

# **Annotation transfer for genomics: assessing the transferability of protein-protein and protein-DNA interactions between organisms**

Haiyuan Yu\*, Nicholas M Luscombe\*, Hao Xin Lu\*, Xiaowei Zhu\*, Yu Xia\*, Jing-Dong J. Han§, Nicolas Bertin§, Sambath Chung\*, Chern-Sing Goh\*, Marc Vidal§ and Mark Gerstein\*¶

\*Department of Molecular Biophysics and Biochemistry  
266 Whitney Avenue, Yale University  
PO Box 208114, New Haven, CT 06520, USA

§Dana-Farber Cancer Institute and Department of Genetics  
Harvard Medical School  
Boston 02115, Massachusetts, USA

¶ To whom correspondence should be addressed.  
Tel: +1 203 432 6105; Fax: +1 360 838 7861;  
Email: [Mark.Gerstein@yale.edu](mailto:Mark.Gerstein@yale.edu)

haiyuan.yu@yale.edu  
nicholas.luscombe@yale.edu  
haoxin.lu@yale.edu  
xiaowei.zhu@yale.edu  
yuxia@bioinfo.mbb.yale.edu  
jackie.han@research.dfc.harvard.edu  
nicolas\_bertin@dfci.harvard.edu  
sambath@bioinfo.mbb.yale.edu  
chernsing.goh@yale.edu  
Marc\_Vidal@dfci.harvard.edu  
mark.gerstein@yale.edu

# Abstract

Proteins function mainly through physical interactions, especially with DNA and other proteins. Large-scale networks of both types of interactions are now available for a number of model organisms, but the experimental generation of these networks is still difficult. Therefore, interolog mapping - the transfer of interaction information from one organism to another using comparative genomics - is of significant value. Here we quantitatively assess the degree to which interologs can be reliably transferred between species as a function of sequence similarity of interacting proteins. Using interaction information from *S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *H. pylori*, we find that protein-protein interactions can be reliably transferred when a pair of proteins has a joint sequence identity greater than 50% or a joint E-value smaller than  $10^{-120}$ . (These "joint" values are the geometric mean of the identities or E-values for the two pairs of interacting proteins.) We generalize our interolog analysis to protein-DNA binding, and find that such interactions are conserved at specific thresholds between 30% and 60% sequence identity depending on the protein family. Furthermore, we introduce the concept of a "regulog" -- a conserved regulatory relationship between proteins across different species. We map interologs and regulogs from yeast to a number of genomes for which there are limited experimental data (e.g. *A. thaliana*) and make these available through an on-line database at <http://genecensus.org/interactions/interolog/>. Specifically, we are able to transfer about 90,000 potential protein-protein interactions to worm. We test a number of these in large-scale two-hybrid experiments. We are able to verify 45 overlaps, which we show to be statistically significant.

Supplementary materials are attached at the end.

# Introduction

The ultimate goal of functional genomics is to determine the functions of all gene products in newly sequenced genomes. Unfortunately, while there is a deluge of sequence data available, only a small fraction has been functionally characterized (Andrade and Sander 1997). Nevertheless, for some genomes belonging to experimentally tractable model organisms, such as *S. cerevisiae*, *C. elegans*, and *H. pylori*, scientists have elucidated the functions of many of their gene products. Given the quantity of sequence and structural data available, a major method for assigning functions is to transfer the existing annotation of a known gene to the newly sequenced gene product. This is based on the concept that sequence and structural similarities between gene products suggest functional similarities (Bork et al. 1998; Bork et al. 1994; Fraser et al. 1995; Fraser et al. 1998; Hegyi and Gerstein 2001; Wilson et al. 2000).

The transfer of structural annotations is well characterized. It has been shown that structural similarity [measured as Root Means Square (RMS) of matching C $_{\alpha}$  backbone atoms] between two proteins decreases exponentially with increased sequence divergence (measured as percent identity) (Chothia and Lesk 1986; Chothia and Lesk 1987). Thus, the reliability of a homology-based structural annotation depends on the level of sequence similarity between homologous proteins.

Several groups have recently examined the dependency of functional similarity on sequence and structural similarity (Bork et al. 1998; Bork et al. 1994; Marcotte et al. 1999). The best matching sequences in a database search are often used as basis for initial annotations (Fraser et al. 1995; Fraser et al. 1998). However, further work has provided potential for more robust annotation transfer, including analyzing patterns of protein family occurrence in different phylogenetic groups (Pellegrini et al. 1999) and associating key sequence motifs with particular functions (Attwood et al. 1997; Bairoch et al. 1996). Other work has also shown that, in general, protein function is conserved for sequence identities down to 40% for single-domain proteins that share the same structural fold; however, for multi-domain proteins, the pattern of functional conservation is more complex: proteins are most likely to share functions if they contain similar domain combinations (Brenner 1999; Hegyi and Gerstein 2001; Wilson et al. 2000).

It is difficult to evaluate the relationship between sequence homology and function, because no clear measure of functional similarity exists between any two proteins, and the definition of 'function' itself is often vague (Bork et al. 1998; Lan et al. 2002; Lan et al. 2003; Wilson et al. 2000). Previous studies, based on hierarchical classification systems, such as ENZYME (Webb 1992), MIPS (Mewes et al. 2000) and GO (Ashburner et al. 2000), determine functional similarity by comparing both proteins' respective levels in the hierarchy. This is a rough definition underlying the difficulties inherent in the earlier work. However, an important aspect of protein function is the physical interactions of proteins with other molecules, in particular, with other proteins or with DNA. No previous work has addressed this issue. With recent genome-wide studies on protein-

protein and protein-DNA interactions (Gavin et al. 2002; Ho et al. 2002; Horak et al. 2002; Ito et al. 2000; Iyer et al. 2001; Lee et al. 2002; Uetz et al. 2000), it is now possible to examine the degree to which protein-protein and protein-DNA interactions are transferred between different organisms as a function of the underlying sequence similarities of the interacting proteins.

To this end, Walhout et al. (2000) introduced the concept of “interologs”: orthologous pairs of interacting proteins in different organisms. In this study, we extend and assess this concept in detail. We present a large-scale quantitative assessment on conservation of protein-protein and protein-DNA interactions between proteins and organisms. Compared to the previous survey, our investigation has greater statistical weight and precision. In our calculations, we use almost all available genome-wide interaction datasets from four model organisms (14,911 interactions total). Moreover, we generalize the interolog concept and propose that there are at least two kinds of interologs: protein-protein interologs and protein-DNA interologs. Based on the latter idea, we also introduce a new concept, “regulog”. Furthermore, we calibrate the ability of interologs to reliably map interactions across different organisms. Combining our interolog and regulog mapping with available large-scale interaction data for yeast, we construct genome-wide interaction maps and regulatory networks for several organisms.

## Results and Discussion

### Definitions and formalism for protein-protein interologs

#### *Homologs and Orthologs*

*Homologs* are proteins with significant sequence similarity. Operationally, this can be defined as having a E-value  $\leq 10^{-10}$  from BLASTP (Altschul et al. 1990). This is a similar cutoff to that used previously (Matthews et al. 2001).

*Orthologs* are proteins in different species that evolved from a common ancestor “by speciation” (Tatusov et al. 1997). Orthologous proteins in different organisms usually have the same functions. Operationally, the ortholog of a protein is usually defined as its best-matching homolog in another organism. Here we define orthologs as:

- (i) candidates with a significant BLASTP E-value ( $\leq 10^{-10}$ );
- (ii) having  $\geq 80\%$  residues in both sequences included in the BLASTP alignment;
- (iii) having one candidate as the best-matching homolog of the other candidate in the corresponding organism;
- (iv) (i), (ii), and (iii) must be true reciprocally.

It is obvious that this operational definition of ortholog by sequence homology is not perfect. Actually, orthologs are not always determined as the best-matching homologs (Tatusov et al. 1997).

## *Interologs*

Based on Walhout et al. (2000), if interacting proteins A and B in one organism have interacting orthologs A' and B' in another species, the pair of interactions A-B and A'-B' are called *interologs* (see Figure 1A).

## *Joint sequence similarity*

A goal of this work is to measure the transferability of interactions based on sequence similarity. In the case of protein-protein interactions, sequence similarities to homologs of both interacting partners are important. We therefore use joint sequence similarity ( $J$ ) between protein pairs. There are many potential ways to define joint sequence similarity, but our results show that different definitions of  $J$  do not matter much. Here, we use two major definitions of  $J$ :

### 1. Joint sequence identity ( $J_I$ ) as the geometric mean of individual percent identities

Percent identity is routinely used to measure the sequence similarity between proteins. Therefore, joint similarity is first defined as the geometric mean of individual percent identities:

$$J_I = \sqrt{I_A \times I_B}$$

Given that protein A is known to bind to protein B,  $I_A$  represents the individual sequence identity of protein A and its homolog. Likewise,  $I_B$  is the individual sequence identity of protein B and its corresponding homolog. We calculate individual sequence identities based on the sequence alignment using the Smith-Waterman algorithm in FASTA (Pearson and Lipman 1988).

### 2. Joint E-value ( $J_E$ ) as the geometric mean of individual E-values

Measuring homology by percent identity has certain disadvantages (Wilson et al. 2000). For instance, the length of the matching sequences is not considered. Naturally, the shorter the sequence is, the higher the chance of randomly finding similar sequences. Furthermore, it has become more common to use statistical scoring schemes, especially E-values in BLAST, to measure the statistical significance of the homology in order to determine the orthologs across organisms (Brenner et al. 1998; Tatusov et al. 1997). Therefore, we also calculate the joint similarity as a joint E-value, i.e. the geometric mean of the individual E-values:

$$J_E = \sqrt{E_A \times E_B}$$

$E_A$  represents the BLASTP E-value of protein A and its homolog.  $E_B$  is the individual BLASTP E-value of protein B and its homolog.

### 3. Joint similarity as the minimal individual similarity

Calculating the joint similarity using the geometric mean of the individual similarities places equal weight on each of the two similarities. However, the joint similarity could also be defined as the smaller of the two individual similarities:

$$J_{AB} = \min (S_A, S_B)$$

$S_A$  and  $S_B$  represent the individual similarities, respectively, of protein A and its homolog and of protein B and its homolog. In this manner,  $J_{AB}$  measures the minimal similarity level necessary for the reliable transfer of interaction information between protein pairs. Individual similarities can also be determined as percent identities by FASTA or E-values by BLASTP.

### *Source and target organisms*

In the *source organism*, there is a set of known interactions. The *target organism* is a fully-sequenced organism, onto which the known interactions in the source organism are mapped (as described below) based on sequence similarities (see Figure 1C).

### *Interolog mapping*

*Interolog mapping* is a process that maps interactions in the source organism onto the target organism to find possible interactions (i.e. interologs) in that organism (see Figure 1A). To assess the performance of mapping methods, one can use known interacting and non-interacting protein pairs (positives and negatives) in the target organism as benchmarks.

#### 1. Original interolog mapping method: best-match mapping

Previously, Matthews et al. (2001) proposed a best-match mapping method to transfer yeast interactions onto the worm proteome. Simply put, their method selects all best-matching homologs between two organisms (E-value  $< 10^{-10}$ ). In worm, all pairs of best-matching homologs of interacting yeast proteins are considered as potential interologs. Using two-hybrid systems, they tested 216 worm protein pairs and 72 yeast protein pairs. Their results showed that only 16% to 32% of interologs predicted experimentally determined interactions correctly.

#### 2. A new method: reciprocal best-match mapping

A more stringent derivative of this original method would be to use only the reciprocal best-matches in mapping interologs between organisms. In this paper, we present results from both approaches.

## Generalized interolog mapping

Both interolog mapping methods, using only the best matches, suffer from low coverage of the total interactome and low prediction accuracy. This will be discussed further in the next section. To address the problem of low coverage, we introduce a new *generalized interolog mapping* method using all possible homologs of interacting proteins. For any given protein in one organism, all of its homologs in another organism are considered as a homolog family (or simply family). Two families of two interacting proteins are called interacting families, that is, at least a member of one family interacts with a member of the other family. All possible protein pairs between the two interacting families are called *generalized interologs* (see Figure 1B). This method has the advantage of sidestepping some of the ambiguities in defining orthologs.

## Gold standard target datasets

### 1. Set of gold standard positives $P$

To assess the performance of interolog mapping, we need a group of known interactions as positives in the target organism. This set is called gold standard positives and denoted by  $P$ . The total number of elements in this set is  $|P|$ .

