# Identification of Protein Complexes by Comparative Analysis of Yeast and Bacterial Protein Interaction Data

Roded Sharan* Trey Ideker† Brian Kelley‡ Ron Shamir§ Richard M. Karp*

May 2004

**Abstract**

Mounting evidence shows that many protein complexes are conserved in evolution. Here we use conservation to find complexes that are common to the yeast *S. cerevisiae* and the bacteria *H. pylori*. Our analysis combines protein interaction data, that are available for each of the two species, and orthology information based on protein sequence comparison. We develop a detailed probabilistic model for protein complexes in a single species, and a model for the conservation of complexes between two species. Using these models, one can recast the question of finding conserved complexes as a problem of searching for heavy subgraphs in an edge- and node-weighted graph, whose nodes are orthologous protein pairs.

We tested this approach on the data currently available for yeast and bacteria and detected 11 significantly conserved complexes. Several of these complexes match very well with prior experimental knowledge on complexes in yeast only, and serve for validation of our methodology. The complexes suggest new functions for a variety of uncharacterized proteins. By identifying a conserved complex whose yeast proteins function predominantly in the nuclear pore complex, we propose that the corresponding bacterial proteins function as a coherent cellular membrane transport system. We also compare our results to two alternative methods for detecting complexes, and demonstrate that our methodology obtains a much higher specificity.

*International Computer Science Institute, 1947 Center St., Berkeley, CA 94704. {roded,karp}@icsi.berkeley.edu.

†Dept. of Bioengineering, U. C. San-Diego, 9500 Gilman Drive, La Jolla, CA 92093. trey@bioeng.ucsd.edu.

‡Whitehead Institute for Biomedical Research, 9 Cambridge Ctr., Cambridge, MA 02142. bkelley@wi.mit.edu.

§School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel. rshamir@tau.ac.il.

# 1 Introduction

With the sequences of dozens of genomes at hand, and the accumulating information on the transcriptomes and proteomes of different organisms, a new research paradigm is emerging in molecular biology. At the core of this paradigm is the comparative analysis of biological properties of two or more species, using the wealth of organisms to enhance weak relations and to draw conclusions on one species, based on available information on other species. Comparative analysis, in the form of pairwise alignment, is frequently used in predicting protein function and structure. Recently, comparative approaches have been proven useful in diverse domains, including gene finding [5], motif finding [14], cis-regulation [22] and metabolic pathways [17, 26].

Protein interactions are crucial to cellular function, both in assembling protein machinery and complexes and in signaling cascades, where protein interactions enable one protein to modify another to transmit biological information. Recent technological advances enable us for the first time to characterize networks of protein interactions. Among the most direct and systematic methods for measuring protein interactions are co-immunoprecipitation [15] and the two-hybrid system [10] that have defined large protein-protein interaction networks for several organisms, including *S. cerevisiae* [9], *H. pylori* [19], *C. elegans* [6, 13] and *D. Melanogaster* [8].

The growing information on protein networks for different organisms naturally lends itself to comparative analysis, which tries to determine the extent to which protein networks are conserved among species and make use of the discovered conservation to predict novel networks or parts of networks. Mounting evidence suggests that conserved protein interaction pathways indeed exist and may be ubiquitous: For example, proteins in the same pathway are typically present or absent in a genome as a group [18], and several hundred protein-protein interactions in yeast have also been identified for the corresponding protein orthologs in worm [13, 16].

Previously, we have studied the conservation of pathways between the budding yeast *S. cerevisiae* and the bacterial pathogen *H. pylori* by performing a whole-network-based comparison between the protein-protein interaction maps of the two species [12]. Our analysis suggested many similarities between the two networks, some of which correspond to well-known biological pathways.

Here we study another aspect of the similarity between the yeast and bacterial protein networks, namely, the conservation of complexes between the two species. In both the present study and the earlier one the goal is to find two sets of proteins, one in yeast and one in bacteria, such that many of the proteins in each set have orthologous counterparts in the other, there is a high level

of interaction among the proteins in each set, and the patterns of interaction in the two sets are similar. But while the methods of the earlier study were aimed at finding chain-like patterns of interaction of the type that might occur in signal transduction pathways, the present study focuses on protein complexes and looks for dense, clique-like interaction patterns that are conserved in the two species.

At the heart of our analysis is a novel probabilistic model for a protein interaction network in a single species, and a two-species model that combines, in a unified manner, the interaction model of each species as well as information on the similarity of protein pairs between the two species. This model is used to construct an *orthology graph*, whose nodes correspond to pairs of putative orthologous proteins (a protein may appear in multiple nodes with different orthologs), and whose edges correspond to protein interactions. The edges of the graph are assigned weights with probabilistic meaning, so that high weight subgraphs correspond to conserved protein complexes. We propose a practical method to search the orthology graph for complexes of densely interacting proteins, which is based on forming high weight seeds and extending them using local search.

