Pacific Symposium on Biocomputing 14:190-202 (2009)

REVERSE ENGINEERING THE EVOLUTION OF PROTEIN INTERACTION NETWORKS

TODD A. GIBSON[1] AND DEBRA S. GOLDBERG[1,2]

[1] *Computational Bioscience Program*
University of Colorado Denver, CO, USA

[2] *Computer Science Department*
University of Colorado at Boulder, CO, USA

Protein interaction network analyses have moved beyond simple topological observations to functional and evolutionary inferences based on the construction of putative ancestral networks. Evolutionary studies of protein interaction networks are generally derived from network comparisons, are limited in scope, or are theoretic dynamic models that aren't contextualized to an organism's specific genes. A biologically faithful network evolution reconstruction which ties evolution of the network itself to the actual genes of an organism would help fill in the evolutionary gaps between the gene network "snapshots" of evolution we have from different species today. Here we present a novel framework for reverse engineering the evolution of protein interaction networks of extant species using phylogenetic gene trees and protein interaction data. We applied the framework to Saccharomyces cerevisiae data and present topological trends in the evolutionary lineage of yeast.

1. Introduction

Among the tasks utilizing protein interaction networks (PIN) are inferences of how the network itself evolved. With the availability of large-scale protein interaction datasets[1-6], the properties of these networks have been analyzed and theoretical models of evolving networks have been produced which incorporate some aspect of biological evolution in an effort to reproduce properties observed in experimentally-derived networks[7,8].

Evolutionary PIN analysis has been widely explored in network comparisons which are based on the premise that disparate organisms share a common ancestor[9]. Such comparisons generally elucidated protein function and predicted interactions. Protein interaction evolution has also been studied through structures of protein complexes[10-12]. These methods have identified a number of important properties of protein interaction evolution including the role self-interacting proteins play in forming complexes of paralogous proteins and the development of homologous protein complexes through a stepwise progression of duplications of complex constituents rather than via a single monolithic duplication event. Presser, et al.[13] studied PIN evolution using a probabilistic model of motifs formed by paralogous protein pairs born of the Whole Genome Duplication (WGD) event in Saccharomyces cerevisiae's evolutionary history. With the model they found that self-interactions were prevalent in the motifs formed by pre-WGD protein interactions. This study provides a probabilistic view of pre- and post-WGD motifs. A few studies have provided an evolving view of organism-contextualized protein interaction networks. Notably Vernon, et al.[14] reconstructed the interaction evolution of MADS domain MIKC-type proteins, and Pinney, et al.[15] reconstructed the bZIP family of transcription factors.

Here we extend the ability to analyze evolutionary trajectories with a novel framework which incorporates interaction data and phylogenetic gene trees into an evolving view of the protein interaction networks of species sharing the gene trees' phylogeny. The framework meets three criteria required to measure the evolutionary trajectory of some aspect of protein interaction network evolution:

(1) The reconstruction evolves. It includes a modern network, an ancestral network, and transitions between them.
(2) The evolving PIN is associated with a specific organism. That is, each vertex in the evolving network can be traced to an extant gene.
(3) The PIN encompasses the entire interactome.

Existing studies fail to meet all three criteria. Theoretical models are decontextualized–each vertex in the model is simply a generic protein. These models have no power to elucidate organismal evolution beyond identifying broad evolutionary processes which produce network characteristics consistent with those of empirical networks. Comparative studies allow functional and topological inferences, but are less informative of the evolution of the interaction network itself. The structural and stochastic studies have identified important factors in interaction evolution, but do not incorporate them into the larger, network-level evolutionary context. Those studies which have formed organism-specific, multi-step evolutionary views of protein interaction networks have been isolated to individual protein families.

In the following sections we introduce the framework, present some initial findings having applied it to the Saccharomyces cerevisiae lineage, and discuss possible future research the framework readily accommodates.