As the most extensive and reliable interaction datasets exist for *S. cerevisiae*, we use it first as the target organism. In *S. cerevisiae*, the MIPS complex catalogs, which contains 8,250 unique interacting protein pairs, has previously been used as a standard reference for known interactions (Edwards et al. 2002; Jansen et al. 2003; Mewes et al. 2000; von Mering et al. 2002). Therefore, we consider the MIPS interactions as gold standard positives in the next section. In order to compile a reference dataset with the lowest false positive rate, we consider two proteins as interaction partners if and only if they are in the same complex of the highest level in the catalog. At the end of the paper, we reverse this situation and use *S. cerevisiae* as the source organism and map its reliable interaction information (from the complex catalog) onto other eukaryotes (such as *A. thaliana*) to build an interolog database.

It should also be noted that proteins in the same complex do not necessarily interact with each other directly. Here, we use the term “interaction” to signify “complex association”, i.e., two protein subunits may belong to the same quaternary complex but not physically interact. Therefore, the number of complex associations of a protein may be larger than the number of its pair-wise physical associations.

In order to probe the direct physical interactions more closely, we constructed a refined, smaller dataset comprising 1,867 interactions between 1,391 proteins. In parallel to our “gold standard” nomenclature, we call this the “platinum standard” dataset. Briefly, the dataset contains: physical interactions from complex protein structures in the Protein Data Bank (Westbrook et al. 2003), verified interactions from small-scale experiments (Bader et al. 2003; Mewes et al. 2000; Xenarios et al. 2002), and protein pairs from small MIPS catalog complexes ( $\leq 4$  subunits). The dataset and detailed explanation of its

construction is available from our website. The platinum-standard dataset is of equally high quality as the gold standard set, but differs as it describes physical pair-wise interactions between proteins rather than complex associations. As shown below, the two datasets yield very similar results, indicating a good correspondence between physical interactions and complex associations. However, because better statistics are obtained from a larger dataset, we perform the bulk of the analysis in this paper using the gold standard interactions.

## 2. Set of gold standard negatives $N$

We also need a set of negatives (i.e. non-interacting proteins) in the target organism to assess our method. This set is called gold standard negatives and denoted by  $N$ .

Previously, Jansen et al. (2003) considered pairs of proteins in different sub-cellular compartments as good estimates for non-interacting pairs (Kumar et al. 2002). In total, there are 2,708,746 such protein pairs.

However, sometimes not all interolog features could be defined for each of the pairs in the gold standard. In this case, we use alternate sets  $P'$  and  $N'$ , subsets of  $P$  and  $N$  with defined features.

## *Source datasets*

To assess the interolog mapping method, we need source organisms with known interaction data. In this paper, *C. elegans*, *D. melanogaster*, and *H. pylori* are used as source organisms. We then map the interactions in these organisms onto *S. cerevisiae* genome. These are the only three organisms, besides *S. cerevisiae*, in which large-scale interaction datasets are available.

### 1. *C. elegans* interaction dataset

For *C. elegans*, we have a dataset containing 410 high-quality interactions from individual small-scale experiments (Boulton et al. 2002; Davy et al. 2001; Walhout et al. 2000). It comprises the most reliable dataset for all eukaryotes apart from yeast (M. Vidal, personal communication), and we emphasize its use in the calculations here. Because of its small size, however, we cannot always obtain sufficiently good statistics with this dataset. Results of high-throughput two-hybrid experiments have recently been published for *C. elegans* (5,686 interactions) (Li et al. 2004). We use this larger dataset, together with those for *D. melanogaster* and *H. pylori*, to improve the statistical precision in some of our calculations.

### 2. *D. melanogaster* interaction dataset

For *D. melanogaster*, there are 4,786 interaction pairs from two-hybrid experiments (Giot et al. 2003).

### 3. *H. pylori* interaction dataset

For *H. pylori*, there are 1,465 interaction pairs from two-hybrid experiments (Rain et al. 2001).

## **Assessment parameters**

As shown in Figure 1C, based on interactions in the source organisms, all generalized interologs with joint similarities larger than a certain cutoff ( $J$ ) are considered possible interactions in the target organism. We then assess these predictions (thin red solid lines) against gold standard positives (thick black solid lines) and negatives (dashed lines) in the target organism. The assessment parameters are as follows:

### 1. $G(J)$

The set of generalized interologs in the target organism at a certain joint similarity level ( $J$ ) is denoted by  $G(J)$ .

### 2. $T(J)$

The set of the true positives in  $G(J)$  is denoted by  $T(J)$ , i.e.  $T(J) = G(J) \cap P$ . We define the number of true positives at a given  $J$  as  $TP = |T(J)|$ .

### 3. $F(J)$

The set of the false positives in  $G(J)$  is denoted by  $F(J)$ , i.e.  $F(J) = G(J) \cap N$ . We define the number of false positives at a given  $J$  as  $FP = |F(J)|$ .

### 4. $V(J)$

We denote  $V(J)$  as the percentage of verified predictions among generalized interologs at a certain joint similarity level  $J$ , which is calculated as:

$$V(J) = \frac{|T(J)|}{|G(J)|} \times 100\%$$

We also call  $V$  a level of verification (or loosely an accuracy). Please note that  $V$  calculated here may be a lower bound estimate because the MIPS complex catalog is not complete.

### 5. $L(J)$

We denote  $L(J)$  as the likelihood ratio for a generalized interolog, with a certain joint similarity ( $J$ ), to be a true prediction.  $L(J)$  can be calculated by a Bayesian approach. This is a straightforward extension of the formalism described previously (Jansen et al. 2003). If we know the number of positives ( $Np$ ) among the total number of protein pairs ( $Nt$ ), the probability of finding a interacting pair in the genome,  $P(pos)$ , can be defined as  $Np/Nt$ . Therefore, the "prior" odds of finding a positive are:

$$O_{prior} = \frac{P(pos)}{P(neg)} = \frac{P(pos)}{1 - P(pos)}$$

In contrast, the "posterior" odds are the odds of finding a positive given that, in another organism, its generalized interolog with a joint similarity  $J$  is a known interaction:

$$O_{post} = \frac{P(pos | J)}{P(neg | J)}$$

The likelihood ratio  $L$  defined as

$$L(J) = \frac{P(J | pos)}{P(J | neg)} = \frac{\frac{TP}{|P|}}{\frac{FP}{|N|}}$$

relates prior and posterior odds according to Bayes' rule:

$$O_{post} = L(J)O_{prior}$$

As  $O_{prior}$  is fixed for a given organism,  $O_{post}$  is proportional to  $L(J)$ , i.e. the higher the likelihood ratio, the more likely the prediction is true. In a naive Bayesian network where there are no correlations between features, this procedure can be iterated. Specifically,  $O_{post}$  can be multiplied again by another  $L$  for a different feature. In doing so, one could combine many different features within a uniform framework of likelihood ratios. In particular, it would allow us to combine our likelihood ratios from interologs with the other features in Jansen et al. (2003).

## Assessment of interologs on current interaction datasets

### *Conservation of generalized interologs*

#### 1. Relationships between $V$ and $J$

To measure the conservation of interactions between homologous protein pairs, we assessed the chance ( $V$ ) that two proteins interact with each other as a function of their joint sequence identities ( $J_I$ ) with other known interacting pairs. First, we mapped only high-quality worm interactions onto the yeast genome. As there are not many data points, we grouped all the generalized interologs into three bins based on their joint identities: low, medium and high. Figure 2A shows a clear monotonic relationship between  $V$  and

$J_I$ . This confirms that the higher the joint identity, the more likely the predicted interolog is true.

To get better statistics, we mapped all interactions in *C. elegans*, *D. melanogaster*, and *H. pylori* onto the *S. cerevisiae* genome, assessing them against our gold standards described above. In Figure 2B, the relationship between  $V$  and  $J_I$  is the weighted average (based on the total number of true positives in each dataset) of the relationships in all four mapping processes. The plot exhibits a sigmoidal relationship with a sharp decrease around 50%  $J_I$ . This indicates that all protein pairs having  $J_I \geq 50\%$  with a known interacting pair will interact with each other; whereas few pairs interact at  $J_I < 30\%$ . These results confirm that pairs of proteins with sufficient sequence similarity tend to share the annotation of protein-protein interactions.

Furthermore, we performed a similar analysis using joint E-values ( $J_E$ ). Figure 2C shows the same monotonic relationship, as that in Figure 2A, when we mapped the high-quality worm interactions onto yeast genome. In Figure 2D, the weighted average curve also has a sigmoidal characteristic. Overall, more than half of the protein pairs with  $J_E \leq 10^{-120}$  indeed bind to each other. Therefore,  $J_E$  of  $10^{-120}$  could be used as a good threshold to reliably transfer the annotation of interactions.

## 2. Relationships between $L$ and $J$

The above approach (i.e. assessing the transferability of a property between organisms by calculating the fraction sharing the property with certain similarity) has been generally used for similar purposes (Hegyí and Gerstein 2001; Wilson et al. 2000). Here, we apply a Bayesian network approach to further evaluate the transferability of interactions. Likelihood ratios ( $L$ ) are more directly related to probabilities and are, therefore, more quantitative and precise in describing the transferability of the interactions.

As we did for  $V$  above, we calculated the relationships between  $L$  and  $J_E$  for two mappings: worm-to-yeast and a weighted average of all three organisms to yeast (Figure 2E and F, respectively). Both figures exhibit positive relationships between  $L$  and  $J_E$ , suggesting that the better the joint E-values, the higher the likelihood ratios. This further confirms the relationships found in Figure 2A-D and the validity of using joint similarities.

Conservatively, the total number of interactions in yeast genome is approximately 30,000 (Kumar and Snyder 2002). Given that there are approximately 18 million yeast protein pairs in total, the prior odds ( $O_{prior}$ ) would be roughly 1/600. Therefore, only protein pairs with  $L > 600$  would have a greater than 50% chance of interaction. As shown in Figure 2F, protein pairs with  $J_E < 10^{-100}$  have  $L > 600$ . The  $J_E$  threshold ( $10^{-120}$ ), determined previously, easily satisfies this criterion. If we were to use  $L$  to perform the mapping methods, cross-validation could be applied in choosing the optimal  $L$  cutoff as described previously (Jansen et al. 2003).

We examine the correspondence between direct, physical interactions and complex associations, by repeating the calculations for Figures 2B, D, and F using the platinum standard dataset. The results show similar trends to the gold standard dataset, indicating the high correspondence between the two datasets. Due to its smaller size, the statistics for the platinum-standard dataset is not as good as the gold standard. Owing to the similarity of results, and better statistics, we therefore use the MIPS complex catalog as the main reference dataset in this paper.

### 3. Results of $J$ as the minimal sequence similarity remain the same

As discussed above, we could also use the minimal individual similarity instead of the geometric mean to calculate  $J$ . We repeated all calculations in Figure 2 using this new definition of  $J$ . The results show that the new definition has little effect (supplementary Figure 2). Therefore, for the remaining discussion  $J$  is defined as the geometric mean of the individual E-values (i.e.  $J_E$ ).

### *Comparison of different interolog mapping methods*

In order to compare different mapping methods, *C. elegans* was used as the source organism and its interactions were mapped onto *S. cerevisiae* genome by three different mapping methods as discussed above. We compared the predicted interologs produced by the different methods above against the gold standard positives and negatives. The results are as follows:

#### 1. Best-match mapping method

From 410 interacting pairs in worm, we found 84 corresponding interolog candidates in yeast. Only 25 of these pairs overlapped with gold standard positives, corresponding to  $V \approx 30\%$  (i.e. loosely 30% accuracy). This agrees with previous results (Matthews et al. 2001).

#### 2. Reciprocal best-match mapping method

In total, we determined 33 interolog candidates based on the 410 worm interactions, among which 18 pairs (54%) were true positives.

#### 3. Generalized interolog mapping method

Based on the 410 interacting pairs, we found 92 pairs of interacting families in yeast, 91 of which contain at least one true interaction. In total, we predicted 9,317 interactions (i.e. generalized interologs), among which 162 pairs (2%) are true positives. In Figure 3, it is evident that the fraction of true positives clearly increases as  $J_E$  decreases. When only the top 5% pairs with the best  $J_E$  values are selected,  $V$  increases to 31% (35 true positives out of 112 predictions), resulting in even better accuracy than that of best-match mapping method (30%).