We applied our algorithm to data on protein interactions and protein sequences for yeast and bacteria. The algorithm identified 11 significantly conserved complexes, with sizes ranging between 6 and 20 protein pairs. Several of these complexes match very well with prior experimental knowledge on complexes in yeast only, and serve for validation of our methodology. The identified complexes suggest new functions for a variety of uncharacterized proteins. In particular, by identifying a conserved complex whose yeast proteins function predominantly in the nuclear pore complex, we propose that the corresponding bacterial proteins function as a coherent cellular membrane transport system. We also compare our results to two alternative methods for detecting complexes, and demonstrate that our methodology obtains a much higher specificity. In addition, we show the utility of using our probabilistic model for analyzing interaction data from a single species.

The paper is organized as follows: In Section 2 we present our probabilistic model for comparative protein interaction data. In Section 3 we describe our algorithm for identifying high-scoring, conserved protein complexes. The application to real data and comparison to extant approaches are presented in Section 4.

A preliminary version of this paper has appeared in the Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology [21].

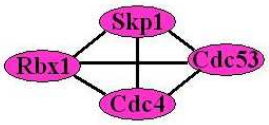


Figure 1: A protein complex in the yeast interaction graph. Shown is the Skp1-Cdc53/cullin-F-box (SCF) complex, which is involved in ubiquitin ligation. Each edge denotes a direct interaction between the corresponding proteins. As shown, this complex induces a clique in the yeast interaction graph.

## 2 A Probabilistic Model for Protein Complexes

In this section we present a probabilistic model for protein interaction data within a single species, and then extend it to two species. Given a dataset of protein interactions for some organism, we translate it into an *interaction graph* $G$, whose vertices are the organism's interacting proteins, and whose edges represent pairwise interactions between distinct proteins. Using this formulation, a protein complex corresponds to a subgraph of $G$ that is typically dense (see, e.g., Figure 1). Hence, surprisingly dense subgraphs in $G$ may be suggested as putative protein complexes.

In the case of perfect data, each edge in the interaction graph represents a known interaction, each non-edge represents a known non-interacting pair, and we are seeking a surprisingly dense subgraph of $G$. To this end we formulate a log likelihood ratio model that is additive over the edges and non-edges of $G$, such that highly scoring subgraphs would correspond to likely protein complexes. This approach requires specifying a null model and a protein-complex model for vertex pairs. Similarly to the probabilistic approach taken in [24, 23], we define the two models as follows: The *protein-complex model*, $M_c$, assumes that every two proteins in a complex interact with some high probability $\beta$. In terms of the graph, the assumption is that two vertices that belong to the same complex are connected by an edge with probability $\beta$ independently of all other information. While our model assumes that a clique is the ideal structure of a protein complex, other reasonable models could be formulated, such as a "hub" model, in which all vertices are connected to a center vertex of high degree.

In contrast, the *null model*, $M_n$, assumes that each edge is present with the probability that one would expect if the edges of $G$ were randomly distributed but respected the degrees of the vertices. More precisely, we let $F^G$ be the family of all graphs having the same vertex set as $G$ and the same

degree sequence, and we define the probability of observing the edge $(u, v)$ to be the fraction of graphs in $F^G$ that include this edge. Note that in this way, edges incident on vertices with higher degrees have higher probability. We estimate these probabilities using a Monte-Carlo approach, as described in Section 4. We assume that all pairwise relations are independent.

A complicating factor in constructing the interaction graph is that we do not know the real protein interactions, but rather have partial, noisy observations of them. Formally, let us denote by $T_{uv}$ the event that two proteins $u, v$ interact, and by $F_{uv}$ the event that they do not interact. Denote by $O_{uv}$ the (possibly empty) set of available observations on the proteins $u$ and $v$, that is, the set of experiments in which an interaction between $u$ and $v$ was, or was not, observed. Using prior biological information one can estimate for each protein pair the probability $Pr(O_{uv}|T_{uv})$ of the observations on this pair, given that it interacts, and the probability $Pr(O_{uv}|F_{uv})$ of those observations, given that this pair does not interact. Also, one can estimate the prior probability $Pr(T_{uv})$ that two random proteins interact.

Given a subset $U$ of the vertices, we wish to compute the likelihood of $U$ under a protein-complex model and under a null model. Denote by $O_U$ the collection of all observations on vertex pairs in $U$. Then

$$
\begin{aligned}
Pr(O_U|M_c) &= \prod_{(u,v)\in U\times U} Pr(O_{uv}|M_c) && (1)\\
&= \prod_{(u,v)\in U\times U} [Pr(O_{uv}|T_{uv}, M_c)Pr(T_{uv}|M_c) + Pr(O_{uv}|F_{uv}, M_c)Pr(F_{uv}|M_c)] && (2)\\
&= \prod_{(u,v)\in U\times U} [\beta Pr(O_{uv}|T_{uv}) + (1-\beta)Pr(O_{uv}|F_{uv})] && (3)
\end{aligned}
$$

Equation 1 follows from the assumption that all pairwise interactions are independent. Equation 2 is obtained using the law of complete probability. Equation 3 follows by noting that given the hidden event of whether $u$ and $v$ interact, $O_{uv}$ is independent of any model.