2. Evolutionary considerations

Gene duplication is readily accepted as a primary mechanism for generating organismal complexity. Two evolutionary mechanisms have been proposed for the fate of gene duplicates: neofunctionalization and subfunctionalization. Neofunctionalization posits that the functional redundancy intrinsic to initially identical gene duplicates releases one duplicate from selective pressure. While under neutral selection one of the duplicates can accumulate random mutations and potentially acquire novel and beneficial functions[16]. Subfunctionalization states that both gene duplicates acquire mutations resulting in each duplicate assuming a complementary subset of the ancestral gene's original functions[17]. Gene duplication and subsequent neofunctionalization and subfunctionalization have straightforward analogs in models of protein interaction network (PIN) evolution. With proteins as nodes, edges between proteins represent physical interactions and serve as an indication of protein function. Proteins with identical sets of interacting partners are presumed to have identical functions. Gene duplication is modeled by copying a protein vertex in the network along with its interactions. Neofunctionalization and subfunctionalization are modeled by the gain and loss of interactions respectively.

The high false negative rate endemic of interaction data sets makes it difficult to distinguish neofunctionalization from subfunctionalization in PINs, even with the availability of PIN data from several related species. We have recently found that the established ubiquity of neofunctionalization was based on three independent flaws: a bias against observing homodimers in both yeast two-hybrid (Y2H) and affinity-purification mass-spectrometry (AP-MS) high-throughput interaction assays, a failure to consider gene duplications of interacting partners subsequent to the duplication event under analysis, and theoretical models unable to produce clustering coefficients found in empirical protein interaction networks and biologically untenable parameter requirements[18]. In the absence of compelling evidence of ubiquitous *de novo* interaction formation, we include only the well-established ubiquity of subfunctionalization[17,19,20] in our framework.

The evolutionary dynamics integrated into the evolutionary framework assume that interactions between paralogs arose via the duplication of a self-interacting protein. If either paralog is homomeric, or if a paralogous interaction exists between them, then the ancestral protein is assumed to

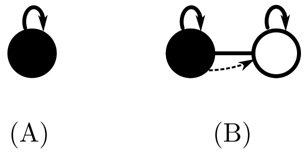

Figure 1. Duplication of self-interacting proteins. (A) A self-interacting protein. (B) When the self-interacting protein duplicates, a paralogous interaction is formed. Either or both of the two self-interactions, or the paralogous interaction may be lost in evolution, but any one of the three interactions is sufficient information to reconstitute the original ancestral gene with a homomeric interaction.

be self-interacting (see Figure 1).

3. Methods

We now introduce our framework for reverse engineering the evolution of an organism's protein interaction network. To assist in the exposition, we refer the reader to Figure 2 which graphically represents an example of the entire framework.

3.1. *Framework inputs*

As the figure shows, there are three inputs (highlighted in yellow) the framework requires. The first is a phylogenetic tree of the species subject to the evolutionary reconstruction. Second are gene family trees reconciled against the phylogenetic tree: trees describing each gene family's speciations, duplications, and losses (not shown in figure). Figure 2 has been color-coded to identify the portion of the phylogeny each phase of the reconstruction is associated with (i.e., extant species, common ancestor, last common ancestor).

The third input is protein interaction data for the extant species present in the phylogenetic tree. Semantically, all of the genes identified in the gene trees comprise the nodes of a single large network. The edges of the network may be drawn from the interaction data of more than one species. The example in Figure 2 shows the genes for all three species, and interaction data for two of the three species.

3.2. *Evolutionary event identification*

Prior to reverse-engineering the network evolution,t his phase simple processes information found from the previously-published inputs. Each speciation, duplication, and loss event in the gene trees are associated with a node of the phylogenetic tree. In Figure 2, the evolutionary events identified from the gene trees are shown within the soft-cornered box running down the center of the figure and are color-coded to the phylogenetic node they are associated with.