Previously, four large-scale experimental interaction datasets in yeast have been combined into a “PIE” (i.e. *Probabilistic Interactome Experimental*), in which each interaction is associated with a particular  $L$  (Jansen et al. 2003). To assess the performance of our method in relation to known standards, we compared our results against the PIE. We show our comparison as a  $TP/|P'|$  vs.  $TP/FP$  graph, a close analogue of the conventional ROC curve. As shown in Figure 4, the coverage and accuracy of interolog mapping are roughly comparable to those of the large-scale experiments.

### *Examples of protein-protein interologs*

The Ste5-MAPK complex is a key 6-subunit complex in yeast mating-pheromone response pathway (Posas et al. 1998). The interaction partners of worm MAPK (F43C1.2a) were determined experimentally (see supplementary table 1). In total, there are 26 known partners for F43C1.2a, none of which is involved in this MAPK signal transduction pathway. However, using the generalized interolog mapping method, we successfully predicted 5 of the 6 subunits in yeast based on only one MAP kinase in worm. This illustrates the power and utility of our method (see supplementary materials).

## **Definitions and formalism for protein-DNA interologs and regulogs**

### *Protein-DNA interologs and mapping*

If transcription factor (TF)  $A$  with binding site  $S_A$  has, in another species, an ortholog  $A'$  with binding site  $S_{A'}$  of identical DNA sequence,  $A'-S_{A'}$  is a *protein-DNA interolog* of  $A-S_A$  (see Figure 5).

We can extend protein-protein interolog mapping to protein-DNA interolog mapping. In this process, we transfer the DNA-binding information of a given TF  $A$  to its ortholog  $A'$  as a function of the sequence similarity between  $A$  and  $A'$ .

### *Regulogs*

TFs bind to DNA to regulate the expression of downstream genes. Therefore, there is a regulatory relationship between a given TF and its target. Suppose that TF  $A$  and its target  $B$  in one organism have orthologs  $A'$  and  $B'$ , respectively, in another organism. Furthermore, suppose that in the second organism,  $A'$  is also a TF regulating  $B'$ , then, we call  $A'=>B'$  a *regulog* of  $A=>B$ .

### *Source datasets*

For practical calculations, we used TF families as described previously (Luscombe and Thornton 2002). Target binding sequences of individual factors were obtained from the TRANSFAC database (Wingender et al. 2001). All known protein-DNA interactions are considered as positives. We do not have negative datasets for protein-DNA interologs and regulogs.

### ***Assessment parameters***

The parameters involved in assessing the conservation of protein-DNA interologs are analogous to those for protein-protein interologs. They are given as follows:

1.  **$G(I)$**

The set of predicted protein-DNA interologs with the sequence identities between TFs larger than a certain cutoff ( $I$ ) is denoted by  $G(I)$ .

2.  **$T(I)$**

The set of the transcription factor pairs that share the same DNA binding sites in  $G(I)$  is denoted by  $T(I)$ .

3.  **$V(I)$**

We denote  $V(I)$  as the percentage of verified predictions among the predicted protein-DNA interologs at a certain sequence identity level,  $I$ . This is calculated as:

$$V(I) = \frac{|T(I)|}{|G(I)|} \times 100\%$$

We calculate  $V$ 's both for TFs within each family separately and for all TFs together (see Figure 5). Due to the relatively small amount of TF binding data, we aggregate all of our predictions. This procedure is described in the supplementary materials.

## **Assessment of protein-DNA interologs and regulogs**

### ***Conservation of protein-DNA interologs***

As shown in Figure 6, the relationship between  $V$  and  $I$  is sigmoidal, with a sharp decrease in target site conservation between 30% to 60% sequence identity. This indicates that all TFs within a certain range of identities invariably share the same target sequence. The specific threshold for the identities is highly family dependent, ranging from 30 to 60%. The hormone receptor and LacI repressor families have a higher threshold of about 60% whereas the other families diverge at lower thresholds of 30%. The C<sub>2</sub>H<sub>2</sub>-zinc finger family is an exception and sequence recognition is barely conserved even for close homologs (threshold identity 80%). The main reason for this is that the binding domains of C<sub>2</sub>H<sub>2</sub>-zinc fingers are often very short (~30-90 amino acids in length) and, therefore, only a few mutations are required to alter its specificity.

The fact that TF families have different thresholds reflects the regulatory diversity of different families. Families with high thresholds contain factors that regulate many different processes; while those with low thresholds regulate only a few different processes (Luscombe and Thornton 2002).

We further assessed the general transferability of protein-DNA binding properties between homologous protein sequences, by calculating the relationship between  $V$  and  $I$  for all TFs. As shown in Figure 6, approximately 60% of homologous TFs share the same binding sites at 30% sequence identity; at 50% sequence identity, 80% of TFs share the same binding sites. Therefore, if two proteins have  $\geq 30\%$  sequence identity, they can be predicted to share the same binding sites. The confidence level of the prediction is shown as a function of sequence identity in Figure 6.

### *Protein-DNA interolog (regulog) mapping method*

When a protein-DNA interaction is transferred across species, the regulatory relationship between the TF and its target is also implicitly transferred. Based on our calculations, at least three conditions are necessary for regulogs to be transferred (see Figure 5):

- (i) TF A and its homolog A' must have  $\geq 30\%$  sequence identity. (Note that formally A and A' should be orthologs. However, practically this is defined here by this sequence similarity criterion.)
- (ii) Target gene B and its homolog B' must be orthologs;
- (iii) The DNA sequence upstream of B' must contain the same binding site as that of B;

Unfortunately, we only have large-scale transcriptional regulatory networks in *S. cerevisiae* for eukaryotes and in *E. coli* for prokaryotes. Because the transcription machinery differs radically between eukaryotes and prokaryotes, the performance of our regulog mapping method cannot currently be assessed on a large-scale. However, we would like to discuss one specific example of regulogs between *S. cerevisiae* and *D. melanogaster* to illustrate the process of regulog mapping and its underlying logic.

In *S. cerevisiae*, Cyc1 is a mitochondrial protein with electron-transport function. The Hap2-Hap3 heteromeric TF complex binds to the UAS2 activation sequence (GTTGG) upstream of *CYC1* and then activates transcription of this gene (Hahn and Guarente 1988; Olesen et al. 1987). Using the above-mentioned three conditions, we define potential regulogs in *D. melanogaster*:

- (i) CG10447 (a TF) and CG17618 (function unknown) are fly homologs of yeast proteins Hap2 and Hap3 with 30% and 40% sequence identities, respectively;
- (ii) CG17903 (CD4) is a fly ortholog of Cyc1. It shows electron-transport activities and is located in the mitochondria (Limbach and Wu 1985);
- (iii) the same UAS2 activation sequence (GTTGG) is also found in the promoter regions of CG17903 at the appropriate position ( $\sim -200\text{bp}$ );

Based on the above, we predict that CG10447 and CG17618 may also regulate the expression of CG17903. This regulatory relationship is the fly regulog of its counterpart involving the yeast proteins Hap2-Hap3, and *CYC1*. Elucidating this allows us to predict the function of an un-annotated fly protein, CG17618. Furthermore, the interactions between the two fly TFs and the UAS2 DNA sequence is the fly protein-DNA interologs of those between Hap2, Hap3, and the UAS2 sequence. More interestingly, because Hap2 and Hap3 interact with each other, their fly homologs CG10447 and CG17618 may also interact. This fly interaction is a potential protein-protein interolog of that between Hap2 and Hap3.

## Database of interologs and regulogs

Finally, having proven the feasibility of the generalized interolog mapping method, we applied this method on the MIPS complex dataset in yeast to predict protein-protein interactions in several other important eukaryotic organisms, including *C. elegans*, *C. albicans*, *D. melanogaster*, and *A. thaliana*. In each organism, the top 1% of predicted generalized interologs with the best  $J_E$ 's are considered as highly reliable interologs. Simple statistics relating to the interolog database are shown in Table 1.

To assess the accuracy of our database, we compared our predicted worm interactions against those from independent and on-going large-scale worm two-hybrid experiments. A total of 3,730 interaction pairs were generated. Because only one splicing form was used for each gene in these experiments, we removed all alternative splicing forms and our prediction of yeast-to-worm interologs decreased from 91,224 (in Table 1) to 55,223 pairs. Among these, 45 pairs were confirmed experimentally. We employ a hypergeometric model (see supplementary material) to evaluate the significance of this overlap. The calculated P value is smaller than  $10^{-10}$ . The P value is the probability of finding a certain overlap between two independent datasets by chance within the whole worm interactome. Therefore, the experimental results support and validate our predictions.

More interestingly, the experimentally-determined interaction pairs can be further divided into different groups involved in different pathways, e.g. 26S proteasome (Davy et al. 2001), DNA-damage repair (DDR) (Boulton et al. 2002), and vulval development (Walhout et al. 2000). The overlaps between these groups and our predictions vary considerably, as shown in Figure 7. For groups known to be well conserved in eukaryotes, such as proteasome and DDR (Davy et al. 2001; Larsen and Finley 1997), the overlaps are much better than those that are not. The non-significant P value for the group "others" is also attributable to the fact that the baits in this group are specially selected to ensure they have no yeast homologs. Thus, Figure 7 further confirms the biological relevance of our database.

We also applied our regulog mapping method to yeast transcriptional regulation datasets (Horak et al. 2002; Lee et al. 2002; Wingender et al. 2001). The results suggest potential regulatory networks in other eukaryotic organisms. Due to variable TF-binding sites and insufficient information on binding sequences, we transferred the yeast regulatory

networks using only the first two conditions, i.e. sequence homology for both TFs and targets. In general, distant organisms share smaller sets of TFs and targets. Using *D. melanogaster* as an example, our regulog method determined 33 TFs, 621 targets, and 2,936 regulatory connections (see Table 1). If the requirement of having the same binding sites is included, we were only able to determine 29 connections between 13 TFs and 5 target genes.

The results of the interolog and regulog mapping are recorded in an interolog/regulog database at <http://genecensus.org/interactions/interolog/> (see Figure 8). To find possible physical or regulatory interaction partners of one's favorite protein, the user simply inputs the names of the organism and the protein. For the protein-protein interolog database, all predicted interaction partners will be shown and ranked by  $J_E$ . Our database also links each protein to an external web resource such as SGD (Christie et al. 2004), WormBase (Harris et al. 2004) and FlyBase (The FlyBase Consortium 2002). For the regulog database, all predicted TFs and their targets are ranked by sequence homologies between query TFs and their yeast homologs. The layout of the webpage is similar to that of the interolog database.

## Conclusion

In this study, we comprehensively assessed the transferability of protein-protein and protein-DNA interactions by analyzing the relationships between sequence similarity and interaction conservation. A total of 14,911 interactions in four organisms are included in our investigation. In general, the conservation of both interaction types shows a sigmoidal relationship with sequence similarity. For these four organisms, protein-protein interactions are well conserved between protein pairs with at least 50%  $J_I$  or  $10^{-120} J_E$ . For protein-DNA interactions, the specific threshold of sequence identity is highly family-dependent. In general, 60% of TFs with 30% or more sequence identity share the same target sites.

Previously, Walhout et al. (2002) proposed an “interolog” concept to transfer protein-protein interactions across species. Here, we develop this concept into a concrete interaction prediction approach, the generalized interolog mapping method. This is readily expandable to any newly completed genomes. Using generalized interolog mapping method, we construct several genome-wide protein-protein interaction maps.

We further introduce a new “regulog” concept to map regulatory relationships between TFs and their targets across organisms. We apply the regulog mapping to produce genome-wide regulatory networks for several eukaryotic organisms. The results of the newly produced interaction maps and regulatory networks are stored in an interolog/regulog database.

## Future directions

There are a number of directions to extend this work. With respect to the conservation of protein-protein interactions, there are many more sequenced genomes without known genome-wide interaction networks. We will apply our method to these genomes to gain insight into their protein-protein interactions, and eventually to shed light on their functions. However, our analysis is still hampered by not having sufficient interaction data for other organisms. Once such large-scale interaction datasets are available, we can repeat our calculations taking into consideration the new information, which will give results with better statistical precision. For the regulog mapping method, we are unable to evaluate its performance at this time. When genome-wide regulatory networks are created in other organisms, we will evaluate the feasibility and accuracy of the regulog mapping method in a similar fashion to that of the protein-protein interolog mapping method.

# **Acknowledgment**

The authors would like to thank the referees for insightful comments that helped improve the manuscript. MG acknowledges support from the NIH grant 5P50GM062413.

