Protein-protein Interaction: Network Alignment *

Lecturer: Roded Sharan Scribers: Ofer Lavi and Lev Ferdinskoif

Lecture 9, December 21, 2006

1 Introduction

In the last few years the amount of available data on protein-protein interaction (PPI) networks have increased rapidly, spanning different species such as yeast, bacteria, fly, worm and Human. The rapid growth is shown in Figure 1. Besides the availability of the data, other incentives to analyze several PPI networks at once are validation of our conclusions over several networks and prediction of unknown protein function and interactions.

The scribe is organized as follows: In section 2 we describe the network alignment and network querying problems. In section 3 we describe network pairwise alignment, and its usage in finding conserved protein paths and complexes in PPI networks. The comparative analysis approach, which allows us for detecting similar functionality by looking at multiple highly conserved interactions between similar proteins from different species, is presented in section 4 using QPath, an efficient algorithm for path queries, based on dynamic programming. In section 5 we show PPI networks can be used to predict functional orthologous genes. In section 6 we describe a model that extends the pairwise alignment model to a multiple alignments. Finally, section 7 contains a brief summary.

2 Network Alignment and Querying

A fundamental problem in molecular biology is the identification of cellular machinery, that is, protein pathways and complexes. PPI data present a valuable resource for this task. But there is a considerable challenge to interpret it due to the high noise levels in the data and the fact that no good models are available to pathways and complexes. Comparative analysis is used to tackle these problems, and improve the accuracy of the predictions.

The main paradigm behind comparison of PPI networks is that evolutionary conservation implies functional significance. Conservation of protein subnetworks is measured both in terms of *protein sequence* similarity, and in terms of similarity in *interaction topology*.

This section describes some basic notions that appear in many previous works that find conserved pathways and complexes in the PPI networks of different organisms.

2.1 Network Alignment

A PPI network is conveniently modeled by an undirected graph $G(V, E)$, where V denotes the set of proteins, and $(u, v) \in E$ denotes an interaction between proteins $u \in V$ and $v \in V$.

*Based on a scribe by Irit Levy and Oved Ourfali, 2005

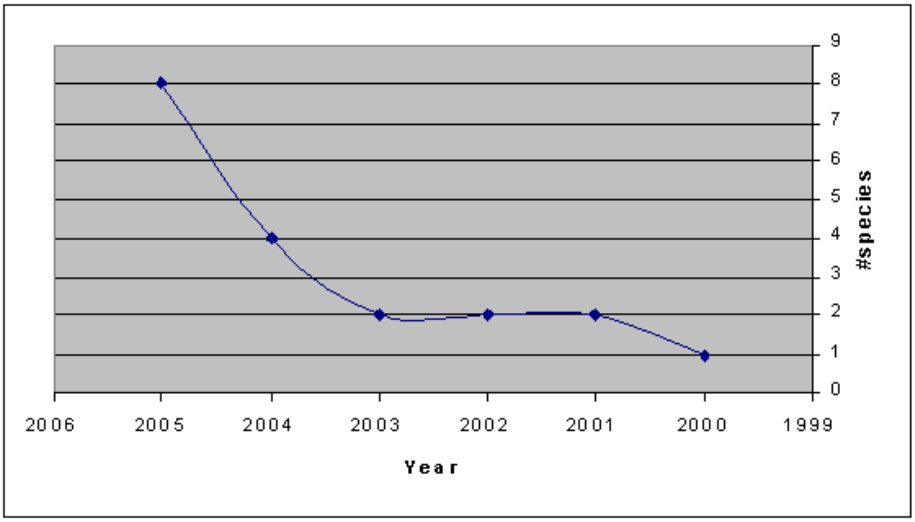

Figure 1: This graph shows the amount of species that their PPI network has been measured. We can clearly see a rapid growth in the number of species since 2003.

The *network alignment* problem: Given k different PPI networks belonging to different species, we wish to find conserved subnetworks within these networks. In order to find these conserved subnetworks an *alignment graph* is built. This graph consists of nodes representing sets of k sequence-similar proteins (one per species), and edges representing conserved interactions between the the species. Illustration of such alignment is shown in Figure 2. This concept was first introduced and used by Ogata et al. [13] and Kelley et al. [10].

Creating an alignment graph from a set of k original networks is one heuristic that enables us to search in all k PPI networks simultaneously. A heuristic approach is required here since the problem of finding conserved subnetworks in a group of networks is NP-Hard, because we can reduce it to subgraph-isomorphism (which is known to be NP-Hard). Other heuristics, or approximation methods are applicable as well.

2.2 Network Querying Problem Definition

Given a PPI network G, and a subnetwork S, we wish to find subnetworks in G that are similar to S. Similarity is measured both in terms of sequence similarity and topological similarity.

The network querying problem can be reduced to a network alignment problem, as shown by Kelley et al. [10], simply by aligning the subnetwork S with the network G. Also, more general formulations are possible, which allow the insertion of proteins into the matched subnetwork, or deletion of vertices from the query subnetwork S.

Network queries can be used to identify conserved functional modules across multiple species, as will be described in the following sections.

2.3 Protein Similarity

In order to build an alignment graph we need to define similarity measure between proteins. First, let us define Homology of proteins (Figure 3 illustrates the *speciation* and *duplication* events, and the described below protein relations):

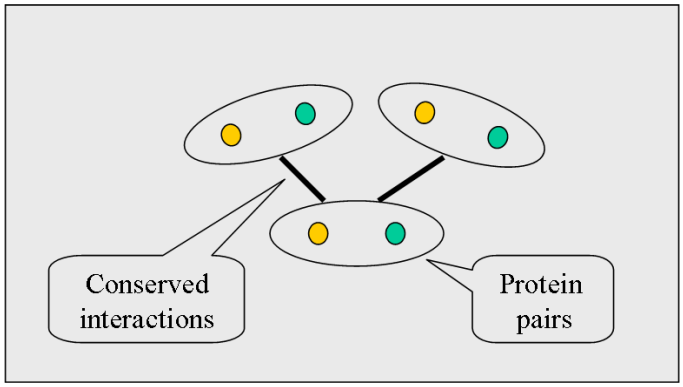

Figure 2: This figure illustrates an alignment graph of two species. Nodes are constructed of pairs of proteins, one per species, which present a high level of sequence-similarity. Edges represent interactions between proteins in the original networks which are conserved, meaning they exist in a high level of confidence in both original networks.

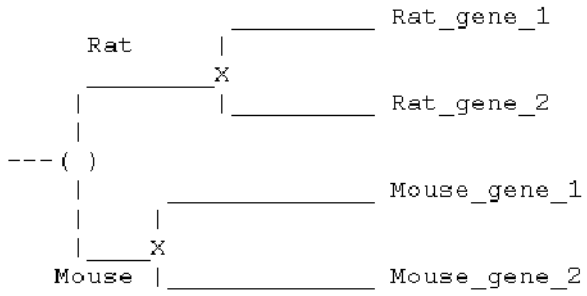

Figure 3: This figure show a gene that diverged after a speciation to a mouse gene and a rat gene. Within the mouse and the rat species the gene has been duplicated to two different genes rat_gene_1 and rat_gene_2 in the rat, and mouse_gene_1 mouse_gene_2 and in the mouse. Each pair of genes are homologous. Each pair of genes that consists of a rat gene and a mouse gene are orthologous, and each pair that consists of genes in the same species are paralogous.

- *Orthologous proteins* - two proteins from different species that diverged after a *speciation* event. In a *speciation* event one species evolves into a different species (*anagenesis*) or one species diverges to become two or more species (*cladogenesis*).
- *Paralogous proteins* - two proteins from the same species that diverged after a *duplication* event, in which part of the genome is duplicated.
- *Homologous proteins* - two proteins that have common ancestry. This is often detected by checking the sequence similarity between these proteins. The proteins can be either from the same species, or from different species (either *orthologous* or *paralogous*).

We define similar proteins as potentially *homologous proteins*, i.e. proteins whose sequences maintain a certain degree of similarity.

3 Pairwise Alignment

In this section we take a closer look on the network alignment problem of two PPI networks.

3.1 PathBLAST

Kelley *et al.* [10] introduced an efficient computational procedure for aligning two PPI networks and identify their conserved interaction pathways, called PathBLAST. This method searches for high-scoring pathway alignments involving two paths, one from each network, in which proteins of the first path are paired with putative homolog proteins occurring in the same order in the second path (Figure 4). Since PPI data are noisy, and in order to overcome evolutionary variations in module structures, both *gaps* and *mismatches* were allowed:

- *Gaps* - A *gap* occurs when a protein interaction in one path skips over a protein in the other path. In the global alignment graph this is shown by one direct protein interaction edge and one indirect protein interaction edge.
- *Mismatches* - A *mismatch* occurs when aligned proteins do not share sequence similarity, and thus are not a pair in the alignment graph. In the global alignment graph this is shown by two indirect protein interaction edges.

3.1.1 Global Alignment and Scoring

In order to build the global alignment graph we need to measure the similarity between proteins in the PPI networks. This similarity is measured using BLAST [2], which quantifies the similarity and assigns it with a p-value, indicating the probability of observing such similarity at random. Protein sequence alignments were computed using BLAST 2.0 with parameters $b = 0$, $e = 1 \times 10^6$, $f =$ "$C; S$", and $v = 6 \times 10^5$. BLAST 2.0 also computes an E-value, or *Expectation Value*, associated with each blast hit, which is the number of different sequence pairs with score equivalent or better than this hit's score that are expected to result by a random search. Unalignable proteins were assigned a maximum E-value of 5. A path through this combined graph represents a conserved pathway between the two networks. A log probability score $S(P)$ for linear paths in the combined graph was formulated as follows:

$$S(P) = \sum_{v \in P} log_{10} \frac{p(v)}{p_{random}} + \sum_{e \in P} log_{10} \frac{q(e)}{q_{random}} \quad (1)$$

where $p(v)$ is the probability of true homology within the protein pair represented by v, and $q(e)$ is the probability that protein-protein interactions represented by e are indeed real, i.e., not false-positive. The background probabilities p_{random} and q_{random} are the expected values of $p(v)$ and $q(e)$ over all possible vertices and edges in the combined graph.

3.1.2 Path Search in PathBLAST

After the alignment graph is built, simple paths of length 4 are searched for. A simple path is one with no repeated nodes, but since the original networks, and the alignment graph are undirected, finding simple paths using DFS and backtracking would be very costly.

In order to efficiently find simple paths, *random acyclic orientation technique* [1] is used, in which acyclic subgraphs are generated by randomly assigning an orientation for each edge. Searching for a

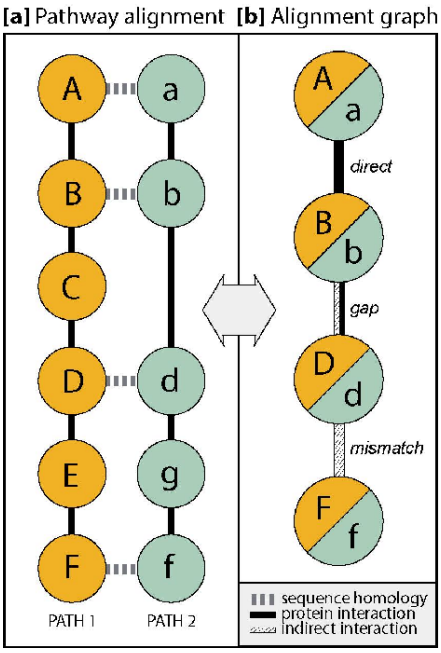

Figure 4: Source [10]. This figure show an example of pathway alignment and merged representation. (a) Vertical solid lines indicate direct proteinprotein interactions within a single pathway, and horizontal dotted lines link proteins with significant sequence similarity($E_{value} \leq E_{cutoff}$). An interaction in one pathway may skip over a protein in the other (protein C), introducing a "gap". Proteins at a particular position that are dissimilar in sequence($E_{value} > E_{cutoff}$, proteins E and g) introduce a "mismatch". The same protein pair may not occur more than once per pathway, and neither gaps nor mismatches may occur consecutively. (b) Pathways are combined as a global alignment graph in which each node represents a homologous protein pair and links represent protein interaction relationships of three types: direct interaction, gap (one interaction is indirect), and mismatch (both interactions are indirect).

maximal-score simple path of length L in a directed acyclic graph can be done in linear time using dynamic programming, and by generating a sufficient number, $5L!$, of acyclic subgraphs, the maximal-score path of length L can be found in linear time.

Every directed path of length L in the acyclic subgraph is simple and corresponds to two paths, one from each of the two original networks. Although the acyclic orientation technique detects only simple paths in the alignment graph, it is possible that the corresponding paths in the original networks are actually not simple, due to one of the following:

1. If a path is not simple in only one of the PPI networks then this path may also not be simple in the alignment graph, due to the use of gaps.

2. Even if a path is simple in both PPI networks, it may not be simple in the alignment graph, due to the use of mismatches.

The probability that a path of length L in the original graph will appear in the acyclic subgraph is $\frac{2}{L!}$ ($\frac{1}{L!}$ in each direction), thus by generating $5L!$ acyclic subgraphs we expect to find the optimal path in one them.

Conserved regions of the network could be highly interconnected (e.g., a conserved protein complex), thus it was sometimes possible to identify a large number of distinct paths involving the same small set of proteins. Rather than enumerating each of these, they were iteratively filtered in PathBLAST. Denote by S_i the average score in the i-th iteration. Thus, for each iteration k, the set of 50 highest-scoring pathway alignments were recorded (with average score $< S_k >$) and then removed their vertices and edges from the alignment graph before the next stage. The p-value of each stage was assessed by comparing $< S_k >$ to the distribution of average scores $< S_1 >$ observed over 100 random global alignment graphs (constructed as per the data in Figure 5) and assigned to every conserved network region resulting from that stage (Figures 6 and 7). The p-values for pathway queries (Figure 8) were computed individually, by comparing each pathway-alignment score to the best scores achieved over 100 random alignment graphs involving the query and target (yeast) network.

3.1.3 Experimental Results

The authors performed three experiments:

1. Yeast (*S. cerevisiae*) vs. Bacteria (*H. pylori*): orthologous pathways between the networks of two species.

2. Yeast vs. Yeast: paralogous pathways within the network of a single species, by aligning the yeast PPI network versus itself.

3. Yeast vs. Yeast: interrogating the protein network with pathway queries, by aligning the yeast PPI network versus simpe pathways.

3.1.4 Yeast vs. Bacteria: Orthologous Pathways Between the Networks of Two Species

Through this experiment a global alignment between the PPI network of the yeast and Bacteria was performed. The yeast network was constructed using the Database of interacting proteins ([29]), as of November 2002, that included interactions from different data sets derived through systematic co-immunoprecipitation and two-hybrid studies. The Bacteria network was also constructed using the Database of interacting proteins and represented a single two-hybrid study (Rain *et al.* [8]).

Figure 5 shows a comparison between the yeast and bacteria global alignment graphs to the corresponding randomized networks obtained by permuting the protein names. As shown, both the graph size,

	Vertices (homologs)	Edges				CPU, min	Score	
		Total	Direct	Gap	Mismatch		Best*	Best 50†
Yeast vs. *H. pylori* ($E_{cutoff} = 10^{-2}$)	829	2,036	7	260	1,769	0.38	8.1	7.5
Random: mean ± SD		509.0 ± 128.0	2.5 ± 1.9	68.8 ± 23.8	437.7 ± 110.3	0.4 ± 0.02	6.1 ± 0.8	4.8 ± 0.7
Yeast vs. yeast ($E_{cutoff} = 10^{-10}$)	5,593	1,389	1,389	N/A	N/A	7.08	11.9	11.0
Random: mean ± SD		62.3 ± 29.4	62.3 ± 29.4	N/A	N/A	6.9 ± 0.2	−4.1 ± 9.5	−15.3 ± 6.5

Figure 5: Source [10]. This figure shows a summary of the results of testing yeast vs. bacteria networks, and the yeast network vs. itself. The results were compared against random graphs that were constructed by permuting the protein names on each network before the creating the alignment graph. The first column presents the number of vertices (homologs) between the two tested networks. The second column presents the amount of different edges constructed by the alignment algorithm. The third column presents the CPU time that was needed to create the alignment graph. The last column presents the average of all the scores vs. the average of top 50 scores.

and the best pathway-alignment scores were significantly larger for the real aligned networks than for the randomized ones. This suggests that both species indeed share conserved interaction pathways, because the alignment results were significantly better compared to randomized networks. Surprisingly, the conservation of a single direct interaction between both networks was rare (but their number was higher in real networks than in randomized networks). However, the fact that ”mismatches” and ”gaps” were permitted, allowed to find much larger regions that were conserved. The use in gaps and mismatches allowed PathBLAST to overcome false negatives in the PPI data.

The top-scoring pathway alignments between bacteria and yeast are described in Figure 6. As validation that the pathway segments found indeed correspond to specific conserved cellular functions, it was observed that the network regions were significantly functionally enriched for particular protein functional categories from the Munich Information Center for Protein Sequences (MIPS - http://mips.gsf.de) for yeast, and the Institute for Genomic Research (TIGR - www.tigr.org) for bacteria.

In addition to recognition of conserved pathways between the two PPI networks, other insights can be observed from the results of this work. For example, due to the certain degree of freedom allowed in the alignment process, a conserved pathway can be found even though it includes a node corresponding two proteins from the original network which are not known to be similar. This might imply that the functionality of these two proteins might be similar, using only their location in the pathway topology, thus we can deduct from the functionality of the protein we know of to the one we don’t, a fact that can be also biologically validated.

Another insight we can deduct, is relation between seemingly unrelated processes, corresponding an aligned pathway of the two processes which are performed by aligned proteins in conserved structures.

