BIOINFORMATICS

Vol. 23 ECCB 2006, pages e170–e176
doi:10.1093/bioinformatics/btl295

Identification of conserved protein complexes based on a model of protein network evolution

Eitan Hirsh and Roded Sharan*

School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

ABSTRACT

Motivation: Data on protein–protein interactions (PPIs) are increasing exponentially. To date, large-scale protein interaction networks are available for human and most model species. The arising challenge is to organize these networks into models of cellular machinery. As in other biological domains, a comparative approach provides a powerful basis for addressing this challenge.

Results: We develop a probabilistic model for protein complexes that are conserved across two species. The model describes the evolution of conserved protein complexes from an ancestral species by protein interaction attachment and detachment and gene duplication events. We apply our model to search for conserved protein complexes within the PPI networks of yeast and fly, which are the largest networks in public databases. We detect 150 conserved complexes that match well-known complexes in yeast and are coherent in their functional annotations both in yeast and in fly. In comparison with two previous approaches, our model yields higher specificity and sensitivity levels in protein complex detection.

Availability: The program is available upon request.

Contact: roded@tau.ac.il

1 INTRODUCTION

Recent technological advances enable the systematic characterization of protein–protein interaction (PPI) networks across multiple species. Procedures such as yeast two-hybrid (Ito *et al*., 2001) and protein co-immunoprecipitation (Mann *et al*., 2001) are routinely employed nowadays to generate large-scale protein interaction networks for human and most model species (Uetz *et al*., 2000; Ito *et al*., 2000; Ho *et al*., 2002; Gavin *et al*., 2002; Stelzl *et al*., 2005; Raul *et al*., 2005). An arising challenge is to organize the accumulating network data into models of cellular machinery. As in other biological domains, a comparative approach provides a powerful basis for addressing this challenge, calling for better understanding of protein network evolution.

Two types of processes have been invoked to explain the evolution of PPI networks (Wagner, 2001; Berg *et al*., 2004): link dynamics and gene duplication. The first consists of sequence mutations in a gene that result in modifications of the interface between interacting proteins. Consequently, the corresponding protein may gain new connections (attachment) or lose (detachment) some of the existing connections to other proteins. The second consists of gene duplication, followed by either silencing of one of the duplicated genes or by functional divergence of the duplicates. The corresponding event in the network is the addition of a protein with the same set of interactions as the original protein, followed by the divergence of their links. Berg *et al*. (2004) estimated the empirical rates of link dynamics and gene duplication in the yeast protein network, finding the former to be at least one order of magnitude higher than the latter. Based on this observation, they proposed a model for the evolution of protein networks in which link dynamics are the major evolutionary forces shaping the topology of the network, while slower gene duplication processes mainly affect its size.

Previous approaches to the problem of identifying protein complexes within PPI data have shown the utility of a comparative analysis that overcomes the high levels of noise characterizing these data (Deng *et al*., 2003). Specifically, Sharan *et al*. (2005, 2004) have compared PPI networks from multiple species to pinpoint network regions that are conserved in evolution and have shown that these regions match well-known protein complexes in yeast. However, their scoring scheme treated the networks being compared as independent of one another and did not take into account the correspondence in interaction patterns between them (see detailed discussion of this issue in Section 4). Another approach by Koyuturk *et al*. (2005) applied an evolutionary based scoring scheme, which takes into account duplication and link turnover events. However, the scoring procedure was empirical with no underlying probabilistic model.

Here we develop a probabilistic model for protein complexes that are conserved across two species, which describes the evolution of conserved protein complexes from an ancestral species through link dynamics and gene duplication events. Pairs of extant complexes are scored by their fit to the model versus the likelihood that they arise at random. We apply our model to search for conserved protein complexes within the PPI networks of yeast and fly, which are the largest networks in public databases. We detect 150 significantly conserved complexes that match well-known complexes in yeast and are coherent in their functional annotations in both yeast and fly. In comparison with the two previous approaches described above, our model displays higher levels of specificity and sensitivity in protein complex detection.

The paper is organized as follows. Section 2 presents the probabilistic model for conserved protein complexes. Section 3 describes the process of searching for high-scoring conserved complexes, and presents measures for quality assessment. Section 4 presents the results of applying our algorithm to detect conserved complexes in the PPI networks of yeast and fly, as well as a comparison of the algorithm's performance to those of two existing approaches.