3.3. *Reverse engineering*

The protein interaction network input is combined with the evolutionary events generated in the previous step to reverse engineer the network evolution. The phylogenetic tree drives the ordering of the process. The extant species leaves of the phylogenetic tree are processed first, followed by the common ancestor nodes. The tree is iteratively "rolled up" until the last common ancestor is reached.

With each phylogenetic node, the evolutionary events associated with the node are applied to the protein interaction network. The evolutionary events are reversed with respect to time. Note that each evolutionary event identified previously describes an action on a single gene. A gene loss is the loss of a single gene. A duplication event produces two genes from one. Similarly, a gene speciation event produces two genes (one for each species) from a single (common ancestor's) gene. During reverse engineering, the opposite action is taken. A gene loss becomes a gene gain. A duplication or speciation event joins two separate genes into a single ancestral gene.

Paralogs which interact with the same neighbor are assumed to be preserving a redundant ancestral function (see Figure 3). An interaction present in only one paralog is presumed to have been genetically silenced in the other paralog after duplication. As evolutionary events are reverse engineered, the ancestor of each gene duplication acquires interactions with all of the neighbors its paralogous progeny pair interact with (Figure 3B).

Running down the left side of Figure 2 is a complete example of reverse engineering network evolution. Starting with the extant species and iterating through the phylogenetic tree, speciation and duplication events are reversed to join nodes from the protein interaction network together[a]. As proteins in the network are joined, the merged protein interacts with the union of the set of neighbors each protein interacted with individually. Later when the evolutionary re-creation is rolled forward and the merged protein is again separated into two separate proteins, it will be necessary to reproduce the sets of neighbors the separate proteins interact with. Therefore as proteins are merged, the neighbors not included in each separate

[a] There are no gene loss evolutionary events in the example.