## Figure Captions:

Figure 1. Schematic illustration of protein-protein interologs and the mapping methods. (A) Original interolog mapping. Theoretically, A-A' and B-B' should be orthologs between the two organisms. Operationally, only best-matching homologs are required. (B) Generalized interolog mapping. Proteins A<sub>1</sub>', A<sub>2</sub>', A<sub>3</sub>', and A<sub>4</sub>' in the target organism are all homologs of protein A in the source organism. These proteins form A' family. Likewise, protein B's homologs (B<sub>1</sub>', B<sub>2</sub>', B<sub>3</sub>') form B' family in the target organism. If we know that protein A interacts with B, we can predict that A' family and B' family are interacting families. All possible pairs between these two families are considered as the generalized interologs (shown as black dashed lines with arrows). (C) Comparison with the gold standards. After the interactions in the source organism are mapped onto the target organism, the predictions (i.e. generalized interologs) are compared with the gold standard positives and negatives. True positives are the predictions that overlap with the gold standard positives. False positives are those that overlap with the gold standard negatives.

Figure 2. Conservation of protein-protein interactions between homologous protein pairs. (A), (B) Relationships between  $V$  and  $J_I$ . (C), (D) Relationships between  $V$  and  $J_E$ . (E), (F) Relationships between  $L$  and  $J_E$ . (A), (C), and (E) were calculated based on the results from worm-yeast mapping. In the mapping process, only the high-quality interactions in worm were included. (B), (D), and (F) are the weighted average obtained when the interactions in all three organisms (i.e. *C. elegans*, *D. melanogaster*, and *H. pylori*, including all high-throughput interactions) were mapped onto yeast. In panel (A), Low:  $J_I \leq 10\%$ ; Medium:  $20\% \leq J_I \leq 30\%$ ; High:  $J_I \geq 40\%$ . In (C) and (D), Low:  $10^{-40} \leq J_E \leq 10^{-10}$ ; Medium:  $10^{-100} \leq J_E \leq 10^{-50}$ ; High:  $J_E \leq 10^{-110}$ . Error bars represent 95% CI calculated by a re-sampling algorithm (see supplementary material).

Figure 3. Distribution of the number of generalized interologs as a function of joint E-value ( $J_E$ ). The dashed line represents the number of all predictions above a given  $J_E$ , i.e.  $G(J)$ . The solid line represents the number of true positives above a given  $J_E$ , i.e.  $TP$ .

Figure 4. Comparison of generalized interolog mapping with PIE. In this figure, the plot ( $TP/|P'|$  vs.  $TP/FP$ ) is analogous to a ROC plot ( $TP/P$  vs.  $FP/N$ ). Based on this curve, the performance of our method is comparable to that of the large-scale experimental datasets.

Figure 5. Schematic illustration of protein-DNA interologs and regulogs. In the source organism, TF A binds to its binding site ( $S_A$ ) and regulates the downstream gene B. To perform the regulog mapping, TF A' in the target organism needs to be the ortholog of A. Proteins B and B' should also be orthologs. The DNA sequence upstream of gene B' needs to contain the same motif ( $S_{A'}$ ) as  $S_A$ . However, practically TF A and A' only need to share  $\geq 30\%$  identity. The interaction between TF A' and  $S_{A'}$  is the protein-DNA interolog of that between A and  $S_A$ . The regulatory relationships between  $A \Rightarrow B$  and  $A' \Rightarrow B'$  are regulogs.

Figure 6. Conservation of protein-DNA interactions between homologous TFs. The conservation is measured as the relationships between  $V$  and  $I$ . The legend appears as an inset on the graph. The red bold curve was calculated for all TFs in the source datasets (see supplementary materials). Error bars represent 95% CI calculated by the re-sampling algorithm.

Figure 7. Percentage of the overlaps between the predictions and different groups. All, all experimentally-determined interaction pairs. Proteasome, interaction pairs involved in 26S proteasome. DDR, interaction pairs involved in DNA-damage repair. Vulval-dev, interaction pairs involved in vulval development. Others, interaction pairs involved in germline, meiosis, metazoan, mitotic machinery, dauer formation, chromosome III, chromatin remodeling, pharynx, and immunity. The P values measuring the statistical significance of the overlaps between different groups and the predictions are given on top of each bar, which are calculated using the hyper-geometric models (see supplementary material).

Figure 8. Screenshot of the interolog/regulog database.

**Table 1. Statistics of the interolog/regulog database.**

Organisms	Total protein-protein interactions	$J_E$ cut-off for highly reliable interologs	Total TFs	Total targets	Total connections*
<i>S. cerevisiae</i>	8250	N/A	148	3380	6765
<i>C. albicans</i>	20470	$10^{-105}$	66	1085	2349
<i>C. elegans</i>	91224	$10^{-75}$	36	601	1625
<i>D. melanogaster</i>	101920	$10^{-90}$	33	621	2936
<i>A. thaliana</i>	201754	$10^{-90}$	19	165	328

\*A connection is a TF-target pair.

## Reference

- Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.
- Andrade, M.A. and C. Sander. 1997. Bioinformatics: from genome data to biological knowledge. *Curr Opin Biotechnol* **8**: 675-683.
- Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25-29.
- Attwood, T.K., M.E. Beck, A.J. Bleasby, K. Degtyarenko, A.D. Michie, and D.J. Parry-Smith. 1997. Novel developments with the PRINTS protein fingerprint database. *Nucleic Acids Res* **25**: 212-217.
- Bader, G.D., D. Betel, and C.W. Hogue. 2003. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* **31**: 248-250.
- Bairoch, A., P. Bucher, and K. Hofmann. 1996. The PROSITE database, its status in 1995. *Nucleic Acids Res* **24**: 189-196.
- Bork, P., T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan. 1998. Predicting function: from genes to genomes and back. *J Mol Biol* **283**: 707-725.
- Bork, P., C. Ouzounis, and C. Sander. 1994. From genome sequences to protein function. *Curr Opin Struct Biol* **4**: 393-403.
- Boulton, S.J., A. Gartner, J. Reboul, P. Vaglio, N. Dyson, D.E. Hill, and M. Vidal. 2002. Combined functional genomic maps of the *C. elegans* DNA damage response. *Science* **295**: 127-131.
- Brenner, S.E. 1999. Errors in genome annotation. *Trends Genet* **15**: 132-133.
- Brenner, S.E., C. Chothia, and T.J. Hubbard. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc Natl Acad Sci U S A* **95**: 6073-6078.
- Chothia, C. and A.M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**: 823-826.
- Chothia, C. and A.M. Lesk. 1987. The evolution of protein structures. *Cold Spring Harb Symp Quant Biol* **52**: 399-405.
- Christie, K.R., S. Weng, R. Balakrishnan, M.C. Costanzo, K. Dolinski, S.S. Dwight, S.R. Engel, B. Feierbach, D.G. Fisk, J.E. Hirschman, E.L. Hong, L. Issel-Tarver, R. Nash, A. Sethuraman, B. Starr, C.L. Theesfeld, R. Andrada, G. Binkley, Q. Dong, C. Lane, M. Schroeder, D. Botstein, and J.M. Cherry. 2004. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res* **32**: D311-314.
- Davy, A., P. Bello, N. Thierry-Mieg, P. Vaglio, J. Hitti, L. Doucette-Stamm, D. Thierry-Mieg, J. Reboul, S. Boulton, A.J. Walhout, O. Coux, and M. Vidal. 2001. A protein-protein interaction map of the *Caenorhabditis elegans* 26S proteasome. *EMBO Rep* **2**: 821-828.

- Edwards, A.M., B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, and M. Gerstein. 2002. Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet* **18**: 529-536.
- Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, and J.M. Kelley. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**: 397-403.
- Fraser, C.M., S.J. Norris, G.M. Weinstock, O. White, G.G. Sutton, R. Dodson, M. Gwinn, E.K. Hickey, R. Clayton, K.A. Ketchum, E. Sodergren, J.M. Hardham, M.P. McLeod, S. Salzberg, J. Peterson, H. Khalak, D. Richardson, J.K. Howell, M. Chidambaram, T. Utterback, L. McDonald, P. Artiach, C. Bowman, M.D. Cotton, and J.C. Venter. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**: 375-388.
- Gavin, A.C., M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon, C.M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.A. Heurtier, R.R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**: 141-147.
- Giot, L., J.S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y.L. Hao, C.E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C.A. Stanyon, R.L. Finley, Jr., K.P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R.A. Shimkets, M.P. McKenna, J. Chant, and J.M. Rothberg. 2003. A protein interaction map of *Drosophila melanogaster*. *Science* **302**: 1727-1736.
- Hahn, S. and L. Guarente. 1988. Yeast HAP2 and HAP3: transcriptional activators in a heteromeric complex. *Science* **240**: 317-321.
- Harris, T.W., N. Chen, F. Cunningham, M. Tello-Ruiz, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, J. Chan, C.K. Chen, W.J. Chen, P. Davis, E. Kenny, R. Kishore, D. Lawson, R. Lee, H.M. Muller, C. Nakamura, P. Ozersky, A. Petcherski, A. Rogers, A. Sabo, E.M. Schwarz, K. Van Auken, Q. Wang, R. Durbin, J. Spieth, P.W. Sternberg, and L.D. Stein. 2004. WormBase: a multi-species resource for nematode biology and genomics. *Nucleic Acids Res* **32**: D411-417.
- Hegy, H. and M. Gerstein. 2001. Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res* **11**: 1632-1640.
- Ho, Y., A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A.R. Willems, H. Sassi, P.A. Nielsen, K.J. Rasmussen, J.R. Andersen, L.E. Johansen, L.H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B.D. Sorensen, J. Matthiesen,

- R.C. Hendrickson, F. Gleeson, T. Pawson, M.F. Moran, D. Durocher, M. Mann, C.W. Hogue, D. Figeys, and M. Tyers. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**: 180-183.
- Horak, C.E., N.M. Luscombe, J. Qian, P. Bertone, S. Piccirillo, M. Gerstein, and M. Snyder. 2002. Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev* **16**: 3017-3033.
- Ito, T., K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki. 2000. Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A* **97**: 1143-1147.
- Iyer, V.R., C.E. Horak, C.S. Scafe, D. Botstein, M. Snyder, and P.O. Brown. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**: 533-538.
- Jansen, R., H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, and M. Gerstein. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**: 449-453.
- Kumar, A., S. Agarwal, J.A. Heyman, S. Matson, M. Heidtman, S. Piccirillo, L. Umansky, A. Drawid, R. Jansen, Y. Liu, K.H. Cheung, P. Miller, M. Gerstein, G.S. Roeder, and M. Snyder. 2002. Subcellular localization of the yeast proteome. *Genes Dev* **16**: 707-719.
- Kumar, A. and M. Snyder. 2002. Protein complexes take the bait. *Nature* **415**: 123-124.
- Lan, N., R. Jansen, and M. Gerstein. 2002. Toward a Systematic Definition of Protein Function That Scales to the Genome Level: Defining Function in Terms of Interactions. *Proceeding of the IEEE* **90**: 1848-1858.
- Lan, N., G.T. Montelione, and M. Gerstein. 2003. Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level. *Curr Opin Chem Biol* **7**: 44-54.
- Larsen, C.N. and D. Finley. 1997. Protein translocation channels in the proteasome and other proteases. *Cell* **91**: 431-434.
- Lee, T.I., N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D.B. Gordon, B. Ren, J.J. Wyrick, J.B. Tagne, T.L. Volkert, E. Fraenkel, D.K. Gifford, and R.A. Young. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**: 799-804.
- Li, S., C.M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.O. Vidalain, J.D. Han, A. Chesneau, T. Hao, D.S. Goldberg, N. Li, M. Martinez, J.F. Rual, P. Lamesch, L. Xu, M. Tewari, S.L. Wong, L.V. Zhang, G.F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H.W. Gabel, A. Elewa, B. Baumgartner, D.J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S.E. Mango, W.M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K.C. Gunsalus, J.W. Harper, M.E. Cusick, F.P. Roth, D.E. Hill, and M. Vidal. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science* **303**: 540-543.