2 A PROBABILISTIC MODEL FOR CONSERVED PROTEIN COMPLEXES

In this section we present a probabilistic model for protein complexes that are conserved across two species. The model is based on

*To whom correspondence should be addressed.

specifying the pattern of interactions in an unobserved common ancestor of the two species and on describing the evolutionary events that have yielded the observed complexes in each of the species.

We first present the model assuming that the interaction data are accurate and complete, that is each interaction is true and each non-interaction is also true. We then generalize the model to account for probability assignment to the interactions to reflect their reliabilities. Our full model consists of a conserved protein complex model, M_C, and a null model, M_N. Candidate conserved complexes are scored by their ratio of likelihoods according to each of the models. In the following sections we describe these models and their underlying assumptions.

2.1. Conserved protein complex model

As suggested by Berg *et al*. (2004) the evolution of a PPI network is shaped by link turnover and gene duplication events. For a conserved protein complex, consisting of a pair of species-specific complexes, we assume the existence of an ancestral complex in the common ancestor of the two species under study, from which its current forms have evolved through duplication and link turnover events.

Let the two species under study be indexed by 0 and 1, respectively. Denote their sets of proteins by P_0 and P_1. Denote the set of proteins of a common ancestor of the two species by P. Let $\phi(\cdot)$ be a mapping from proteins in $P_0 \cup P_1$ to P, where $\phi(x) = \phi(y)$ for $x \neq y$ iff x and y are homologous. In other words, $\phi(x)$ is the ancestral protein from which x originated. Ways to compute ϕ are described in Section 4.1.

Consider a given conserved protein complex, and denote by S_0, S_1 and S the sets of proteins comprising it in species 0, species 1 and the ancestral species, respectively. Our model for the interaction pattern of the ancestral complex is based on the assumption that a protein complex induces a dense subnetwork of PPIs. This assumption is in agreement with known complexes and has already been used successfully by us in a previous work (Sharan *et al*., 2004). Specifically, we assume that within a complex each interaction occurs with high probability β, independently of the other protein pairs in the complex.

The interaction patterns of the extant protein sets, S_0 and S_1, are assumed to have evolved from the ancestral interaction pattern. Let m be the number of protein pairs in the ancestral complex S. For each of these pairs, $p_i = (a_i, b_i)$, let I_i be the set of equivalent pairs in S_0 and S_1 under ϕ: $I_i = \{(x, y) \in S_0 : \phi(x) = a, \phi(y) = b\} \cup \{(x, y) \in S_1 : \phi(x) = a, \phi(y) = b\}$. We assume that each interaction in I_i evolved from p_i, independently of all other events, i.e. interactions are attached with some probability P_A and detached with probability P_D[1].

To handle duplications in extant species, we have to specify separately our assumption regarding interactions between duplicates, since such interactions did not evolve from an ancestral protein pair as the duplication is assumed to have happened after the speciation event. We choose to treat such interaction in the same manner that we treat interactions in the ancestral species and assume that they occur with probability β independently of all other protein pairs.

For two proteins x,y, let us denote by T_{xy} the event that these two proteins interact and by F_{xy} the event that they do not interact. Let $O_{xy} \in \{0,1\}$ denote the observation on whether x and y interact. Let O_S denote the entire set of observations on the members of S. Let D_S be the set of duplicated pairs in S. We can finally state the likelihood of a set of observations on a conserved complex:

$$P(O_{S_0}, O_{S_1} \mid M_C) = \prod_{i=1}^{m} P(O_{I_i} \mid M_C) \cdot \prod_{(x,y) \in D_{S_0} \cup D_{S_1}} P(O_{xy} \mid M_C),$$

where

$$\begin{aligned} P(O_{I_i} \mid M_C) =& P(O_{I_i} \mid T_{a_i b_i}) P(T_{a_i b_i} \mid M_C) \\ &+ P(O_{I_i} \mid F_{a_i b_i}) P(F_{a_i b_i} \mid M_C) \\ =& \beta P(O_{I_i} \mid T_{a_i b_i}) + (1 - \beta) P(O_{I_i} \mid F_{a_i b_i}) \end{aligned}$$