It remains to compute $Pr(O_U|M_n)$. Since our previous null model depended on having the degree sequence of the interaction graph, we cannot use it as is. To overcome this difficulty we approximate the degree sequence of the hidden interaction graph: Let $d_1, \ldots, d_n$ denote the expected degrees of the vertices in $G$, rounded to the closest integer. In order to compute $d_1, \ldots, d_n$ we apply Bayes' rule to derive the expectation of $T_{uv}$ for any pair $u, v$ given our observations on this vertex pair:

$$Pr(T_{uv}|O_{uv}) = \frac{Pr(O_{uv}|T_{uv})Pr(T_{uv})}{Pr(O_{uv}|T_{uv})Pr(T_{uv}) + Pr(O_{uv}|F_{uv})(1 - Pr(T_{uv}))}$$

Hence, $d_i = [\sum_j Pr(T_{ij}|O_{ij})]$, where $[\cdot]$ indicates rounding.

Our refined null model assumes that $G$ is drawn uniformly at random from the collection of all graphs, whose degree sequence is $d_1, \ldots, d_n$. This induces a probability $p_{uv}$ for every vertex pair $(u, v)$. We can now calculate the probability of $O_U$ according to the null model:

$$Pr(O_U|M_n) = \prod_{(u,v)\in U\times U} [p_{uv}Pr(O_{uv}|T_{uv}) + (1 - p_{uv})Pr(O_{uv}|F_{uv})]$$

Finally, the log likelihood ratio that we assign to a subset of vertices $U$ is

$$\begin{aligned} L(U) &= \log \frac{Pr(O_U|M_c)}{Pr(O_U|M_n)} && (4)\\ &= \sum_{(u,v)\in U\times U} \log \frac{\beta Pr(O_{uv}|T_{uv}) + (1-\beta)Pr(O_{uv}|F_{uv})}{p_{uv}Pr(O_{uv}|T_{uv}) + (1-p_{uv})Pr(O_{uv}|F_{uv})} && (5)\\ &= \sum_{(u,v)\in U\times U} \log \frac{\beta Pr(T_{uv}|O_{uv})(1-Pr(T_{uv})) + (1-\beta)(1-Pr(T_{uv}|O_{uv}))Pr(T_{uv})}{p_{uv}Pr(T_{uv}|O_{uv})(1-Pr(T_{uv})) + (1-p_{uv})(1-Pr(T_{uv}|O_{uv}))Pr(T_{uv})} && (6) \end{aligned}$$

where Equation 6 follows by applying Bayes' rule and cancelling common terms in the numerator and denominator.

## 2.1 Two-Species Conservation Model

Consider now the case of data on two species 1 and 2, denoted throughout by an appropriate superscript. Here we wish to score a conserved complex that is defined by two subsets of proteins, one from each species, and a many-to-many correspondence associating proteins in one species with their putative orthologous proteins in the other species. Consider two subsets $U^1 = \{u^1_1, \ldots, u^1_{k_1}\}$ and $V^2 = \{v^2_1, \ldots, v^2_{k_2}\}$ and some many-to-many mapping $\theta : U^1 \to V^2$ between them. Assuming that the interaction graphs of the two species are independent of each other, the log likelihood ratio score for these two sets is simply:

$$L(U^1, V^2) = \log \frac{Pr(O_{U^1}|M^1_c)}{Pr(O_{U^1}|M^1_n)} + \log \frac{Pr(O_{V^2}|M^2_c)}{Pr(O_{V^2}|M^2_n)}$$

However, this score does not take into account the degree of sequence conservation among the pairs of proteins associated by $\theta$. In order to include such information we have to define a conserved complex model and a null model for pairs of proteins from two species. Our conserved complex model assumes that pairs of proteins associated by $\theta$ are orthologous. The null model assumes that such pairs consist of two independently chosen proteins. Let $E_{uv}$ denote the BLAST E-value assigned to the similarity between proteins $u$ and $v$, and let $h_{uv}, \bar{h}_{uv}$ denote the events that $u$ and $v$ are orthologous, or non-orthologous, respectively. The likelihood ratio corresponding to a pair of

proteins $(u, v)$ is therefore

$$\frac{Pr(E_{uv}|M_c)}{Pr(E_{uv}|M_n)} = \frac{Pr(E_{uv}|h_{uv})}{Pr(E_{uv}|h_{uv})Pr(h_{uv}) + Pr(E_{uv}|\bar{h}_{uv})Pr(\bar{h}_{uv})}$$

Using Bayes' rule we can derive a simpler expression, $\frac{Pr(h_{uv}|E_{uv})}{Pr(h)}$, where $Pr(h)$ is the prior probability that two proteins are orthologous.

Thus, the complete score of $U^1$ and $V^2$ under the mapping $\theta$ is:

$$S_\theta(U^1, V^2) = L(U^1, V^2) + \sum_{i=1}^{k_1} \sum_{v_j^2 \in \theta(u_i^1)} \log \frac{Pr(h_{u_i^1 v_j^2}|E_{u_i^1 v_j^2})}{Pr(h)}$$

# 3 Searching for Conserved Complexes

Using the above model for comparative interaction data, the problem of identifying conserved protein complexes reduces to the problem of identifying a subset of proteins in each species, and a correspondence between them, such that the score of these subsets exceeds a threshold. In this section we present our search strategy.