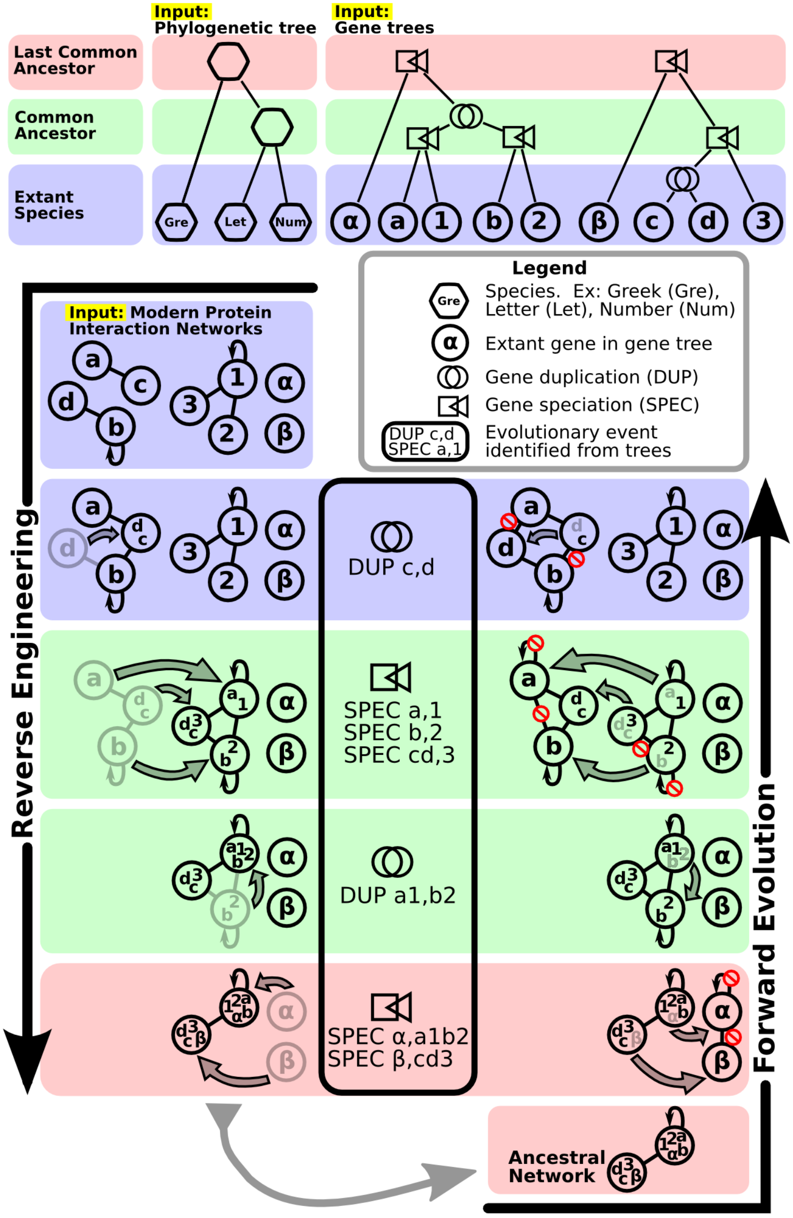

Figure 2. The reverse-engineering and forward evolutionary re-creation framework.

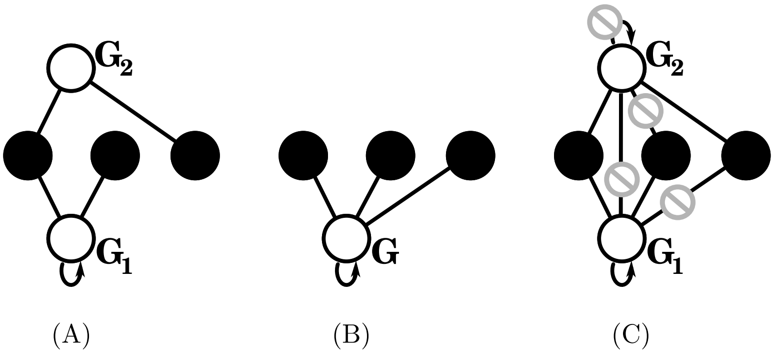

Figure 3. Reverse engineering and forward evolutionary re-creation. For this example assume a duplication event to duplicate gene G into G_1 and G_2 was generated during event identification (Section 3.2). (A) The extant protein interaction network. (B) During reverse engineering, G_1, G_2, and their interactions are collapsed into a single gene G. At this time the interactions to be lost after duplication are recorded. (C) Forward evolutionary re-creation. After the gene and its interactions are duplicated (including a paralogous interaction born of the self-interaction), the interaction losses recorded during reverse engineering are processed. Note that speciation differs from gene duplication in that a speciating self-interacting protein never produces a paralogous interaction between species duplicates. Forward evolutionary re-creation naturally segments the last common ancestor into species-specific components.

protein's interacting neighbor set is recorded. Once reverse-engineering is completed, the resulting network represents the last common ancestor's putative protein interaction network.

3.4. *Forward evolution*

Forward evolution in the framework begins with the last common ancestor's putative protein interaction network. The evolutionary events associated with the last common ancestor are processed first, followed by iteratively moving through the phylogenetic tree until the evolutionary events of the extant species are processed. Network nodes and interactions are duplicated, and the interaction losses are removed (Figure 3C). The lost interactions for each duplication and speciation event are recorded during reverse engineering process. Because redundant interactions have been found to diverge rapidly between paralogous genes[19,20], redundant interactions are removed immediately after the duplication event (i.e., prior to the next speciation event in the phylogenetic tree.

Figure 2 shows a complete example of the forward evolutionary re-creation running up the right side of the figure. The end of the forward evolutionary re-creation recapitulates the protein interaction network used at the beginning of the reverse-engineering process.