- Limbach, K.J. and R. Wu. 1985. Characterization of two *Drosophila melanogaster* cytochrome c genes and their transcripts. *Nucleic Acids Res* **13**: 631-644.
- Luscombe, N.M. and J.M. Thornton. 2002. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol* **320**: 991-1009.
- Marcotte, E.M., M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**: 83-86.
- Matthews, L.R., P. Vaglio, J. Reboul, H. Ge, B.P. Davis, J. Garrels, S. Vincent, and M. Vidal. 2001. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* **11**: 2120-2126.
- Mewes, H.W., D. Frishman, C. Gruber, B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Pfeiffer, C. Schuller, S. Stocker, and B. Weil. 2000. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **28**: 37-40.
- Olesen, J., S. Hahn, and L. Guarente. 1987. Yeast HAP2 and HAP3 activators both bind to the CYC1 upstream activation site, UAS2, in an interdependent manner. *Cell* **51**: 953-961.
- Pearson, W.R. and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85**: 2444-2448.
- Pellegrini, M., E.M. Marcotte, M.J. Thompson, D. Eisenberg, and T.O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**: 4285-4288.
- Posas, F., M. Takekawa, and H. Saito. 1998. Signal transduction by MAP kinase cascades in budding yeast. *Curr Opin Microbiol* **1**: 175-182.
- Rain, J.C., L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne, and P. Legrain. 2001. The protein-protein interaction map of *Helicobacter pylori*. *Nature* **409**: 211-215.
- Tatusov, R.L., E.V. Koonin, and D.J. Lipman. 1997. A genomic perspective on protein families. *Science* **278**: 631-637.
- The FlyBase Consortium. 2002. The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res* **30**: 106-108.
- Uetz, P., L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J.M. Rothberg. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**: 623-627.
- von Mering, C., R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**: 399-403.
- Walhout, A.J., R. Sordella, X. Lu, J.L. Hartley, G.F. Temple, M.A. Brasch, N. Thierry-Mieg, and M. Vidal. 2000. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**: 116-122.
- Webb, E.C. 1992. *Enzyme Nomenclature 1992, Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. New York: Academic.

- Westbrook, J., Z. Feng, L. Chen, H. Yang, and H.M. Berman. 2003. The Protein Data Bank and structural genomics. *Nucleic Acids Res* **31**: 489-491.
- Wilson, C.A., J. Kreychman, and M. Gerstein. 2000. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* **297**: 233-249.
- Wingender, E., X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhauser, M. Pruss, F. Schacherer, S. Thiele, and S. Urbach. 2001. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res* **29**: 281-283.
- Xenarios, I., L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, and D. Eisenberg. 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30**: 303-305.

## Web site references

[http://vidal.dfci.harvard.edu/main\\_pages/All\\_interactions.txt](http://vidal.dfci.harvard.edu/main_pages/All_interactions.txt) Worm Interaction Data from Dr. Marc Vidal's Group as of February 2003.

<http://bioinfo.mbb.yale.edu/proteinchip/interolog/> Interolog/Regulog Database from Dr. Mark Gerstein's Group.

<http://flybase.bio.indiana.edu/> FlyBase-A database of Drosophila genome.

<http://pfam.wustl.edu/> Pfam database of protein families and HMMs.

<http://transfac.gbf.de/TRANSFAC/> TRANSFAC-The Transcription Factor Database.

<http://mips.gsf.de/> MIPS database.

# Supplementary materials

## Platinum standard positives

We manually collected these interactions from three independent sources:

1. Physical interactions recorded in BIND, DIP and MIPS databases (6650, 15113 and 5834 interactions, respectively). Because high-throughput methods are known to be error-prone, all high-throughput interactions are excluded. Furthermore, previous studies have shown that interactions recorded in more than one datasets tend to be more reliable. Therefore, an interaction is considered as a gold standard positive only if it appears in at least two of the three databases. According to these two conditions, 1513 interactions in the three databases are considered as true interactions.
2. Small complexes in the MIPS complex catalog. We consider protein pairs within the same small complexes ( $\leq 4$  subunits) as interaction partners, which results in 308 interactions.
3. Derived interactions in the PDB. 485 PDB entries provided structural information on yeast proteins and complexes. We calculated the contact surface area between any two subunit of each complex. Pairs of proteins (subunits) with sufficient contact surface areas ( $\geq 50 \text{ \AA}^2$ ) are considered as interaction partners. We are able to identify 99 interactions.

In total, we generate a smaller set of platinum standard positives consisting of 1867 interactions among 1391 proteins.

## Supplementary Figure 1 caption

Conservation of protein-protein interactions between homologous protein pairs using platinum standard positives. (A) Relationships between  $V$  and  $J_I$ . (B) Relationships between  $V$  and  $J_E$ . (C) Relationships between  $L$  and  $J_E$ . The results are the weighted average obtained when the interactions in all three organisms (i.e. *C. elegans*, *D. melanogaster*, and *H. pylori*) were mapped onto yeast.

## Supplementary Figure 2 caption

Conservation of protein-protein interactions between homologous protein pairs using minimal similarity as the joint similarity ( $J$ ). (A), (B) Relationships between  $V$  and  $J_I$ . (C), (D) Relationships between  $V$  and  $J_E$ . (E), (F) Relationships between  $L$  and  $J_E$ . (A), (C), and (E) were calculated based on the results from worm-yeast mapping (using only the high-quality worm interactions). (B), (D), and (F) are the weighted average obtained when the interactions in all three organisms (i.e. *C. elegans*, *D. melanogaster*, and *H. pylori*) were mapped onto yeast. In panel (A), Low:  $J_I \leq 10\%$ ; Medium:  $20\% \leq J_I \leq 30\%$ ; High:  $J_I \geq 40\%$ . In (C) and (D), Low:  $10^{-30} \leq J_E \leq 10^{-10}$ ; Medium:  $10^{-60} \leq J_E \leq 10^{-40}$ ; High:  $J_E \leq 10^{-70}$ .

## Re-sampling algorithm for the calculation of the 95% CI

For each bin in the figures, 90% of the population is randomly chosen, in which the number of true positives is determined. This re-sampling step is repeated 100 times. Then, the 95% CI is calculated for the average percentage of true positives.

## Detailed description of the protein-protein interolog example

The yeast mating-pheromone response is one of the best-characterized signal transduction pathways in eukaryotes. In brief, the receptors for the  $\alpha$  and a mating factors (*ste2* and *ste3*, respectively) are G-protein coupled receptors, which, when activated, will transduce signals through *Ste20* and *Ste5* to the mating-pheromone MAPK cascade. *Ste5*, a scaffold protein, forms the core of a signaling complex containing *Ste11* (MAPKKK), *Ste7* (MAPKK), and *Fus3/Kss1* (MAPKs). Activated *Fus3/Kss1* MAPKs then activate the transcription factor *Ste12* to induce the expression of specific genes. Furthermore, activated *Fus3* (but not *Kss1*) will arrest the cell cycle at the G1/S transition by activating the inhibitor of *Cdc28-Cln* kinase, *Far1*. Therefore, *Ste5*-MAPK complex is a key complex in this signal transduction pathway (*Current Opinion in Microbiology* **1**: 175-182).

The interaction partners of worm MAPK (F43C1.2a) were determined by experiments (see supplementary table 1). So far, there are in total 26 partners for F43C1.2a in the database, none of which is involved in this MAPK cascade signal transduction pathway. In these partners, K11E8.1C is a calcium/calmodulin dependent protein kinase, whose best-matching homolog in yeast is *CMK2* (E-value =  $2 \times 10^{-52}$ ), also a calcium/calmodulin dependent protein kinase. They have nothing to do with the MAPK cascade pathway. However, K11E8.1C has two distant homologs in yeast, which are *Ste7* (E-value =  $10^{-11}$ ) and *Ste11* (E-value =  $2 \times 10^{-18}$ ). F43C1.2a also has a non-best-matching homolog in yeast, *Kss1* (E-value =  $3 \times 10^{-93}$ ). Therefore, given only one MAP kinase in worm, we are able to successfully predict 5 subunits of the 6-subunit complex in yeast, which is an excellent example demonstrating the power of our method.

## Supplementary Table 1. Interaction partners of F43C1.2a and their best-matching homologs in yeast.

Interaction partners with F43C1.2a	Gene Name	Description	Best-Matching Homolog in Yeast	Gene Name	E-value	Description
C06A8.5	PCS*		Not found			
C06C3.1	mel-11	ankyrin-like repeats, electron transport	YDR264C	AKR1	1.00E-16	ankyrin-like repeats, endocytosis
C49A9.6	PCS		Not found			
C49C3.7	PCS		YKR095W	MLP1	6.00E-11	involved in translocation of macromolecules between the nucleoplasm and the NPC
F08C6.7	unc-98	Zinc Finger	Not found			
F10E9.3	PCS		Not found			
F14F3.2	PCS	ankyrin motif	Not found			
F29G9.2	PCS		Not found			
F32D1.1	PCS	AAA ATPase	YPL074W	YTA6	4.00E-74	AAA ATPase
F38B2.1	ifa-1	intermediate filament protein	YDR356W	SPC110	2.00E-15	structural constituent of cytoskeleton,
F42A10.2	PCS		Not found			
F42C5.10	PCS		Not found			
F42H10.7	PCS		Not found			
F47B10.2	PCS		Not found			
F54D5.5	PCS		Not found			
K04G2.10	PCS		Not found			
K04G7.1	PCS		Not found			
K11E8.1c	unc-43	calcium/calmodulin-dependent protein kinase	YOL016C	CMK2	2.00E-52	calcium/calmodulin-dependent protein kinase
M6.1	ifc-2	intermediate filament protein A	Not found			
T05C12.6	mig-5	presynaptic density protein (PSD-95) repeat-like domain	Not found			
T08D10.1	PCS	CCAAT-binding transcription factor	YGL237C	HAP2	4.00E-11	CCAAT-binding factor complex
T22A3.3	PCS		Not found			
T23H4.2	nhr-69	Ligand-binding domain of nuclear hormone receptors, Zinc finger, C4 type (two domains)	Not found			
T27F2.2	PCS		Not found			
W10D9.3	PCS		Not found			
W10G6.3	ifa-2	Intermediate filament protein	YDL058W	USO1	3.00E-14	ER to Golgi transport

\* PCS: Predicted Coding Sequence

## Calculation of the $V$ 's for protein-DNA interologs

First we calculate the  $V$  for each TF family. In the protein-protein interolog mapping, we map all the interactions in the source organism onto the target organism. However, due to the relatively small amount of TF binding data, we consider each TF individually in this calculation, that is, all TF pairs within the family are considered as a potential protein-DNA interolog, regardless whether they are in the same genome or not. Suppose that there are  $N$  TFs in the family, the number of pairs is  $N(N-1)/2$ . For each pair, the sequence identity ( $I$ ) between the TFs is measured by FASTA. If they bind to the same DNA sequence, the pair is considered as a true positive. After all the TF pairs are measured,  $T(I)$  and  $G(I)$  for a certain identity cutoff ( $I$ ) have been determined. Then, the percentage of verified predictions,  $V(I)$ , is calculated as:

$$V(I) = \frac{|T(I)|}{|G(I)|} \times 100\%$$

We also calculated the relationship between  $V$  and  $I$  for all TFs in the source dataset including all different TF families. The curve (the red bold curve in Figure 6) was generated by averaging all the curves for different families.

## Hyper-geometric model

Suppose that the total sample space is  $N$ . In the first round, a sample of size  $S_1$  is randomly selected without replacement from  $N$ . Then, the entire initial sample ( $S_1$ ) is subsequently returned to the sample space. In the second round, another sample of size  $S_2$  is randomly selected without replacement from  $N$ . The size of the overlap (denoted as  $i$ ) between the two samples ( $S_1$  and  $S_2$ ) is a hyper-geometric random variable. The probability of observing an overlap of a given size  $X$  or greater is calculated by the formula:

$$P(i \geq X) = 1 - \sum_{i=0}^{X-1} \frac{\binom{S_1}{i} \binom{N-S_1}{S_2-i}}{\binom{N}{S_2}}$$

To calculate the statistical significance of the overlap between the experimentally-determined interactions and our predictions, we first need to determine the total sample space ( $N$ ). As we discussed in the text, splicing forms of the same gene are removed, which cut the number of genes in worm genome to 19485, 3697 of which have at least one yeast homolog. Only 2816 baits, 785 of which have at least one yeast homolog, have been used in the two-hybrid experiments. Therefore, a pair of interacting proteins will not be identified by the experiments, if neither of them are baits. Likewise, a pair of interacting proteins will not be identified by the family interaction mapping method, if neither of them have at least one yeast homolog. Therefore, the sample space ( $N$ ) is  $2.6 \times 10^6$ ;  $S_1 = 1982$ ;  $S_2 = 26875$ ;  $X = 45$ . The numbers for specific groups are given in supplementary table 2.