Figure 6.

3.1.5 Yeast vs. Yeast: Paralogous Pathways Within the Network of a single Species

In addition to identifying homologous features between the protein networks of yeast and bacteria, a search was also performed within each network individually to identify its potentially paralogous pathways, that is, pathways with proteins and interactions that have been duplicated one or more times in the course of evolution.

Such an approach is similar to performing an ”all vs. all” BLAST of sequences encoded by a single genome in order to find gene families. This procedure was explored in the context of yeast, by constructing a global alignment graph merging the yeast protein interaction network with an identical copy of itself.

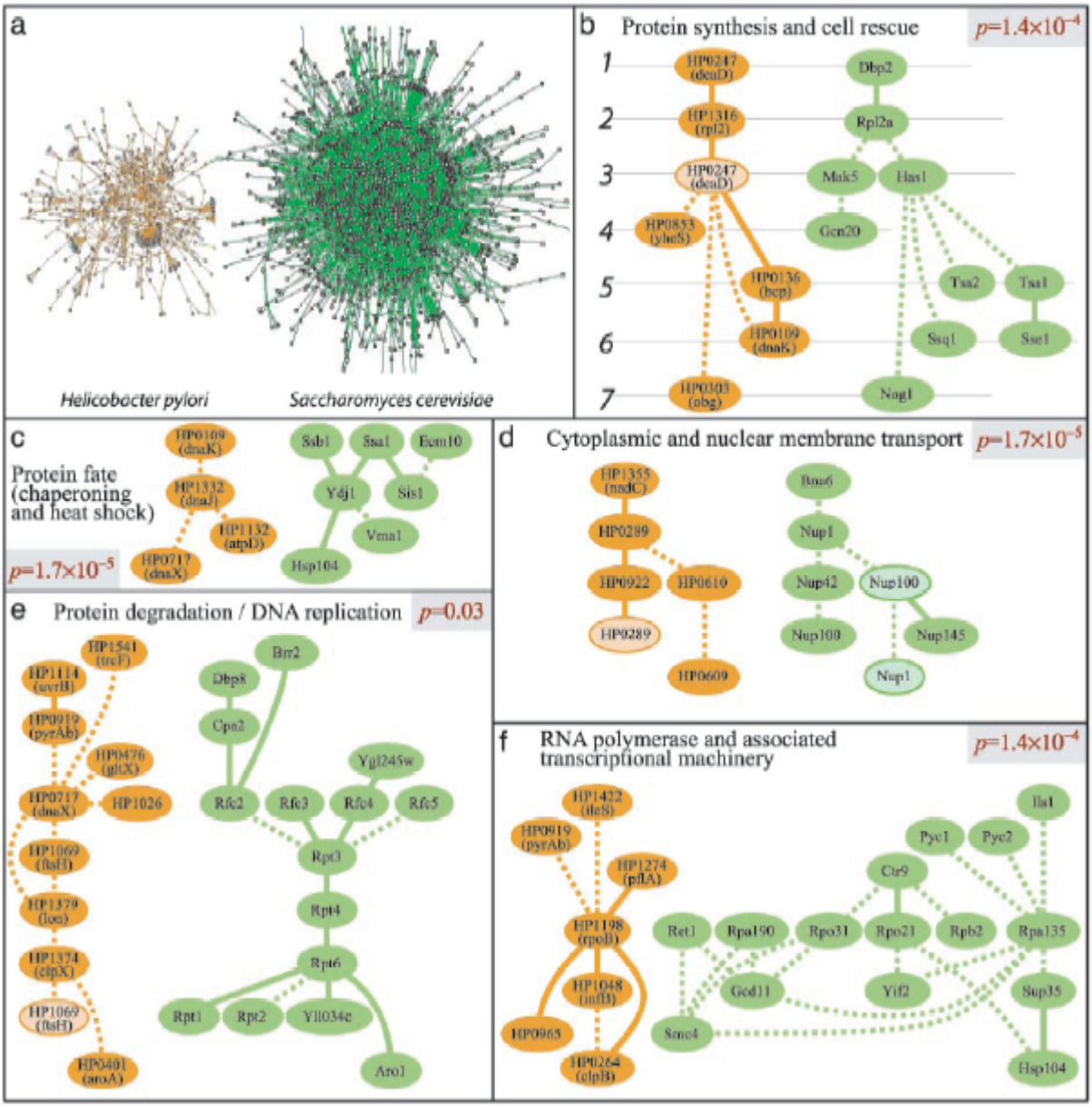

Figure 6: Source [10]. This figure shows the top scoring alignments between yeast (green vertices) vs. bacteria (orange vertices) networks, their functional annotation and their p-value. (a) Both PPI networks. (b) Protein synthesis and cell rescue functionality. (c) Protein fate (chaperoning and heat shock) functionality. (d) Cytoplasmic and nuclear membrane transport. (e) Protein degradation / DNA replication. (f) RNA polymerase and associated transcriptional machinery.

To ensure that pathway alignments occurred between two distinct network regions and to avoid aligning a path with its exact copy, proteins were not allowed to pair with themselves or their network neighbors. The resulting graph wa analyzed to obtain the 300 highest-scoring pathway alignments of length four, corresponding to a level of significance of $p \leq 0.0001$. Several regions involve alignments between protein complexes with related functions, confirming that the approach is capable of identifying paralogous network structures (Figure 7).

3.1.6 Yeast vs. Yeast: Interrogating the Protein Network with Pathway Queries

The last experiment was to query a single protein network with specific pathways of interest. Using of PathBLAST in this mode is similar to using BLAST to interrogate a sequence database with a short nucleotide or amino acid sequence query.

The yeast protein network was queried with a classic MAPK pathway associated with the filamentation response, consisting of a MAPK (Ste11), a MAPK kinase (Ste7), and a MAPK kinase (Kss1). MAPK pathways transmit incoming signals to the nucleus through activation cascades in which each kinase phosphorylates the next one downstream. PATHBLAST identified two other well known MAPK pathways as the highest-scoring hits (the low-and high-osmolarity response pathways Bck1-Mkk1-Slt2 and Ssk2-Pbs2-Hog1), indicating that the algorithm was sufficiently sensitive and specific to identify known paralogous pathways.

This strategy was repeated to search for new components of the cellular ubiquitin and ubiquitin-like conjugation machinery. Ubiquitin targets proteins for degradation by the proteasome and modifies different sets of proteins through distinct pathways, some of which are unknown. These tests showed that pathway-based queries using PathBLAST are capable of identifying both known and potentially novel paralogous pathways within an organism. The pathways searched and the results are shown in Figure 8.

3.2 Identifying Conserved Protein Complexes

The previous section handled the problem of finding conserved linear pathways. It is not uncommon for such pathways to overlap, the following heuristic deals with those overlaps ending up identifying more complex conserved structures. First, and PathBLAST is used to find conserved paths and then overlapping paths are merged into complexes. An example of this is shows in Figure 9, where a conserved complex is found using two conserved intersecting pathways.

This section describes a direct approach for identifying conserved complexes. Sharan et al. [18] introduced a method for finding conserved complexes by comparative analysis of two PPI networks. This work assumed *protein complexes* to be manifested as dense subgraphs (*Clusters*). Indeed, in order for a complex to act as single mechanism, all it's proteins should be connected between themselves. Moreover, the average density of currently known complexes is around 0.4 (40% of all possible interactions exist).

3.2.1 A Probabilistic Model for Protein Complexes

To measure how good a complex is, a likelihood ratio is used. The measure looks at the ratio between the likelihood of the complex to exist assuming all its proteins interact with each other, and the likelihood of the complex to exist assuming a random distribution of the protein interactions in the graph.

The two models are defined as follows:

1. The protein-complex model, M_c - assumes that every two proteins in a complex interact with some high probability p (0.8 is used in this work). In terms of the graph, the assumption is that two vertices that belong to the same complex are connected by an edge with probability p, independently of all other information.

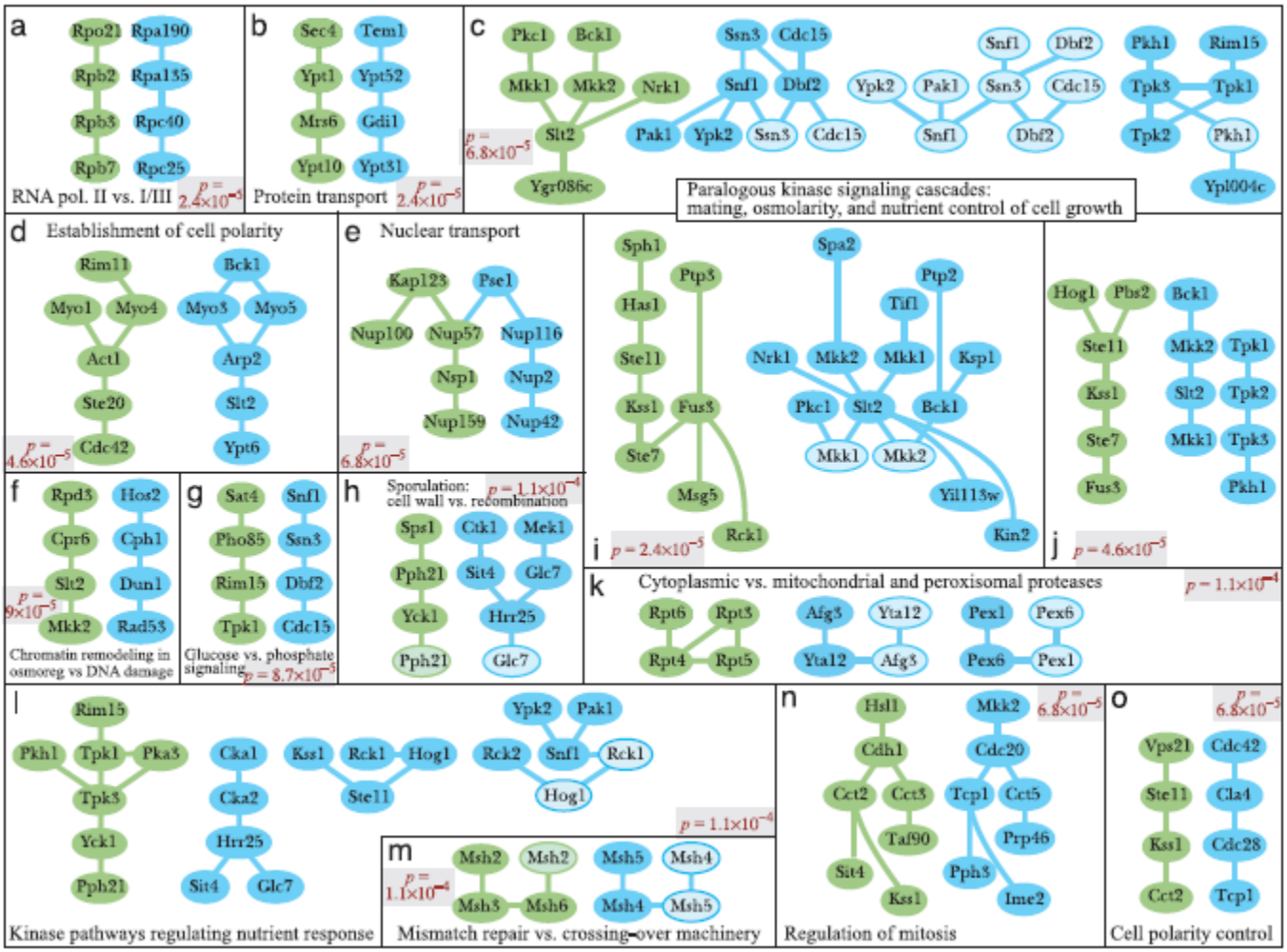

Figure 7: Source [10]. This figure shows paralogous pathways within the yeast, by merging the yeast network with itself, and searching pathways. Each side pathway is drawn in a different color(green/blue). The different regions in the figure shows top scoring alignments, their functional annotation, and their p-value. (a) RNA polymerase II vs. I/III. (b) Protein transport. (c+i+j) Paralogous kinase signaling cascades: mating, osmolarity, and nutrient control of cell growth. (d) Establishment of cell polarity. (e) Nuclear transport. (f) Chromatin remodeling in osmoreg vs. DNA damage. (g) Glucose vs. phosphate signaling. (h) Sporulation: cell wall vs. recombination. (k) Cytoplasmic vs. mitochondrial and peroxisomal proteases. (l) Kinase pathways regulating nutrient response. (m) Mismatch repair vs. crossing-over machinery. (n) Regulation of mitosis. (o) Cell polarity control.

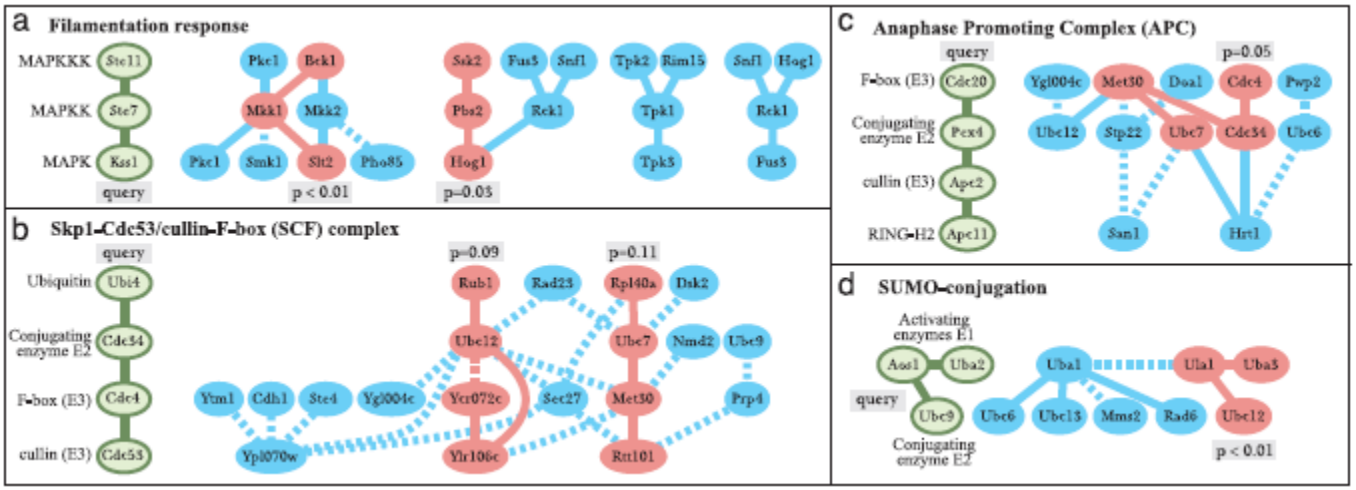

Figure 8: Source [10]. This figure shows the results of querying the yeast network with specific pathways. The different regions in the figure shows top scoring alignments (each sub-figure for each pathway queried). The high-scoring alignments are indicated in red. The p-value is also shows for each alignment. (a) Filamentation response (b) Skp1-Cdc53/cullin-F-Box (SCF) complex. (c) Anaphase Promoting Complex (APC). (d) SUMO-conjunction.

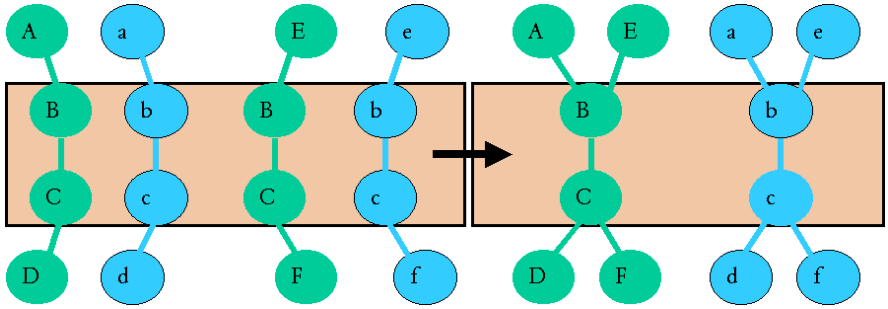

Figure 9: This figure shows two pairs of aligned paths from two networks on the left side, and the way they are joined into aligned complexes on the right.

2. The random 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, let F^G represent the family of all graphs having the same vertex set as G and the same degree sequence. The probability of observing the edge (u, v), $p(u, v)$, is defined 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 assume that all pairwise relations are independent.

Given a protein complex $C = (V', E')$, a naive approach could be to define this complex score as follows:

$$L(C) = \prod_{(u,v)\in E'} \frac{p}{p(u,v)} \times \prod_{(u,v)\notin E'} \frac{1-p}{1-p(u,v)} \tag{2}$$

It can be easily seen that complexes with higher density will have more edges and thus higher scores. However, such a score ignores information on the reliability of interactions. A more rigorous scoring would treat data of interactions as noisy observations of interactions. In other words, we will incorporate the edge confidence scores into our complex score to deal with noisy data.