and

$$\begin{aligned} P(O_{I_i} \mid T_{a_i b_i}) &= \prod_{x,y \in I_i} P(O_{xy} \mid T_{a_i b_i}) \\ &= \prod_{x,y \in I_i} P_D^{[O_{xy}=0]} (1 - P_D)^{[O_{xy}=1]} \\ P(O_{I_i} \mid F_{a_i b_i}) &= \prod_{x,y \in I_i} P_A^{[O_{xy}=1]} (1 - P_A)^{[O_{xy}=0]}. \end{aligned}$$

2.2 The null model

The null model assumes that each edge in the PPI networks of the two species is present with probability that one would expect if the edges were randomly distributed, but respected vertex degrees. Formally, for a given PPI network G and a given protein pair (x, y), the probability that x and y interact is defined as the fraction of graphs with the same degree sequence as G that contain an edge between x and y. We estimate these probabilities using a Monte-Carlo approach as suggested by Sharan *et al*. (2004). This allows us to compute

$$P(O_{S_0}, O_{S_1} \mid M_N) = \prod_{x,y \in S_0} P(O_{xy} \mid M_N) \cdot \prod_{x,y \in S_1} P(O_{xy} \mid M_N).$$

2.3 Putting it all together

The above description assumed that interactions and non-interactions are known. In practice, we have partial, noisy observations on PPIs. As in Sharan *et al*. (2004), we tackle this problem by generalizing our model to consider the interaction data as noisy observations. To this end, we redefine O_{xy} as the set of experimental observations on whether x and y interact (rather than denoting their status of interaction, which is unknown). As before, let T_{xy} and F_{xy} denote the hidden events of whether x and y interact or not, respectively. We can now use Bayes theorem to compute the likelihood of the observations on an interaction given some model M as follows:

$$\begin{aligned} P(O_{xy} \mid M) = \; & P(O_{xy} \mid T_{xy}) P(T_{xy} \mid M) \\ & + P(O_{xy} \mid F_{xy}) P(F_{xy} \mid M). \end{aligned}$$

$P(T_{xy} \mid M)$ and $P(F_{xy} \mid M)$ are computed as described above [where T_{xy} (F_{xy}) corresponds to the event $O_{xy} = 1$ ($O_{xy} = 0$) in the previous

[1]Note that P_A and P_D are related: empirical evidence suggests that the overall rate of interaction attachment equals that of interaction detachment (Berg *et al*., 2004).

notation]. $P(O_{xy} \mid T_{xy})$ and $P(O_{xy} \mid F_{xy})$ can be computed from interaction reliabilities as shown in Sharan *et al*. (2005).

Finally, the likelihood ratio score that we assign to a putative conserved complex is

$$L(O_{S_0}, O_{S_1}) = \frac{P(O_{S_0}, O_{S_1} \mid M_C)}{P(O_{S_0}, O_{S_1} \mid M_N)}.$$

3 SEARCHING FOR CONSERVED COMPLEXES AND VALIDATION CRITERIA

A common approach to the problem of identifying conserved complexes, which we adopt in this work, is the use of an alignment graph in which the two studied networks are compared (Kelley *et al*., 2003; Sharan *et al*., 2005). We describe this approach briefly below.

Given PPI data for two species, we translate it into two separate interaction graphs, one for each species. In an interaction graph, each node is a protein and each edge corresponds to a PPI. These two graphs are then combined into a network alignment graph in which each node represents a pair of sequence-similar proteins, one from each species, and each edge represents a conserved interaction between the corresponding protein pairs within each species. More precisely, in our setting two nodes (u, u') and (v, v') are linked if at least one of the pairs (u, v),(u', v') is observed to interact and the other pair spans proteins of distance at most two in the corresponding interaction graph.

