Conserved patterns of protein interaction in multiple species

Roded Sharani∗†, Silpa Suthram‡, Ryan M. Kelley‡, Tanja Kuhn§, Scott McCuine‡, Peter Uetz§, Taylor Sittler‡, Richard M. Karp∗‡, and Trey Ideker∗‡

∗Computer Science Division, University of California, and International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704; †Department of Bioengineering, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093; and §Institute of Genetics, Research Center Karlsruhe, Postfach 3640, D-76021 Karlsruhe, Germany

Contributed by Richard M. Karp, December 22, 2004

To elucidate cellular machinery on a global scale, we performed a multiple comparison of the recently available protein–protein interaction networks of Caenorhabditis elegans, Drosophila melanogaster, and Saccharomyces cerevisiae. This comparison integrated protein interaction and sequence information to reveal 71 complexes and pathways conserved across all three species. We tested 60 interaction predictions for yeast by an exclusive to the metazoans. We used this conservation, and network regions that were conserved across all three species and found statistically significant support for 4,645 previously undescribed protein interactions. We tested 60 interaction predictions for yeast by two-hybrid analysis, confirming approximately half of these. Significantly, many of the predicted functions and interactions would not have been identified from sequence similarity alone, demonstrating that network comparisons provide essential biological information beyond what is gleaned from the genome.

A major challenge of postgenomic biology is to understand the complex networks of interacting genes, proteins, and small molecules that give rise to biological form and function. Advances in whole-genome approaches are now enabling us to characterize these networks systematically, by using procedures such as the two-hybrid assay (1) and protein coimmunoprecipitation (2) to screen for protein–protein interactions. To date, these technologies have generated large interaction networks for bacteria (3), yeast (4–7), and, recently, fruit fly (8) and nematode worm (9).

The large amount of protein interaction data now available presents opportunities and challenges in understanding evolution and function. Such challenges involve assigning functional roles to interactions (10), separating true protein–protein interactions from false positives (11), and, ultimately, organizing large-scale interaction data into models of cellular signaling and regulatory machinery. As is often the case in biology, an approach based on evolutionary cross-species comparisons provides a valuable framework for addressing these challenges. However, although methods for comparing DNA and protein sequences have been a mainstay of bioinformatics over the past 30 years, development of similar tools at other levels of biological information, including protein interactions (12–14), metabolic networks (15–17), or gene expression data (18–20), is just beginning.

Recently, we devised a method called PATHBLAST (13) for comparing the protein interaction networks of two species. Just as BLAST performs rapid pairwise alignment of protein sequences (21), PATHBLAST is based on efficient alignment of two protein networks to identify conserved network regions. Here, we extend this approach to present a computational framework for alignment and comparison of more than two protein networks. We apply this multiple network alignment strategy to compare the recently available protein networks for worm, fly, and yeast, and show that although any single network contains false-positive interactions, embedded beneath this noise are a repertoire of protein interaction complexes and pathways conserved across all three species.

Methods

We developed a general framework for comparison and analysis of multiple protein networks. Full details are provided in Supporting Text, Figs. 5–11, and Tables 3–6, which are published as supporting information on the PNAS web site. Briefly, this process integrates interactions with sequence information to generate a network alignment graph. Each node in the graph consists of a group of sequence-similar proteins, one from each species; each link between a pair of nodes represents conserved protein interactions between the corresponding protein groups (Fig. 1). A search over the network alignment is performed to identify two types of conserved subnetwork structures: short linear paths of interacting proteins, which model signal transduction pathways, and dense clusters of interactions, which model protein complexes.

The search is guided by reliability estimates for each protein interaction (computed based on a method by Bader et al., ref. 22), which are combined into a probabilistic model for scoring candidate subnetworks. Under the model, a log likelihood ratio score is used 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 that (i) in a real subnetwork, each interaction should be present independently with high probability, and (ii) in a random subnetwork, the probability of an interaction between any two proteins depends on their total number of connections in the network.

The search algorithm exhaustively identifies high-scoring subnetwork seeds and expands them in a greedy fashion. The significance of the identified subnetworks is evaluated by comparing their scores to those obtained on randomized data sets, in which each of the interaction networks is shuffled along with the protein similarity relationships between them.

Results

We applied the multiple network alignment framework (Fig. 1) to perform a three-way alignment of the protein–protein interaction networks of Caenorhabditis elegans, Drosophila melanogaster, and Saccharomyces cerevisiae. These species span the largest sets of protein interactions in the public databases to-date and, along with mouse, comprise the major model organisms used to study cellular physiology, development, and disease. Protein interaction data were obtained from the Database of Interacting Proteins (23) (February 2004 download) and contained 14,319 interactions among 4,389 proteins in yeast, 3,926 interactions among 2,718 proteins in worm, and 20,720 interactions among 7,038 proteins in fly. Protein sequences obtained from the Saccharomyces Genome Database (24), WormBase
Sharan et al.