Let T_{uv} denote the event that two proteins u, v interact, and F_{uv} denote the event that they do not interact. O_{uv} denote the (possibly empty) set of available observations on the proteins u and v , that is, the set of experiments in which u and v were tested for interaction and the outcome of these tests. Using prior biological information (see Section 4.1 of [18]), 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, the likelihood of U under a protein-complex model (M_c) and a random model (M_n) is computed. Denoting by O_U the collection of all observations on vertex pairs in U, the probability that this collection of observations will occur under the complex model can be computed as follows:

$$Pr(O_U|M_c) = \prod_{(u,v)\in U\times U} (pPr(O_{uv}|T_{uv}) + (1-p)Pr(O_{uv}|F_{uv})) \tag{3}$$

and the probability that this collection of observations will occur in the random model can be computed as follows:

$$Pr(O_U|M_n) = \prod_{(u,v)\in U\times U} (p(u,v)Pr(O_{uv}|T_{uv}) + (1-p(u,v))Pr(O_{uv}|F_{uv})) \tag{4}$$

therefore, the log likelihood score of a complex C can be calculated as follows:

$$L(C) = \prod_{(u,v)\in U\times U} \frac{Pr(O_U|M_c)}{Pr(O_U|M_n)} = \frac{pPr(O_{uv}|T_{uv}) + (1-p)Pr(O_{uv}|F_{uv})}{p(u,v)Pr(O_{uv}|T_{uv}) + (1-p(u,v))Pr(O_{uv}|F_{uv})} \tag{5}$$

3.2.2 Scoring for Two Species

Consider C and $C^{'}$ two network subsets, one for each species, and a mapping θ between them. Then, we can compute the likelihood score as follows:

$$L(C, C^{'}) = L(C)L(C^{'}) \tag{6}$$

But, this does not take into account the degree of sequence conservation among the pairs of proteins mapped by θ. In order to include such information, a conserved complex model and a random model for pairs of proteins from two species were defined. Let E_{uv} denote the BLAST E-value assigned to the

similarity between proteins u and v, and let h_{uv}, h'_{uv} denote the events that u and v are orthologous, or nonorthologous, respectively. The likelihood ratio corresponding to a pair of proteins (u, v) is therefore:

$$L(C, C') = L(C)L(C') \prod_{u,v-matched} \frac{Pr(E_{uv}|h_{uv})}{Pr(E_{uv}|h_{uv})Pr(h_{uv}) + Pr(E_{uv}|h'_{uv})Pr(h'_{uv})} \quad (7)$$

A downside of this scoring method is that it treats the aligned complexes independently, meaning that it ignores the preservation of interactions between the complexes. Nevertheless, because most of currently available PPI networks originate from evolutionary distant species, this scoring produces similar results as other, more complex methods, which do incorporate interaction preservation scores.

3.2.3 Searching Conserved Protein Complexes

Using the model explained above, the problem of identifying conserved protein complexes reduces to the problem of finding heavy sub-graphs in the alignment graph.

3.2.4 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 (Shamir *et al.* [16]). Thus, heuristic strategies for searching the alignment graph for conserved complexes were proposed.

A bottom-up search for heavy subgraphs in the alignment graph is performed, by starting from high weight seeds, refining them by exhaustive enumeration, and then expanding them using local search. An edge in the alignment graph is defined as *strong* if the sum of its associated weights (the edge weights within each species graph) is positive.

The search proceeds as follows:

1. Compute a seed around each node v, which consists of v and all its neighbors u such that (u, v) is a strong edge
2. If the size of this set is above a threshold (e.g.,10), iteratively remove from it the node whose contribution to the subgraph score is minimum, until we reach the desired size.
3. Enumerate all subsets of the seed that have size at least 3 and contain v . Each such subset is a refined seed on which a local search heuristic is applied.
4. Local search: 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 the original refined seed is preserved and nodes are not deleted from it.
5. For each node in the alignment graph record up to k (e.g. 5) heaviest subgraphs that were discovered around that node.

Notice that the resulting subgraphs may overlap considerably. In order to solve that a greedy algorithm is used to filter subgraphs whose percentage of intersection is above a threshold as follows:

1. Iteratively find the highest weight subgraph.
2. Add that subgraph to the final output list.

			Yeast enrichment		Bacterial enrichment	
ID	*Score*	*Size*	*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 (1)
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)

Figure 10: Source [18]. This figure shows the conserved protein complexes identified between the yeast and bacteria. Table columns: ID (the ID of the complex), Score ($-ln(p_{value})$ adjusted for multiple testing), size (with the number of distinct bacterial and yeast proteins in parentheses), purity and complex category for the yeast, and purity and functional category for the bacteria.

3. Remove all other highly intersecting subgraphs (large overlap between two complexes was disallowed by filtering complex that has 60 percent or more than other complex, and its p-value is worse than the other complex). This p-value of a complex measures the fraction of random runs in which the output complex had higher score, as explained next.

3.2.5 Evaluation of Complexes

The statistical significance of identified complexes was tested in two ways:

1. The first is based on the z-scores that are computed for each subgraph and assumes a normal approximation to the likelihood ratio of a subgraph. The approximation relies on the assumption that the subgraphs nodes and edges contribute independent terms to the score. The latter probability is Bonferroni ([6]) corrected for multiple testing, according to the size of the subgraph.

2. The second 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, a heuristic search is used to find the highest-scoring conserved complex of a given size. Then a p-value is estimated for a suggested complex of the same size, as the fraction of random runs in which the output complex had higher score.

3.2.6 Complex Identification and Validation

The algorithm was applied to yeast and bacteria networks, identifying 11 nonredundant complexes, with significant p-value < 0.05 (after the correction for multiple testing). The score was also compared against empirical runs on randomized data ($p < 0.05$). The results are listed in Figure 10.

In order to validate these results the MIPS database was used (www.mips.gsf.de) that contains assignment of yeast genes to known complexes. The *purity* of a complex was defined as follows: denote by x the highest number of proteins from a single complex category in MIPS. Denote by y all the proteins in the complex that are categorized members in MIPS; thus, the purity was calculated as x/y.

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 indicate either an incorrect complex or a previously unidentified correct one. Note that most of the predicted complexes also contain proteins that are not known to belong to any complex in yeast. Thus, the results could be used in order to suggest additional members in known complexes.

For bacteria, since experimental information on complexes is unavailable, functional annotations were used in order to calculate the purity of the complex. The functional annotations which were taken from the TIGR database (www.tigr.org).

The significant complexes that have been identified exhibit a nice correspondence between the protein complex annotation in yeast and the functional annotation in bacteria, as presented in Figure 10 and further visualized in Figure 11. 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.

The conserved protein complexes that were found imply new functions for a variety of uncharacterized proteins. For instance, complex 17 (Figure 11(a)) defines a set of conserved interactions for the cells protein degradation machinery. Bacterial proteins HP0849 and HP0879 (emphasized in Figure 11(a)) are uncharacterized, but their appearance within yeast and bacterial complexes involved in proteolysis suggests that they also play an important role in this process. Another example is the yeast proteins Hsm3 and Rfa1 (with known functional roles in DNA-damage repair) that may also be associated with the yeast proteasome (see Figure 11(b,d)).

As another example of protein functional prediction, Figure 11(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. The 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, or perhaps are part of some sort of ancient predecessor of the yeast NUP complex.

3.2.7 Comparison to Previous Methods

The authors performed a comparison between three methods/applications:

- Yeast vs. Bacteria: the algorithm was applied to the yeast bacteria alignment graph in search of conserved complexes.
- Yeast only: this method is a noncomparative variant of the algorithm that uses the protein-protein interactions in yeast only. 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.
- A variant of the algorithm that relies on the previous probabilistic model for protein interactions by Kelley et al. [10].

And the following measures were used in order to compare the experiments:

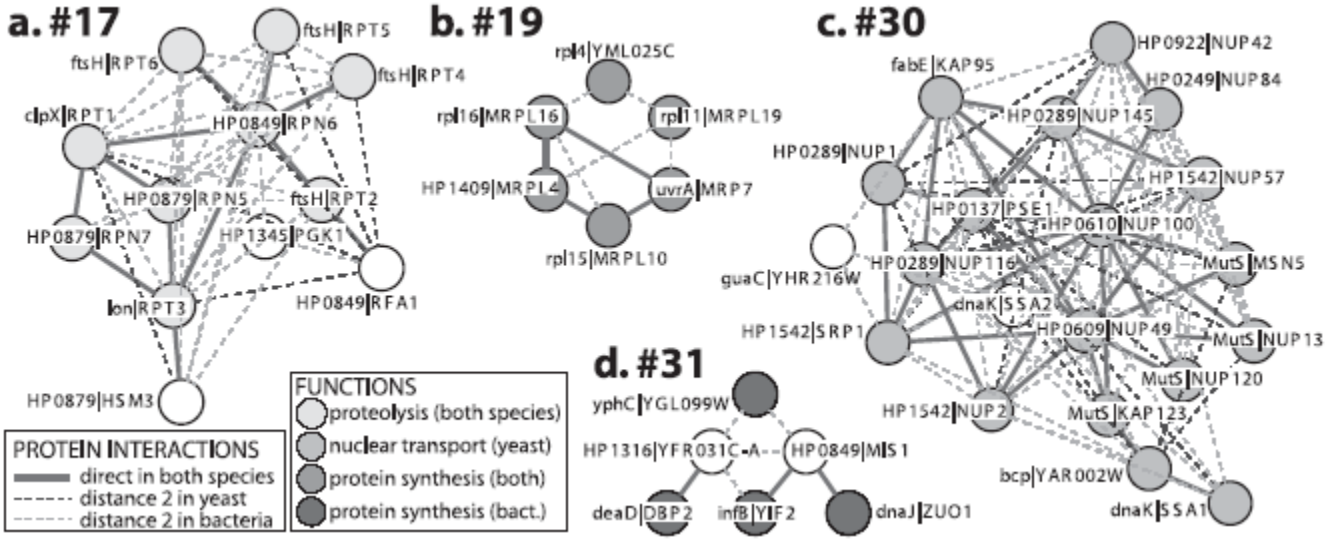

Figure 11: Source [18]. This figure shows conserved protein complexes: (a) proteolysis complexes, (b,d) protein synthesis complexes, (c) nuclear transport complexes. Conserved complexes are connected subgraphs within the bacteria-yeast alignment graph, whose nodes represent orthologous protein pairs and edges represent conserved protein interactions of three types: direct interactions in both species (solid edges); direct in bacteria but distance 2 in the yeast interaction graph (dark dashed edges); and distance 2 in the bacterial interaction graph but direct in yeast (light dashed edges). The number of each complex indicates the corresponding complex ID listed in figure 10.

Algorithm	*Jaccard*	*Sensitivity*	*Specificity*
This study	0.32	0.33	0.7
Kelley *et al.* (2003)	0.22	0.44	0.4
Yeast only	0.33	0.67	0.48

Figure 12: Source [18]. This figure shows a comparison between the three experiments using three comparison measures: the *Jaccard* measure, the *Sensitivity* measure and the *Specificity* measure.

- The *Jaccard* measure: Two proteins are called mates in a solution if they appear together in at least one complex in that solution. Given two solutions, let n_{11} be the number of pairs that are mates in both, and let n_{10} (n_{01}) 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 (for more on Jaccard score see [9]).
- 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 of the three approaches is shown in Figure 12. Analysis of the results shows that:

- The Jaccard score is significantly better in the current approach than in Kelley et al. [10].
- The sensitivity is lower, as fewer categories are captured, but the specificity is much higher, so the predicted complexes are much more accurate.
- Using data on yeast only, Sharan et al. 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 the new probabilistic model can be effectively used, even for detecting complexes using interaction data from a single species.
- The results of the yeast vs. bacteria experiment were evaluated 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.

3.3 Evolutionary Based Scoring

The methods above for scoring and searching for conserved complexes do not take into account the evolutionary process shaping protein interaction. Koyuturk et al. introduce a scoring method which is based on the *duplication/divergence* model (see [11]).

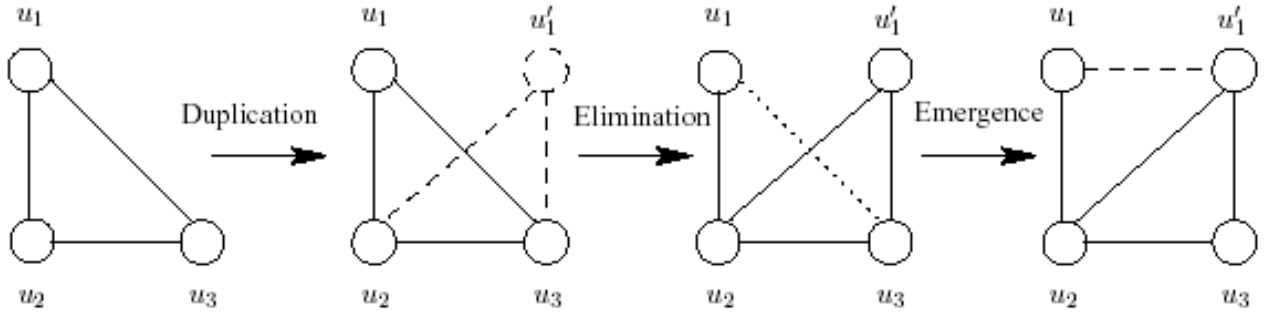

Figure 13: Source [11]. This figure shows duplication, elimination and emergence events on a PPI network. Starting with three interactions between proteins u_1,u_2 and u_3. Then, node u_1 is duplicated to node $u_1^{'}$, together with its interactions (dashed circle and lines). Then, node u_1 loses its interaction with u_3 (elimination - dotted line). Finally, an interaction between u_1 and $u_1^{'}$ is added to the network (emergence - dashed line).

3.3.1 The Duplication/Divergence Model

The *duplication/divergence* model is a common model used to explain the evolution of protein interaction networks via preferential attachment. According to this model, when a gene is duplicated in the genome, the node corresponding to the product of this gene is also duplicated together with its interactions. An example of protein duplication is shown in Figure 13. A protein loses many aspects of its functions rapidly after being duplicated. This translates to divergence of duplicated (paralogous) proteins in the interactome through *elimination* and *emergence* of interactions:

- *Elimination* of an interaction in a PPI network implies the loss of an interaction between two proteins due to mutations int their interface.
- *Emergence* of an interaction in a PPI network implies the introduction of a new interaction between two non-interacting proteins, again caused by mutations that change protein surfaces.

Examples of elimination and emergence of interactions are also illustrated in Figure 13.

If an elimination or emergence is related to a recently duplicated protein, it is said to be *correlated*; otherwise, it is *uncorrelated* ([14]). Since newly duplicated proteins are more tolerant to interaction loss because of redundancy, correlated elimination is generally more probable than emergence and uncorrelated elimination ([24]). In the context of duplication two types of pairs of proteins are defined as follows:

- A pair of proteins from different species will be called *in-paralogs*, if they are the result of duplication that occurred before a speciation event.
- A pair of proteins from different species will be called *out-paralog*, if they are the result of a duplication that occurred after a speciation event.

The interaction profiles of duplicated proteins tend to almost totally diverge in about 200 million years, as estimated on the yeast interactome. On the other hand, the correlation between interaction profiles of duplicated proteins is significant for up to 150 million years after duplication, with more than half of interactions being conserved for proteins that are duplicated less than 50 million years back. Thus, while comparatively analyzing the proteome and interactome, it is important to distinguish in-paralogs from out-paralogs since the former are more likely to be functionally related. This, however, is a difficult task since out-paralogs also show sequence similarity.

3.3.2 Local Alignment of the PPI Network

Given two PPI networks $G(U, E)$ and $H(V, F)$, a protein subset pair $P = \{\tilde{U}, \tilde{V}\}$ is defined as a pair of protein subsets $\tilde{U} \subseteq U$ and $\tilde{V} \subseteq V$. Any protein subset P induces a local alignment $A(G, H, S, P) = \{M, N, D\}$ of G and H with respect to S, which is the similarity function between each pair of proteins in $U \cup V$:

- M - set of matches. A *match* corresponds to a conserved interaction between two orthologous protein pairs, which is rewarded by a match score that reflects the confidence in both protein pairs being orthologous.
- N - set of mismatches. A *mismatch* is the lack of an interaction in the PPI network of one organism between a pair of proteins whose orthologs interact in the other organism. A mismatch may correspond to the emergence of a new interaction or the elimination of a previously existing interaction in one of the species after the split, or an experimental error. Thus, mismatches are penalized to account for the divergence from the common ancestor.
- D - set of duplications. A *duplication* is the duplication of a gene in the course of evolution. Each duplication is associated with a score that reflects the divergence of function between the two proteins, estimated using their similarity.

Let functions $\triangle_G(u, u')$ and $\triangle_H(v, v')$ denote the distance between two corresponding proteins in the interaction graphs G and H, respectively. Given a pairwise similarity function S, a *distance cutoff* $\bar{\triangle}$, and the set P from above, we get:

$$M = \{u, u' \in \tilde{U}, v, v' \in \tilde{V} : S(u, v) > 0, S(u', v') > 0, \\ ((uu' \in E \wedge \triangle_H(v, v') \leq \bar{\triangle}) \vee (vv' \in F \wedge \triangle_G(u, u') \leq \bar{\triangle}))\} \tag{8}$$

$$N = \{u, u' \in \tilde{U}, v, v' \in \tilde{V} : S(u, v) > 0, S(u', v') > 0, \\ ((uu' \in E \wedge \triangle_H(v, v') > \bar{\triangle}) \vee (vv' \in F \wedge \triangle_G(u, u') > \bar{\triangle}))\} \tag{9}$$

$$D = \{u, u' \in \tilde{U} : S(u, u') > 0\} \cup \{v, v' \in \tilde{V} : S(v, v') > 0\} \tag{10}$$

Following the definition of match and mismatch we see that not only direct but also indirect interactions are allowed. If two proteins directly interact with each other in one organism, and their orthologs are reachable from each other via at most $\bar{\triangle}$ interactions in the other (the value $\bar{\triangle} = 2$ is used), it is considered as a match. Conversely, a mismatch corresponds to the situation in which two proteins are not reachable via $\bar{\triangle}$ interactions in one network while their orthologs directly interact in the other.