4. Results

We implemented our framework in Python and ran it on a standard PC running Linux. For the protein interaction framework input we combined multi-validated protein interactions[21] and high confidence interactions culled from a genome-wide in vivo screen using a protein-fragment complementation assay (PCA)[22]. Both data sets are Saccharomyces cerevisiae interactions

The phylogenetic tree and gene tree inputs were drawn from a study which generated 25,408 gene trees reconciled against a phylogeny of 19 species of Ascomycota fungi[23] [b].

The Evolutionary Event Identification phase generated 103,091 gene loss events, 7,711 gene duplication events, and 84,167 speciation events among the 19 species. The protein interaction network included 117,286 individual genes spread across 19 yeast species. Protein interaction input included only Saccharomyces cerevisiae so the vertices in the protein interaction network associated with the other species were not connected. The Saccharomyces cerevisiae vertices contain 5,780 genes, 4,052 of them in the largest component, and 12,341 edges, of which 12,241 are contained in the largest component. From this network, and the evolutionary events created previously, the evolving network was reconstituted.

The evolutionary reconstruction requires no computationally difficult algorithms. With the stated inputs and implementation the entire reconstruction runs in under half an hour.

Figure 4 illustrates the effect the forward evolutionary re-creation has on the evolutionary trajectories of the average degree and clustering coefficient. The average degree does not include self-interactions in the count. Self-interactions are absent from the Vázquez, et al. model, and omitting them from degree calculations is consistent with other protein interaction network analyses.

The plot has been highlighted between timesteps 5 and 6, indicating the time period within which the whole-genome duplication event occurred during Saccharomyces cerevisiae's evolution[24]. The curves are flat beyond timestep 6 indicating an absence of further duplication or loss events.

[b] We utilized an updated and expanded data set acquired from Ref. 23's companion Web site.

Figure 4A shows that the average degree decreases during evolution. Intuitively we might expect that as network vertices become less well-connected with their neighbors, they would become less clustered as well. However, Figure 4B indicates that the clustering coefficient increases.

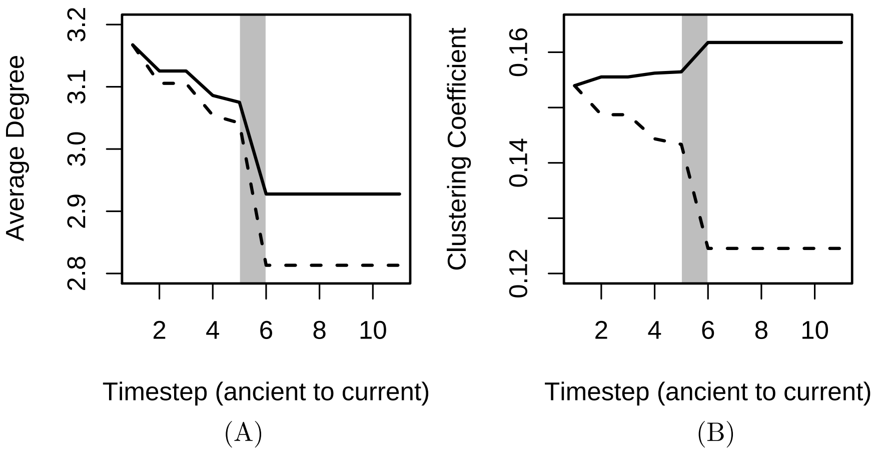

Figure 4. Basic topological properties of the evolving protein interaction network. The solid line is the evolutionary re-creation of Saccharomyces cerevisiae. The dashed line is the Vázquez, et al[4].model of protein interaction network evolution. The grey bar between timepoints 5 and 6 represent the evolutionary period within which the whole-genome duplication event occurred. (A) Average degree of the large component. (B) Clustering coefficient of the large component.

We next tracked the same topological values in an implementation of the Vázquez, et al.[8] theoretical model. Vázquez, et al. was selected because it reflects the principles of gene duplication, subfunctionalization, and homomeric duplication central to our framework. It is also appropriate as the model generates high clusterings consistent with empirical networks. Other theoretical models, in particular that of Solé, et al.[7] are unable to attain high clusterings and do not specifically address evolution of self-interactions[18].