(25), and FlyBase (26) were combined with the protein interaction data to generate a network alignment of 9,011 protein similarity groups and 49,688 conserved interactions for the three networks.

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 Fig. 2. The identified conserved clusters and paths, along with their graphical layouts, are available from the authors upon request.

Fig. 3 shows a global map of all clusters and paths conserved among the yeast, worm, and fly protein networks. 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. The largest numbers of conserved clusters were involved in protein degradation (green boxes at lower right), RNA polyadenylation and splicing (blue boxes at lower left), and protein phosphorylation and signal transduction (red boxes at upper right). Other significant conserved clusters were involved in DNA synthesis, nuclear-cytoplasmic transport, and protein folding. The map also reveals conserved links between different biological processes, for instance linking kinase signaling (red) to protein catalysis (green; lower right) or to regulation of transcription (yellow; upper middle).

To validate our results, we compared these conserved clusters to known complexes in yeast as annotated by the Munich Information Center for Protein Sequences (MIPS) (27). We 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. We defined a cluster to be pure if it contained three or more annotated proteins and at least half of these shared the same annotation. Ninety-four percent of the conserved clusters were pure, indicating the high specificity of our approach, compared to a lower percentage of 85% 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).

We further checked whether the conserved clusters were biased by spurious interactions, resulting from “sticky” proteins that lead to positive two-hybrid tests without interaction. Of 39 proteins with >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.

Three-Way Versus Two-Way Network Alignments. In addition to the three-way comparison, we also performed all possible pairwise network alignments: yeast/worm, yeast/fly, and worm/fly. This process identified 220 significant conserved clusters for yeast/worm, 835 for yeast/fly, and 132 for worm/fly. Several examples of these are shown in Fig. 9. Global overviews of the pairwise conserved clusters (similar to Fig. 3) are provided in Figs. 6–8.

Analysis of the proteins shared among the different pairwise and three-way network comparisons led to two general 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 (see Table 6 and Fig. 11). Second, the worm/fly conserved clusters were largely distinct from the clusters arising from the other analyses. For example, only 29% 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 and the particular selection of protein “baits” for the C. elegans protein-protein interaction screen: roughly one-quarter were specifically chosen to be metazoan specific, and almost two-thirds had no clear yeast ortholog (9).

Prediction of Protein Functions. Conserved subnetworks that contain many proteins of the same known function suggest that their remaining proteins also have that function. Based on this concept, we predicted protein functions whenever the set of proteins in a conserved cluster or path (combined over all species) was significantly enriched for a particular Gene Ontology (GO) (28) 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 (see Supporting Text).

This process resulted in 4,669 predictions of previously undescribed GO Biological Process annotations spanning 1,442 distinct proteins in yeast, worm, and fly; and 3,221 predictions of GO Molecular Function annotations spanning 1,120 proteins. We estimated the specificity of these predictions by 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 (see Supporting Text). As shown in Table 1, depending on the species, 58–65% of our predictions of GO Processes agreed with the known annotations (see also Tables 3 and 4). This analysis outperformed a sequence-based method of annotating proteins based on the known functions of their best sequence matches, for which the accuracy ranged between 37% and 53% (see Supporting Text). The complete list of protein function predictions is provided in Table 7, which is published as supporting information on the PNAS web site.
Prediction of Protein Interactions. We also used the multiple network alignment to predict protein–protein physical interactions. We predicted an interaction between a pair of proteins based on (i) evidence that proteins with similar sequences interact within other species (directly or by a common network neighbor) and, optionally, (ii) cooccurrence of these proteins in the same conserved cluster or
The accuracy of these predictions was evaluated by using 5-fold cross validation, as described in Supporting Text. In cross validation, strategy i achieved 77–84% specificity and 23–50% sensitivity, depending on the species for which the predictions were made (see Tables 2 and 5). 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 (see Table 2). Given the elevated specificity of the combined strategies, we were able to predict 176 previously undescribed interactions for yeast, 1,139 for worm, and 1,294 for fly with high confidence. Thus, although protein interactions have been used previously to predict interactions among the orthologous proteins of other species (9, 29), screening these against conserved paths and clusters markedly improves the specificity of prediction. The complete list of predicted protein interactions is provided in Table 8, which is published as supporting information on the PNAS web site.