There are two observations that explain the use of the distance cutoff :

1. Proteins that are linked by a short alternate path are more likely to tolerate interaction loss because of relaxation of evolutionary pressure ([11]).
2. High-throughput methods such as TAP ([7]) identify complexes that are associated with a single central protein and these complexes are recorded in the interaction database as star networks with the central protein serving as a hub.

3.3.3 Scoring Match, Mismatch and Duplications

For scoring the matches and mismatches, the similarity between two protein pairs is defined as follows:

$$S(uu^{'}, vv^{'}) = S(u, v)S(u^{'}, v^{'}) \tag{11}$$

The similarity value is calculated using Inparanoid ([15]), which is a sequence-based method for finding orthology relations. It uses clustering in order to derive orthology families, leaving some of the orthology relations ambiguous. $S(uu^{'}, vv^{'})$ quantifies the likelihood that the interactions between u and v, and $u^{'}$ and $v^{'}$ are orthologous. Consequently, a match that corresponds to a conserved pair of orthologous interactions is rewarded as follows:

$$\mu(uu^{'}, vv^{'}) = \bar{\mu}S(uu^{'}, vv^{'}) \tag{12}$$

Here, $\bar{\mu}$ is the match coefficient that is used to tune the relative weight of matches against mismatches and duplications, based on the evolutionary distance between the species that are being compared.

A mismatch may correspond to the functional divergence of either interacting partner after speciation. It might also be due to a false positive or negative in one of the networks that is caused by incompleteness of data or experimental error. However, this problem was already solved by considering indirect interactions as matches. According to Wagner ([26]), after a duplication event, duplicate proteins that retain similar functions in terms of being part of similar processes are likely to be part of the same complex. Moreover, since conservation of proteins in a particular module is correlated with interconnectedness ([28]), we expect that interacting partners that are part of a common functional module will at least be linked by short alternative paths. Based on these observations, mismatches are penalized for possible divergence in function as follows:

$$v(uu^{'}, vv^{'}) = -\bar{v}S(uu^{'}, vv^{'}) \tag{13}$$

As for match score, mismatch penalty is also normalized by a coefficient $\bar{v}$, that determines the relative weight of mismatches w.r.t. matches and duplications.

A duplication has an evolutionary significance. Since duplicated proteins rapidly lose their interactions, it is more likely that in-paralogs, i.e., the proteins that are duplicated after a speciation event, will share more interacting partners than out-paralogs do ([26]). Furthermore, sequence similarity is employed as a means for distinguishing in-paralogs from out-paralogs. This is based on the observation that sequence similarity provides a crude approximation for the age of duplication ([27]). Moreover, recently duplicated proteins are more likely to be in-paralogs, and thus show more significant sequence similarity than older paralogs. Therefore, duplicate score is defined as follows:

$$\delta(u, u^{'}) = \bar{\delta}(S(u, u^{'}) - \bar{d}) \tag{14}$$

Here $\bar{d}$ is the cutoff for being considered in-paralogs. If $S(u, u^{'}) > \bar{d}$, suggesting that u and $u^{'}$ are likely to be in-paralogs, the duplication is rewarded by a positive score. If, on the other hand, $S(u, u^{'}) < \bar{d}$, the proteins are considered out-paralogs, thus the duplication is penalized.

3.3.4 Alignment Score and the Optimization Problem

Given PPI networks G and H, the score of alignment $A(G, H, S, P) = M, N, D$ is defined as:

$$\sigma(A) = \sum_{m \in M} \mu(m) + \sum_{n \in N} v(n) + \sum_{d \in D} \delta(d) \tag{15}$$

The PPI network alignment problem is one of finding all maximal protein subset pairs P such that $\sigma(A(G, H, S, P))$ is locally maximal, i.e. the alignment score cannot be improved by adding individual proteins to or removing proteins from P. The aim is to find local alignments with locally maximal score.

The information regarding matches, mismatches and duplications of the two PPI networks is represented using a single weighted alignment graph: Given $G(U, E), H(V, F)$, and protein similarity function S, the corresponding weighted alignment graph $G(\bar{V}, \bar{E})$ is computed as follows:

$$\bar{V} = \{\bar{v} = \{u, v\} : u \in U, v \in V \text{ and } S(u, v) > 0\} \tag{16}$$

In other words, there is a node in the alignment graph for each pair of putatively ortholog proteins. Each edge $\bar{v}\bar{v}^{'}$, where $\bar{v} = \{u, v\}$ and $\bar{v}^{'} = \{u^{'}, v^{'}\}$, is assigned a weight:

$$w(\bar{v}, \bar{v}^{'}) = \mu(uu^{'}, vv^{'}) + v(uu^{'}, vv^{'}) + \delta(u, u^{'}) + \delta(v, v^{'}) \tag{17}$$

Here, $\mu(uu^{'}, vv^{'}) = 0$ if $(uu^{'}, vv^{'}) \notin M$ and the same for mismatches and duplications.

They used a greedy search heuristic in order to find the conserved complexes in the alignment graph. For more information on this heuristic refer to section 3.3 in [11].

3.3.5 Significance Evaluation

To evaluate the statistical significance of discovered high-scoring alignments, a comparison is made between the alignments and a reference model generated by a random source. In the reference model, it is assumed that the interaction networks of the two species are independent of each other. In order to assess the significance of conservation of interactions between orthologous proteins rather than the conservation of proteins itself, it is assumed that the orthology relationship between protein is already established, i.e., is not generated by a random source. Other interactions are generated randomly while preserving the degree sequence.

Given proteins u and $u^{'}$, that are interacting with d_u and $d_{u'}$ proteins, respectively, then the probability $p_{uu'}$ can be estimated as:

$$p_{uu'} = \frac{d_u d_{u'}}{\sum_{v \in U} d_v} \tag{18}$$

Recall that the weight of a subgraph of the alignment graph is equal to the score of the corresponding alignment, therefore, in the reference model, the expected value of the score of an alignment induced by $\tilde{V} \subseteq V$ is :

$$E[W(\tilde{V})] = \sum_{v, v^{'} \in \tilde{V}} E[w(vv^{'})] \tag{19}$$

where

$$E[w(vv^{'})] = \bar{\mu} S(uu^{'}, vv^{'}) p_{uu'} p_{vv'} - \bar{v} S(uu^{'}, vv^{'})(p_{uu'}(1 - p_{vv'}) + (1 - p_{uu'})p_{vv'}) + \delta(u, u^{'}) + \delta(v, v^{'}) \tag{20}$$

is the expected weight of an edge in the alignment graph. With the simplifying assumption of independence of interactions, they have

$$Var[W(\tilde{V})] = \sum_{v, v^{'} \in \tilde{V}} Var[w(vv^{'})] \tag{21}$$

enabling them to compute the z-score to evaluate the statistical significance of each discovered high-scoring alignment, under the normal approximation that is assumed.

Species	No. of proteins	No. of Interactions
S. Cerevisiae	5157	18192
C. Elegans	3345	5988
D. Melanogaster	8577	28829

Table 1: Source [11]. Number of proteins and interactions for yeast (S. Cerevisiae), worm (C. Elegans) and fly (D. Melanogaster).

Organism pair	# Nodes	# Matched nodes		# Matches		# Mismatches	# Duplications	
		$\bar{\Delta}=1$	$\bar{\Delta}=2$	$\bar{\Delta}=1$	$\bar{\Delta}=2$	$\bar{\Delta}=1$	Org. 1	Org. 2
SC vs CE	2746	312	1230	412	3007	40262	6107	6886
SC vs DM	15884	1730	8622	2061	42781	1054241	6107	32670
CE vs DM	11805	491	3391	455	6626	205593	6886	32670

Figure 14: Source [11]. The number of nodes, matched nodes, matches, mismatches and duplications, for each experiment done: SC vs. CE (yeast vs. worm), SC vs. DM (yeast vs. fly) and CE vs. DM (worm vs. fly). It shows the data both for $\bar{\triangle}=1$ and $\bar{\triangle}=2$.

3.3.6 Experimental Results

The interaction data that was used was downloaded from BIND ([3]) and DIP ([29]) molecular interaction databases. The statistics for the PPI networks of yeast (S. Cerevisiae), worm (C. Elegans) and fly (D. Melanogaster) are shown in Table 1.

They performed pairwise alignments of the three pairs of PPI networks, using the following alignment parameters: $\bar{\mu}=1.0$, $\bar{v}=1.0$ and $\bar{\delta}=0.1$. The alignment was done between yeast-worm, yeast-fly, and worm-fly, for both for $\bar{\triangle}=1$ and $\bar{\triangle}=2$. The results are shown in Figure 14.

Alignment of yeast PPI network with fly PPI network results in identification of 412 conserved subnetworks. Ten of the conserved subnetworks with highest alignment scores are shown in Figure 15. In total, 83 conserved subnetworks are identified on yeast-worm alignment, and 146 are identified on worm-fly alignment.

While most of the conserved subnetworks are dominated by one particular processes and the dominant processes are generally consistent across species, there also exist different processes in different organisms that are mapped to each other by the discovered alignments. This illustrates that the comparative analysis of PPI networks is effective in not only identifying particular functional modules, pathways, and complexes, but also in discovering relationships between different processes in separate organisms and crosstalk between known functional modules and pathways. Moreover, alignment results provide a means for discovery of new functional modules in relatively less studied organisms through mapping of functions at a modular level rather than at the level of single protein homologies. These significant use of the experiments results was also noticed by Both Kelley *et al.* with PathBLAST ([10]), and Sharan *et al.* ([18]).

A selection of interesting conserved subnetworks is shown in Figure 16. The alignments in the figure illustrate that the alignment algorithm takes into account the conservation of interactions in addition to sequence similarity while mapping orthologous proteins to each other. In all of the alignments shown in

Rank	Score	z-score	# Proteins	# Matches	# Mismatches	# Duplications
1	15.97	6.6	18 (16, 5)	28	6	(4, 0)
	protein amino acid phosphorylation (69%) /JAK-STAT cascade (40%)					
2	13.93	3.7	13 (8, 7)	25	7	(3, 1)
	endocytosis (50%) / calcium-mediated signaling (50%)					
5	8.22	13.5	9 (5, 3)	19	11	(1, 0)
	invasive growth (sensu Saccharomyces) (100%) / oxygen and reactive oxygen species metabolism (33%)					
6	8.05	7.6	8 (5, 3)	12	2	(0, 1)
	ubiquitin-dependent protein catabolism (100%) / mitosis (67%)					
8	6.83	12.4	6 (4, 4)	12	6	(0, 1)
	protein amino acid phosphorylation (50%, 50%)					
10	6.75	13.7	10 (7, 3)	24	12	(0, 1)
	ubiquitin-dependent protein catabolism (100%)					
14	5.69	8.7	11 (11, 2)	10	1	(0, 0)
	regulation of progression through cell cycle (9%, 50%)					
21	4.36	6.2	9 (5, 4)	18	13	(0, 5)
	cytokinesis (100%, 50%)					
22	4.22	3.9	7 (6, 6)	9	5	(1, 1)
	protein folding (67%, 17%)					
30	3.76	39.6	6 (3, 5)	5	1	(0, 6)
	DNA replication initiation (100%, 80%)					

Figure 15: Source [11]. This figure shows the representative top-scoring subnetworks identified by the alignment of yeast and fly. The dominant biological process/functionality for each species, in which the majority of proteins in the conserved subnetwork participate is also shown in the second raw of each subnetwork. For each subnetwork we it also show the z-score, number of proteins (for each species in parenthesis), number of matches, mismatches and duplications (for each species in parenthesis).

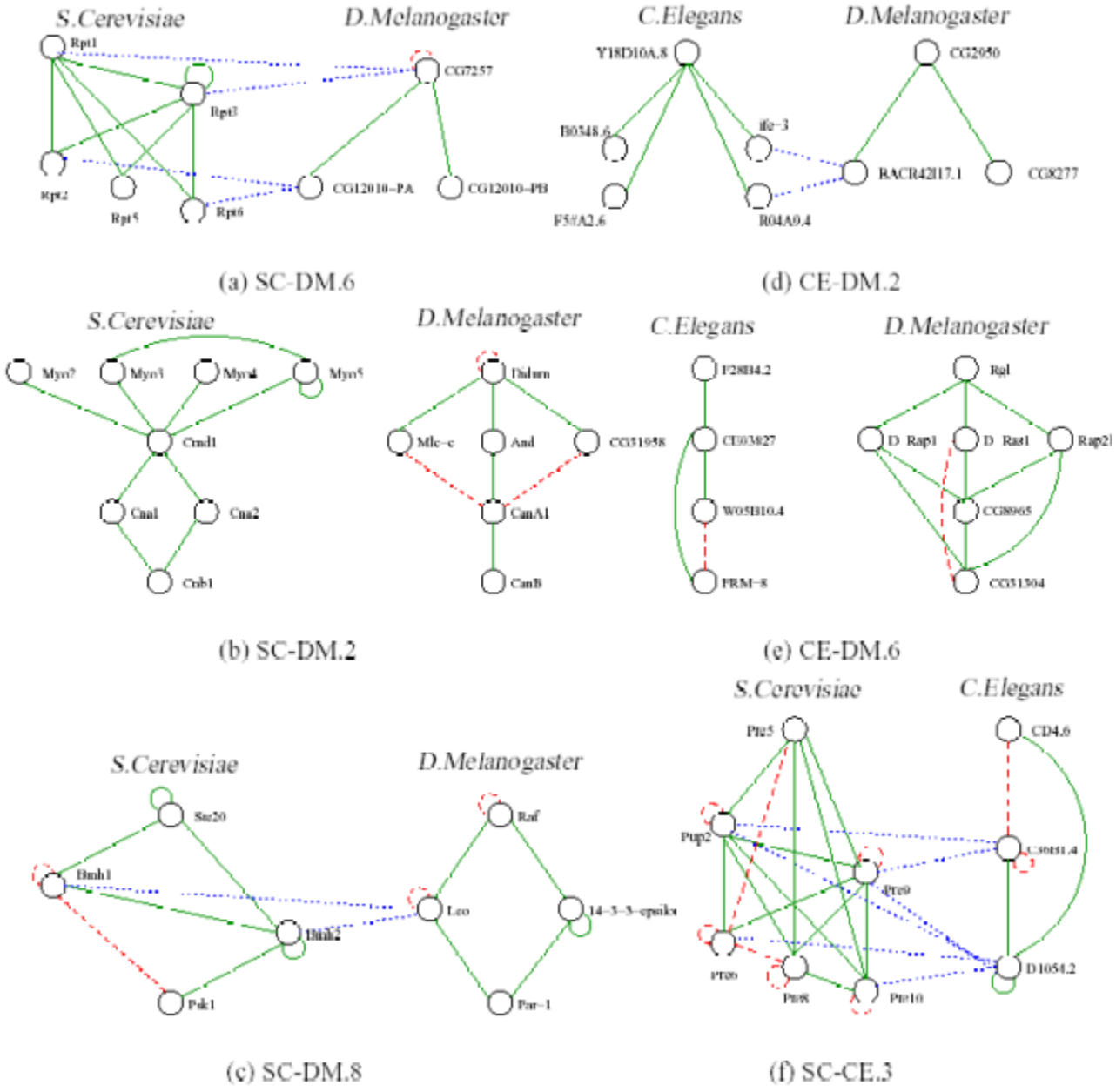

Figure 16: Source [11]. This figure shows a sample of conserved subnetworks identified by the alignment algorithm. Orthologous and paralogous proteins are either vertically aligned, or connected by blue dotted lines. Existing interactions are shown by green solid lines, and missing interactions that have an orthologous counterpart are shown by red dashed lines. The organisms aligned and the rank of the alignment are shown in the label. (a,b,c) yeast vs. fly. (d,e) worm vs. fly (f) yeast vs. worm.

the figure, the interactions of proteins that belong to the same orthologous group are highly conserved, suggesting relatively recent duplications.

4 Path Queries

Sequence comparison is a basic tool in biological research, widely used for nucleotide sequences comparison and search (as in RNA and DNA molecules), and for amino acids sequence comparisons (as in protein homology discovery). It is used both for evolutionary and functional research of both genes and proteins. The availability of PPI Networks allows us to extend the use of sequence comparison methods to more complex functional units, such as protein pathways and modules, and thus elevate homology detection from the level of single protein homology to the level of functional protein pathways and modules homology.

This section describes another method for comparing and aligning PPI networks, QPath ([20]), which overcomes some fundamental drawbacks of the PathBLAST algorithm introduced in section 3:

1. In a PathBLAST result, a matched pathway may contain the same protein more than once, which is biologically implausible.

2. The resulted matched pathways must be very close to each other, while we might want to allow a higher degree of freedom, and support more than a single consecutive insertion or a single consecutive deletion difference between the paths, which is the maximum PathBLAST allows.

3. The running time of the algorithm involves a factorial function of the pathway length, limiting its applicability to short pathways (in practice, it was applied to paths of up to 5 proteins).

4.1 The Path Query Problem

The problem setting is defined as follows: the input is a target network, represented as an undirected weighted graph $G(V, E)$, with a weight function on the edges $w : E \times E \longrightarrow R$, and a path query $Q = (q_1, \ldots, q_k)$. Additionally, a scoring function $H : Q \times V$ is given. The output is a set of best matching pathways $P = (p_1, \ldots, p_k)$ in G, where a good match is measured in two respects:

1. Each node in the matched pathway and its corresponding node in the query are similar with respect to the given scoring function H.

2. The reliability of edges in the matched pathway is high.

If we don't force the size of the query and matching paths to be equal, we can still measure the match between a query $Q = (q_1, \ldots, q_k)$ and a pathway $P = (p_1, \ldots, p_l)$ by introducing *dummy nodes* which allow for deletions, if inserted in the matching path and for insertions, if inserted in the query.