## 3.1 The Orthology Graph

To allow efficient search for conserved protein complexes, we define a complete edge- and node-weighted *orthology graph* (extending [12]). We focus on yeast and bacteria. Denote by superscripts $p$ and $y$ the model parameters corresponding to bacteria and yeast, respectively. For two yeast proteins $y_1$ and $y_2$ define

$$w^y_{(y_1,y_2)} = \log \frac{\beta^y Pr(O_{y_1y_2}|T_{y_1y_2}) + (1-\beta^y)Pr(O_{y_1y_2}|F_{y_1y_2})}{p^y_{y_1y_2} Pr(O_{y_1y_2}|T_{y_1y_2}) + (1-p^y_{y_1y_2})Pr(O_{y_1y_2}|F_{y_1y_2})}$$

Similarly, for two bacterial proteins $p_1$ and $p_2$ define

$$w^p_{(p_1,p_2)} = \log \frac{\beta^p Pr(O_{p_1p_2}|T_{p_1p_2}) + (1-\beta^p)Pr(O_{p_1p_2}|F_{p_1p_2})}{p^p_{p_1p_2} Pr(O_{p_1p_2}|T_{p_1p_2}) + (1-p^p_{p_1p_2})Pr(O_{p_1p_2}|F_{p_1p_2})}$$

Every pair $(y_1, p_1)$ of yeast and bacterial proteins is assigned a *node*, whose weight reflects the similarity of the proteins, that is,

$$w_{(y_1,p_1)} = \log \frac{Pr(h_{y_1p_1}|E_{y_1p_1})}{Pr(h)} .$$

Every two distinct (but possibly overlapping) nodes $(y_1, p_1), (y_2, p_2)$, are connected by an edge, which is associated with a pair of weights $(w^y_{(y_1,y_2)}, w^p_{(p_1,p_2)})$. If $y_1 = y_2$ $(p_1 = p_2)$ we set the first (second) weight to 0.

By construction, an induced subgraph of the orthology graph corresponds to two subsets of proteins, one from each species, and a many to many correspondence between them. We define the *z-score* of an induced subgraph with vertex sets $U^1$ and $V^2$ and a mapping $\theta$ between them, as the log likelihood ratio score $S_\theta(U^1, V^2)$ for the subgraph, normalized by subtracting its mean and dividing by its standard deviation. In computing the $z$-score we assume that node- and edge-weights are independent, so the mean and variance of $S_\theta(U^1, V^2)$ are obtained by summing the sample means and variances of the corresponding nodes and edges. In detail, the sample mean of a node is the average node weight in the orthology graph; the sample mean of an edge is the average weight of an edge in the yeast interaction graph plus the average weight of an edge in the bacterial interaction graph; variances are similarly computed. High-scoring induced subgraphs in the orthology graph correspond to putative conserved protein complexes.

In order to reduce the complexity of the graph and focus on biologically plausible conserved complexes, we filter nodes from the graph as follows: We start with an initial set of yeast-bacterial protein pairs, whose BLAST E-value is smaller than $10^{-2}$. We filter from this set pairs for which at least one of the proteins has no interactions with other proteins in our data. Let $S$ be the resulting set of pairs. Pairs of proteins that do not belong to $S$ are considered to have low likelihood to take part in a conserved complex [12]. Hence, we consider them only if they satisfy the following condition: For every node $(p, y) \notin S$ we check whether there exist two nodes $(p_1, y_1), (p_2, y_2) \in S$ such that $p$ interacts with $p_1$ and $p_2$ and $y$ interacts with $y_1$ and $y_2$. If no such nodes exist, we remove $(p, y)$ from the graph. Otherwise, we retain it. The added nodes serve as 'bridges' in the orthology graph between protein pairs, whose members in each species are not known to directly interact.

### 3.2 The Search Strategy

The problem of searching for heavy induced subgraphs in a graph is NP-hard even when considering a single species where all edge weights are 1 or -1 and all vertex weights are 0 [20]. Thus, in the following we propose heuristic strategies for searching the orthology graph.

We perform a bottom-up search for heavy subgraphs in the orthology graph. We start from high weight seeds, refine them by exhaustive enumeration, and then expand them using local search. A similar approach based on local search was shown to work well in analyzing high-throughput genomic data [24]. Define an edge in the orthology graph as *strong* if the sum of its associated weights (the edge weights within each species' graph) is positive. In the first phase of the search we

compute a seed around each node $v$, which consists of $v$ and all its neighbors $u$ such that $(u, v)$ is a strong edge. If the size of this set is above a threshold (e.g., 10) we iteratively remove from it the node whose contribution to the subgraph score is minimum, until we reach the desired size. Next, we enumerate all subsets of the seed that have size at least 3 and contain $v$. Each such subset is a refined seed on which we apply a local search heuristic. During the local search we iteratively add a node, whose contribution to the current seed is maximum, or remove a node, whose contribution to the current seed is minimum, as long as this operation increases the overall score of the seed. Throughout the process we preserve the original refined seed and do not delete nodes from it. For each node in the orthology graph we record up to $k$ (e.g., 5) heaviest subgraphs that were discovered around that node.