Table 1. Cross-validation results for protein cellular process prediction

<table>
<thead>
<tr>
<th>Species</th>
<th>No. correct</th>
<th>No. of predictions</th>
<th>Success rate, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td>114</td>
<td>198</td>
<td>58</td>
</tr>
<tr>
<td>Worm</td>
<td>57</td>
<td>95</td>
<td>60</td>
</tr>
<tr>
<td>Fly</td>
<td>115</td>
<td>184</td>
<td>63</td>
</tr>
</tbody>
</table>

For each species, the number of correct predictions, the total number of predictions, and the success rate in 10-fold cross-validation are listed.

Table 2. Cross-validation results for protein interaction prediction

<table>
<thead>
<tr>
<th>Species</th>
<th>Sensitivity, %</th>
<th>Specificity, %</th>
<th>P value</th>
<th>Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td>50</td>
<td>77</td>
<td>1.1e-25</td>
<td>i</td>
</tr>
<tr>
<td>Worm</td>
<td>43</td>
<td>82</td>
<td>1e-13</td>
<td>i</td>
</tr>
<tr>
<td>Fly</td>
<td>23</td>
<td>84</td>
<td>5.3e-5</td>
<td>i</td>
</tr>
<tr>
<td>Yeast</td>
<td>9</td>
<td>99</td>
<td>1.2e-6</td>
<td>i + ii</td>
</tr>
<tr>
<td>Worm</td>
<td>10</td>
<td>100</td>
<td>6e-4</td>
<td>i + ii</td>
</tr>
<tr>
<td>Fly</td>
<td>0.4</td>
<td>100</td>
<td>0.5</td>
<td>i + ii</td>
</tr>
</tbody>
</table>

For each species, the specificity and sensitivity of the predictions in 5-fold cross-validation, the significance of the results, and the prediction strategy (see text) are listed.
Discussion

Comparison to Existing Methods. We have developed pairwise network alignment algorithms that were used to detect linear paths (13) or dense clusters (14) that are conserved between yeast and the bacteria Helicobacter pylori. The multiple network alignment scheme that we have presented here is an extension of our earlier approaches to handle more than two species. Additional advantages of the current approach over the previous approaches are: (i) it is a unified method to detect both paths and clusters, which generalizes to other network structures; (ii) this approach incorporates a refined probabilistic model for protein interaction data; and (iii) it includes an automatic system for laying out and visualizing the resulting conserved subnetworks.

A related method that uses cross-species data for predicting protein interactions is the interolog approach (12, 18); a pair of proteins in one species is predicted to interact if their best sequence matches in another species were reported to interact. In comparison, our proposed scheme 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. Because conservation is evaluated in the context of a protein interaction subnetwork and not independently for each interaction, the high specificity of the resulting predictions can be maintained (see below).

Best blast Hits 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. For instance, the conserved network region in Fig. 2h suggests that the worm protein exc-7 plays the same functional role as yeast Pab1 and fly CG33070 (BLAST e-value $\approx 10^{-42}$) based on the conserved interactions with Asc1/F08G12.2/Rack1 (yeast/worm/fly), Rna15/Unc-75 (yeast/worm), and T01D1.2/Tbp (worm/fly). However, CG33070 is only the fifth best blast match in fly overall (the best being CG3151 at E value $\approx 10^{-70}$).

Overall, of the 679 protein triples aligned at the same position within a three-way conserved cluster, only 177 contained at least one pair of best sequence matches; of the 129 additional triples in conserved paths, only 31 contained best sequence matches. 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.

Functional Links Within Conserved Networks. Conserved network regions enriched for several functions point to cellular processes that may work together in a coordinated fashion. 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 conserved network region. For example, Fig. 2h 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-Rack1-17p). These findings are consistent with previously documented association of proteosomes with mRNA-binding proteins, although the exact nature of this association has been controversial (30, 31). A related functional link between the proteasome and nucleic acid synthesis was detected in our earlier network comparison of yeast and the bacteria H. pylori (13).

As another example, Fig. 9 shows a worm/fly conserved cluster for which ~40% of the proteins have no significant yeast ortholog (BLAST E value > 0.01). The cluster links functions such as proteolysis (Pros25, Pros28.1, Pas1–7), actin binding (Cher, W04D2.1), ion transport (CG32810, C40A11.7, C52B11.2), and axon guidance (Fra). Taken together, these functions suggest a role for this cluster in growth cone formation during axon guidance. Guidance of axons to their synaptic targets is an initial step in the development of the central nervous system (32) and is mediated by special compartments called growth cones at the tips of the extending neurites. Formation of growth cones is induced by elevated levels of Ca2+ ions, which trigger local proteolysis and restructuring of the actin cytoskeleton (33). Thus, as implicated by our findings, axon guidance requires synergy between proteolysis, actin binding, and ion transport within an intricate network of protein interactions.