In the PPI Network framework, as described in section 2, the target graph is a PPI Network of species 1, where the vertices are proteins, and edges' weights represent the interaction probability between two proteins. The query pathway Q is a pathway extracted from a PPI Network of species 2, and the function H is a similarity measure between proteins in the two species.

4.2 The QPath algorithm

First, in order to allow more flexibility in deletions and insertions, deletions of nodes in the target network are allowed by introducing a mapping M form Q to $P \cup \{0\}$ where deleted query nodes are mapped to 0 by M. The total score of an alignment reflects the measures of protein homology, and the interaction

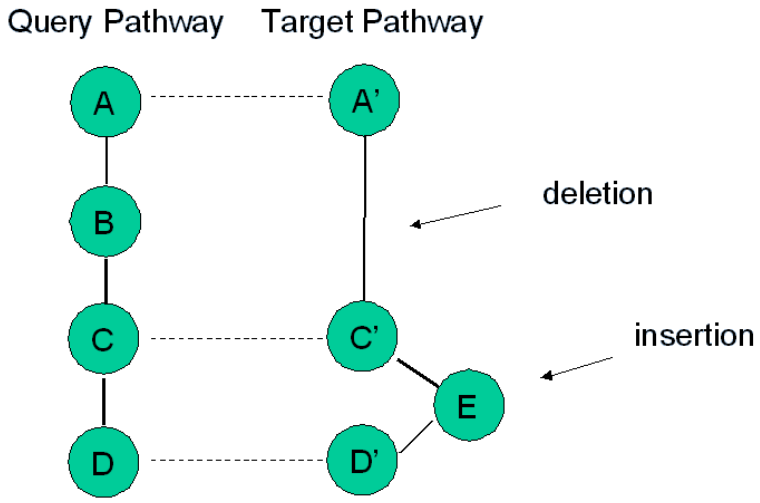

Figure 17: An example of an alignment that induces insertions (F') and deletions (C).

probabilities of the path, while keeping the path similarity with a certain degree of freedom for insertions and and deletions, and is set to be:

$$\sum_{i=1}^{l-1} w(p_i, p_{i+1}) + \sum_{i=1, p_i \neq 0}^{k} h(q_i, M(p_i))$$

Where the first summation is the *interaction score* and the second is the *sequence score*. Edge weights represent the logarithm of reliability of interaction between two proteins, and the protein similarity scoring function H is set to be the BLAST E-value for the two proteins, normalized by the maximal E-value over all pairs of proteins from the two networks.

4.3 Avoiding cycles (non-trivial paths)

In order to find only simple paths, QPath uses the color coding technique (Alon et al. [1]). The method allows finding simple paths of size k by randomly choosing a color out of k colors for every vertex in the graph, and looking only for subgraphs that do not contain more one vertex of the same color. Since a particular path may be assigned non-distinct color, the method requires choosing many random colorings, and running the search for each of them separately.

4.4 Finding the best matching paths

QPath sets in advance two parameters - N_{ins} and N_{del}, which are the number of insertions and deletion allowed in the matched path. When looking for a path of size k, QPath assigns $k + N_{ins}$ colors for the vertices.

The following dynamic programming recursion is then used for dynamically building the best path:

$$W(i, j, S, \Theta_{del}) = max_{m \in V} \begin{cases} W(i-1, m, S - c(j), \theta_{del}) + w(m, j) + h(q_i, j) & (m, j) \in E \\ W(i, m, S - c(j), \theta_{del}) + w(m, j), & (m, j) \in E \\ W(i-1, m, S, \theta_{del} - 1), & \theta_{del} < N_{del} \end{cases}$$

$W(i, j, S, \Theta_{del})$ is the maximum weight of an alignment for the first i nodes in the query that ends at vertex $j \in V$, induces θ_{del} deletions, and visits a vertex of each color in S. The first case is the case where q_i is aligned with vertex j, and thus we add to the best alignment so far the score $h(q_i, j)$, and remove the color of j from the set of available colors S. In the second case, q_i is not aligned with j, meaning q_i is an insertion, and the score does not change. The third case is a deletion case, and therefore we decrease the number of allowed deletions from this point on by one.

The best alignment score will be $max_{j \in V, S \subseteq C, \theta_{del} < N_{del}} W(k, j, S, \theta)$, and the alignment itself can be find by backtracking. The running time for each coloring choice is $2^{O(k+N_{ins})mN_{del}}$. For a choice of $\varepsilon \in (0, 1)$ such that the probability to find the optimal match is at least $1 - \varepsilon$ we would need to choose $ln(n/\varepsilon)$ random colorings, which will give a total running time of $ln(n/\varepsilon)2^{O(k+N_{ins})mN_{del}}$.

In order to use QPath for searching homologous paths between two given PPI networks, it is first required to extract good candidates from the first network, and then search for these paths in the target network. QPath can find good candidates by searching the first network for a dummy path query, consisting of dummy proteins that have the same similarity score H to all vertices in the network. Such a search yields pathways with high interaction scores in the first network, regardless of the path query itself.

4.5 Running QPath on yeast and fly PPI networks

The yeast (S. cerevisiae) PPI network contains 4,726 proteins and 15,166 known interaction between them. The fly (D. melanogaster) PPI network contains 7,028 proteins and 22,837 interactions, but in spite of its larger size, it is much less complete than the yeast network.

The algorithm was tested first on the more complete yeast PPI network, finding good candidates for querying the fly PPI network next. It discovered 271 pathways which were better than 99% of randomly chosen pathways obtained by setting all interaction scores to be equal and running the query on the tweaked data. The 271 pathways were then used as queries for the fly PPI network.

The results of running the algorithm on the yeast PPI network were assessed by looking at the functional enrichment of the found paths. 80% of the paths found were functional enriched, implying their biological significance. In comparison, running dummy queries on the less complete fly network resulted with only 39% of the 132 fly paths found to be functionally enriched.

Running the 271 paths found in the yeast PPI network as queries on the fly network discovered that 63% of them had a match in the fly network (Figure 18).

The results show that pathway similarity can be used for identification of functionally significant pathways, and that those query pathways can help us to infer the actual function of matched pathways. a first annotation map of protein pathways in fly that are conserved from yeast was obtained this way by QPath.

4.6 Scoring the paths

After setting the scoring framework, there is a need to set the weighs parameters, and define the actual contribution of the different scoring components - the *interaction score*, the *sequence score* and the cost of insertions and deletions.

The target is to find a weight function that will maximize the probability that a path with high score is indeed functionally enriched. This was done by using logistic regression on the path attributes - interactions reliability, sequences similarity, number of insertions, and number of deletions, using known functionally enriched paths in the yeast network for training.

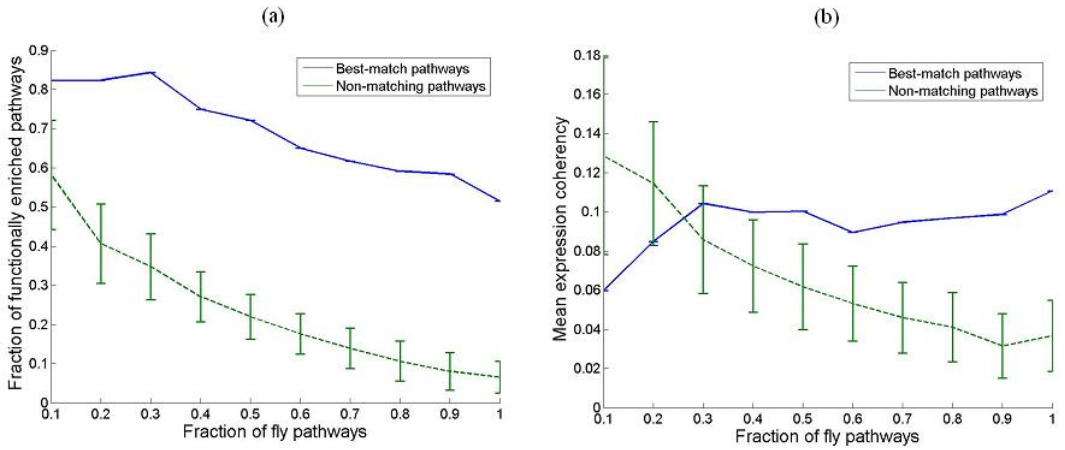

Figure 18: Source [20]. Functional significance of best-match pathways in fly. Functional enrichment (a) and expression coherency (b) of fly best-match pathways obtained by QPath compared to fly pathways that are not the result of a query

4.7 Is the insertion and deletion flexibility really required?

As mentioned before, one of the most important features QPath introduced is the ability to align sequences with a high number of subsequent insertions or deletions. Figure 19 illustrates that this feature is indeed important, as most of the conserved paths between the yeast and the fly, required more than one insertion and deletion.

In the same manner, discovering functionally enriched paths was also found to be strongly depended on the fact that a high number of insertions and deletions is required (See Figure 20)

4.8 Functional conservation

Results of running QPath on the yeast and fly PPI networks, yielded that for 64% of the conserved paths, the matched paths in the fly network conserved one or more functions of the yeast query pathways. In

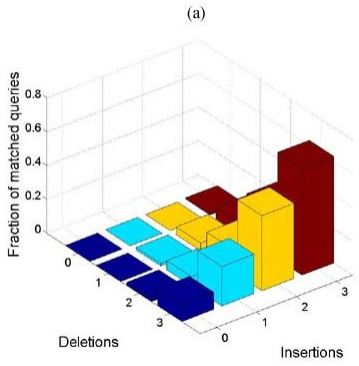

Figure 19: Source [20]. Fraction of matched queries between yeast and fly networks in respect to the number of deletions and insertions in the conserved paths

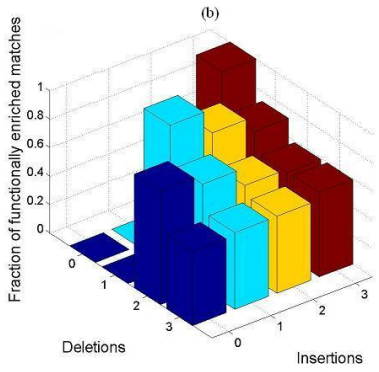

Figure 20: Source [20]. Fraction of functionally enriched matches in respect to the number of deletions and insertions in the conserved paths

contrast, a random shuffling of the matches was tested and resulted to in conservation rate of only 31%. Interestingly, the functional conservation was much lower when limiting the protein homology only to the best pairs, one from each species. This implies that pathway homology can be used to predict function. More explicit methods for such a prediction, that make use of the networks homology on top of the straight forward sequence alignment will be presented in the next section.

5 Orthology Mapping

Annotating protein function across species is an important task which is often complicated by the presence of large paralogous gene families. Most of the methods of dealing with this problem are sequence-based models, thus sequence of proteins from different species was compared, in order to find a group of proteins that have the same functional annotation. Two such methods are COG (Clusters of Orthologous Groups) (Tatusov *et al.* [23]) and Inparanoid ([15]).

The COG approach defines orthologs using sets of proteins that contain reciprocal best BLAST matches across a minimum of three species. The Inparanoid approach is a sequence-based method of finding functional annotation. It uses clustering in order to derive orthology families, leaving some of the orthology relations ambiguous. For more information see [15].

Based on the concept that a protein and its functional ortholog are likely to interact with proteins in their respective networks that are themselves functional orthologs, Bandyopadhyay *et al.* in [4] introduced a novel strategy for identifying functionally related proteins that supplements sequence-based comparisons with information on conserved protein-protein interactions.

While the tools we described in the previous sections used orthology to identify conserved protein information, the approach shown here reverse that logic and use conserved protein interactions to predict functional orthology.

5.1 Functional Orthology

Ambiguities in the functional annotation process arise when the protein in question has similarity to not one but many paralogous proteins, making it harder to distinguish which of these is the true ortholog that is, the protein that is directly inherited from a common ancestor. Especially in the genomes of mammals and other higher eukaryotes, large protein families are typically not the exception but the rule.

Moreover, as we saw in section 3.3, the assignment of protein orthology depends largely on the evolutionary history. Protein families for which speciation predates gene duplication (out-paralogs) are particularly challenging. In these cases, every cross-species protein pair is technically orthologous but it is still necessary to distinguish which protein pairs play functionally equivalent roles, i.e. are functional orthologs ([15]).

Conversely, when gene duplication predates speciation (out-paralogs), the family can often be subdivided into orthologous pairs which have higher sequence similarity to each other than to other members. However, evolutionary processes such as gene conversion serve to homogenize paralogous sequences over time, making these cases problematic as well. Even more complicated, protein function may be lost between distant organisms or conserved across multiple proteins within a single species.

As opposed to ambiguous functional orthologs, *definite functional orthologs* are defined as proteins that are functionally equivalent as a result of direct ancestry.

5.2 Model Review

The model consists of the following steps:

1. The protein interaction networks of two species are aligned by assigning proteins to sequence homology groups using the Inparanoid algorithm.
2. Networks are aligned into a merged graph representation.
3. Probabilistic inference is performed on the aligned networks to identify pairs of proteins, one from each species, that are likely to retain the same function based on conservation of their interacting partners.
4. A logistic function is used to compute the probability of functional orthology for a protein pair i given the states of functional orthology for its network neighbors.
5. The previous probability is updated for each pair over successive iterations of Gibbs sampling.

An overview of the method is seen in Figure 21. Also, the probabilistic model, the logistic function and Gibbs sampling will be explained in the next sections.

5.2.1 Conservation Index

Consider an alignment graph, G, with nodes representing sequence-similar protein pairs, and edges linking nodes (a, b) and $(a^{'}, b^{'})$ if one of $(a, a^{'})$ or $(b, b^{'})$ directly interacts, and the other interacts via a neighbor, which is directly connected to him (i.e. interaction of distance ≤ 2). An edge is *strongly conserved* if its endpoints are true functional orthologs.

The conservation index c of a node i (representing protein pair $(a, a^{'})$ is defined as twice its number of strongly conserved interactions, divided by its total number of interactions over both species:

$$c(i) = \frac{2d(i)}{d(a) + d(a')} \tag{22}$$

where $d(i)$ denoted the number of strongly conserved links involving node i, while $d(a)$ and $d(a^{'})$ denoted the degrees (number of interactions) of proteins a and $a^{'}$ in their respective single-species networks.

An example of the values of the conservation index in the yeast vs. fly experiment is shown in Figure 22(a).

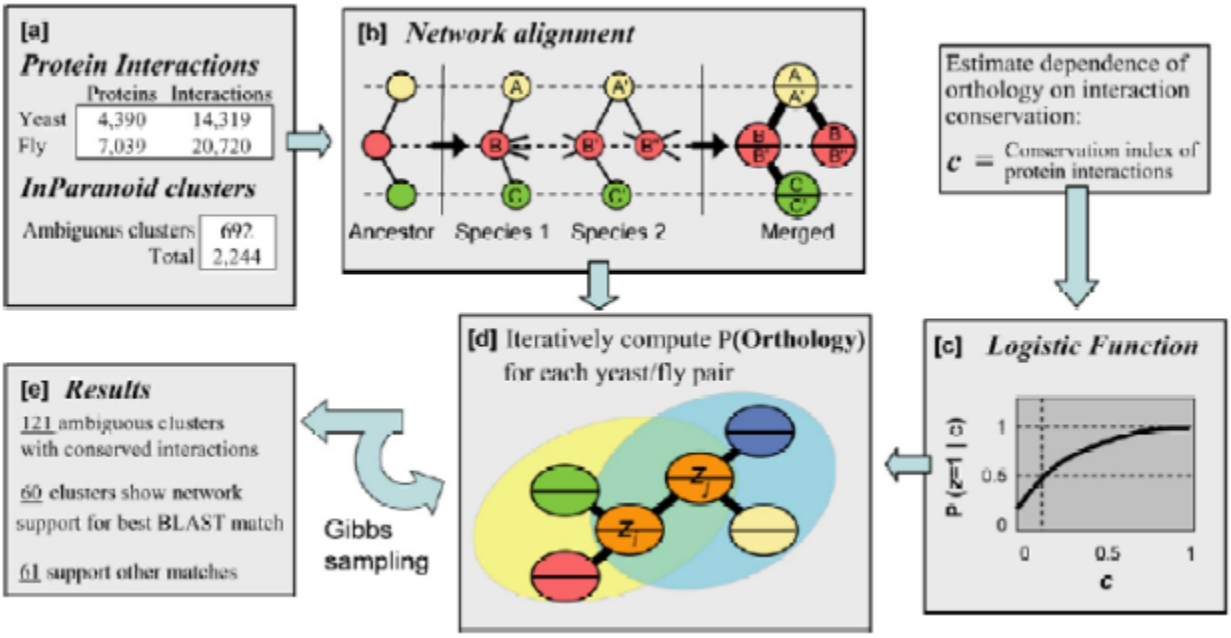

Figure 21: Source [4]. This figure shows an overview of the method, as an example on yeast/fly PPI networks. (a) PPI networks for yeast and fly are combined with clusters of orthologous yeast and fly protein sequences as determined by the Inparanoid algorithm. (b) Networks are aligned into a merged graph representation. (c) The logistic function is used to compute the probability of functional orthology for a protein pair i given the states of functional orthology for its network neighbors. (d) This probability is updated for each pair over successive iterations of Gibbs sampling. (e) The final probabilities confirm 60 of the best BLAST match pairings. The network supports a different hypothesis for 61 pairings.

5.2.2 Probabilistic Model

The probabilistic model is based on the assumption that the probability of functional orthology for a pair of proteins is influenced by the probabilities of functional orthology for their network neighbors, which in turn depend on their network neighbors, and so on. This type of probabilistic model is known as a Markov random field (see [5]).

