stefan Wuchty is a postdoctoral fellow in the department of physics at the University of Notre Dame, Notre Dame, Indiana. He earned a M.S. in Chemistry and a Ph.D. in theoretical biochemistry and bioinformatics both from the University of Vienna, Vienna, Austria. His current research focuses on the investigation of networks in areas as diverse as molecular biology, sociology and business.



Danny Z. Chen received the B.S. degrees in Computer Science and in Mathematics from the University of San Francisco, California, in 1985, and the M.S. and Ph.D. degrees in Computer Science from Purdue University, West Lafayette, Indiana, in 1988 and 1992, respectively. He has been on the faculty of the Department of Computer Science and Engineering at the University of Notre Dame, Indiana since 1992, and is currently a professor. His main research interests are in algorithm design, analysis, and implementation, computational geometry, parallel and distributed computing, computational medicine, data mining, robotics, and VLSI design. He has published over 130 journal or conference papers in these areas. In 1996, Dr. Chen received the Faculty Early Career Development (CAREER) Award of the National Science Foundation.

Jesus A. Izaguirre is an assistant professor of computer science and engineering at the University of Notre Dame, Notre Dame, Indiana. His current research is on efficient methods in chemistry and biology, particularly molecular dynamics, Monte Carlo methods, cellular automata, and analysis of biological networks. He is also interested in the portable implementation of high performance software for scientific computing. He received a Ph.D. in computer science from the University of Illinois at Urbana-Champaign in 2000. Dr. Izaguirre received a CAREER Award of the National Science Foundation in 2001.