The resulting subgraphs may overlap considerably, so we use a greedy algorithm to filter subgraphs whose percentage of intersection is above a threshold. The algorithm iteratively finds the highest weight subgraph, adds it to the final output list, and removes all other highly intersecting subgraphs (see Section 4 for precise details on the filtering criterion).

In order to evaluate the statistical significance of identified complexes we compute two kinds of $p$-values. The first is based on the $z$-scores that were computed (and optimized) for each subgraph, and assumes a normal approximation to the likelihood ratio of a subgraph. The approximation relies on the assumption that the subgraph's nodes and edges contribute independent terms to the score. The latter probability is Bonferroni corrected for multiple testing, according to the size of the subgraph. The second $p$-value is based on empirical runs on randomized data. The randomized data are produced by random shuffling of the input interaction graphs of the two species, preserving their degree sequences, as well as random shuffling of the orthology relations, preserving the number of orthologs associated with each protein. For each randomized dataset we use our heuristic search to find the highest-scoring conserved complex of a given size. We then estimate the $p$-value of a suggested complex of the same size, as the fraction of random runs in which the output complex had larger score.

# 4 Experimental Results

## 4.1 Building the Orthology Graph and Parameter Estimation

We applied our method to publicly available data on *S. cerevisiae* and *H. pylori*. We first constructed a protein-protein interaction network for each species. The yeast network contained 14,848 pairwise

interactions among 4,716 proteins, and is based on several systematic studies using protein co-immunoprecipitation and the yeast two-hybrid system. The bacterial network contained 1,403 pairwise interactions among 732 proteins, which come from a single two-hybrid study [19]. All interactions were extracted from DIP [29] (August 2003 version).

Protein sequences for both species were obtained from PIR [28]. Alignments and associated E-values were computed using BLAST 2.0 [3] with parameters $b = 0$; $e = 1E6$; $f =$"C;S"; $v = 6E5$. Unalignable proteins were assigned a maximum E-value of 5. Altogether, 1,909 protein pairs had E-value below 0.01, out of which 822 pairs contained proteins with some measured interaction. Adding 1,242 additional pairs with weak homology (as described in Section 3), and removing nodes with no incident strong edge, resulted in a final orthology graph $G$ with 866 nodes and 12,420 edges. In total, 248 distinct bacterial proteins and 527 yeast proteins participated in $G$.

A good estimation of the probabilistic parameters in the model is a precondition to obtaining meaningful results. We used the maximum likelihood method of Deng et al. [7] for estimating the reliability of observed interactions in yeast. The method provides reliability estimates that depend on the experimental method used to detect the interaction, and on the number of times each interaction was observed. The reliability of the interactions in *H. pylori* was estimated at 0.53, which was the same reliability as that assigned by Deng et al. [7] to the yeast two-hybrid experiment of Uetz et al. [27], and was supported by the estimations given in [19]. We did not use negative information on interactions that were tested but were not observed, as such data were not readily available.

For each species, the probabilities of observing each particular edge in a random graph with the same degree sequence, was computed by Monte-Carlo simulations as follows. Starting from the original interaction graph, we performed a long series of random edge crosses, each time picking at random two edges $(a, b), (c, d)$, and replacing them with $(a, c), (b, d)$ (disallowing self-loops), provided that the latter two edges were not present in the current graph. This process shuffles the edges in the graph while preserving the degrees of the vertices. The percentage of simulations in which an edge was observed was the estimate of its probability.

Other parameters related to the edge weights were set as follows: The probability of observing an interaction in a complex model was set in both yeast and bacteria to $\beta = 0.95$. The prior probability of a true interaction was set to 0.001 in both species based on an estimate of 6,334 proteins and 20,000 true protein-protein interactions in yeast [4]. The probability of observing a true interaction was estimated by the ratio of expected number of true interactions that were

observed and the total number of observed interactions. The prior probability of observing some interaction for a given pair of vertices was set according to the number of interactions and proteins in our data: $Pr(O^y_{y_1,y_2} \neq \emptyset) = 0.00135$ and $Pr(O^p_{p_1,p_2} \neq \emptyset) = 0.00524$.

As for the node weights, the parameters of the model were derived as follows: The prior probability that a yeast protein and a bacterial protein are orthologous was computed as the frequency of protein pairs from both species that are in the same COG (cluster of orthologous genes) [25], with a value of $Pr(h) = 0.001611$ (see also [12]). The conditional probability that a pair of proteins are orthologous, given their BLAST E-value, was computed as in [12].

### 4.2 Protein Complex Identification and Validation

We applied our algorithm to the yeast-bacteria orthology graph in search of conserved complexes. Altogether, the algorithm identified 11 non-redundant complexes, whose $p$-values were smaller than 0.05, after correction for multiple testing. These complexes were also found to be significant when scored against empirical runs on randomized data ($p < 0.05$). The complexes are listed in Table 1. Complex sizes varied between 6 to 20 protein pairs (20 was the maximum allowed size). Redundant complexes were filtered by disallowing large overlap between two complexes. Precisely, if 60% of the nodes or 60% of the distinct proteins in each species were common to two complexes, the one with the poorer $p$-value was removed.