This model is specified by an undirected graph $G = (V, E)$ corresponding to a network alignment, and conditional probability distributions which relate the event that a given node represents a functionally orthologous pair with those events for its neighbors. A Markov random field model is specified in terms of potential functions on the cliques in the graph:

$$P(\overrightarrow{z}) = \frac{1}{Z} exp\{-U(\overrightarrow{z})\} \tag{23}$$

where $\overrightarrow{z}$ is some assignment to the states of all nodes in the graph, U is an "energy" function which integrates the potentials over all cliques in the graph, and Z is a normalizing constant. It is not necessary to compute the normalization constant, since all that is required are the conditional probabilities for each node given its neighbors (rather than the joint distribution). For computational efficiency, the common auto-logistic model ([5]) which assigns zero potential to cliques of size > 2 was used. Under this model, the energy takes the form:

$$U(\overrightarrow{z}) = -\sum_{i} \alpha_i z_i - \sum_{(i,j)\in E} \beta_{ij} z_i z_j \tag{24}$$

which, when substituted into the equation for $P(\overrightarrow{z})$ above, reduces to a logistic function.

Based on the initial observation that the functional orthology of a node is a function of its conservation index (well approximated by a logistic function see the next section), they set $\alpha_i = \alpha$, $\beta_{ji} = \beta_i = \frac{2\beta}{d(a_i + d(a'))}$ to obtain the following:

$$P(z_i | Z_{N(i)}) = \frac{1}{1 + exp\{-\alpha_i - \sum_{j\in N(i)} \beta_{ij} z_j\}} = \frac{1}{1 + exp\{-\alpha - \beta c(i)\}} \tag{25}$$

where $N(i)$ is the set of neighbors of node i, and $z_{N(i)}$ denotes the set of all z_j such that $j \in N(i)$. Note that α_i and β_{ij} could be set to accommodate other equations for conservation index, as long as they are linear in the number of strongly conserved neighbors $d(i)$.

5.2.3 Fitting the Logistic Function

In order to provide a set of training data for fitting the parameters α and β of the logistic function, a fraction (about a half) of the definite functional orthologs having at least one conserved interaction is chosen randomly, as positive examples, and their states are set to $z = 1$. Negative examples of "non-orthologs" are also generated by randomly selecting a fraction (about a half) of the proteins in one species and pairing each with its best BLAST E-value matching protein in the other species not in the same cluster; their states are set to $z = 0$ (ideally, the negative training data would consist of orthologs that are not functional orthologs, but few such examples exist). Parameters α and β are optimized by maximizing the product of $P(z_i | z_{N(i)})$ over all positive and $(1 - P(z_i | z_{N(i)}))$ over all negative training data using the method of conjugate gradients. The logistic function obtained for yeast vs. fly is shown in Figure 22(b).

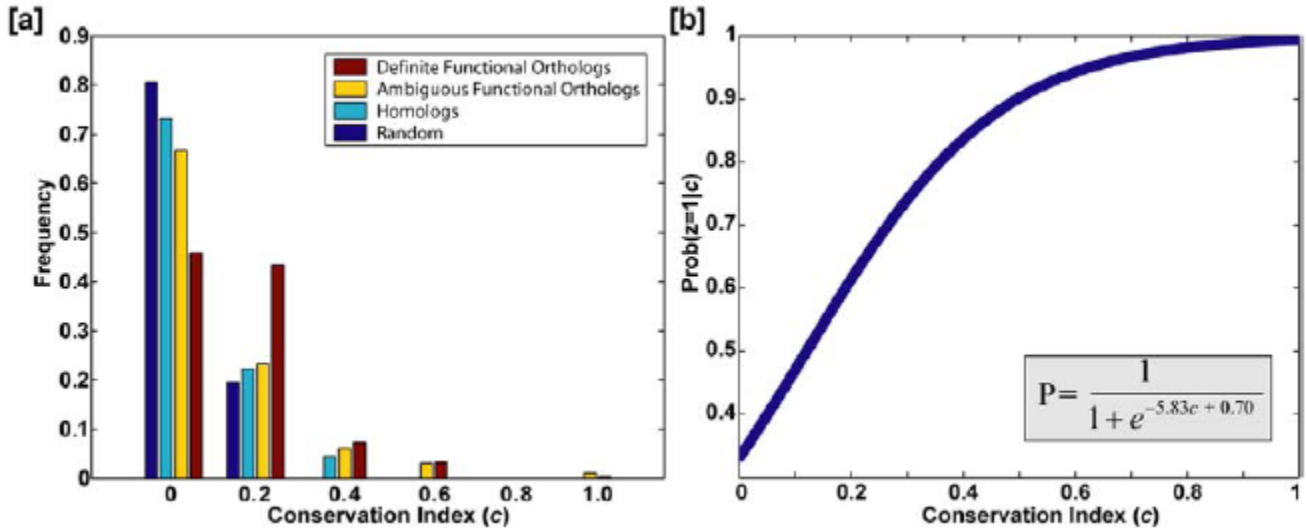

Figure 22: Source [4]. This figure shows two graphs: (a) Network neighborhood conservation for definite orthologs versus other yeast/fly protein pairs. The distribution of the conservation index is shown for definite functional orthologs (sole members of an Inparanoid group); ambiguous functional orthologs (in a group with multiple members); homologs (different groups but similar sequences); and random protein pairs. Definite functional orthologs show a shift towards higher conservation of protein interactions between the yeast and fly protein networks. Mean c=0.1512, 0.1171, 0.0870, 0.0615 for definite functional orthologs, ambiguous functional orthologs, homologs and random pairs respectively. (b) Logistic function relating conservation index to probability of functional orthology. Logistic regression was performed using the "definite functional ortholog" and "homolog" pairs as positive vs. negative training data, respectively. The resulting function is shown.

Species	No. of proteins	No. of interactions
yeast	4,389	14,319
fly	7,038	20,720

Table 2: Source [4]. Number of proteins and interactions used in the experiment.

5.2.4 Orthology Inference

In order to identify the functional orthologs the above model was used to estimate the final posterior probabilities $P(z_i)$ using the method of Gibbs sampling ([21]). In this approach, nodes representing ambiguous functional orthologs are each assigned a temporary state $z = 0$ or $z = 1$, initially at random. At each iteration, a node i is sampled (with replacement) and its value of z_i is updated given the states of its neighbors, $z_{N(i)}$. The new value of z_i is set to 0 or 1 with probability $P(z_i|z_{N(i)})$. Over all iterations, the nodes designated as definite functional orthologs and "non-orthologs" are forced to states of 1 and 0, respectively.

5.3 Experimental Results

The method was applied on yeast and fly, and PPI data was downloaded from DIP ([29]). The statistics for the PPI networks of yeast and fly are shown in Table 2.

A total of 2,244 clusters were generated, covering 2,834 proteins in yeast and 3,881 proteins in fly. Of these, 1,552 clusters contained only a single yeast and fly protein pair and were assumed to represent unambiguous or "definite" functional orthologs. The remaining 692 clusters contained multiple proteins from yeast and/or fly, leaving the functional orthologs ambiguous.

To determine the extent to which proteins and their functional orthologs had conserved protein interactions, the network neighborhoods of definite functional orthologs were examined and compared to the neighborhoods of less related protein pairs (Figure 22). As a measure of local network conservation, the conservation index of each protein pair was computed as proportional to the fraction of interactions that were conserved across the two species.

For example, in Figure 21(b) the orthologous pairing $B/B^{'}$ has a higher conservation index $(4/9)$ than the alternative pairing $B/B^{''}$ $(2/9)$. Figure 22(a) shows the set of conservation indices for definite functional orthologs versus those of ambiguous functional orthologs, non-orthologous homologs (best cross-species BLAST matches not assigned to the same Inparanoid group), and random pairs of proteins chosen independent of sequence similarity.

The set of definite functional orthologs had the highest occurrence of conserved interactions. Moreover, the mean conservation index was related to the stringency of the pairing: definite functional orthologs tended to have higher conservation indices than ambiguous functional orthologs, ambiguous functional orthologs higher indices than homologs, and homologs higher indices than random protein pairs.

Beyond the mean conservation index, there were also significant differences among the four distributions. These findings confirm that yeast/fly proteins classified as definite functional orthologs are more likely to have equivalent functional roles in the protein network and, conversely, that conserved network context could be used to help discriminate functional orthology from general sequence similarity.

They applied their approach to resolve ambiguous functional orthology relationships in the yeast and fly protein networks. Of the 692 ambiguous Inparanoid clusters, 121 contained protein pairs for which at least one pair had conserved interactions between networks. Application of their Gibbs sampling procedure yielded estimates of probability of functional orthology for each protein pair in these 121 ambiguous clusters. In 60 of these clusters, the highest probability was assigned to the protein pair that was also the most sequence-similar via BLAST. These cases reinforced the intuition that the best sequence matches are also the most functionally similar.

The remaining 61 clusters showed the opposite behavior, i.e., the highest probability pair was not the most sequence similar pair. Of these 61 cases, 15 were supported by two or more conserved interactions. Because the yeast and fly networks are incomplete (i.e., they contain false negatives because of the ”noisy” data of the PPI networks), in some of these cases they could not rule out the possibility that conserved interactions with the best BLAST matches have been missed.

A complete listing of the results can be found on their website (http://bioinf.ucsd.edu/ sbandyop/GR). For some examples of the clusters found see Figure 23.

5.3.1 Validation

One approach to validate their results would be to compare them against databases of functional annotations (such as GO). However, such databases are based directly on sequence similarity, thus they are lack the specificity to discriminate among subtle functional differences across large gene families. Therefore, cross-validation was used in order to test the ability of their approach to reclassify protein pairs in the definite functional ortholog set (positive test data), against the non-orthologs homolog set (negative test data).

In each cross-validation trial, 1% of these assignments were hidden (declassified) and monitored during Gibbs sampling to obtain probabilities of functional orthology for positive and negative examples. Reclassification was judged successful if the probability of functional orthology exceeded a particular cutoff value. These statistics were compiled over 100 trials. Figure 24(a) charts cross-validation performance over a range of probability cutoffs.

At a probability cutoff of 0.5, they observed a 50% true positive rate and a 15% false positive rate. This shows marked improvement over a random predictor where we would expect to see the same true positive rate as false positive rate.

Declassifying 1% of the known functional orthologous and non-orthologous pairs reduces the amount of information available to the algorithm and, thus, can reduce its predictive ability. Therefore, the cross–validation analysis was repeated at varying percentages of declassification of positive and negative data (ranging from 1% to 100%) (Figure 24(b)). For instance, changing the amount of declassification of available training data from 1% to 25% reduced the maximum precision from 83% to 75%. Further declassification yielded more marked reductions in precision and recall.

6 Multiple Network Alignment

All the methods we saw in the previous sections of this scribe ([10, 18, 11]) were limited to the alignment of two networks. Sharan *et al.* in [19] introduced a method of performing comparison and analysis of multiple PPI networks.

6.1 The Alignment Graph

The method that Sharan *et al.* introduce in [19] is similar to the one introduced in [18] (which we described earlier). The main difference is that instead of aligning only two networks the interactions are integrated with the sequence information in order to generate a multiple network alignment graph. Each node in this graph consists of a group of sequence-similar proteins, one from each species. Each link between a pair of nodes in the alignment graph represent conserved protein interactions between the corresponding protein group. The sequence-similarity is calculated, as before, using BLAST ([2]). See Figure 25. A search over the alignment graph is performed to identify two types of conserved subnetworks:

1. Short linear paths of interacting proteins, which model signal transduction pathways.
2. Dense clusters of interactions, which model protein complexes.

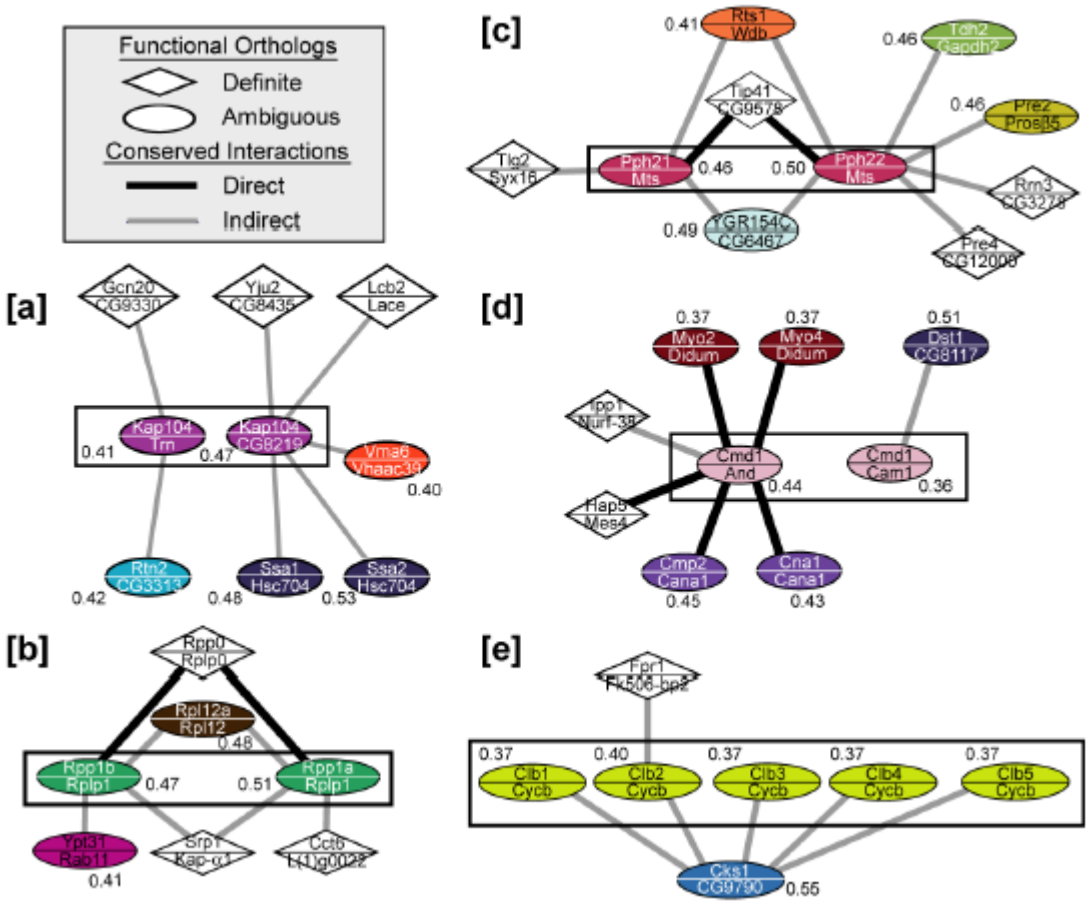

Figure 23: Source [4]. Example orthologs resolved by network conservation. Each node represents a putative functional match between a yeast/fly protein pair (with names shown above/below the line,respectively). Links between nodes denote conserved interactions (thick black, direct interactions in both species; thin gray, indirect interaction in one of the species). Diamond vs. oval nodes represent definite vs. ambiguous functional orthologs. Oval nodes of the same color represent ambiguous protein pairs belonging to the same Inparanoid cluster. The mean probability of functional orthology is given next to each ambiguous pair. (a)-(e) show examples of clusters that were disambiguated by conserved network information;the cluster resolved in each panel is outlined by a black rectangle.

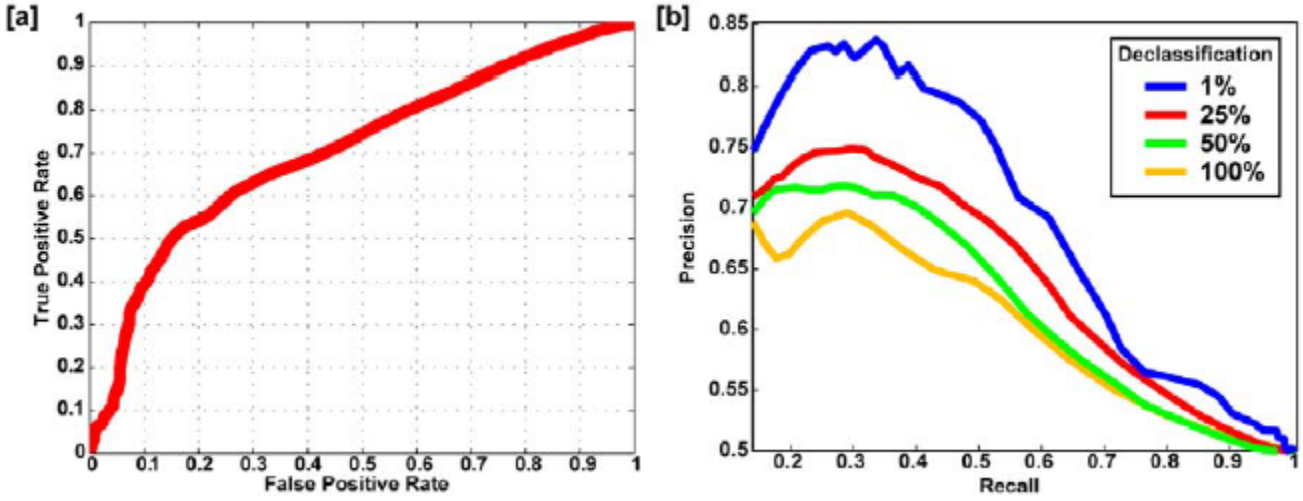

Figure 24: Source [4]. (a) The Receiver Operating Characteristic (ROC) curve shows the true positive rate (percent of true data predicted correctly as positive) vs. the false positive rate (percent of false data predicted incorrectly, i.e. as positives) of the method. (b) Dependence of predictions on number of available training examples. Percent recall (true positive rate) vs. precision (percent of positive predictions that were correct) is plotted as the probability cutoff ranges from [0-1]. Different color plots correspond to different percents of declassification of training examples.