In a more simple minded way, one could conceivably think in terms of the 3730 experimentally-determined interactions and our 55223 predictions being sampled from the potential  $2 \times 10^8$  pairs in the worm genome. Thus, the sample space ( $N$ ) is  $2 \times 10^8$ ;  $S_1 = 3730$ ;  $S_2 = 55223$ ;  $X = 45$ . The calculated P value is  $9.2 \times 10^{-7}$ , which is slightly better than the one that we calculated above ( $P < 10^{-5}$ ).

## Supplementary table 2. Parameters for hyper-geometric models

Group	$N$	$S_1$	$S_2$	$X$
All	2.59E+06	1982	26875	45
Proteasome	1.14E+05	123	1644	20
DDR	1.54E+05	107	455	6
Vulval-dev	4.06E+04	98	1092	4
Others	2.33E+06	1645	22309	16

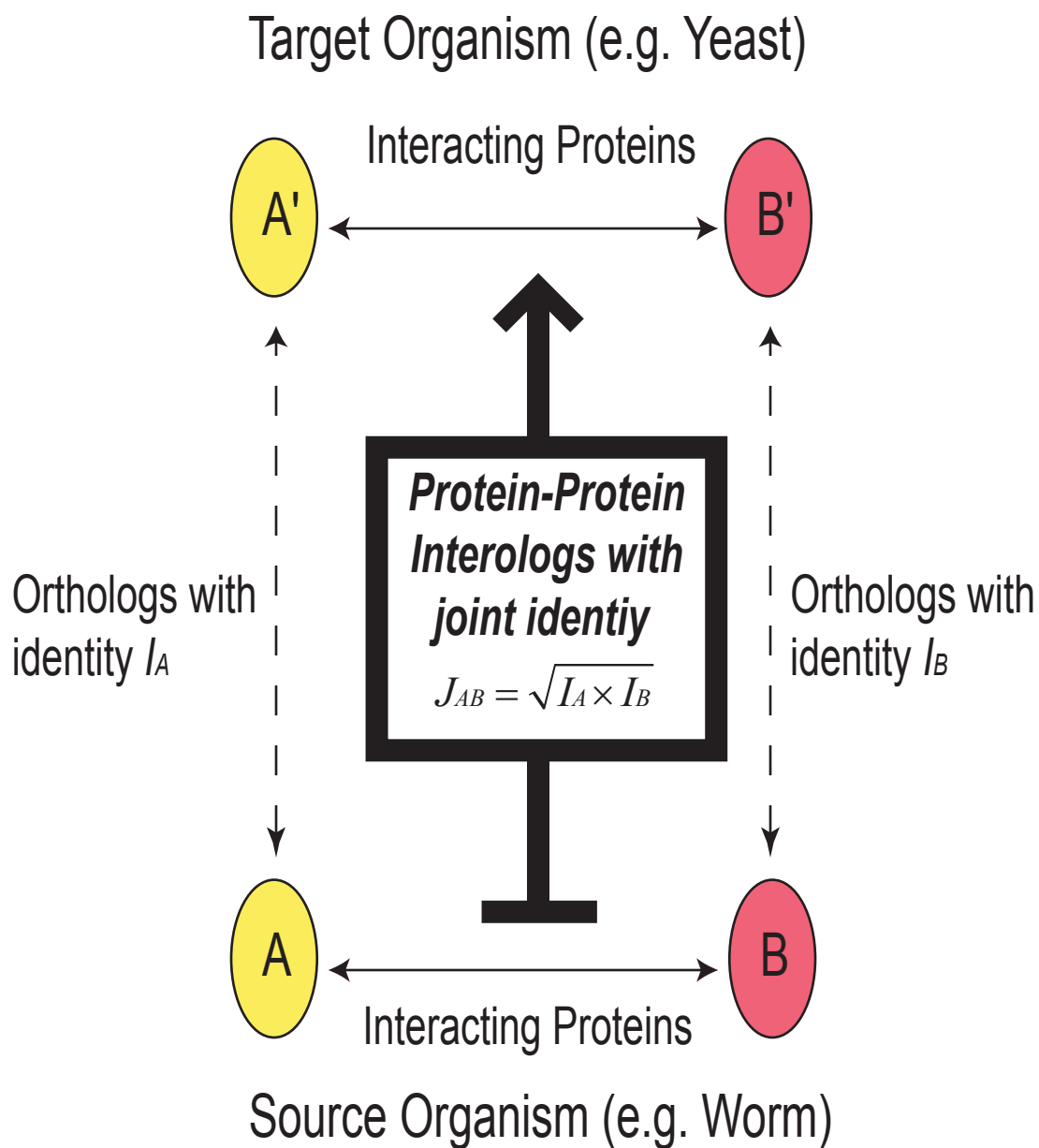
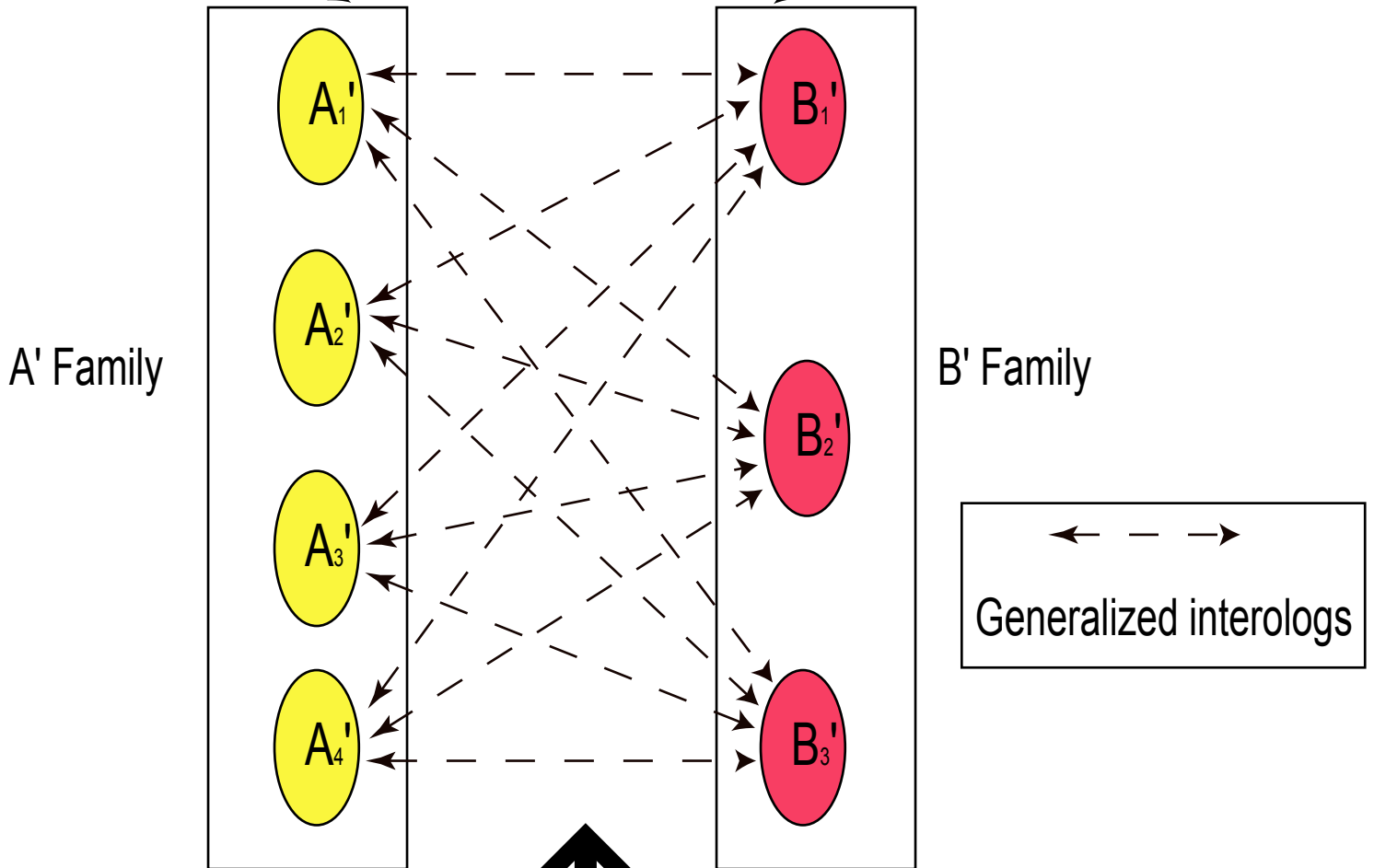


Figure 1A

Target Organism (e.g. Yeast)

Interacting Families



Homologs

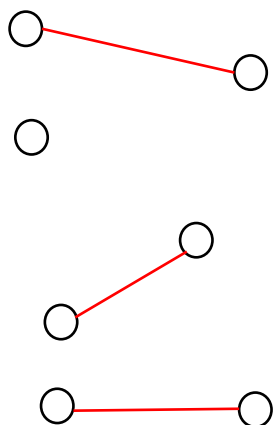
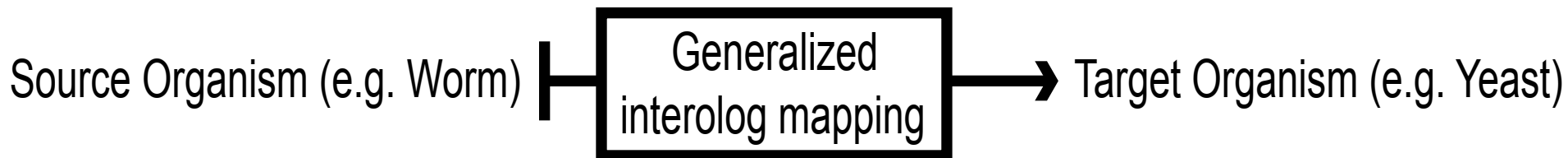
**Generalized  
Interologs with  
many possible  $J_{AB}$**

Homologs

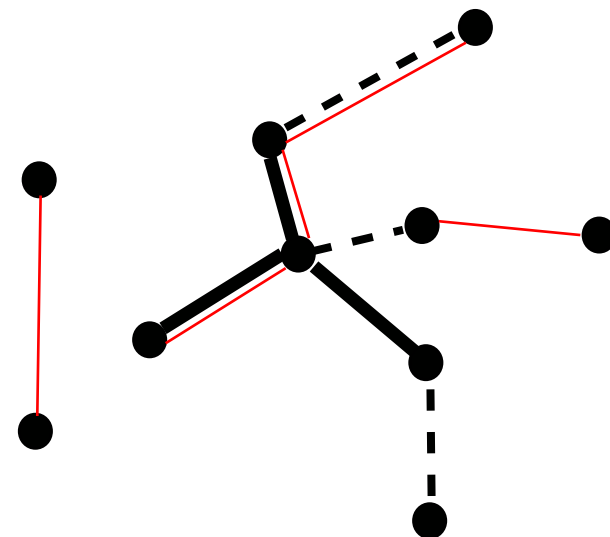
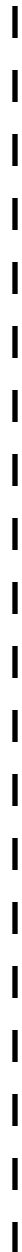


Source Organism (e.g. Worm)

Figure 1B



— Known interactions (e.g. determined by two-hybrid experiments in worm)



— Predicted generalized interologs based on the interactions in the source organism

— Gold standard positive interactions in the target organism (from MIPS complex catalog)

- - - Gold standard negative interactions in the target organism (from localization dataset)

— True positives (e.g. predictions that overlap with the gold standard positives)

- - - False positives (e.g. predictions that overlap with the gold standard negatives)