To validate the results, we first used information about known protein complexes in yeast. We extracted assignments of yeast genes to complexes from the MIPS database [1] (August 2003 version). 285 nodes in the orthology graph had such assignments. We used complex categories at level 3 of the MIPS complex hierarchy. In total, 18 categories had at least three genes from the orthology graph, and six categories had at least five. For each of our complexes we computed the largest number of proteins from a single category, as a fraction of all its categorized members. This fraction is called the *purity* of the complex. High purity indicates a conserved complex that corresponds to a known complex in yeast, and serves as a validation for the result. Low purity may either indicate an incorrect complex or a previously unidentified correct one. Note that most complexes also contain proteins that are not known to belong to any complex in yeast, and so our results suggest additional members in known complexes.

For bacteria, since experimental information on complexes is unavailable, we used functional annotations instead. We extracted 864 functional annotations of bacterial genes from the TIGR database [2]. We used a categorization of *H. pylori* genes into 14 broad functional classes, spanning

757 nodes in our orthology graph. Purity was computed in the same manner.

The significant complexes that we have identified exhibit a nice correspondence between the protein complex annotation in yeast and the functional annotation in bacteria, as presented in Table 1 and further visualized in Figure 2. For instance, complex 17 contains proteins from both yeast and bacteria that are involved in protein degradation; complexes 19 and 28 consist predominantly of proteins that are involved in translation; and complex 30 includes proteins that are involved in membrane transport.

Our conserved protein complexes suggest new functions for a variety of uncharacterized proteins. For instance, complex 17 (Figure 2(a)) defines a set of conserved interactions for the cell's protein degradation machinery. Bacterial proteins HP0849 and HP0879 are largely uncharacterized, but their appearance among yeast and bacterial proteins involved in proteolysis suggests that they also play an important role in this process. Furthermore, it appears that the yeast proteins Hsm3 and Rfa1 (with known functional roles in DNA-damage repair) may also be associated with the yeast proteasome. Complexes 19 and 31 (Figure 2(b,d)) suggest that their component proteins, some of which are uncharacterized, are involved in protein synthesis.

As another example of protein functional prediction, Figure 2(b) shows a conserved complex which contains yeast proteins that function in the nuclear pore (NUP) complex. The NUP complex is integral to the eukaryotic nuclear membrane and serves to selectively recognize and shuttle molecular cargos (e.g., proteins) between the nucleus and cytoplasm. Unlike the yeast proteins, the corresponding bacterial proteins are less well characterized, although three have been associated with the cell envelope due to their predicted transmembrane domains. Our results therefore indicate that the bacterial proteins may function as a coherent cellular membrane transport system in bacteria, similar to the nuclear pore in eukaryotes. Although further experimentation will be necessary to explore this hypothesis, it is possible that these proteins comprise the ancestral prokaryotic machinery from which the NUP transport system evolved.

### 4.3 Comparisons to Extant Approaches

In order to assess the advantage of our approach, we searched for complexes in the data using two other methods: First, we formed a variant of the algorithm that uses only the protein-protein interactions in yeast. That is, this variant searches for heavy subgraphs in the yeast interaction graph, where the edges of the graph are weighted according to the log-likelihood ratio model. Using this variant to search for protein complexes in yeast is aimed at seeing what is gained (and lost) by

| ID | Score | Size | Yeast enrichment | | Bacterial enrichment | |
|---|---|---|---|---|---|---|
| | | | Purity | Complex Category | Purity | Functional Category |
| 1 | 16.16 | 12 (12,10) | 0.17 (1/6) | Translation (1) | 0.56 (5/9) | DNA-metabolism (19) |
| 8 | 3.31 | 6 (6,6) | 1.00 (4/4) | Respiration (4) | 0.33 (2/6) | Energy-metabolism (19) |
| 17 | 141.31 | 12 (6,12) | 0.90 (9/10) | Proteasome (9) | 0.50 (2/4) | Protein-fate (11) |
| 18 | 37.31 | 13 (9,13) | 0.45 (5/11) | Proteasome (9) | 0.25 (2/8) | DNA-metabolism (19) |
| 19 | 19.09 | 6 (6,6) | 1.00 (6/6) | Translation (10) | 0.80 (4/5) | Protein-synthesis (31) |
| 25 | 40.16 | 10 (8,10) | 0.67 (4/6) | Replication (4) | 0.20 (1/5) | DNA-metabolism (19) |
| 28 | 9.39 | 9 (9,9) | 0.60 (3/5) | Translation (10) | 0.50 (4/8) | Protein-synthesis (31) |
| 30 | 383.52 | 20 (12,20) | 0.55 (6/11) | NUP (6) | 0.43 (3/7) | Cell-envelope (7) |
| 31 | 7.21 | 6 (6,6) | 0 | - | 1.00 (4/4) | Protein-synthesis (31) |
| 32 | 3.05 | 7 (6,7) | 0.67 (2/3) | Transcription (3) | 0.25 (1/4) | Transcription (4) |
| 33 | 15.68 | 13 (12,12) | 0.40 (2/5) | RNA-processing (2) | 0.33 (3/9) | DNA-metabolism (19) |