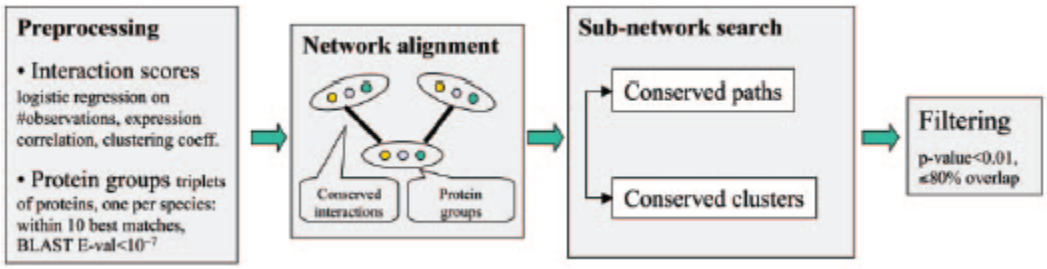

Figure 25: Source [19]. This figure shows a schematic illustration of the comparison process in [19]. Raw data are preprocessed to estimate the reliability of the available protein interactions and identify groups of sequence-similar proteins, using BLAST. The protein group contains one protein from each species and it is required that each protein has a significant sequence match to at least one other protein in the group (BLAST $E_{value} < 10^7$); considering the 10 best matches only). Next, protein networks are combined to produce a network alignment that connects protein similarity groups whenever the two proteins within each species directly interact or are connected by a common network neighbor. Conserved paths and clusters identified within the network alignment are compared to those computed from randomized data, and those at a significance level of $P < 0.01$ are retained. A final filtering step removes paths and clusters with more that 80 percent overlap.

6.2 The Search and the Subnetwork Score

The search and scoring methods used here are very similar to the ones used by Sharan *et al.* in [18], thus we add a brief explanation of these methods. For further information about the methods see [19].

The search is guided by reliability estimates for each protein interaction, calculated using a logistic regression method (see Lecture 7), which are combined into a probabilistic model for scoring candidate subnetworks. This probabilistic model is similar to the one introduced by Sharan *et al.* in [18].

As for the way that the complex that was identified is scored, it is similar to that in equation 6. For example, given three complexes C, $C^{'}$ and $C^{''}$ the likelihood score is:

$$L(C, C^{'}, C^{''}) = L(C)L(C^{'})L(C^{''}) \tag{26}$$

and similar to that goes for n complexes.

Note that in [18] the likelihood score was also multiplied by a value that represents the similarity between pairs of proteins. The reason this part is missing here is that there is no convention about similarity between more than two protein sequences. Also note that $L(C)$, $L(C^{'})$ and $L(C^{''})$ are calculated as before, using equation 5. The score of a path is calculated in a similar way.

As before, a log–likelihood ratio score is used in order to compare the fit of a subnetwork to the desired structure (path or cluster) versus its likelihood given that each species' interaction map was randomly constructed. The underlying model assumptions are (same as in the statistical model in [18], recall the Complex model(M_c) and the Null model(M_n)):

1. In a real subnetwork, each interaction should be present independently with high probability.

2. In a random subnetwork, the probability of an interaction between any two proteins depends on their degree of interactions in the network.

The significance of the identified subnetworks is evaluated by comparing their scores to those obtained on randomized data sets, in which each of the PPI networks is shuffled along with the protein similarity relationships between them.

6.2.1 Experimental Results Visualization

The layout of conserved pathes and clusters between all three networks, as it can be seen in figure 27 were automatically generated via a plug–in to Cytoscape [17]. The layout was produced using a "spring" method, commonly used to present graphs. According to this method, a "spring" is attached to every edge (a certain attraction force is applied between every two nodes connected by edges) and a certain rejection force is a applied between all nodes. Then, all is left to do, is to bring the entire system to equilibrium.

In this case, except the usual attraction force on all PPI edges, other forces were distributed as following:

- Repulsive force was applied between intra–species nodes.

- Attractive force was applied between sequence–similar nodes.

After the system stabilized, the different species networks were separated and placed side by side. A vivid demonstration can be seen on Figure 26.

6.3 Experimental Results

Sharan *et al.* applied the multiple network alignment framework (see Figure 25) to three PPI networks:

1. Saccharomyces cerevisiae (yeast), including 14,319 interactions among 4,389 proteins.

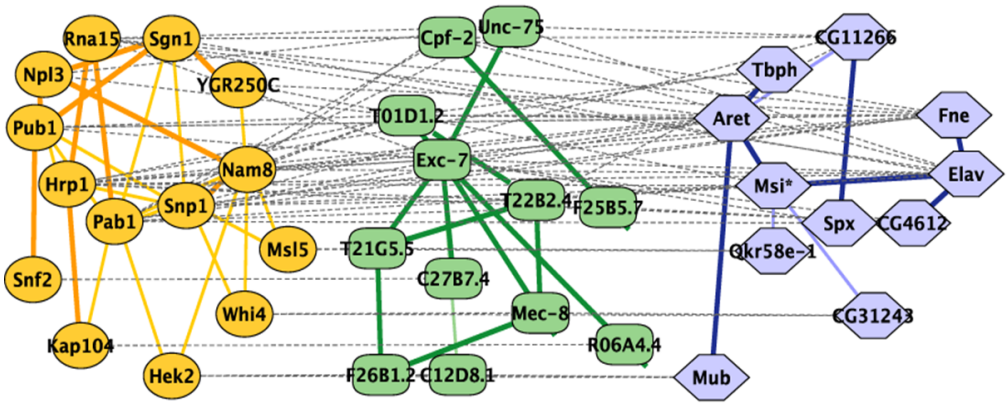

Figure 26: Source [19]. Demonstration of the automated layout of the network alignments using a Cytoscape ([17]) software plug–in.

2. Caenorhabditis elegans (worm), including 3,926 among 2,718 proteins. (note that this network is a very partial network).

3. Drosophila melanogaster (fly), including 20,720 among 7,038 proteins. interactions among 4,389 proteins.

A search over the network alignment identified 183 protein clusters and 240 paths conserved at a significance level of $P < 0.01$. These covered a total of 649 proteins among yeast, worm, and fly. Representative examples of conserved clusters and paths are shown in Figure 27. Figure 28 shows a global map of all clusters and paths conserved among the three species protein networks, and the biological functions of each cluster and path. The map shows evidence of modular structure, groups of conserved clusters overlap to define 71 distinct network regions most enriched for one or more well defined biological functions.

To validate the results, these conserved clusters were compared to known complexes in yeast as annotated by the Munich Information Center for Protein Sequences (MIPS - http://mips.gsf.de). They only considered MIPS complexes that were manually annotated independently from the Database of Interacting Proteins interaction data (i.e., excluding complexes in MIPS category 550 that are based on high–throughput experiments).

Overall, the network alignment contained 486 annotated yeast proteins spanning 57 categories at level 3 of the MIPS hierarchy. They defined a cluster to be pure if it contained three or more annotated proteins and at least half of these shared the same annotation. 94% of the conserved clusters were pure, indicating the high specificity of their approach, compared to a lower percentage (83%) of when applying a noncomparative variant of our method to data from yeast only (i.e., applying the same methodology to search for high–scoring clusters within the yeast network only).

Another test was performed to find whether the conserved clusters were biased by spurious interactions, resulting from proteins that lead to positive two–hybrid tests without interaction (resulting from bait signals). Of 39 proteins with more than 50 network neighbors, only 10 were included in conserved clusters. These 10 proteins were involved in 60 intracluster interactions, 85% of which were supported by coimmunoprecipitation experiments. This finding indicates that the clusters were not biased because of artifacts of the yeast two–hybrid assays.

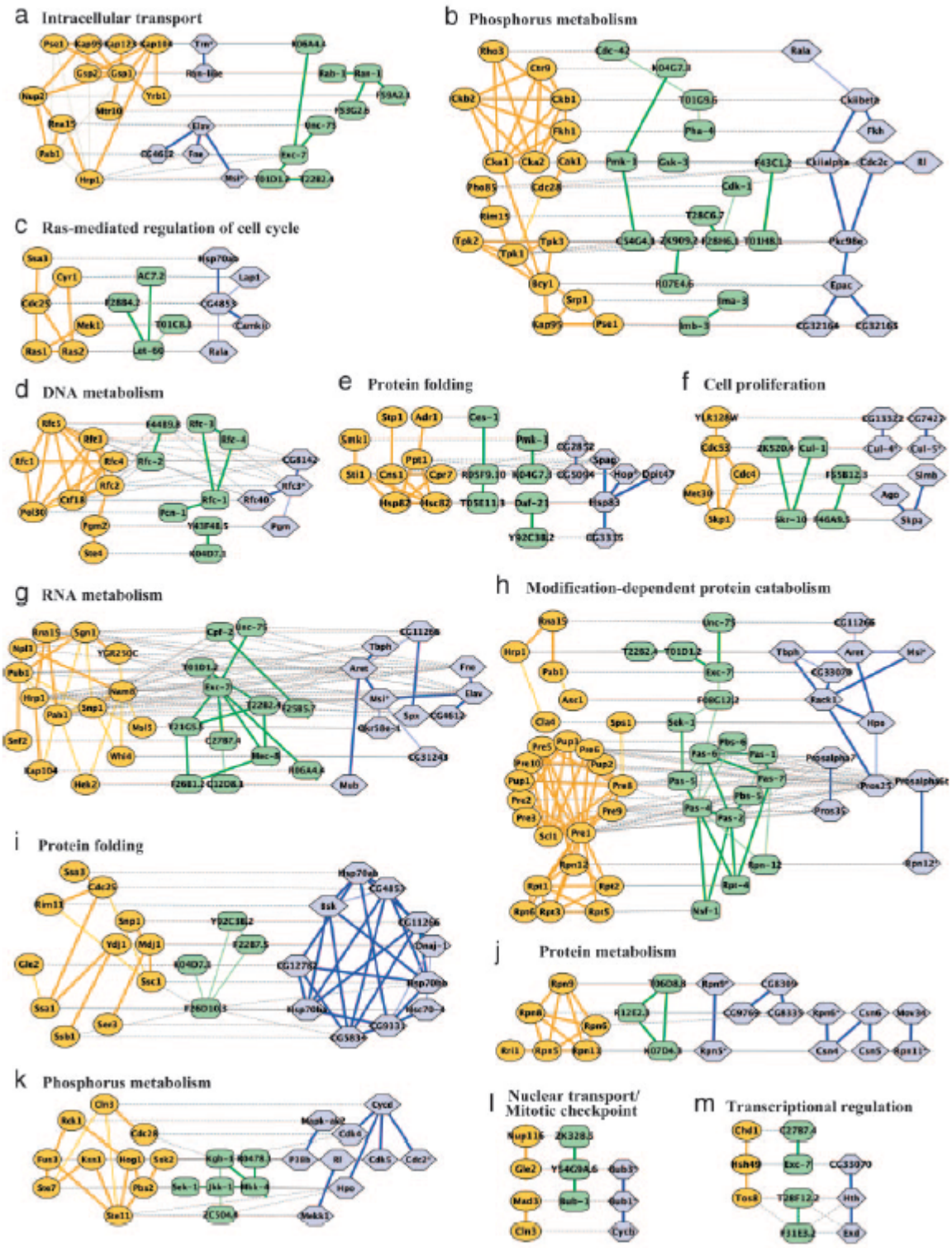

Figure 27: Source [19]. This figure shows the conserved networks regions identified by the yeast (orange ovals), worm (green rectangles) and fly (blue hexagons) network comparison. Direct interactions are drawn by thick line. Indirect interactions (connections via a common neighbor) are drawn by thin line. Horizontal dotted gray lines show cross-species sequence similarity between the proteins. Each sub-figure indicated the biological function of the region. Regions (a-k) refer to clusters, and regions (l-m) refer to paths.

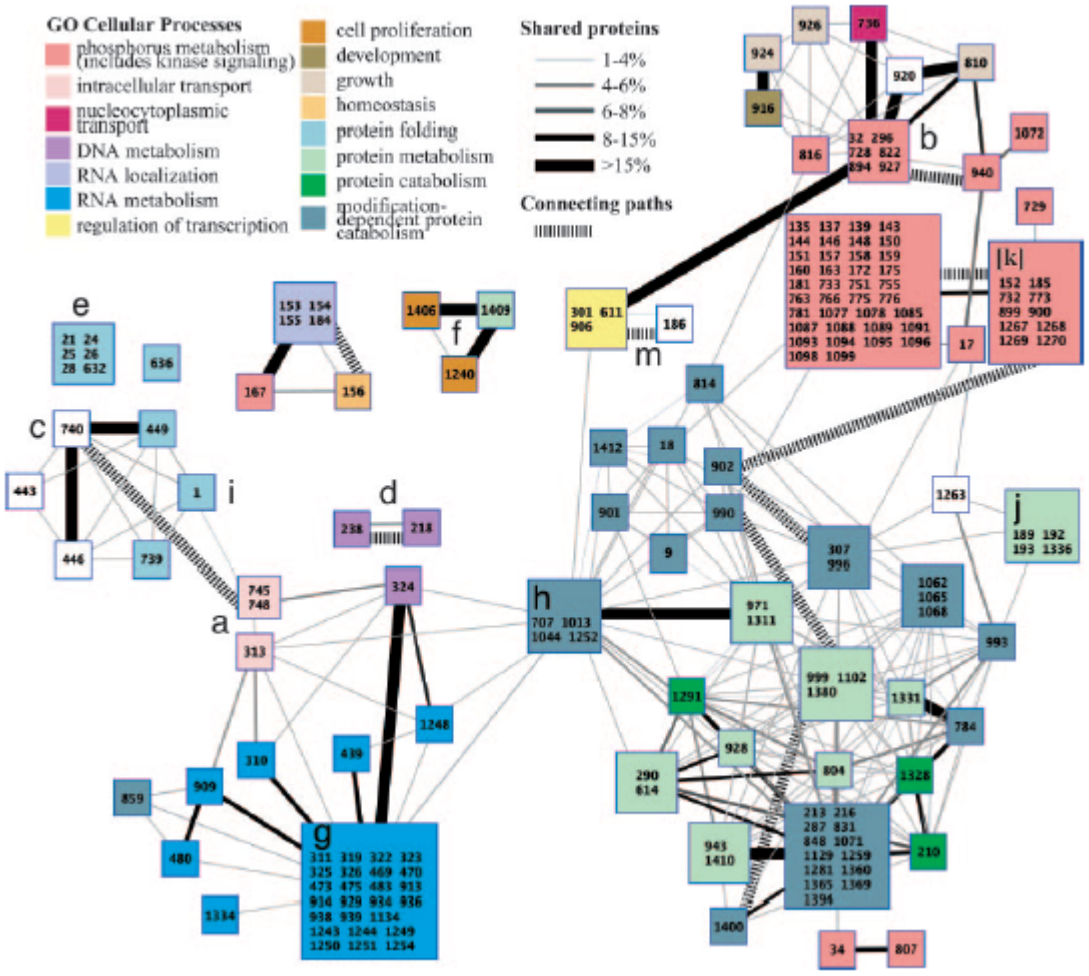

Figure 28: Source [19]. This figure shows the modular structure of conserved clusters among yeast, worm, and fly. Multiple network alignment revealed 183 conserved clusters, organized into 71 network regions represented by colored squares. Regions group together clusters that share more than 15 percent overlap, with at least one other cluster in the group, and are all enriched for the same GO ([30]) cellular process ($P < 0.05$ with the enriched processes indicated by color). Cluster ID numbers are given within each square;numbers are not sequential because of filtering. Solid links indicate overlaps between different regions, where thickness is proportional to the percentage of shared proteins (intersection size divided by the union size). Hashed links indicate conserved paths that connect clusters together. Labels $a - k$ and m mark the network regions exemplified in Figure 27.

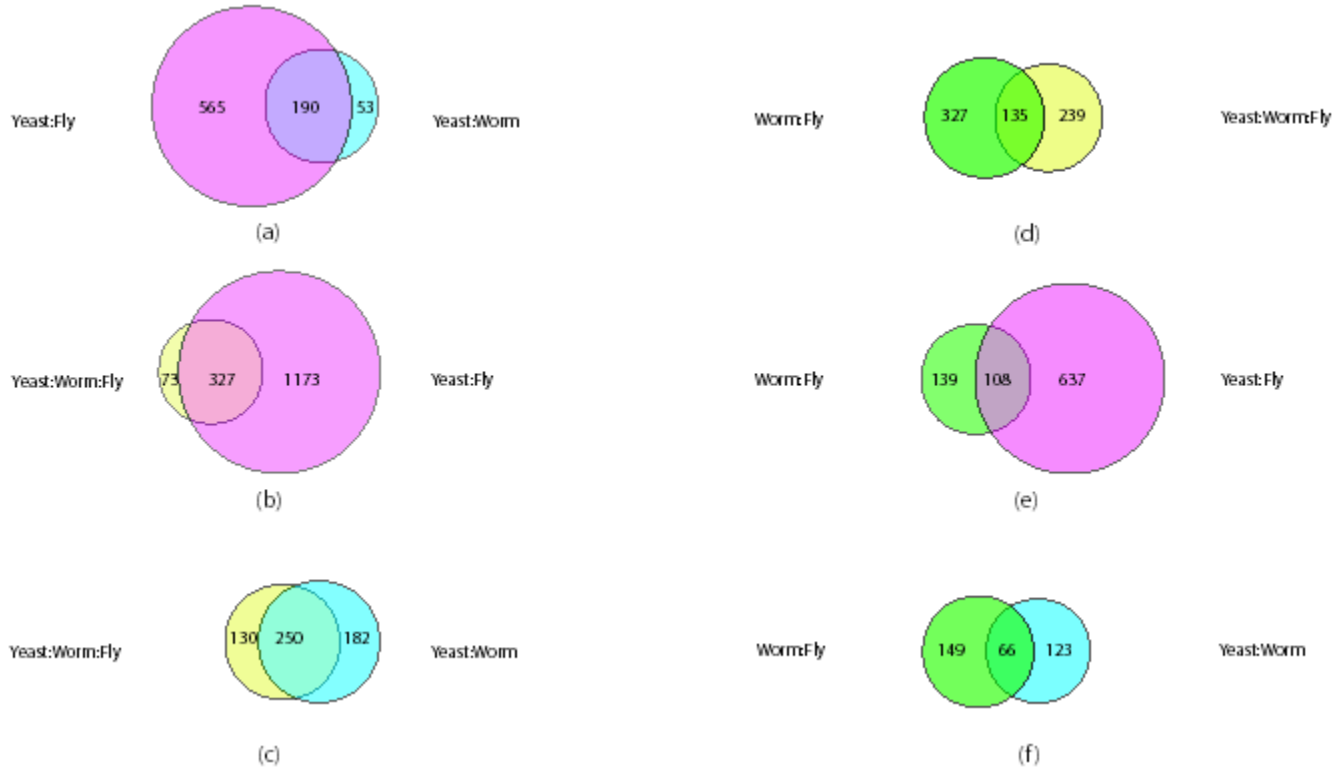

Figure 29: Source [19]. This figure shows a comparison of two and three-way clusters, via Venn diagrams depicting the relationships between the computed two and three-way clusters in terms of the number of distinct proteins that are included in each set of clusters.