Figure 1C

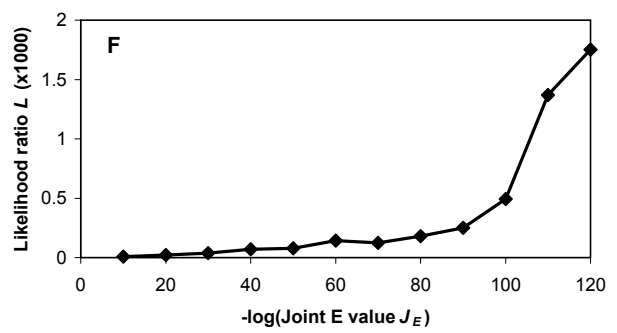
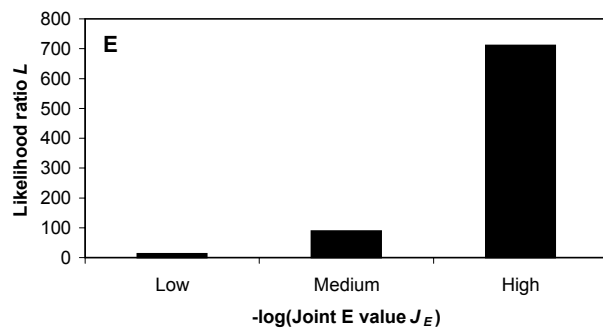
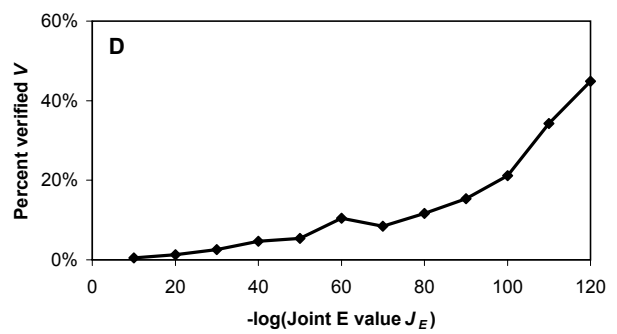
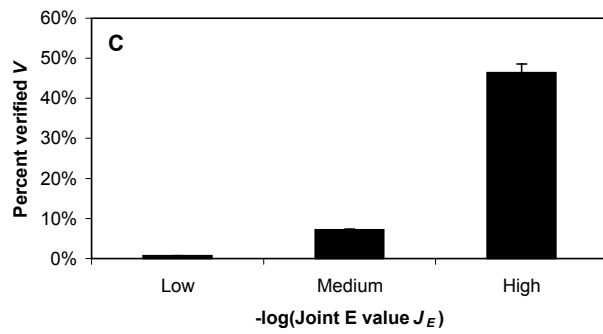
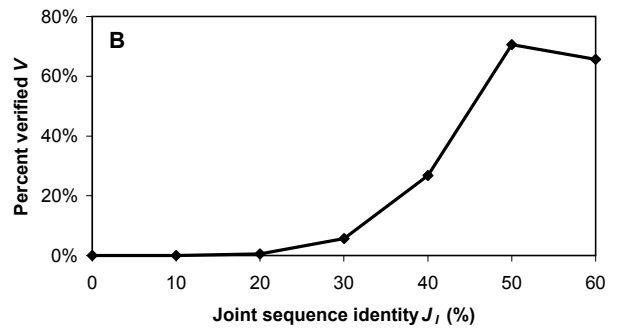
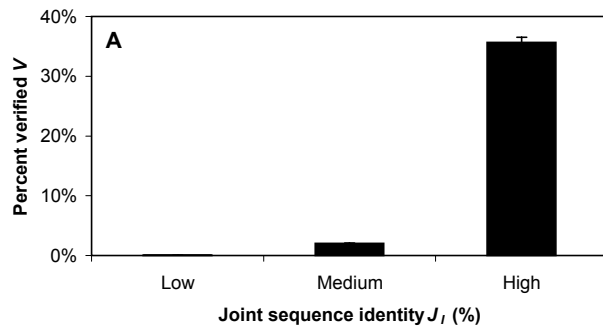


Figure 2

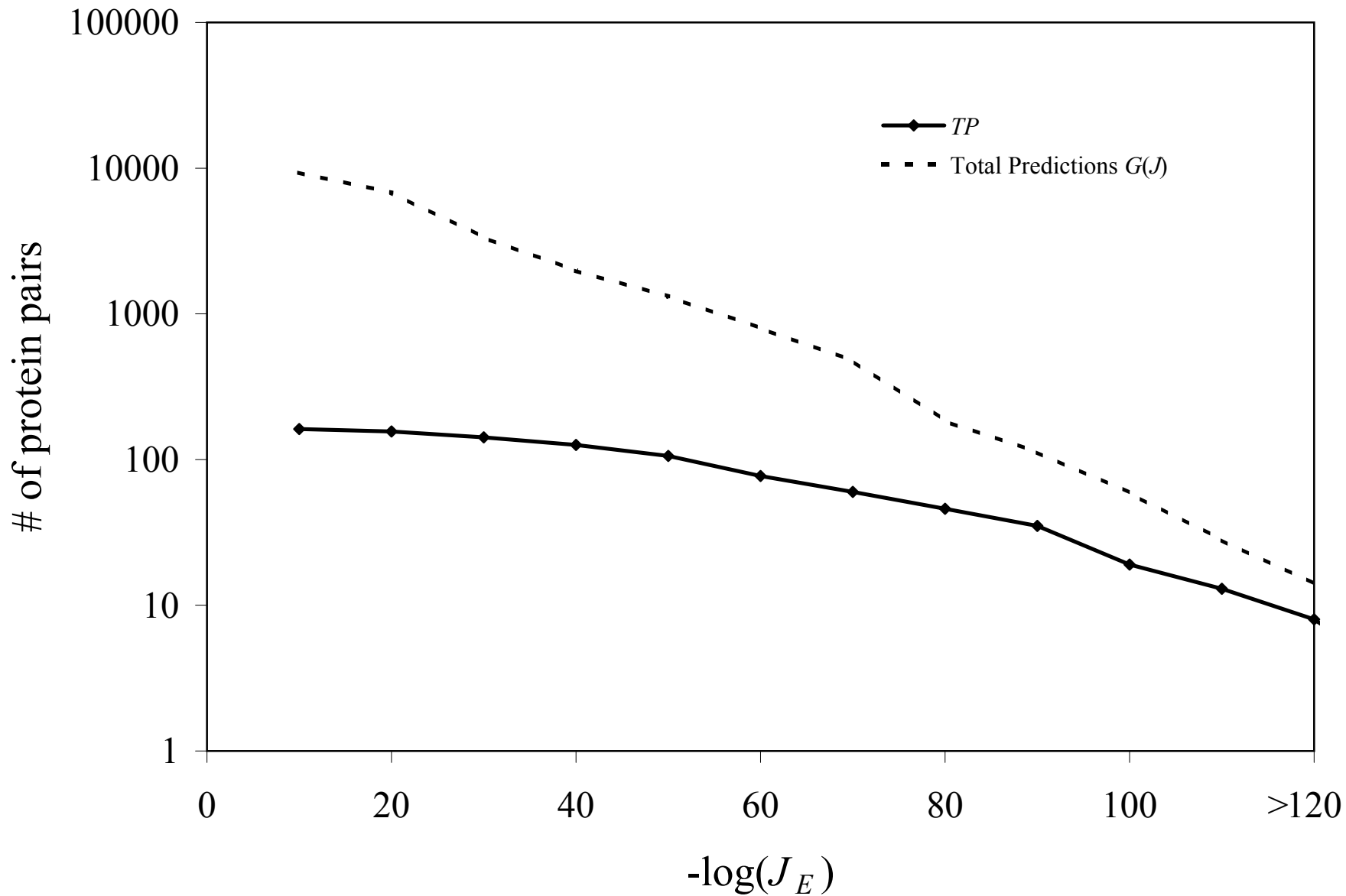


Figure 3

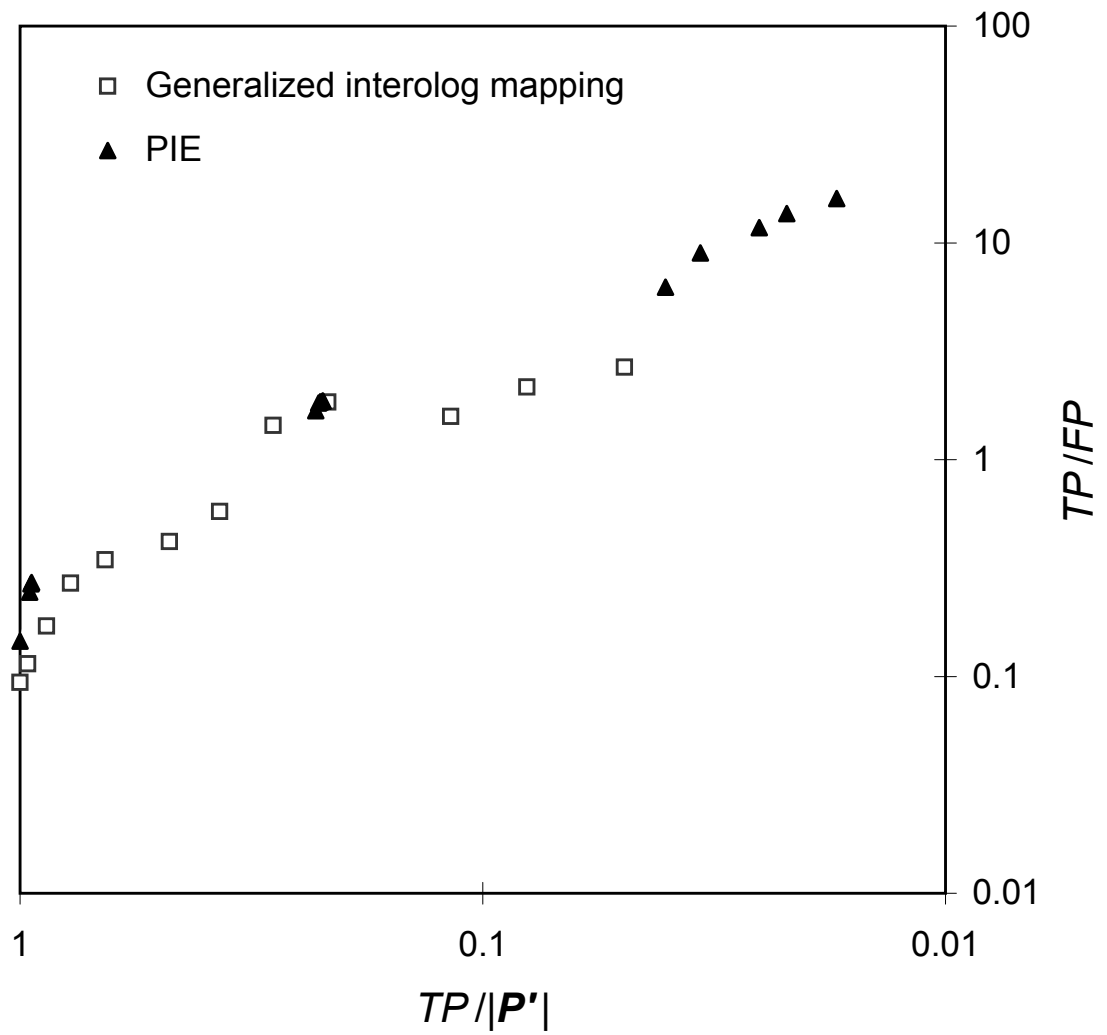
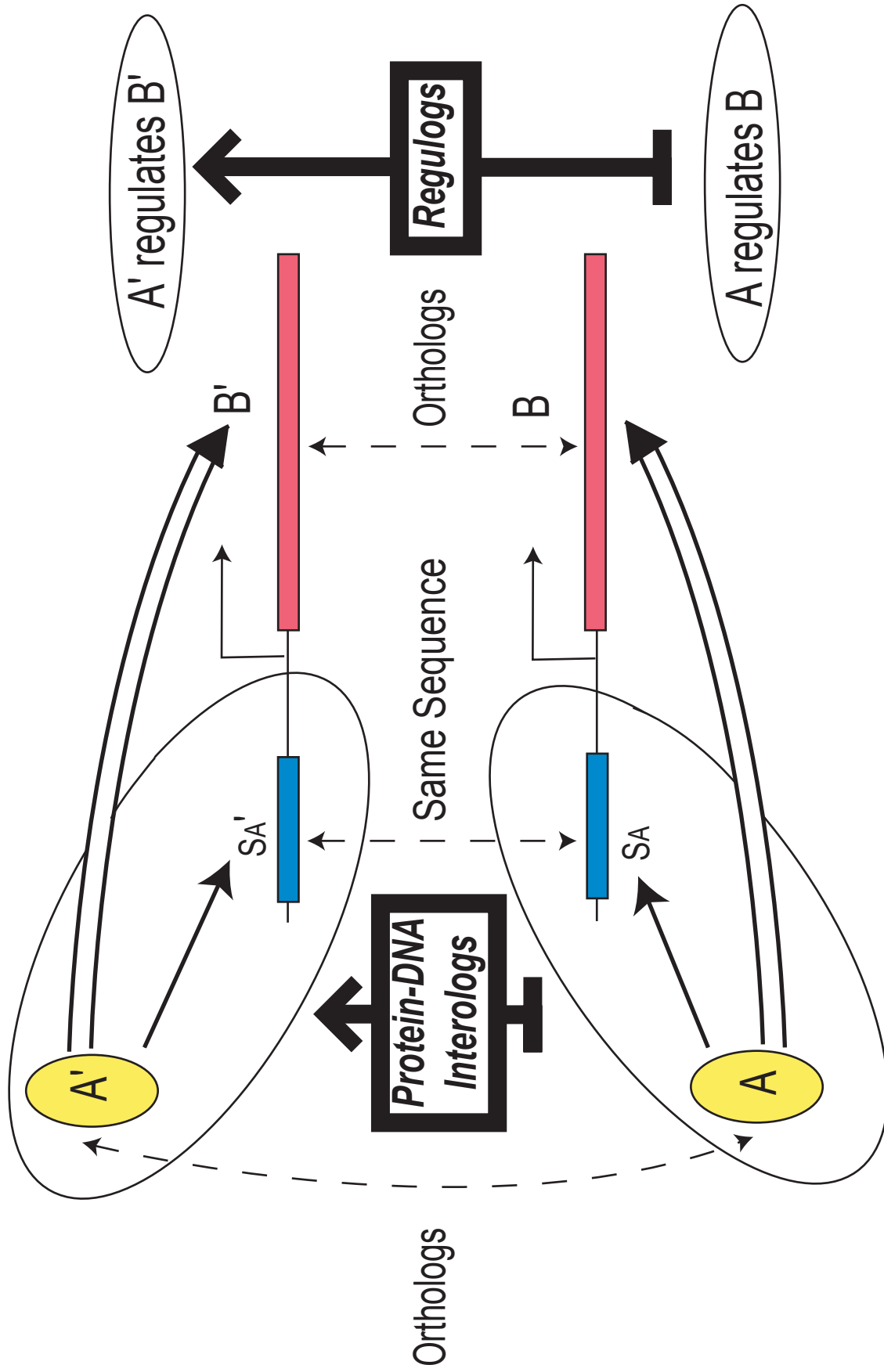