Table 1: Conserved protein complexes identified between yeast and bacteria. For each complex the table lists its score (-ln $p$-value, adjusted for multiple testing); its size (with the numbers of distinct bacterial and yeast proteins in parentheses); purity, as measured using MIPS level 3 categorization of complexes in yeast (with the number of proteins from the most abundant category, and the total number of categorized proteins in the complex, in parentheses); the most abundant category (and its size in parentheses); functional purity, as measured using functional annotation in bacteria; and the most abundant class (with its size among the graph's annotated proteins in parentheses). A zero enrichment for a complex indicates that there is at most one annotated member of the complex. Abbreviations: NUP (nuclear pore complex).

adding the constraint of cross-species conservation. Second, we tested our previous probabilistic model for protein interactions [12]. That latter model is much less involved than the current one: The weight of each vertex in the orthology graph is set to the logarithm of the probability that the member proteins are orthologous. The weight of an edge is set to the logarithm of the probability that it represents a true interaction.

We used three measures to compute the quality of the results. All three quantify the similarity between a given solution and a reference, putatively true, solution. In our case, we used the known complex categories in yeast as the reference solution, since no knowledge on *conserved* complexes is available. The *Jaccard* measure, which is often used in clustering (cf. [11]), uses the notion of mates. Two proteins are called *mates* in a solution if they appear together in at least one complex

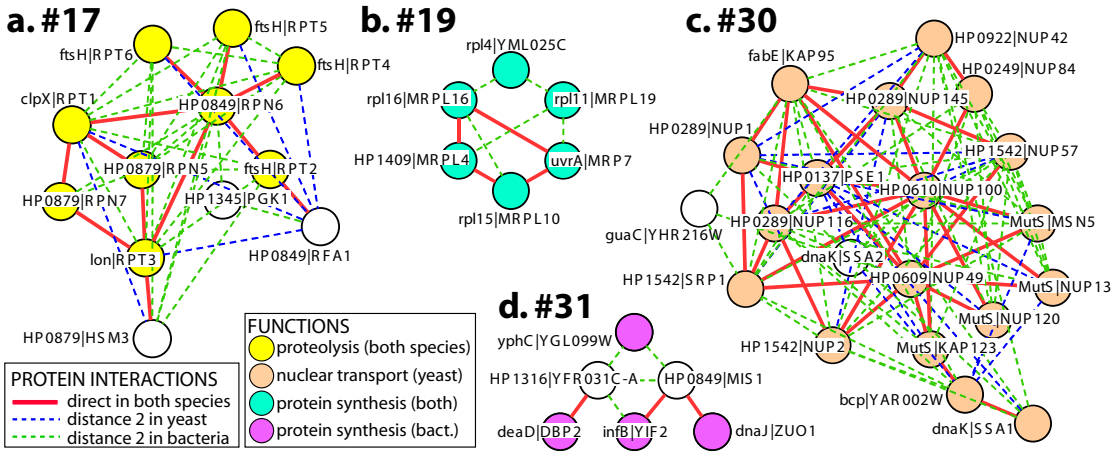


Figure 2: Conserved protein complexes for proteolysis (panel a), protein synthesis (panels b and d), and nuclear transport (panel c). Conserved complexes are connected subgraphs within the bacteria-yeast orthology graph, whose nodes represent orthologous protein pairs and edges represent conserved protein interactions of three types: Direct interactions in both species (red edges); direct in bacteria but distance 2 in the yeast interaction graph (blue edges); and distance 2 in the bacterial interaction graph but direct in yeast (green edges). In the algorithm, both nodes and edges are assigned weights according to the probabilistic model. The number of each complex indicates the corresponding complex ID listed in Table 1.

in that solution. Given two solutions, let $n_{11}$ be the number of pairs that are mates in both, and let $n_{10}$ ($n_{10}$) be the number of pairs that are mates in the first (second) only. The Jaccard score is $n_{11}/(n_{11}+n_{10}+n_{01})$. Hence, it measures the correspondence between protein pairs that belong to a common complex according to one or both solutions. Two identical solutions would get a score of 1, and the higher the score the better the correspondence. The *sensitivity* measure quantifies the extent to which a solution captures complexes from the different yeast categories. It is formally defined as the number of categories for which there was a complex with at least half its annotated elements being members of that category, divided by the number of categories with at least three annotated proteins. The *specificity* measure quantifies the accuracy of the solution. Formally, it is the fraction of predicted complexes whose purity exceeded 0.5.

A comparison of the performance the three approaches is presented in Table 2. The Jaccard score is significantly better in our current approach than in [12]. The sensitivity is lower, as

we capture fewer categories, but the specificity is much higher, so our predicted complexes are much more accurate. Interestingly, when applying our algorithm using only data on yeast we get even higher sensitivity, although again at the cost of specificity. The Jaccard score of this run is comparable to that of the comparative algorithm. This shows that our new probabilistic model can be effectively used, even for detecting complexes using interaction data from a single species. Note that we evaluated the results using data on yeast complexes only, not all of which are expected to be conserved. Still, the use of the bacterial data significantly improved the specificity of the results.

| Algorithm | Jaccard | Sensitivity | Specificity |
|---|---|---|---|
| This study | 0.32 | 0.33 | 0.7 |
| Kelley et al. [12] | 0.22 | 0.44 | 0.4 |
| Yeast only | 0.33 | 0.67 | 0.48 |

Table 2: Performance comparison of three algorithms for complex detection.

## Conclusions

We have presented a novel probabilistic model for the detection of conserved complexes among two species, and an algorithm to search for significant complexes. We applied our approach to study the conservation between yeast and bacterial protein interaction networks. We identified highly specific complexes that were validated using known complexes in yeast and functional annotation in bacteria. Although the present work has already revealed several conserved biological structures that may have functional significance, many refinements and extensions to our method should be explored. Our model can be readily extended to allow interactions between two domains of the same protein (manifested as self-loops in the interaction graph). Models in which the primitive elements are domains within proteins, rather than entire proteins, may be of value. We have used a dense subgraph model that tends to find clique-like patterns of interaction; variations of the model oriented towards other kinds of interaction patterns are also of interest. Protocols such as the two-hybrid system detect directed interactions between proteins, suggesting the use of a directed or mixed interaction graph instead of the current undirected model. In order to find complexes conserved in $k$ species, where $k > 2$, our models should be extended to $k$-species orthology graphs, in which each node specifies a protein from each of $k$ species; the scoring of such nodes is an open question. Negative data, indicating the absence of protein-protein interactions, should be used

to supplement the positive data presently used. Co-expression of genes, as measured in microarray experiments, can also provide indirect evidence for the interaction between the corresponding proteins. Finally, the current protein-protein interaction data are sparse and unreliable; as the abundance and quality of the data improve, the predictive power of our methods and their future refinements will be greatly enhanced.

## Acknowledgments

R. Sharan was supported by a Fulbright grant. T. Ideker was supported by NCRR grant 1P41RR018627-01. R. M. Karp and R. Shamir were supported in part by a grant from the US-Israel Binational Science Foundation (B.S.F.). This research was supported in part by NSF ITR Grant CCR-0121555.

## References

[1] The MIPS database. http://mips.gsf.de/.

[2] The TIGR database. http://www.tigr.org/.

[3] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997.

[4] G.D. Bader and C.W.V. Hogue. Analyzing yeast protein-protein data obtained from different sources. *Nature Biotechnology*, 20:991–997, 2002.

[5] S. Batzoglou, L. Pachter, J.P. Mesirov, et al. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, 10(7):950–958, 2000.

[6] S.J. Boulton et al. Combined functional genomic maps of the C. elegans DNA damage response. *Science*, 295:127–131, 2001.

[7] M. Deng, F. Sun, and T. Chen. Assessment of the reliability of protein-protein interactions and protein function prediction. In *Eighth Pacific Symposium on Biocomputing*, pages 140–151, 2003.

[8] L. Giot et al. A protein interaction map of Drosophila melanogaster. *Science*, 302(5651):1727–1736, 2003.

[9] Y. Ho et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415:180–3, 2002.

[10] T. Ito, T. Chiba, and M. Yoshida. Exploring the protein interactome using comprehensive two-hybrid projects. *Trends Biotechnol.*, 19:S23–S27, 2001.

[11] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley and Sons, Inc., 1990.

[12] B.P. Kelley, R. Sharan, R.M. Karp, et al. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl. Acad. Sci. USA*, 100:11394–11399, 2003.

[13] S. Li et al. A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657):540–543, 2004.

[14] G.G. Loots, I. Ovcharenko, L. Pachter, et al. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.*, 12(5):832–9, 2002.

[15] M. Mann, R.C. Hendrickson, and A. Pandey. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem*, 70:437–473, 2001.

[16] L.R. Matthews et al. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or 'interologs'. *Genome Res.*, 11:2120–6, 2001.

[17] H. Ogata, W. Fujibuchi, S. Goto, and M. Kanehisa. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res.*, 28:4021–8, 2000.

[18] M. Pellegrini et al. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, 96:4285–8, 1999.

[19] J.C. Rain et al. The protein-protein interaction map of Helicobacter pylori. *Nature*, 409:211–215, 2001.

[20] R. Shamir, R. Sharan, and D. Tsur. Cluster graph modification problems. In *Proceedings of the 27th International Workshop Graph-Theoretic Concepts in Computer Science (WG)*, pages 379–390, 2002.

[21] R. Sharan, T. Ideker, B. P. Kelley, R. Shamir, and R. M. Karp. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. In *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology*, pages 282–289, 2004.

[22] R. Sharan, I. Ovcharenko, A. Ben-Hur, and R.M. Karp. CREME: A framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, 19(Suppl 1):I283–I291, 2003.

[23] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. USA*, 101(9):2981–6, 2004.

[24] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18, Supplement 1:136–144, 2002.

[25] R.L. Tatusov, M.Y. Galperin, A. Darren, A. Natale, and E.V. Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):3336, 2000.

[26] Y. Tohsato, H. Matsuda, and A. Hashimoto. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In *Proc. of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 376–383, 2000.

[27] P. Uetz et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–7, 2000.

[28] C.H. Wu et al. The protein information resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res.*, 30:35–37, 2002.

[29] I. Xenarios et al. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, 30:303–305, 2002.