To eliminate biases introduced by seed graphs[25], the theoretical model was seeded with the ancestral network which begins the forward evolutionary re-creation. The topological values of the theoretical model were measured each time the number of proteins in the model equaled that of the evolutionary re-creation at the same timepoint.

The Vázquez, et al. model takes two parameters: p, the probability of adding an edge between protein duplicates (i.e., probability of duplicating a self-interacting protein), and q, the probability of a redundant edge being lost from either the progeny or progenitor genes due to subfunctionalization. The parameter values were selected based on the number of paralogous interactions which survived duplication of a self-interacting protein in the evolutionary re-creation itself ($p = 0.98, q = 0.046$). The theoretical model was run 1,000 times and the mean of their topological values plotted.

Figure 4A shows that the theoretical model produces a similar reduction in the average degree as in the evolutionary re-creation. However, as Figure 4B shows, the clustering coefficient decreases with the addition of new genes. The parameter values published by Vázquez, et al.[8] ($p = 0.7, q = 0.1$) produced similar curves (data not shown).

5. Discussion

The evolutionary framework provides a novel, dynamic view of an organism's protein interaction network. Previous efforts have identified the influence of evolution on topologically-relevant factors (e.g., self-interacting proteins), and the few which have measured the evolution of protein interaction networks across several evolutionary periods have been isolated to small subnetworks. Despite the dynamic nature of theoretical models, theoretical model validation has commonly involved post hoc analysis of the generated network. A comparison of the evolutionary trajectories of Vázquez, et al.[8] and the yeast evolutionary re-creation reveal that while in both models the proteins on average interact with fewer neighbors over time, the yeast evolution generates higher clustering while the clustering is reduced in the theoretical model. This is of particular interest due to the high clustering levels found in empirically-derived protein interaction networks[19,7,26–28]. These results suggest that surviving gene duplicates are not distributed randomly throughout the interactome.

The framework may also provide some inferential power to species for which interaction data is not available but are nonetheless represented in the re-creation by phylogenetic gene trees. Under these conditions the species' genes are represented in the network as zero-degree vertices. As the gene duplication, speciation, and loss events are reverse-engineered, the ancestors of these genes interact with other ancestral genes based on interaction data available from other species. For example, the ancestral genes α and β in Figure 2 acquire interactions based on interaction data of other species.

As the evolutionary re-creation moves forward in evolutionary time, these genes speciate, duplicate and ultimately lose all ancestral interactions to return to their non-interacting empirical state. To construct initial inferences on genes or entire species which are lacking interaction data, interaction losses associated with gene duplication and speciation events can be selectively suppressed. For species in the phylogenetic tree without interaction data, this amounts to creating a first-pass putative interaction network based on homologous data sets. Though the inferred network surely contains false positives, the candidate set of interactions have novel derivation. Specifically, the putative network is derived directly from the last common ancestor rather than solely, for example, sequence similarity with homologous genes[29]. Putative interactions produced from the re-creation are a product of the gene duplication and loss events. This effectively combines the homology-based prediction of interactions[29] with parameters derived from networks of species included in the phylogeny input. Indeed, subsequent development of the framework should include a comparison with homology-based predictions.

Another area for subsequent exploration is the framework's assumption that all interactions arise through subfunctionalization. Although it has been suggested that the de novo acquisition of interactions is not a common occurrence[18], de novo interactions can not be completely ruled out. One possibility is the development of an an error model to handle this uncertainty.

As with all network analyses, the utility of the framework relies in part on the quality of the interaction data. Generally, high-confidence data sets eliminate false-positives at the expense of reduced coverage[30]. However, as disparate protein interaction datasets from different species are combined into common ancestors, ancestral networks benefit from wider coverage than the individual data sets provide.