By construction, an induced subgraph of the alignment graph corresponds to two species-specific sets of proteins, S_0 and S_1, and is assigned the score $L(S_0, S_1)$. We perform a bottom-up search for heavy subgraphs in the alignment graph, starting with seeds around each of the nodes, constructed as follows: for each node i in the alignment graph we identify a neighbor p_i of i such that the weight of this pair is maximum among all pairs (i, v), where v is adjacent to i. We use as seeds around node i high weight 4-node subsets that consist of i, $p(i)$ and two of their neighbors. These seeds are then expanded by local search, each time adding or deleting a node whose modification increases the weight of the current subgraph the most (Sharan *et al*., 2005). The resulting subgraphs may overlap considerably, and we use a greedy algorithm to filter them, so that the intersection of any two subgraphs in their node sets and in their species-specific protein sets is below a threshold (80%; computed w.r.t. the smaller set). The algorithm iteratively finds the highest scoring subgraph, adds it to an output list and removes all the subgraphs that (sufficiently) intersect it from consideration.

3.1 Significance evaluation

The output of the previous stage undergoes further filtering to remove non-significant findings. The statistical significance of the subgraphs is evaluated by comparing their scores with those obtained on randomized instances of the data. These instances are created by shuffling the edges of the two interaction graphs while preserving vertex degrees, as well as shuffling the pairs of sequence-similar proteins while preserving the number of homologs per protein. This process yields empirical P-values for the output subgraphs; only significant results with $P < 0.05$ are retained. Henceforth, we call the significant subgraphs detected conserved clusters.

3.2 Quality assessment

We used four measures to evaluate the biological significance of the results. The first three quantify the similarity between a given collection of conserved clusters and a reference, putatively true, catalog of protein clusters. As a reference we used known yeast clusters cataloged in the MIPS Database (2005, http://mips.gsf.de/) (we excluded category 550, which was obtained from high throughput experiments, and retained only manually annotated clusters). The fourth measure assesses the functional coherency of the conserved clusters based on the gene ontology (GO) annotation (The Gene Ontology Consortium, 2000). These measures are described below.

Specificity and sensitivity To measure the level of correspondence between conserved clusters and true complexes, we computed statistically significant matches between the two collections and used these matches to evaluate the specificity and sensitivity of the suggested solution. Specifically, for each conserved cluster we found a true complex with which its intersection was the most significant according to a hypergeometric score. Significance levels were compared with those obtained for 10 000 random sets of proteins of the same size, and empirical P-values were calculated for each of the conserved clusters. These P-values were further FDR corrected for testing multiple conserved clusters. Let C be the initial set of conserved clusters, and let $C^* \subseteq C$ be the subset of clusters that had a significant match ($P < 0.05$; only clusters with at least one annotated protein are considered). The specificity of the solution is defined as $|C^*|/|C|$. Let M be the set of true complexes, and let $M^* \subseteq M$ be the subset of complexes with a significant match by a conserved cluster. The sensitivity of the solution is defined as $|M^*|/|M|$.

Purity This is an alternative measure for the specificity of the solution. A conserved cluster is called pure if there exists a true complex whose intersection with the cluster covers at least 75% of the MIPS annotated proteins in the cluster (considering only clusters with at least 3 MIPS annotated proteins). Let C be the set of all clusters with at least 3 MIPS annotated proteins, and let C^* be a subset of pure clusters. The purity of the solution is defined as $|C^*|/|C|$.

Functional enrichment We used the GO process annotation for yeast and fly to evaluate the functional coherency of the conserved clusters returned by the algorithm. For each cluster and each GO term, we computed the enrichment of the term in the cluster using a specially designed hypergeometric score, which takes into account ontology relations between terms. Specifically, since the GO terms are not independent but are rather connected by an ornithology of parent–child relationship, we computed the enrichment of each term conditioned on the enrichment of its parent term, as done in Sharan *et al*. (2005) (see also Grossmann *et al*., 2006). For each cluster we chose the term that yielded the highest significance level. We compared this significance level with those obtained for random sets of proteins of the same size as the cluster and derived an empirical P-value for the cluster. These P-values were further FDR corrected for multiple testing. Finally, we report the fraction of functionally enriched clusters ($P < 0.05$; only clusters with at least one GO annotated protein are considered). This procedure is done separately for yeast and fly.