6.3.1 Three-Way vs. Two-Way Network Alignments

In addition to the three–way comparison described above, two–way network alignments were performed between all three network pairs (yeast/worm, yeast/fly, and worm/fly). The clusters and paths identified by these comparisons are described in [19].

Interestingly, analysis of the proteins shared among the different pairwise comparisons led to the following findings:

- First, the density and number of conserved clusters found in the yeast/fly comparison were considerably greater than for the other comparisons, because of the large amounts of interaction data for these species relative to worm (Recall that the worm data is very partial). You can see a comparison between the different experiments in Figure 29 and in Table 3.

- Second, the worm/fly conserved clusters were largely distinct from the clusters arising from the other analysis. For example, only 29 percent of the proteins in the worm/fly clusters were assigned to conserved clusters in the three-way analysis (135 of 462). This observation is consistent with the closer taxonomic relationship of worm and fly compared to yeast.

Species	No. of proteins	No. of proteins in subnetworks	Coverage, %
Yeast-worm	765	271	35
Yeast-**worm**	536	204	38
Yeast-fly	1,494	790	53
Yeast-**fly**	1,559	778	50
Worm-fly	852	246	29
Worm-**fly**	1,131	291	26
Yeast-worm-fly	801	219	27
Yeast-**worm**-fly	551	190	34
Yeast-worm-**fly**	911	240	26

Table 3: Source [19]. Protein coverage by clusters and paths. In each raw the data refer to the **bolded** species.

Species	No. of correct	No. of predictions	Success rate, %
Yeast	114	198	58
Worm	57	95	60
Fly	115	184	63

Table 4: Source [19]. Cross–validation results for protein cellular process prediction, using 10-fold cross–validation.

6.3.2 Prediction of Protein Functions

As we saw earlier, conserved subnetworks that contain many proteins of the same known function suggest that their remaining proteins also have that function. Based on this concept, protein functions were predicted whenever the set of proteins in a conserved cluster or path (combined over all species) was significantly enriched for a particular Gene Ontology (GO) ([30]) annotation ($p < 0.01$) and at least half of the annotated proteins in the cluster or path had that annotation. When these criteria were met, all remaining proteins in the subnetwork were predicted to have the enriched GO annotation. They estimated the specificity of these predictions using cross validation, in which one hides part of the data, uses the rest of the data for prediction, and tests the prediction success by using the held–out data. (For more information on cross–validation see scribe of Lecture 6.) Results are shown in Table 4 (cellular process prediction), Table 5 (biological processes) and Table 6 (molecular functions).

Species	No. of correct	No. of predictions	Success rate, %
Yeast-worm	93	216	43
Yeast-**worm**	54	121	45
Yeast-fly	280	637	44
Yeast-**fly**	208	517	40
Worm-fly	22	55	40
Worm-**fly**	34	67	51
Yeast-worm-fly	114	198	58
Yeast-**worm**-fly	57	95	60
Yeast-worm-**fly**	115	184	63

Table 5: Source [19]. Cross–validation results for predicting protein Gene Ontology (GO) biological processes. In each raw the data refer to the **bolded** species.

Species	No. of correct	No. of predictions	Success rate, %
Yeast-worm	61	179	34
Yeast-**worm**	40	118	33
Yeast-fly	171	488	35
Yeast-**fly**	156	402	39
Worm-fly	37	64	58
Worm-**fly**	31	61	51
Yeast-worm-fly	79	162	49
Yeast-**worm**-fly	51	103	49.5
Yeast-worm-**fly**	77	149	52

Table 6: Source [19]. Cross–validation results for predicting protein Gene Ontology (GO) molecular functions. In each raw the data refer to the **bolded** species.

Species	Sensitivity %	Specificity %	P-value	Strategy
Yeast	50	77	1.1e-25	1
Worm	43	82	1e-13	1
Fly	23	84	5.3e-5	1
Yeast	9	99	1.2e-6	1+2
Worm	10	100	6e-4	1+2
Fly	0.4	100	0.5	1+2

Table 7: Source [19]. Cross-validation results for protein interaction prediction, using 5-fold cross-validation.

6.3.3 Prediction of Protein Interactions

The multiple network alignment, rather than predict functionality, can also be used to predict protein–protein physical interactions, based on the following strategies:

1. Evidence that proteins with similar sequences interact within other species (directly or by a common network neighbor)

2. (optionally) Cooccurrence of these proteins in the same conserved cluster or path.

The accuracy of these predictions was evaluated by using 5-fold cross validation. In cross-validation, strategy 1 achieved 77%-84% specificity and 23%–50% sensitivity, depending on the species for which the predictions were made (see Table 7). These results were highly significant for the three species. Combining both strategies resulted in eliminating virtually all false positive predictions (specificity $> 99\%$), while greatly reducing the number of true positives, yielding sensitivities of 10% and lower (again, see Table 7).

To further evaluate the utility of protein interaction prediction based on network conservation, 65 of the interactions that were predicted for yeast were tested by using the combined strategies 1 and 2 from above (see Figure 30a). The tests were performed by using yeast-two hybrid, using a bait and a prey (see Figure 30b). Five of the tests involved baits that induced reporter activity in the absence of any prey (Figure 30c). Of the remaining 60 putative interactions, 31 tested positive (more conservatively, 19 of 48, see Figure 30), yielding an overall success rate in the range of 40%-52%. (For more on yeast-two hybrid see scribe of lecture 7).

Two hybrid tests of predicted interaction were performed in order to validate the predicted interactions. Those tests yielded a success range of 40%–52%. These are satisfactory results for three reasons:

1. The performance is clearly significant compared to the chance of identifying protein interactions in random.

2. Two-hybrid analysis is known to miss a substantial portion of true interactions([25]), especially in this case, in which the protein pairs were checked in only one orientation of the bait and the pray.

3. Predicting interactions by using a multiple network alignment approach compares favorably to previous approaches based on conservation of individual protein interactions.

6.4 Further Issues

6.4.1 Comparison to Existing Methods

In [18] and in [10], Sharan *et al.* introduce pairwise network alignment algorithms, which he uses to identify conserved paths and complexes among the PPI networks of yeast and bacteria. In this article Sharan *et al.* extend this approach to handle more than two species, using the multiple alignment graph. Additional advantages of this approach are:

- It is a unified method to detect both paths and clusters, which generalizes to other network structures.
- It incorporates a refined probabilistic model for protein interaction data.
- It includes an automatic system for laying out and visualizing the resulting conserved subnetworks.

Also, a related method called *interolog* ([12],[22]) uses cross-species data for predicting protein interactions: a pair of proteins in one species is predicted to interact if their best sequence matches in another species were reported to interact. In contrast, the approach in [19] can associate proteins that are not necessarily each other's best sequence match. This advantage confers increased flexibility in detecting conserved function by allowing for paralogous family expansion and contraction or gene loss.

Another observation is that best BLAST values may not imply functional conservation. Frequently, the network alignment associates sequence-similar proteins between species even though they are not each other's best sequence match. Clearly, in some cases, the best matches are not present within conserved clusters because of missing interactions in the protein networks of one or more species. However, it is unlikely that true interactions with the best-matching proteins would be missed repeatedly across multiple proteins in a cluster and across multiple species. These observations suggest that protein network comparisons provide essential information about function conservation.

6.4.2 Functional Links Within Conserved Networks

Conserved network regions that are enriched for several functions point to cellular processes that may work together in a coordinated fashion. But, because of the appreciable error rates inherent in measurements of protein-protein interactions, an interaction in a single species linking two previously unrelated processes would typically be ignored as a false positive.

However, an observation that two or three networks reinforce this interaction is considerably more compelling, especially when the interaction is embedded in a densely connected conserved network region. For example, Figure 27 h links protein degradation to the process of poly(A)RNA elongation. Although these two processes are not connected in this region of the yeast network, several protein interactions link them in the networks of worm and fly (e.g., Pros25-Rack1-Msi or Pros25-Rack1-Tbph).

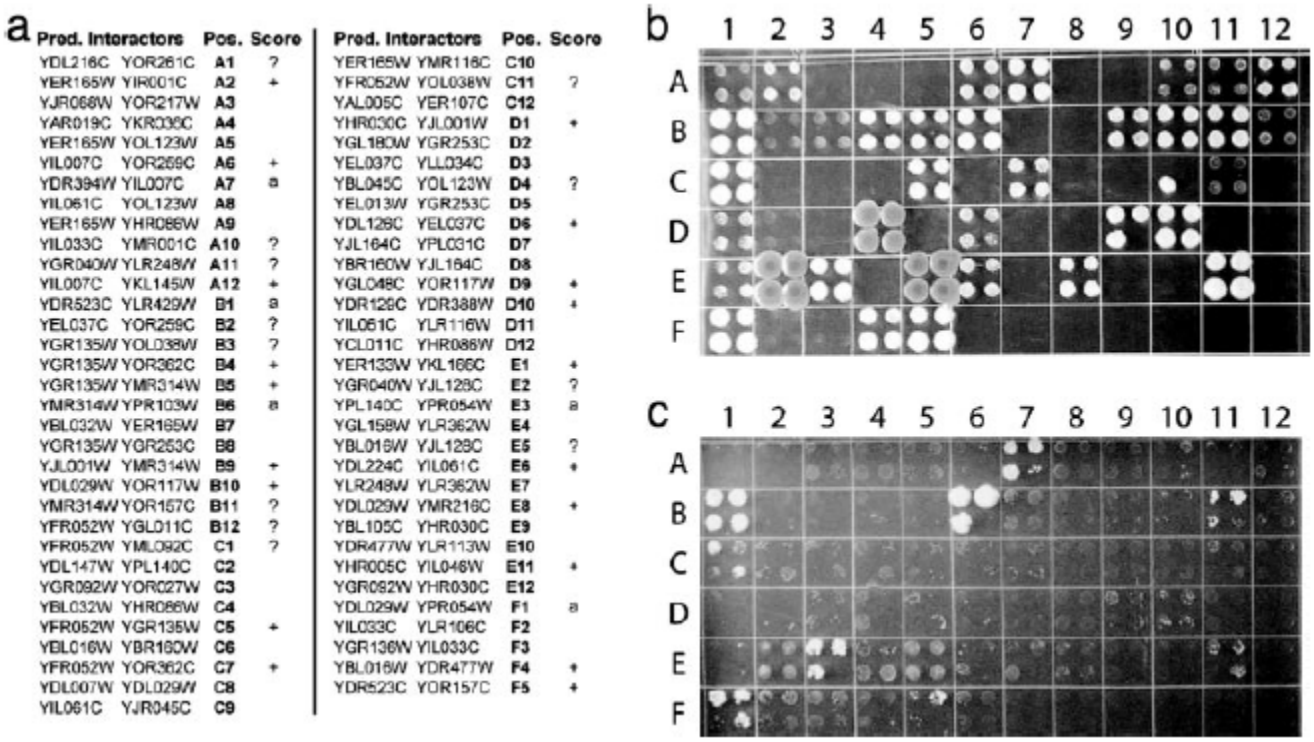

Figure 30: Source [19]. This figure shows verification of predicted interactions by two-hybrid testing. (a) Sixty-five pairs of yeast proteins were tested for physical interaction based on their co-occurrence within the same conserved cluster and the presence of orthologous interactions in worm and fly. Each protein pair is listed along with its position on the agar plates shown in b and c and the outcome of the two-hybrid test. (b) Raw test results are shown, with each protein pair tested in quadruplicate to ensure reproducibility. Protein 1 vs. 2 of each pair was used as prey vs. bait, respectively. (c) This negative control reveals activating baits, which can lead to positive tests without interaction. Protein 2 of each pair was used as bait, and an empty pOAD vector was used as prey. Activating baits are denoted by ”a” in the list of predictions shown in a. Positive tests with weak signal (e.g., A1) and control colonies with marginal activation are denoted by ”?” in a; colonies D4, E2, and E5 show evidence of possible contamination and are also marked by a ”?”. Discarding the activating baits, 31 of 60 predictions tested positive overall. A more conservative tally, disregarding all results marked by a ”?” yields 19 of 48 positive predictions.

7 Summary

We started with section 2 where we described the network alignment problem, and the network querying problem. In sections 3 and 4 we described the pairwise alignment, and its advantages in finding conserved protein paths and complexes in the PPI networks using two different algorithms. In section 5, as opposed to the previous sections, we showed how we can use the PPI networks in order to predict functional orthology. In section 6 we described a model that extends the pairwise alignment model to a multiple alignment.

Network alignment is obviously a very powerful tool for identifying functional modules and predicting protein function, interactions between proteins and proteins orthology.

References

[1] N. Alon, R. Yuster, and U. Zwick. Color–coding. *J. ACM*, 42(4):844–856, 1995.

[2] 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 Research*, 25:3389–3402, 1997.

[3] G.D. Bader, I. Donaldson, C. Wolting, B. Ouellette, T. Pawson, and C. Hogue. Bind–the biomolecular interaction network database. *Nucleic Acids Research*, pages 242–245, 2001.

[4] S. Bandyopadhyay, R. Sharan, and T. Ideker. Systematic identification of functional orthologs based on protein network comparison. *Genome Research*, 16(3):428–435, 2006.

[5] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, B:192–236, 1974.

[6] C.E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60. Rome, 1935.

[7] A.C. Gavin et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.

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

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

[10] R.B. Kelley, R. Sharan, R. Karp, T. Sittler, D. Root, B. Stockwell, and T. Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *PNAS*, 100(20):11394–11399, 2003.

[11] M. Koyuturk, Y., Kim, U. Topkara, S. Subramaniam, W. Szpankowski, and A. Grama. Pairwise alignment of protein interaction networks. *RECOMB*, 2005.

[12] R. Matthews, P. Vaglio, J. Reboul, H. Ge, B.P. Davis, J. Garrels, S. Vincent, and M. Vidal. Identification of potential interaction networks using sequence–based searches for conserved protein–protein interactions or "interologs". *Genome Research*, 11:2120–2126, 2001.

[13] 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.*, 28:40214028, 2000.

[14] R. Pastor-Satorras, E. Smith, and R.V. Sole. Evolving protein interaction networks through gene duplication. *Journal of Theoretical Biology*, 222:199–210, 2003.

[15] M. Remm, C.E. Storm, and E.L. Sonnhammer. Automatic clustering of orthologs and in–paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314:1041–1052, 2001.

[16] R. Shamir, R. Sharan, and D. Tsur. Cluster graph modification problems. *In Proceedings of 28th WG*, 2573:379–390, 2002.

[17] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Decker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:2498–2504, 2003.

[18] 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. *Journal of Computational Biology*, 12(6):835–846, 2005.

[19] R. Sharan, S. Suthram, R.M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R.M. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *PNAS*, 102(6):1974–1979, 2004.

[20] T. Shlomi, D. Segal, E. Ruppin, and R. Sharan. Qpath: a method for querying pathways in a protein–protein interaction network. *BMC Bioinformatics*, 7:199–208, 2006.

[21] A. Smith and G. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society*, B55:3–23, 1993.

[22] J.M. Stuart, E. Segal, D. Koller, and S.K. Kim. A gene–coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003.

[23] 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.

[24] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction netwokrs. *Complexus Journal*, 1:38–44, 2003.

[25] C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, and P. Bork. Comparative assessment of large–scale data sets of protein–protein interactions. *Nature*, 417:399–403, 2002.

[26] A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular Biology and Evolution*, 18(7):1283–1292, 2001.

[27] A. Wagner. How the global structure of protein interaction networks evolves. *Proceedings Royal Society of London Biological Sciences*, 270(1514):457–466, 2003.

[28] S. Wuchty, Z.N. Oltvai, and A.L. Barabasi. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, 35(2):176–179, 2003.

[29] I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, and D. Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30:303–305, 2002.

[30] H. Xie, A. Wasserman, Z. Levine, A. Novik, V. Grebinskiy, A. Shoshan, and L. Mintz. Large–scale protein annotation through gene ontology. *Genome Research*, 12:785–794, 2002.