The framework presented here reinvigorates the study of network evolution. Differences in topological properties, development of motifs, and the development of functional modules are just a few of areas that may now be analyzed in the context of their evolutionary trajectories. The framework provides new opportunities to analyze both the evolutionary trajectory of a single species as well as processes through which network features diverge between different species. As phylogenetic gene trees come available for a wider range of species, and additional interaction sets are published, evolutionary network re-creations contextualized to specific species will become increasingly valuable.

Acknowledgments

Funding. This material is based upon work supported by the National Science Foundation under Grant MCB 0630250.

References

1. Uetz P, et al. (2000) A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature* 403:623–627.
2. Ito T, et al. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 98:4569–4574.
3. Gavin A C, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415:141–147.
4. Gavin A C, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440:631–636.
5. Krogan N J, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 440:637–643.
6. Reguly T, et al. (2006) Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol* 5:11.
7. Solé R V, Pastor-Satorras R, Smith E, Kepler T B (2002) A model of large-scale proteome evolution. *Advances in Complex Systems* 5:43.
8. Vázquez A, Flammini A, Maritan A, Vespignani A (2003) Modeling of protein interaction networks. *ComPlexUs* 1:38–44.
9. Sharan R, Ideker T (2006) Modeling cellular machinery through biological network comparison. *Nat Biotechnol* 24:427–433.
10. Pereira-Leal J, Levy E, Kamp C, Teichmann S (2007) Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol* 8:R51.
11. Pereira-Leal J B, Levy E D, Teichmann S A (2006) The origins and evolution of functional modules: lessons from protein complexes. *Philos Trans R Soc Lond B Biol Sci* 361:507–517.
12. Pereira-Leal J B, Teichmann S A (2005) Novel specificities emerge by stepwise duplication of functional modules. *Genome Res* 15:552–559.
13. Presser A, Elowitz M B, Kellis M, Kishony R (2008) The evolutionary dynamics of the saccharomyces cerevisiae protein interaction network after duplication. *Proc Natl Acad Sci U S A* 105:950–954.
14. Veron A S, Kaufmann K, Bornberg-Bauer E (2007) Evidence of interaction network evolution by whole-genome duplications: A case study in mads-box proteins. *Mol Biol Evol* 24:670–678.
15. Pinney J W, Amoutzias G D, Rattray M, Robertson D L (2007) Reconstruction of ancestral protein interaction networks for the bzip transcription factors. *Proc Natl Acad Sci U S A* 104:20449–20453.
16. Ohno S (1970) *Evolution by Gene Duplication* (Springer, New York).
17. Force A, et al. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.

18. Gibson T, Goldberg D (2008) Questioning the ubiquity of neofunctionalization. PLoS Computation Biology. In Revision.
19. Wagner A (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* 18:1283–1292.
20. He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169:1157–1164.
21. Batada N N, et al. (2007) Still stratus not altocumulus: further evidence against the date/party hub distinction. *PLoS Biol* 5:e154.
22. Tarassov K, et al. (2008) An in vivo map of the yeast protein interactome. *Science* 320:1465–1470.
23. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449:54–61.
24. Kellis M, Birren B W, Lander E S (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature* 428:617–624.
25. Hormozdiari F, Berenbrink P, Pržulj N, Sahinalp S C (2007) Not all scale-free networks are born equal: The role of the seed graph in PPI network evolution. *PLoS Comput Biol* 3:e118.
26. Goldberg D S, Roth F P (2003) Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A* 100:4372–4376.
27. Yook S H, Oltvai Z N, Barabási A L (2004) Functional and topological characterization of protein interaction networks. *Proteomics* 4:928–942.
28. Han J D J, Dupuy D, Bertin N, Cusick M E, Vidal M (2005) Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol* 23:839–844.
29. Yu H, et al. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res* 14:1107–1118.
30. von Mering C, et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417:399–403.