Table 1. High-scoring conserved clusters

Cluster ID	Size	MIPS category	*P*-value	Yeast GO process	*P*-value	Fly GO process	*P*-value
#114	5	Cytoskeleton	0.01	Structural constituent of cytoskeleton	0.0048	Structural constituent of cytoskeleton	0.015
#225	7	RNA processing	0.037	pre-mRNA splicing factor activity	0.0206	RNA-binding	0.0029
#342	7	Proteasome	0.0002	Proteasome endopeptidase activity	0.0001	Endopeptidase activity	0.0035
#479	7	Replication	0.0001	DNA clamp loader activity	0.0001	Nucleotidyl transferase activity	0.007199

High-scoring conserved clusters identified by our algorithm. For each cluster, shown are its size, best matching MIPS complex (and *P*-value), and most enriched GO annotations in yeast and fly (and *P*-values). MIPS identifiers for the categories mentioned above are as follows: Cytoskeleton, 140.20.20, RNA processing, 440.30.10, Proteasome, 360.10.10 and Replication, 410.40.30.

4 EXPERIMENTAL RESULTS

We applied our method to search for conserved protein complexes in the PPI networks of yeast (*Saccharomyces cerevisiae*) and fly (*Drosophila melanogaster*), which are the two largest networks in public databases. In the following sections we describe our results and present a comparison with two existing methods for conserved complex detection by Sharan *et al*. (2005) and Koyuturk *et al*. (2005).

4.1 Data description and parameter estimation

We downloaded protein interaction data for yeast and fly from the database of interacting proteins [DIP Database (July 2005 download, http://dip.doe-mbi.ucla.edu/)]. The yeast network contained 15 147 interactions, spanning 4738 proteins; the fly network contained 23 484 interactions, spanning 7165 proteins. We used a previously published logistic regression method (Sharan *et al*., 2005) to assign reliabilities to the PPIs. The reliabilities were based only on the experimental evidence for each interaction.

The network alignment graph was constructed over pairs of proteins with some interaction information whose BLAST *E*-value $\leq 10^{-10}$. Overall, it contained 890 nodes and 1070 edges, spanning 482 and 453 distinct proteins in yeast and fly, respectively.

To determine duplicates and cluster extant proteins according to their ancestral origin, we used the InParanoid algorithm (Remm *et al*., 2001). InParanoid clusters sequence-similar proteins from two species, so that each cluster corresponds to one ancestral protein and contains its present-day descendants and their in-paralogs (duplications after the speciation event). We used an Inparanoid clustering for yeast and fly from InParadoid Database (December 2005 download, http://inparanoid.cgb.ki.se/), containing 1128 clusters over the proteins in the yeast and fly PPI networks. Note that nodes of the alignment graph may contain pairs of proteins that do not map to the same InParanoid cluster. The inclusion of these nodes reflects our previous observations that functional orthology does not necessarily imply sequence orthology (Sharan *et al*., 2005; Bandyopadhyay *et al*., 2006).

While previous works have tried to estimate the probabilities of edge attachment and detachment (Wagner, 2001; Berg *et al*., 2004), these computations were limited to mean estimates over the entire PPI network and do not directly apply to estimating the rate of these events within conserved complexes. Hence, we set these values empirically as follows: P_D, the probability that an interaction within a conserved complex is removed in an extant species, was set to 0.01. The algorithm had similar performance when varying P_D from 0.01 to 0.1. P_A, the probability that an interaction is introduced into a conserved complex, was computed from P_D assuming that the rate of interaction attachment within conserved complexes is equal to that of interaction detachment, as is the case over the entire network (Berg *et al*., 2004). P_A attained a value of $\sim$0.001. The last parameter, β, was set to 0.8 as in our previous work Sharan *et al*. (2005), and similar results were obtained when varying β from 0.7 to 0.9.

For validation purposes, we downloaded the MIPS complex catalog (December 2005 download, http://mips.gsf.de/). We used complexes at level 3 or lower with at least one protein in the yeast PPI network. Overall, there were 113 such complexes spanning 697 proteins; 68 of these complexes had at least 3 proteins in the network. We also extracted 4818 and 6140 GO process annotations for yeast and fly, respectively (December 2005 download).

4.2 Application to yeast–fly PPI data

We applied our algorithm to the yeast–fly network alignment graph in search for conserved protein clusters. The algorithm identified 150 significant, non-redundant conserved clusters spanning 224 proteins in yeast and 196 proteins in fly. The sizes of the clusters ranged from 4 to 14, with an average size of 7. Four representative, high-scoring conserved clusters are detailed in Table 1 and depicted in Figure 1.

We assessed the biological significance of the conserved clusters by comparing them with known MIPS complexes and testing their functional enrichment (see Section 3.2 for a description of the measures we used). Of the clusters, 94 significantly matched a MIPS complex, yielding a specificity level of 76% and a sensitivity level of 19%. Moreover, 78% of the clusters had an enriched GO annotation in yeast, and 43% were enriched for fly annotations. The enriched annotations in the two species matched in the majority of the cases, as exemplified by the clusters in Table 1. Further information on the identified clusters is given in Table 1.

4.3 Comparison with extant methods

We compared our approach with two previously published methods: (1) NetworkBLAST by Sharan *et al*. (2005), which is based on a similar probabilistic model, but treats the two species independently in its score and (2) MaWish by Koyuturk *et al*. (2005), which is based on evolutionary principles but has no underlying probabilistic model.

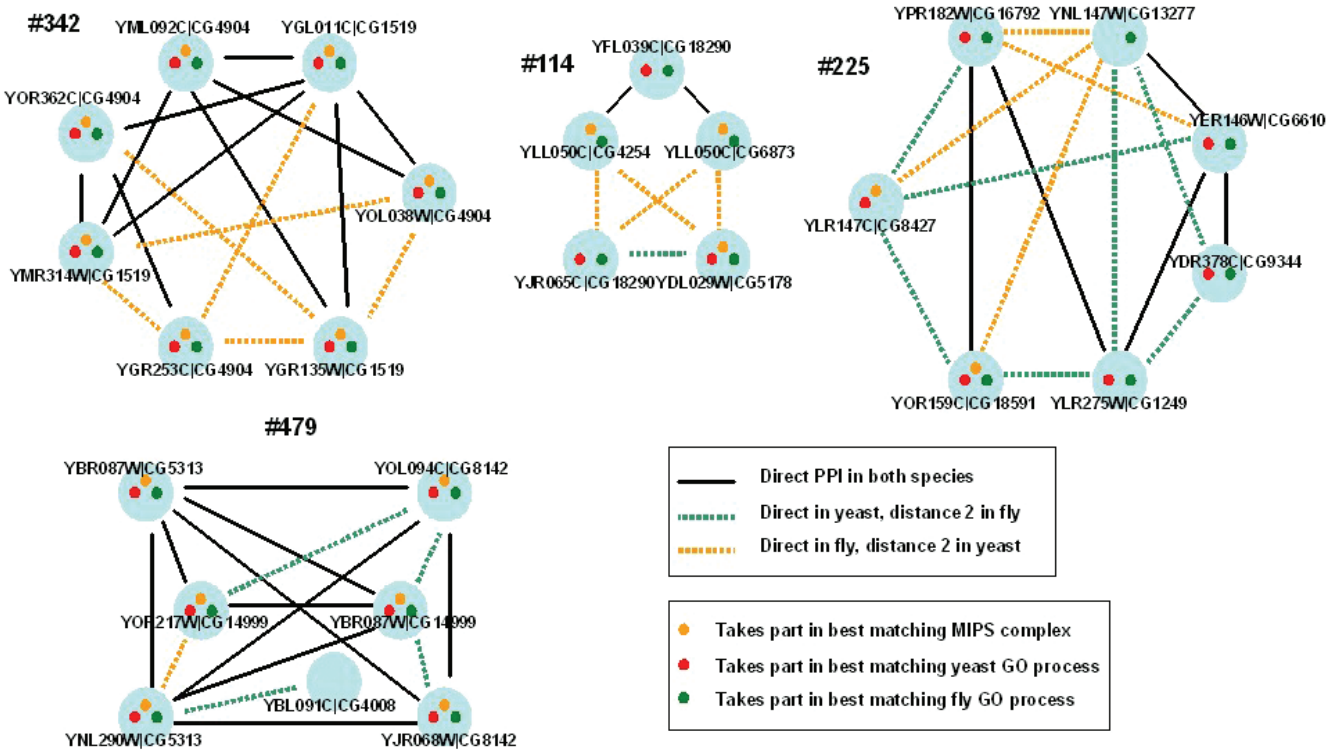

Fig. 1. Illustration of four high-scoring conserved clusters presented in Table 1. Shown are the alignment subgraphs corresponding to each conserved cluster. Nodes represent pairs of proteins, one from each species. Edges represent conserved (solid) or semi-conserved (dashed; direct in one species and distance 2 in the other) interactions. Edges spanning a direct interaction in one species and the same protein in the other species also appear solid. Colors within nodes indicate whether they participate in the best matching MIPS complex or GO term.

Table 2. A comparison of our algorithm with two existing approaches for conserved complex detection

Algorithm	No. of complexes	% Intersection	Specificity (%)	Purity (%)	Sensitivity (%)	Functional enrichment	
						Yeast (%)	Fly (%)
This study	150	—	76	70	19	78	43
NetworkBLAST (Sharan, 2005)	146	89	74	65	19	79	46
MaWish (Koyuturk, 2005)	97	25	69	55	13	67	38

Performance measures of our algorithm, NetworkBLAST and MaWish when applied to the yeast–fly alignment graph. Details on all measures can be found in Section 3.2. The third column specifies the percentage of overlapping clusters with our solution ($\geq$80% overlap).

In order to allow a fair comparison we used the same PPI networks, alignment graph, search heuristic and validation methods, thus emphasizing the scoring component of each method. Table 2 summarizes the performance of the three methods when applied to the yeast–fly alignment graph. It can be seen that our algorithm outperforms MaWish by a significant margin in all measured parameters and manages to discover 1.5-fold more significant conserved clusters. Overall, the solutions are very different with only 25% intersection. In comparison with NetworkBLAST, our algorithm has an overall similar performance, which is reflected also in the high overlap between the two solutions (89%). Nevertheless, our algorithm exhibits better correspondence with the MIPS catalog, with higher specificity and purity levels than those attained by NetworkBLAST.

Due to the overall similarity between the solutions of our algorithm and NetworkBLAST, we conducted a more refined analysis of the differences between the two approaches. Intuitively, if we consider two species-specific clusters spanning matching sets of proteins, NetworkBLAST will not distinguish between the case that the interaction sets of the two clusters identify and the case that the interactions sets are randomly distributed w.r.t. each other (Fig. 2). Thus, the key difference between the two approaches is the way they treat conserved interactions within conserved clusters. While the scoring of NetworkBLAST depends only on the total

Table 3. A comparison of our algorithm and NetworkBLAST on clusters that contain conserved interactions

Collection	No. of complexes	Specificity (%)	Purity (%)	Sensitivity (%)	Functional Enrichment	
					Yeast (%)	Fly (%)
This study, $k = 1$	105	72	57	18	75	60
NetworkBLAST, $k = 1$	103	70	48	18	74	55
This study, $k = 4$	21	78	58	8	76	95
NetworkBLAST, $k = 4$	19	56	38	6	59	74

Performance measures of our algorithm and NetworkBLAST w.r.t. conserved clusters containing at least k conserved interactions, for $k = 1,4$. Columns are as in Table 2.

Table 4. A comparison of our algorithm with two existing approaches on the conserved core of the yeast–fly alignment graph

Algorithm	No. of complexes	Specificity (%)	Purity (%)	Sensitivity (%)	Functional enrichment	
					Yeast (%)	Fly (%)
This study	94	85	60	13	82	45
NetworkBLAST	78	79	56	12	81	44
MaWish	30	77	53	7	83	52

Performance measures of our algorithm, NetworkBLAST and MaWish w.r.t. the conserved core of the yeast–fly alignment graph. Columns are as in Table 2.

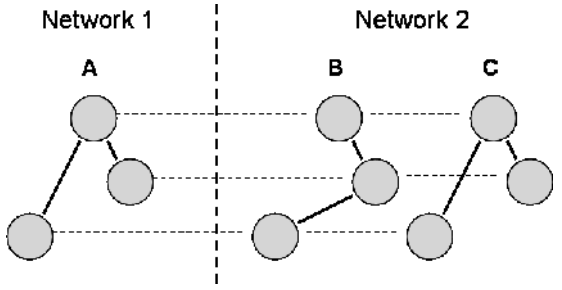

Fig. 2. A toy example demonstrating the difference between our model and NetworkBLAST. Shown are three clusters: A in network 1; B and C in network 2. Solid lines represent interactions, and dotted lines represent sequence similarity. Our model will favor a conserved cluster containing A and C, while NetworkBLAST will not distinguish between the pairs (A,B) and (A,C).

number of interactions within each species, our model distinguishes between a conserved interaction and a pair of species-specific interactions with no match in the other species.

In the light of the discussion above, we focused the comparison with NetworkBLAST on clusters containing conserved interactions. We recomputed the quality measures of the two solutions when restricting the computations to conserved clusters that contain at least k conserved interactions, for $k = 1,4$. The existence and biological significance of such clusters are supported by empirical observations on the tendency of interaction conservation across species (Matthews *et al*., 2001). The results, summarized in Table 3, demonstrate the superiority of our algorithm in this setting.

Moreover, we also applied the two algorithms to a conserved core of the network data, obtained by considering only proteins that participate in nodes of the alignment graph that are involved in a conserved interaction. Again, our new algorithm is shown to outperform NetworkBLAST (Table 4). For comparison purpose, we also detail the performance of MaWish on these data. Evidently, it is less aligned with the MIPS complex data, although displaying high functional enrichment levels.

5 CONCLUSIONS

We have presented a probabilistic model for the detection of conserved complexes across two species based on the evolutionary processes shaping their networks. Our model has relatively few parameters related to the density of protein complexes and to the determination of gene duplications and link turnover rates. We applied our approach to study the conservation between the PPI networks of yeast and fly. We successfully identified putatively conserved complexes that matched well-known complexes in yeast and displayed functional coherency in both species. Moreover, we have shown that our model aligns with the biological data better than previous approaches. We expect our model to be more advantageous when comparing evolutionarily closer PPI networks as those become available. The probabilistic framework we have devised is extensible to more than two species and such extension is expected to assist in overcoming the high noise rates in current network data.

ACKNOWLEDGEMENTS

The authors thank Trey Ideker for helpful discussions about the work and Tomer Shlomi for suggesting the specificity and sensitivity measures. R.S. is supported by an Alon Fellowship and by grant 3-2589 from the Ministry of Science and Technology, Israel.

REFERENCES

Bandyopadhyay,S. *et al*. (2006) Systematic identification of functional orthologs based on protein network comparison. *Genome Res*., **16**, 428–435.

Berg,J. *et al*. (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol. Biol*., **4**, 51.

Deng,M. *et al*. (2003) Assessment of the reliability of protein–protein interactions and protein function prediction. *Pac. Symp. Biocomput*., 140–151.

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

Grossmann,S. *et al*. (2006) An improved statistic for detecting over-represented gene orthology annotations in gene sets. *Proc. RECOMB*.

Ho,Y. *et al*. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.

Ito,T. *et al*. (2001a) Exploring the protein interactome using comprehensive two-hybrid projects. *Trends Biotechnol*., **19** (Suppl. 10), 23–27.

Ito,T. *et al*. (2001b) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Kelley,B.P. *et al*. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc. Natl Acad. Sci. USA*, **100**, 11394–11399.

Koyuturk,M. *et al*. (2005) Pairwise alignment of protein interaction networks. *J. Comput. Biol*., **13**, 182–199.

Mann,M. *et al*. (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem*., **70**, 437–473.

Matthews,L.R. *et al*. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or 'interologs'. *Genome Res*., **11**, 2120–2126.

Remm,M. *et al*. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Mol. Biol*., **314**, 1041–1052.

Raul,J.F. *et al*. (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature,*, **437**, 1173–1178.

Sharan,R. *et al*. (2004) Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J. Comput. Biol*., **12**, 835–846.

Sharan,R. *et al*. (2005) Conserved patterns of protein interaction in multiple species. *Proc. Natl Acad. Sci. USA*., **102**, 1974–1979.

Stelzl,U. *et al*. (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.

The Gene Ontology Consortium (2000), Gene ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Uetz,P. *et al*. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Wagner,A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol*., **18**, 1283–1292.