Validation of Predicted Interactions. Our two-hybrid tests of predicted interactions yielded a success rate in the range of 40–52%. These results are satisfactory for three reasons. First, the performance is clearly significant compared to the chance of identifying protein interactions at random (0.024%), estimated from an earlier two-hybrid screen (4) of 192 baits identifying protein interactions at random (0.024%, estimated). These results are satisfactory for three reasons. First, the performance is clearly significant compared to the chance of identifying protein interactions at random (0.024%), estimated from an earlier two-hybrid screen (4) of 192 baits identifying protein interactions at random (0.024%, estimated). These results are satisfactory for three reasons. First, the performance is clearly significant compared to the chance of identifying protein interactions at random (0.024%), estimated from an earlier two-hybrid screen (4) of 192 baits identifying protein interactions at random (0.024%, estimated).

Second, 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-Rack1-17p). These findings are consistent with previously documented association of proteosomes with mRNA-binding proteins, although the exact nature of this association has been controversial (30, 31). A related functional link between the proteasome and nucleic acid synthesis was detected in our earlier network comparison of yeast and the bacteria H. pylori (13).

As another example, Fig. 9 shows a worm/fly conserved cluster for which ~40% of the proteins have no significant yeast ortholog (BLAST E value > 0.01). The cluster links functions such as proteolysis (Pros25, Pros28.1, Pas1–7), actin binding (Cher, W04D2.1), ion transport (CG32810, C40A11.7, C52B11.2), and axon guidance (Fra). Taken together, these functions suggest a role for this cluster in growth cone formation during axon guidance. Guidance of axons to their synaptic targets is an initial step in the development of the central nervous system (32) and is mediated by special compartments called growth cones at the tips of the extending neurites. Formation of growth cones is induced by elevated levels of Ca2+ ions, which trigger local proteolysis and restructuring of the actin cytoskeleton (33). Thus, as implicated by our findings, axon guidance requires synergy between proteolysis, actin binding, and ion transport within an intricate network of protein interactions.

Validation of Predicted Interactions. Our two-hybrid tests of predicted interactions yielded a success rate in the range of 40–52%. These results are satisfactory for three reasons. First, the performance is clearly significant compared to the chance of identifying protein interactions at random (0.024%), estimated from an earlier two-hybrid screen (4) of 192 baits × 6,000 preys that yielded 281 interacting pairs). Second, two-hybrid analysis is known to miss a substantial portion of true interactions (11); this is particularly likely in our case where protein pairs were checked in only one orientation of bait and prey. For instance, two of the pairs that tested negative (YJR068W-YOR217W and YBL105C-YHR030C) have been shown to interact genetically in synthetic-lethal screens (27), suggesting a possible physical interaction. Third, predicting interactions by using a multiple network alignment approach compares favorably to previous approaches based on conservation of individual protein interaction. For instance, in ref. 12, the interolog approach was applied to a set of 72 reported interactions in yeast, predicting 71 previously undescribed interactions in worm. Seven of the predicted worm interactions tested positive by using a two-hybrid assay (10%), whereas 19 of the previously reported yeast interactions (26%) retested positive. Considering only the worm interactions that were predicted based on the 19 confirmed interactions in yeast, six of these tested positive, upper bounding the prediction accuracy at 31%. In tests of 145 additional predictions, 28 were confirmed, obtaining an overall accuracy of 16%. Similar results were obtained in a subsequent study by Yu et al. (29), where the accuracies of the interolog approach and an extension of it were estimated at 30–31%.

Conclusion

Nearly all comparative genomic studies of multiple species have been based on DNA and protein sequence analysis. Here, we transcribed that framework by presenting a comparative study of the protein–protein interaction networks of three model eukaryotes. These comparisons show that many circuits embedded within the protein networks are conserved over evolution, and that these circuits cover a variety of well defined functional categories. Because measurements of protein interactions tend to be noisy and incomplete, it would have been difficult, if not impossible, to find these mechanisms by looking at only a single species. Moreover, many of these similarities and the network connections they imply would not have been suggested by sequence similarity alone, as the proteins involved are frequently not best sequence matches. The multiple network alignment also allows us to ascribe unique functions to many proteins and predict previously unobserved protein–protein interactions. Therefore, comparative network analysis is a powerful approach for elucidating network organization and function.

We thank Jacob Scott and Shen-Hong Chen for help with the analyses and Steven Brenner, Jess Gramm, and Markus Herrgard for critical reading of the manuscript. This research was supported by grants from the National Science Foundation and National Institutes of Health.