Figure 4

Target Organism (e.g. Fly)



Source Organism (e.g. Yeast)

Figure 5

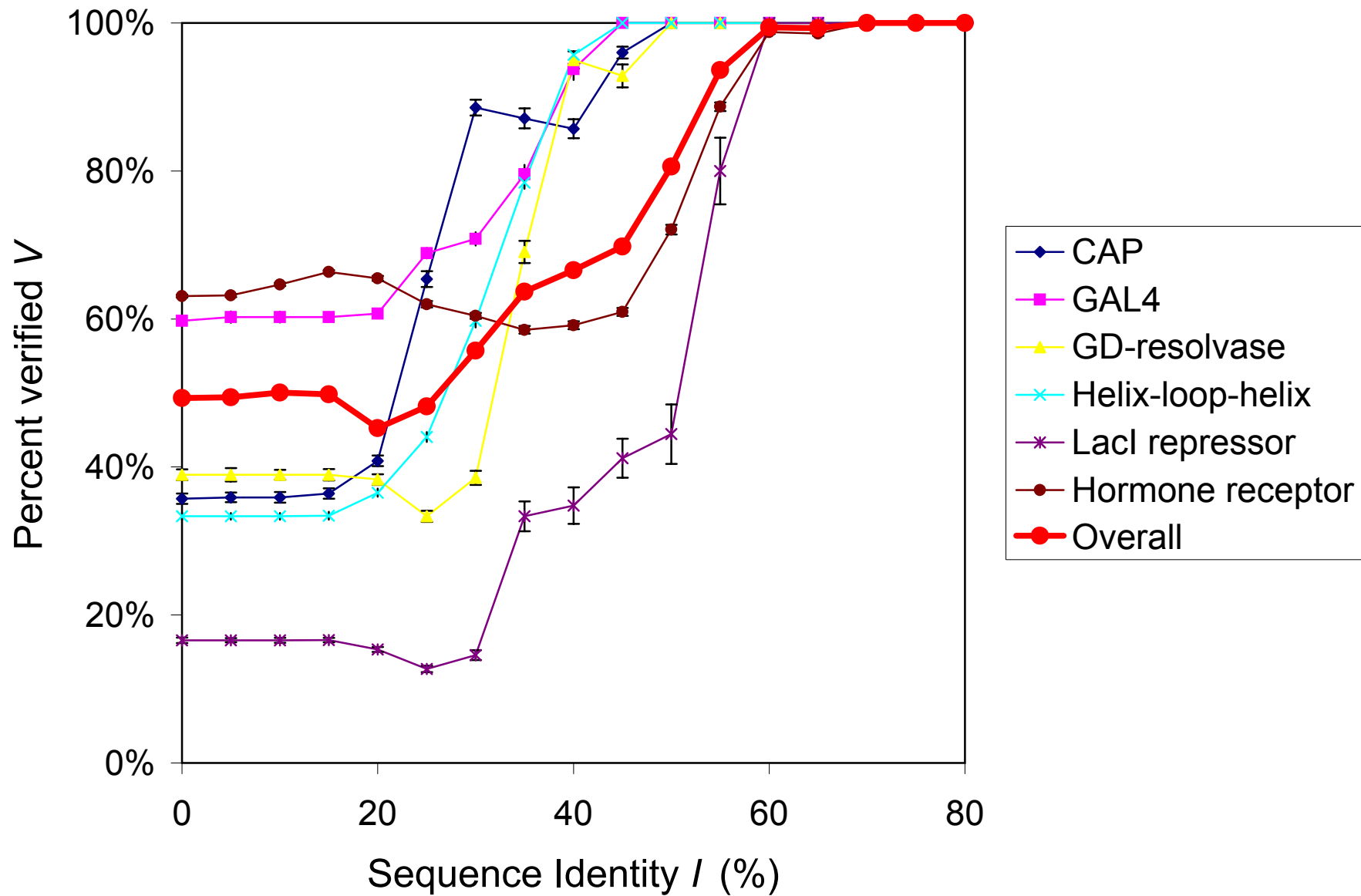


Figure 6

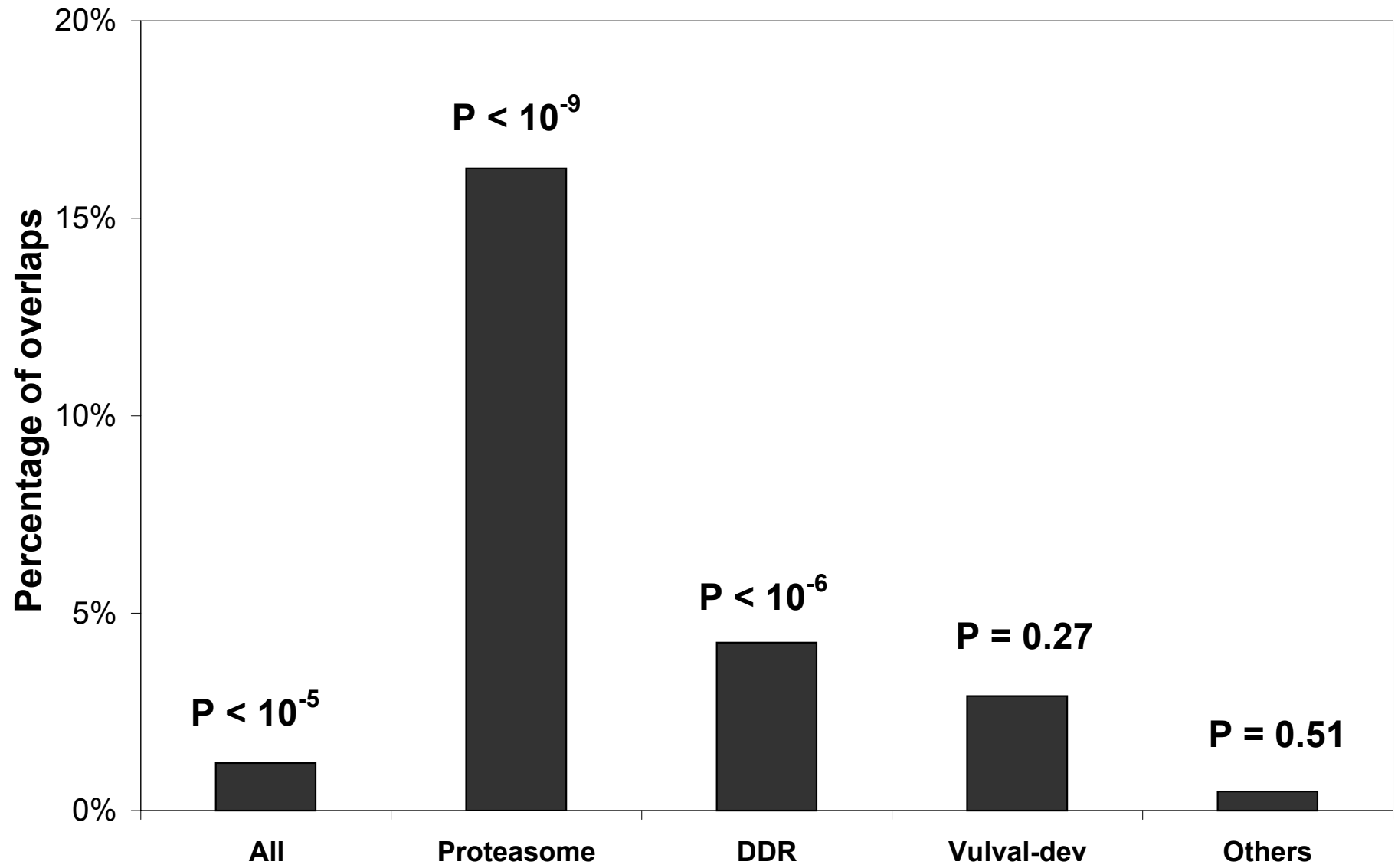


Figure 7

# Interolog/Regulog Database

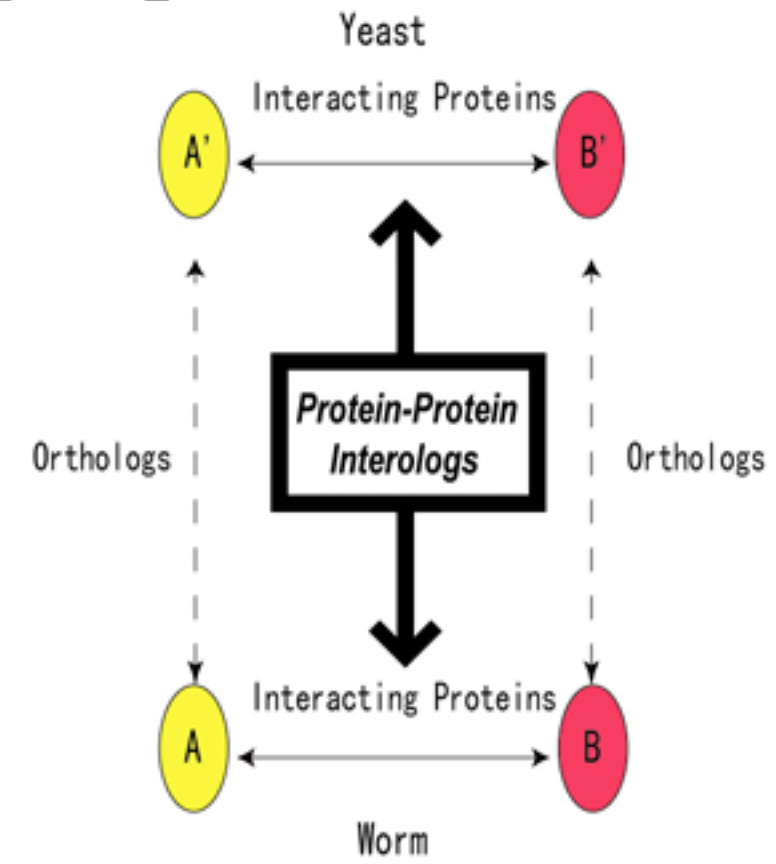
- [Home](#)
- [Download](#)
- [Document](#)
- [Help](#)

## 1. Interologs

### Example

Choose organism:

Input protein:



## 2. Regulogs

### Example

Choose organism:

Input protein:

