Protein Interaction Networks

Kai Tan and Trey Ideker

Department of Bioengineering

University of California at San Diego

9500 Gilman Drive

La Jolla, CA 92093

USA

Table of Contents

Introduction

Every living cell is governed by a vast network of interacting proteins, RNA, DNA, metabolites, and other molecules. Interactions among proteins are especially crucial to a wide variety of cellular processes: assembly of the structural compartments of a cell such as the cytoskeleton and nuclear pore; signal transduction pathways such as the classical mitogen-activated protein kinase (MAPK) cascade involved in pheromone signaling; enzyme-protein substrate interactions; and assembly of large molecular machines such as DNA polymerase and the proteasome.

Knowledge of the stable and transient protein interactions in a cell facilitates functional annotation of novel genes and provides insight into its higher-order organization. Considered individually, protein interactions stimulate the formulation of hypotheses that can be tested experimentally. For example, a membrane protein found to interact with a transcription factor might seem at first to be a "false positive", but such findings have also led to unexpected new insights into signal transduction, as in the case of Notch and Suppressor of hairless (Artavanis-Tsakonas et al. 1999) or the SREBPS transcription factors that localize to the ER membrane (Edwards et al. 2000). Further, when combined with diverse large-scale data such as microarray gene expression profiles (DeRisi et al. 1997) or genomic phenotypes (Begley et al. 2002, 2004; Deutschbauer et al. 2002), protein interaction networks provide a more complete picture of cellular pathways and responses than has ever before been available. Such an integrated network is useful because it provides a lucid means of summarizing existing biological knowledge about molecular behavior.

Recent years have witnessed an explosive growth of research on protein interactions and networks, with new experimental techniques, data sets, analyses, and modeling methods being published at an ever increasing rate. In particular, the last two years have seen the arrival of large-scale protein interaction data sets from the multi-cellular organisms fruit fly (Giot et al. 2003) and round worm (Li et al. 2004). Biologists are now faced with the challenge of deciphering these complex metazoan networks with the ultimate goal of describing the network of protein interactions in humans.

In this chapter, we summarize current technologies for generating large-scale protein interaction data, as well as visualizing and modeling protein interaction networks together with complementary large-scale data of various types. We cover recent work to extend and compare these models across different species or biological conditions, and we describe efforts to understand the evolution and dynamics of protein interaction networks.

Methodologies to obtain protein interaction data

Traditionally, protein interactions have been studied individually by genetic, biochemical, and biophysical techniques. However, the speed with which protein sequences are now discovered (or predicted) has created a need for high-throughput methods for interaction detection also. Consequently, a variety of experimental and computational approaches have been introduced in the past several years that can tackle the problem at large scale, resulting in a vast amount of interaction data in the public domain. As described in the following text, yeast two-hybrid and mass spectrometry (MS) technologies aim to detect physical binding between proteins, whereas genetic interactions and computational

methods seek to predict protein functional associations. Such functional associations may or may not result from physical binding.

Experimental technologies to identify protein-protein interactions

A variety of methods are now available for measuring protein-protein interactions, such as co-immunoprecipitation (Lane and Crawford, 1979), the two-hybrid system (Fields and Song 1989), and the glutathione-s-transferase (GST) pull-down assay (Kaelin et al. 1991)- the former two being the most widespread. Many of the protein-protein interactions that occur *in vivo* are maintained when a cell is lysed under non-denaturing conditions. Co-immunoprecipitation takes advantage of this fact to detect and identify physiologically relevant protein-protein interactions. The principle is straightforward: if protein X is immunoprecipitated with an antibody to X, then protein Y, which is stably associated with X *in vivo*, may also precipitate *in vitro*. To identify novel associated proteins after immunoprecipitation, mass spectrometry has become the method of choice because of its sensitivity, speed, and ability to identify post-translational modifications (Aebersold and Mann, 2003).

Tandem mass spectrometry (MS/MS) is typically used to identify proteins from complex mixtures (Figure 1b). The protein mixture is digested to form peptides which are introduced into the first mass spectrometer to separate them according to mass (detected as a mass-to-charge ratio). Peptides of a fixed size are selected and directed towards a so-called "collision cell" in which the peptides collide with molecules of an inert gas (such as argon) and break apart into fragments. The resulting fragments are analyzed by a second mass spectrometer which measures the mass of each fragment to produce a peptide "fragmentation profile". The peptides then serve as surrogate markers

for the protein sequence. Proteins are identified by searching the resulting peptide mass fingerprint through sequence databases. To identify novel protein interactions, co-immunoprecipitation can be used initially to collect a mixture of interacting proteins followed by protein identification by MS/MS.

In the two-hybrid system (Figure 1a), a protein "bait" of interest (B) is fused to the DNA binding domain (DB) of a transcription factor such as Gal4p. A second "prey" protein (P) is fused to the transcriptional activation domain (AD) of the same transcription factor. A physical interaction between B and P reconstitutes a functional transcription factor that can activate expression of a reporter gene. Usually, multiple reporter genes that allow growth selection on different media are used to increase the specificity of detection. Because the two-hybrid system is carried out *in vivo* and only requires the manipulation of DNA, it is amenable to automation and high-throughput methods.

Currently, MS/MS is the most practicable way to identify the components of a protein complex but typically does not provide information about interaction topology. In this regard, the two-hybrid system can provide complementary information about direct interactions, revealing which specific proteins bind to which others within a protein complex or signaling pathway. Both yeast two-hybrid technology (Uetz et al. 2000, Ito et al. 2001) and co-immunoprecipitation followed by MS (Gavin et al. 2002, Ho et al. 2002) were initially applied in the yeast *Saccharomyces cerevisiae* (Baker's yeast, a model eukaryotic cell) to generate large-scale protein interaction data. More recently, yeast two-hybrid technology has also been used to generate large-scale protein interaction data

in the multicellular organisms *Drosophila melanogaster* (fruit fly) (Giot et al. 2003) and *Caenorhabditis elegans* (round worm) (Li et al. 2004).

Protein interactions do not always represent physical binding events. For example, genetic interaction, in which two gene mutations have a combined effect not exhibited by either mutation alone, constitutes yet another interaction type that is being measured at high throughput. Two major types of genetic interactions are synthetic lethal interactions, in which mutations in two nonessential genes are lethal when combined; and suppressor interactions, in which one mutation is lethal but combination with a second restores cell viability. Screens for genetic interactions have been used extensively to shed light on pathway organization in model organisms (Avery and Wasserman 1992; Guarente 1993; Hartman et al. 2001; Thomas 1993), while in humans, genetic interactions are critical in linkage analysis of complex diseases (Sham 2001) and in high-throughput drug screening (Dolma et al. 2003). For species such as yeast, recent experiments have defined large genetic networks cataloguing thousands of such interactions (Hartman et al. 2001; Huang and Sternberg 1995; Tong et al. 2001; Tong et al. 2004).

Computational approaches to predict protein-protein interactions

In the late 1990s, several related methods were proposed for predicting protein interactions from DNA sequence information which received much attention due to the increasing number of complete genomes becoming available. These methods relied on the exploitation of "genomic context" in the form of structural or evolutionary constraints. One form of genomic context is the co-occurrence of orthologous genes

across entire genomes which defines a phylogenetic profile (Ouzounis and Kyrpides 1996, Pellegrini et al. 1999). Such a profile associates each gene with a binary representation of the presence/absence of its orthologs in different genomes. Genes that "travel" together during evolution are assumed to be involved in similar cellular processes. It is then possible to predict the functional association of genes that possess similar profiles. This method becomes more powerful with an increasing number of genomes because this allows more accurate profiles to be constructed. However, evolutionary processes such as gene duplication, loss, and horizontal gene transfer could hamper accurate construction of phylogenetic profiles (Galperin and Koonin 2000). Another genomic context based approach (Enright et al. 1999, Marcotte et al. 1999) exploits the notion of gene fusion, in which several genes in one species are merged or concatenated in other species into a single gene which encodes a multifunctional, multidomain protein. This event is maintained by selection, possibly due to the selective advantage of decreased regulational load (Enright et al. 1999). Proteins that are fused in one genome are likely to interact, physically or at least functionally, in other genomes.

An approach analogous to the gene fusion method includes analysis of gene neighborhoods in genomes (Dandekar et al. 1998, Overbeek et al. 1999). The basic assumption is that genes which interact or are functionally associated tend to be located in physical proximity to each other on the genome. The most apparent case of this phenomenon occurs in prokaryotes in which related genes are often co-localized into so-called "operons". Although operons do not generally occur in eukaryotic systems, it is still possible to infer functional association of a pair of genes if their homologs tend to be close in many genomes.

A new trend in *de novo* protein interaction prediction is to search for coordinated mutations between the sequences of interacting proteins, e.g., as has been observed for ligand-receptor interactions (Goh et al. 2000, 2002). The assumption is that the interacting proteins must co-evolve to preserve the interaction over time and thus the functional activity mediated by the interaction. Pazos and Valencia (2001) have used such a method to perform large-scale predictions of interactions with high statistical significance, resulting in 2,742 putative protein interactions for *E. coli*. Ramani and Marcotte (2003) introduced further methods to align phylogenetic trees of interacting protein families to define specific interaction partners. They suggest a model for the evolution of interacting protein families in which interaction partners are duplicated in coupled processes.

Other computational methods have been developed for predicting novel protein interactions through analysis of examples of known interactions. The common theme here is to transfer the existing annotation of a known gene to a newly sequenced gene product. This is based on the concept that sequence and structural similarities between gene products suggest functional similarities. One type of annotation transfer is based on structural data of known interacting proteins. New interactions can be inferred between pairs of proteins for which the sequences are compatible with known crystal structures of heterodimers (Russell et al. 2004, Aloy et al. 2004) and between pairs of proteins with domains that are often observed in interacting proteins (Ng et al. 2003). Another type of annotation transfer is the "interolog" approach where a pair of proteins in one species is predicted to interact if their best sequence matches in another species were reported to interact (Matthews et al. 2001, Yu et al. 2004).

Large protein-protein interaction data sets are now available for a variety of species (Table 1) including *S. cerevisiae* (Gavin et al. 2002; Ho et al. 2002; Ito et al. 2001; Lee et al. 2002; Uetz et al. 2000), *H. pylori* (Rain et al. 2001), *E. coli* (Butland et al 2005), *D. melanogaster* (Giot et al. 2003), *C. elegans* (Li et al. 2004; Walhout et al. 2000), and *H. sapiens* (Peri et al. 2003). In light of these vast scientific resources made available through experimental and computational analyses, several databases storing interaction data are now in wide usage (Table 2). Most of these databases contain interaction data derived from both high-throughput analyses and small-scale experiments. Besides being data warehouses, some of these databases have developed new methods for data exchange and visualization to facilitate the study of molecular interaction networks.

Computational modeling of protein networks

Visualization of protein interaction networks

Numerous articles and textbooks include figures showing different types of molecules and interactions between them. However, these figures typically invoke a limited number of components to describe an isolated biochemical process or signaling pathway, are carefully tailored to illustrate a predetermined concept, and rely heavily on accompanying textual descriptions (Pirson et al. 2000). In contrast, there is a pressing need for visual representations that can systematically present and organize the extremely large amounts of protein-interaction and expression data rapidly accumulating in the wake of two-hybrid screens, DNA microarray technology, and high-throughput proteomics. Such displays are not hand-tailored to illustrate a foregone conclusion, but

should ideally stimulate the discovery of new protein functions and biological relationships. As the raw data become increasingly complex with each type of supplemental information, tools that are both visual and interactive become increasingly important for emphasizing and extracting the key features.

Although protein-protein interactions were originally reported as lists of protein pairs (e.g. Uetz et al. 2000), more and more often they are represented graphically as two-dimensional networks. Figure 2 illustrates the difference on a small set of protein-protein interactions in yeast: while both representations reflect identical information, the network representation (called layout) has fundamental advantages with respect to human perception. Hand-formatted maps (such as those in Michal 1998; Kohn 1999) are usually of high quality, but available for very limited datasets due to the large amount of work involved to construct them. Accordingly, the large numbers of protein interactions in public databases (Table 2) have stimulated a range of automated layout algorithms to visualize them.

Several software tools are available for visualizing physical or genetic interaction networks. Examples of network visualization tools include: Cytoscape (Shannon et al. 2003), Osprey (Breitkreutz et al. 2003), Pajek (Batagelj and Mrvar 1998), ProViz, and WebInterViewer. These are software packages that have either been designed to visualize protein interactions or can be customized for that task (Table 3 gives a side-by-side comparison).

Such software enables a variety of routine operations on the network: automated network layout; association of data attributes (such as gene expression profile and gene ontology) with different network components; mapping of data attributes to visual

properties (such as node and edge color, shape and size), and network filtering. Specific features of each available program are listed in Table 3.

Topological properties of protein interaction networks

Along with other types of cellular networks, such as metabolic, regulatory and genetic networks, the topological properties of protein interaction networks have been intensely studied since the first large-scale data sets were published. In the past few years, the rapidly developing theory of complex networks has led to the discovery that the architectural features of molecular interaction networks within a cell are shared to a large extent by other complex systems, such as the Internet, US power grid and even social networks (Barabasi and Oltvai 2004). This unexpected universality indicates that similar laws may govern most complex networks in nature, which allows the expertise from large scale, non-biological systems to be used to characterize the organizing principles of cellular networks.

Several recent studies have indicated that protein interaction networks in diverse species also have the features of a so-called scale-free network which means the connectivity distribution of the network follows a power-law function (see Chapter X for a detailed discussion) (Jeong et al. 2001, Wagner 2001, Rain et al. 2001, Giot et al. 2003, Li et al. 2004, Butland et al. 2005). This topological feature is illustrated in Figure 3, which shows the protein interaction map of *S. cerevisiae* generated by a systematic two-hybrid screen. Whereas most proteins in the network participate in only a few interactions, a few proteins participate in many interactions (hubs) – a typical feature of scale-free networks. Protein interaction networks also exhibit another common

architectural feature of all complex networks: the so called "small world effect"- any two nodes can be connected with a path of a few links only. Within the cell, this effect was first observed with metabolic networks, in which paths of only three to four reactions can link most pairs of metabolites (Jeong et al. 2000, Wagner and Fell 2001). Although both "scale-free topology" and "small world connectivity" have clear mathematical definitions, the biological consequences of these topological properties remain to be studied. The presence of hubs seems to be a general feature of all cellular networks and they fundamentally determine the network's global behavior (in terms of both scale-free and small world connectivity). The biological importance of hubs is supported by the over-representation of genetic interactions between hubs in protein interaction networks (Ozier et al. 2003) and by the over-representation of hub genes among all lethal genes revealed by genome-wide deletion study (Jeong et al. 2001).

In addition to the aforementioned global topological features, protein interaction networks also possess recurring local topological features known as "network motifs". Network motifs are defined as particular patterns of interaction (i.e., isomorphic subgraphs) that are over-represented compared to randomized versions of the same network. Significant motifs were first shown to exist in transcriptional regulatory networks (Shen-Orr et al. 2002) and subsequently in a variety of biological networks (Milo et al. 2002, Wuchty et al. 2003, Yeger-Lotem et al. 2004). The high degree of evolutionary conservation of motif constituents within the yeast protein interaction network (Wuchty et al. 2003) further indicate that motifs are indeed of direct biological relevance. Many network motifs, for instance, feed-forward loop and single input motif (Figure 4), are also well known in circuit design and other engineering fields and thus can

be studied in detail using similar approaches from these fields. Indeed, as a first step in this direction, the highly significant feed-forward loop has been shown to function as a sign-sensitive delay element in transcriptional regulatory networks, a circuit that responds rapidly to step-like stimuli in one direction and as a delay to steps in the opposite direction (Mangan et al. 2003).

Integrating protein interaction networks with complementary data

Just as BLAST has been proven instrumental for querying sequence databases to identify genes, new pathway discovery and search tools enable us to query a protein interaction network to identify particular interaction pathways in a systematic fashion. For example, several groups (Ge et al. 2001; Hanisch et al. 2002; Ideker et al. 2002; Jansen et al. 2002) have applied "co-clustering" approaches to identify groups of proteins that are co-expressed and also closely connected by interactions in the network. In many cases, these "expression-activated networks" correspond to well known protein complexes, regulatory pathways, or metabolic reaction pathways, such as the 26S proteasome complex (Jansen et al. 2002), the core galactose-induction circuit (Ideker et al. 2002), and the glycolysis pathway (Hanisch et al. 2002). Other methods (Bar-Joseph et al. 2003; Lee et al. 2002; Pe'er et al. 2002; Yeang and Jaakkola 2003) use probabilistic approaches to match changes in gene expression with transcriptional and/or protein signaling interactions that are most likely to regulate them directly. These methods start with a cluster of differentially expressed genes and incrementally choose a small set of transcription factors which, by virtue of their levels and/or protein-DNA interactions in the network, can maximally predict the observed levels of differential expression in the

cluster. All of these approaches serve to reduce network complexity by pinpointing just those regions whose gene/protein states are perturbed by the conditions of interest, while removing false positive interactions and interactions not involved in the perturbation response. Software is available for several of these approaches, such as the GRAM approach by Bar-Joseph et al. (2003). Others are implemented as extensions to existing network visualization software, such as MCODE (Bader et al. 2003) and the ActiveModules approach (Ideker et al. 2002) which are implemented as plug-ins to Cytoscape.

The key concept behind the more advanced queries is that, by interrogating a protein interaction network with other (complementary) large-scale data such as gene expression profiles, it is possible to condense and partition the enormous quantity of data into a small number of relevant pieces suitable for lower-level investigation and modeling. Such an approach reinforces the common signal present in both data sets while filtering out some of the independent noise.

As an example application, Begley et al. (Begley et al. 2002,, 2004) performed a series of network queries to screen for protein pathways and complexes important for cellular recovery to DNA damage. Begley et al. used a systematic phenotyping approach in which growth phenotypes were recorded for a set of 1,615 yeast single-gene knockout strains exposed to MMS. Of the knockout strains, 416 grew more slowly in the presence of MMS and showed less than 67% the growth rate of a wild type strain exposed to identical MMS conditions. These strains were assigned an “MMS sensitive” phenotype, and the genes deleted from each of these were designated as “MMS essential”.

To elucidate protein networks involved in the DNA damage response, the MMS phenotypic state data were integrated with a large combined protein-protein and protein-DNA interaction network for yeast (Figure 5a). In a preliminary step, proteins were removed from the network whose distance from MMS-essential proteins was greater than one interaction. Within this filtered network, ActiveModules was used to search for connected subnetworks having a higher-than-expected proportion of MMS essential proteins. This search identified four significant modules associated with MMS sensitivity. Figure 5b shows three of these: in addition to proteins already known to be associated with the damage response, the modules contained significant numbers of proteins involved in protein degradation (e.g., Vma6, Pep12, and Snf7) and several proteins of unknown function. These likely occur because toxins such as MMS also cause damage to proteins, activating protein degradation and turnover machinery as an integral part of the cellular response.

Network alignment and comparison

A major emerging challenge of protein network biology is to systematically compare and contrast biological networks over different species, conditions, cell types, disease states, or points in time. For this purpose, methods are being developed to compare/contrast protein interaction networks to predict protein interactions (Pazos and Valencia 2001); to assess the specificity of protein interactions (Ramani and Marcotte 2003); and to identify conserved interaction complexes and pathways (Kelley et al. 2003; Sharan et al. 2004).

Recently, we have developed pairwise network alignment algorithms that are used to detect linear interaction paths (Kelley et al. 2003) or dense clusters of interactions (Sharan et al. 2004) that are conserved between networks. For instance, the algorithm

PathBLAST searches for high-scoring "pathway alignments" involving a pair of paths, one from each network, in which proteins of the first path are paired with putative orthologs occurring in the same order in the second path (Figure 6a). We have also developed a similar algorithm to search for dense interacting clusters of proteins rather than linear paths (Sharan et al. 2004). These two network structures, paths versus dense clusters, attempt to capture different biological mechanisms that may be conserved. Very approximately, paths model signal transduction pathways while dense clusters of interactions model protein complexes. PathBLAST is available as a web-based query at http://www.pathblast.org. Target protein-protein interaction networks are currently available for *H. pylori*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *M. musculus*, and *H. sapiens*. A related method that uses cross-species data for predicting protein interactions is the interolog approach (Matthews et al. 2001, Yu et al. 2004): a pair of proteins in one species is predicted to interact if their best sequence matches in another species were reported to interact.

As an example of network evolutionary comparison, a protein network alignment was performed among the protein-protein networks of the budding yeast *S. cerevisiae* and the human gastric pathogen *H. pylori* (Kelley et al. 2003). Both the yeast network (14,489 interactions among 4,688 proteins, assembled from mass spectrometry and two-hybrid studies), and the *H. pylori* network (1,465 interactions among 732 proteins from a single two-hybrid study (Rain et al. 2001)) were extracted from the DIP database (Xenarios et al. 2002). The yeast and bacterial networks were analyzed to select the 150 highest-scoring pathway alignments of length four (four proteins per path), corresponding to a level of significance of $p \leq 0.05$. By combining all overlapping pathway alignments,

each of the 150 fell into one of five conserved network regions, two of which are shown in Figure 6 [b-c]. Interestingly, although the putative yeast-bacterial orthologs in these regions generally had significant sequence homology (i.e., having BLAST E-values $<10^{-10}$), over 50% of these orthologs were in fact not the overall best BLAST matches possible between the two species' genomes. Rather, they were identified by their close proximity to other orthologous proteins in the protein network.

Although an entire network vs. network comparison is invaluable for cataloguing all of the homologous pathways between and within organisms, it is also desirable to query a single protein network with specific pathways of interest. This procedure is similar to using BLAST to interrogate a sequence database with a short nucleotide or amino-acid sequence query. As an example of this approach, we queried the *S. cerevisiae* protein network with a classic mitogen activated protein kinase (MAPK) pathway associated with the filamentation response, consisting of a MAPK (Kss1), a MAPK kinase or MAPKK (Ste7), and a MAPKK kinase or MAPKKK (Ste11). MAPK pathways transmit incoming signals to the nucleus through activation cascades in which each kinase phosphorylates the next one downstream. As shown in Figure 6d, the pathway query 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). Such methods will be instrumental in extending comparative molecular biology from the level of DNA and protein sequences to the level of the protein network.

Robustness of protein interaction networks

Robustness is a property that allows a system to maintain its functions despite external and internal perturbations. This property has been widely observed in many biological systems, such as chemotaxis (Alon et al. 1999), circadian rhythms (Morohashi et al. 2002), and segmental pattern formation in embryogenesis (von Dassow et al. 2000). Understanding the origin and principles of robustness in biological networks enables us to put various observations about the networks into perspective and to facilitate the discovery of principles at the systems level.

A prominent feature of all cellular networks studied so far is their scale-free nature. Unlike random networks, scale-free networks are highly resistant to random failures. By simulation studies, Albert and colleagues (Albert et al. 2000) showed that even if 80% of randomly selected nodes fail, the remaining 20% still form a compact cluster with a path connecting any two nodes. This is because random failure mainly affects nodes with few network connections, the absence of which does not disrupt the network's overall integrity. On the other hand, removal of hubs rapidly disintegrates the network into small isolated node clusters. These computational simulations suggest hub proteins have an important role in cellular fitness. In fact, deletion analyses indicate that in *S. cerevisiae* only about 10% of the proteins with fewer than five interactions are essential, but this fraction increases to over 60% for proteins with more than 15 interactions. This indicates that the protein's number of interactions plays an important role in determining its deletion phenotype (Jeong et al. 2001). The importance of hubs is further supported by their evolutionary conservation: highly connected *S. cerevisiae* proteins have a smaller evolutionary distance to their orthologs in *C. elegans* (Fraser et al. 2002) and are more likely to have orthologs in higher organisms (Krylov et al. 2003).

Although hubs are essential for protecting protein networks from accidental failures, their attack vulnerability makes them ideal targets for manipulating and controlling the network. For instance, from a therapeutic point of view, hub proteins can be used to screen against small molecule libraries to identify potential drug targets.

In addition to global topological features that ensure the robustness of protein networks, local topological features, i.e., network motifs, is also used to maintain robustness. Negative feedback loops are a principle mode of control to enable robust response to perturbations (Kitano 2004). Alon and colleagues (Alon et al. 1999) have shown that bacteria use negative feedback in signal transduction systems to attain the perfect adaptation that allows chemotaxis to occur in response to a wide range of stimuli. Positive feed back contributes to robustness by amplifying stimuli so that the activation level of downstream pathways can be clearly distinguished from non-stimulated states and these states can be maintained. The best-documented example of a positive feedback loop functioning in signal transduction is the Mos-mitogen-activated protein kinase (MAPK) cascade in Xenopus oocytes (Ferrell J. Jr. 2002). This cascade is activated when oocytes are induced to mature by the steroid hormone progesterone. The positive feedback loop in the signal transduction process ensures that oocyte converts a graded, reversible triggering stimulus into an all-or-none, irreversible cell-fate decision.

Evolution of protein interaction networks

As the most prominent of feature of protein interaction networks and other cellular networks, the origin of scale-free topology has attracted the attention of many researchers. Two kinds of evolutionary processes have been invoked to explain this

topological feature of protein interaction networks. The first kind of process consists of gene duplications followed by either silencing of one of the duplicated genes or by functional divergence of the duplicates. In terms of the protein interaction network, a gene duplication corresponds to the addition of a node with links identical to the original node, followed by the divergence of some of the redundant links between the two duplicate nodes. Barabasi and Albert (1999) were the first to suggest that gene duplication is the major mechanism for generating the scale-free topology of protein interaction networks. According to their growth and preferential attachment model, duplicated genes produce identical proteins that interact with the same protein partners. Therefore, each protein that is in contact with a duplicated protein gains an extra link. Highly connected proteins are more likely to have a link to a duplicated protein than their sparsely connected cousins, and therefore they are more likely to gain new links if a randomly selected protein is duplicated. A mathematical model of the growth of networks based on this principle produces scale-free topologies with parameters comparable to those of real-world networks (Barabasi and Albert 1999). Two lines of empirical evidence support this model: An analysis of metabolic networks shows that metabolites of some of the most ancient pathways, such as glycolysis and the tricarboxylic acid (TCA) cycle, are among the most connected substrates of the network (Wagner and Fell, 2001). In terms of protein interaction networks, comparative genomics analyses have revealed that, on average, evolutionarily older proteins have higher connectivity than their younger counterparts (Wagner 2003, Eisenberg and Levanon 2003). The preferential attachment model aims to capture a general mechanism of network evolution capable of producing the observed scale-free topology. But it is

likely to operate under functional constraints, as protein function determines types of binding partners, the degree of connectivity, and time of origin of the network (Kunin et al. 2004).

The second type of evolutionary process consists of point mutations in a gene resulting in modifications of the interface between interacting proteins (Jones and Thornton 1996). Consequently, the corresponding protein may gain new connections (attachment) or lose (detachment) some of the existing connections to other proteins. Berg et al. (2004) refer to these attachment and detachment processes collectively as link dynamics. They estimate the empirical rates of link dynamics and gene duplication in the yeast protein network and find the former to be at least one order of magnitude higher than the latter. Based on this observation, they propose a new model for the evolution of protein networks in which link dynamics due to point mutations are the major evolutionary forces shaping the scale-free topology of the network while slower gene duplication processes mainly affect its size. According to this model, the fast link turnover rate leads to the fast loss of connectivity of proteins encoded by duplicate genes. This is consistent with an earlier observation that the majority of duplicate pairs have few or no interaction partners in common (Wagner 2001).

All of this previous research on protein network evolution has been directed towards understanding the origin of its global structural features. In contrast, little is known about the evolutionary process(es) that shape the network's local wiring diagrams, i.e., network motifs, although it is often implied that the local properties reflect solely evolutionary selection towards desirable functional traits (Shen-Orr et al. 2002, Mangan et al. 2003, Milo et al. 2004). A recent study (Vazquez et al. 2004) demonstrates that a

network's global and local structures mutually define and predict each other, raising intriguing questions about how the evolution of network motifs shape a network's overall structure and *vice versa*.

Perspectives

Although significant advances have been made in the past few years, protein network biology is still in its infancy. Future progress is expected in several directions. First and most importantly, to further expand our knowledge about protein interaction networks, we need to improve our data-gathering capabilities. This means development of highly sensitive and accurate methods to allow data collection under various cellular functional and temporal states as well as in different cell types in the case of metazoans. These new data sets will not only improve coverage of the networks but also enable us to ask questions about the dynamics of protein interaction networks.

In contrast to the yeast protein network, the human network is largely unexplored. Based on the existing data for yeast proteins, a conservative estimate puts the total number of protein interactions in human at roughly 40,000-200,000 (Bork et al. 2004). Currently, about 20,000-30,000 total interactions are recorded in the literature (Peri et al. 2004), mostly from small-scale studies with a few medium-scale studies centered on particular pathways (Bouwmeester et al. 2004) or cellular machineries (Andersen et al. 2003). Thus, there is a pressing need for experimental methods to be scaled up to the size of human proteome.

Meanwhile, novel computational approaches need to be developed to transfer as many interactions as possible from model organisms to human. Also, just as theoretical advances in sequence evolution were essential for the development of modern sequence

analysis algorithms, further advances in our understanding of network evolution will surely benefit many aspects of network analysis, such as cross-species network comparisons.

Interaction data provide a high-level representation of the key molecular components and interactions of a biological system. Queries against this interaction network highlight particular pathways and complexes of interest, which are then prime candidates suitable for low-level verification and modeling as important signaling and compensatory pathways. Over successive iterations of modeling and experiment, the network model becomes annotated with increasingly low-level and pathway-specific parameters such as physico-chemical reaction rates, binding constants, and diffusion and transport coefficients. The promise of this approach is that ultimately, protein network models may provide a comprehensive "wiring diagram" lending global insight into normal and diseased cell function.

References

Aebersold, R. and Mann, M. 2003. Mass spectrometry-based proteomics. *Nature* **422:** 198-207.

Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Avery, L. and S. Wasserman. 1992. Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet* **8:** 312-316.

Bar-Joseph, Z., G.K. Gerber, T.I. Lee, N.J. Rinaldi, J.Y. Yoo et al. 2003. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* **21:** 1337-1342.

Barabasi, A-L and Z.N. Oltvai 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5:** 101-113.

Barrow, H.G. and R.M. Burstall. 1976. Subgraph isomorphism, matching relational structures and maximal cliques. *Inform. Process. Left.* **4:** 83-84.

Batagelj, V. and A. Mrvar. 1998. Pajek - Program for Large Network Analysis. *Connections* **21:** 47-57.

Begley, T.J., A.S. Rosenbach, T. Ideker, and L.D. Samson. 2002. Damage Recovery Pathways in Saccharomyces cerevisiae Revealed by Genomic Phenotyping and Interactome Mapping. *Mol Cancer Res* **1:** 103-112.

Begley, T.J., A.S. Rosenbach, T. Ideker, and L.D. Samson. 2004. Hot spots for modulating toxicity identified by genomic phenotyping and localization mapping. *Mol Cell* **16:** 117-125.

Berg, J., M. Lassig, and A. Wagner. 2004. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evol Biol 4: 51.

Breitkreutz, B.J., C. Stark, and M. Tyers. 2003. Osprey: a network visualization system. *Genome Biol* **4:** R22.

Bouwmeester, T., A. Bauch, H. Ruffner, P.O. Angrand, G. Bergamini, K. Croughton, C. Cruciat, D. Eberhard, J. Gagneur, S. Ghidelli, C. Hopf, B. Huhse, R. Mangano, A.M. Michon, M. Schirle, J. Schlegl, M. Schwab, M.A. Stein, A. Bauer, G. Casari, G. Drewes, A.C. Gavin, D.B. Jackson, G. Joberty, G. Neubauer, J. Rick, B. Kuster, and G. Superti-Furga. 2004. A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nat Cell Biol* **6:** 97-105.

Butland, G., J.M. Peregrin-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, and A. Emili. 2005. Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature* **433:** 531-537.

DeRisi, J.L., V.R. Iyer, and P.O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278:** 680-686.

Deutschbauer, A.M., R.M. Williams, A.M. Chu, and R.W. Davis. 2002. Parallel phenotypic analysis of sporulation and postgermination growth in Saccharomycescerevisiae. *Proc Natl Acad Sci U S A* **99:** 15530-15535.

Dolma, S., S.L. Lessnick, W.C. Hahn, and B.R. Stockwell. 2003. Identification of genotype-selective antitumor agents using synthetic lethal chemical screening in engineered human tumor cells. *Cancer Cell* **3:** 285-296.

Gavin, A.C., M. Bosche, R. Krause, P. Grandi, M. Marzioch et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415:** 141-147.

Ge, H., Z. Liu, G.M. Church, and M. Vidal. 2001. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* **29:** 482-486.

Giot, L., J.S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang et al. 2003. A protein interaction map of Drosophila melanogaster. *Science* **302:** 1727-1736.

Guarente, L. 1993. Synthetic enhancement in gene interaction: a genetic tool come of age. *Trends Genet* **9:** 362-366.

Hanisch, D., A. Zien, R. Zimmer, and T. Lengauer. 2002. Co-clustering of biological networks and gene expression data. *Bioinformatics* **18 Suppl 1:** S145-S154.

Hartman, J.L., B. Garvik, and L. Hartwell. 2001. Principles for the buffering of genetic variation. *Science* **291:** 1001-1004.

Ho, Y., A. Gruhler, A. Heilbut, G.D. Bader, L. Moore et al. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415:** 180-183.

Huang, L.S. and P.W. Sternberg. 1995. Genetic Dissection of Developmental Pathways. In *Methods in Cell Biology* (eds. H.F. Epstein and D.C. Shakes), pp. 99-122. Academic Press, San Diego.

Ideker, T., O. Ozier, B. Schwikowski, and A.F. Siegel. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18 Suppl 1:** S233-240.

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori et al. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98:** 4569-4574.

Jansen, R., D. Greenbaum, and M. Gerstein. 2002. Relating whole-genome expression data with protein-protein interactions. *Genome Res* **12:** 37-46.

Jeong, H., S.P. Mason, A.L. Barabasi, and Z.N. Oltvai 2001. Lethality and centrality in protein networks. *Nature* **411:** 41-2.

Kelley, B.P., R. Sharan, R.M. Karp, T. Sittler, D.E. Root et al. 2003. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci U S A* **100:** 11394-11399.

Lee, T.I., N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph et al. 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298:** 799-804.

Li, S., C.M. Armstrong, N. Bertin, H. Ge, S. Milstein et al. 2004. A map of the interactome network of the metazoan C. elegans. *Science* **303:** 540-543.

Ng, S.K., Z. Zhang, and S.H. Tan. 2003. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics* **19:** 923-929.

Ozier, O, N. Amin, and T. Ideker. 2003. Global architecture of genetic interactions on the protein network. *Nat Biotechnol* **21:** 490-1.

Pe'er, D., A. Regev, and A. Tanay. 2002. Minreg: Inferring an active regulator set. *Bioinformatics* **18 Suppl 1:** S258-S267.

Peri, S., J.D. Navarro, T.Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T.K. Gandhi, K.N. Chandrika, N. Deshpande, S. Suresh, B.P. Rashmi, K. Shanker, N. Padma, V. Niranjan, H.C. Harsha, N. Talreja, B.M. Vrushabendra, M.A. Ramya, A.J. Yatish, M. Joy, H.N. Shivashankar, M.P. Kavitha, M. Menezes, D.R. Choudhury, N. Ghosh, R. Saravana, S. Chandran, S. Mohan, C.K. Jonnalagadda, C.K. Prasad, C. Kumar-Sinha, K.S. Deshpande, and A. Pandey. 2004 *Nucleic Acids Res* **32:** D497-501.

Rain, J.C., L. Selig, H. De Reuse, V. Battaglia, C. Reverdy et al. 2001. The protein-protein interaction map of Helicobacter pylori. *Nature* **409:** 211-215.

Sham, P. 2001. Shifting paradigms in gene-mapping methodology for complex traits. *Pharmacogenomics* **2:** 195-202.

Sharan, R., T. Ideker, B.P. Kelley, R. Shamir, and R. Karp. 2004. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *RECOMB.*

Sharan, R., T. Ideker, B.P. Kelley, R. Shamir, and R. Karp. 2004. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data.

*Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology--RECOMB***:** 282-289.

Thomas, J.H. 1993. Thinking about genetic redundancy. *Trends Genet* **9:** 395-399.

Tong, A.H., M. Evangelista, A.B. Parsons, H. Xu, G.D. Bader et al. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294:** 2364-2368.

Tong, A.H., G. Lesage, G.D. Bader, H. Ding, H. Xu et al. 2004. Global mapping of the yeast genetic interaction network. *Science* **303:** 808-813.

Uetz, P. 2002. Two-hybrid arrays. *Curr Opin Chem Biol* **6:** 57-62.

Uetz, P., L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson et al. 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403:** 623-627.

Walhout, A.J., R. Sordella, X. Lu, J.L. Hartley, G.F. Temple et al. 2000. Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science* **287:** 116-122.

Xenarios, I., L. Salwinski, X.J. Duan, P. Higney, S.M. Kim et al. 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30:** 303-305.

Yeang, C.-H. and T. Jaakkola. 2003. Physical network models and multi-source data integration. *The Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB)*.

Table 1: Current estimates of the volume of experimental protein-protein interaction data available in the public domain.

	Number of Proteins	Number of Interactions
	H. pylori	
Two-hybrid assays	710 [Rain et al. 2001]	1425
	E. coli	
Co-immunoprecipitation/ Mass spectrometry	530 [Butland et al. 2005]	5420 (spoke)
	S. cerevisiae	
Two-hybrid assays	934[Uetz et al. 2000] 4131[Ito et al. 2001]	854 3986
Co-immunoprecipitation/ Mass spectrometry	1361[Gavin et al. 2002] 1560[Ho et al. 2002]	3221 (spoke) 31304(matrix) 3589(spoke) 25333(matrix)
Synthetic lethal assays	1029[Tong et al. 2004]	3627
DIP (small scale experiments)	1629	5068
	C. elegans	
Two-hybrid assays	2898[Li et al. 2004]	4027
	D. melanogaster	
Two-hybrid assays	7048[Giot et al. 2003]	20405
	H. sapiens	
Co-immunoprecipitation/ Mass spectrometry	32[Bouwmeester et al. 2004]	221
HPRD (small scale experiments)	2750[Peri et al. 2004]	10534

Table 2: Brief overview of protein interaction databases

Protein interactions	
ADVICE	http://advice.i2r.a-star.edu.sg
BIND	http://bind.ca/
Bioverse	http://bioverse.compbio.washington.edu/
Curagen	http://curatools.curagen.com/pathcalling_portal/index.htm
CYGD/MIPS	http://mips.gsf.de/services/ppi
DIP	http://dip.doe-mbi.ucla.edu/
GRID	http://biodata.mshri.on.ca/grid/servlet/Index
HPRD	http://www.hprd.org/
Hybrigenics/PIMRider	http://pim.hybrigenics.com/pimriderext/common/
MINT	http://mint.bio.uniroma2.it/mint/
PLEX	http://apropos.icmb.utexas.edu/plex/plex.html
STRING	http://string.embl.de/
Protein networks/pathways	
Biobase/Transpath	http://www.biobase.de/pages/products/transpath.html
Biocarta	http://www.biocarta.com/genes/index.asp
Genmapp	http://www.genmapp.org/links.html
Reactome	http://www.reactome.org/

Table 3: Protein network visualization and analysis tools.

		Cytoscape V2.0	Osprey V1.2.0	Pajek V1.01	ProViz V1.0	WebInterViewer
General	**Website**	http://www.cytoscape.org	http://biodata.mshri.on.ca/osprey	http://vlado.fmf.uni-lj.si/pub/networks/pajek	http://cbi.labri.fr/eng/proviz.htm	http://interviewer.inha.ac.kr
	License	Free	Free for educational, research, and non-for-profit	Free for non-commercial use	Free	Free
	Platform	Linux, Mac, Windows	Linux, Mac, Windows	Windows	Linux	Linux, Mac, Windows
Data Exchange	**Import** *Files*	Flat file (space-delimited interactions, node and edge attributes, gene functional annotations), GML	Flat file (tab-delimited gene names, interactions, experimental system, source, literature evidence)	Flat file (space-delimited gene names, interactions), Vega graphs, Gedcom, Ucinet DL…	Tulip, PSI-MI (XML)	Flat file (tab-delimited gene names and interactions), GML, XML
	Databases	-	GRID interaction data	-	IntAct interaction data	DB on InterViewer3 server or local data server
	Additional	Expression data, arbitrary data attributes on nodes and edges	-	-	-	-
	Export *Text files*	Flat file (space-delimited, genes, interactions), GML	Flat file (tab-delimited, genes and interactions)	Flat file (space delimited, node and edge attributes), Vega graphs, Gedcom, Ucinet DL…	Tulip, PSI-MI	Flat file (tab-delimited, genes and interactions), XML, EdgeCnt, IG1
	Image files	EPS, JPEG, PDF, PNG, PS, SVG	JPEG, PNG, SVG	BMP, EPS/PS, Kinemage, MDL, SVG, VRML	PNG	BMP (with copyright note)
Visualization	**Graph layout**	5 algorithms	7 algorithms	7 algorithms	3 algorithms	2 algorithms
	Data attributes *Proteins*	All imported properties	GO terms	All imported properties	GO terms	-
	Interactions	All imported properties	Source, experimental system (e.g. two-hybrid), literature evidence	All imported properties	PSI-MI terms	-
	Visual mappings *Proteins*	Color, shape, line type, size, label, font	Color	Color, line type, size	-	-
	Interactions	Color, line type, arrow, label, font	Color	Color, line type, arrow	-	-
Analysis	**Filters** *Proteins*	Attribute values	GO terms	Attribute values	GO terms	-
	Interactions	Type (e.g. protein-DNA)	Experimental system, source	Attribute values	PSI-MI terms	-
	Network	Node degree, distance	Node degree, distance	Node degree, distance	Node distance	Node distance
	Multiple data superposition	-	+	+	+	+

	Subnetwork identification	MCODE, ActiveModules plug-ins	-	-	-	-
	Group and collapse nodes	-	-	-	-	Group cliques, nodes with same interactions
	Network comparison	PathBLAST plug-in	-	Intersection, union, difference	Find shared nodes and edges	Find shared nodes and edges
	Extras	Many plug-ins for extended analysis, e.g. netwk comparison via PathBLAST	-	Many operations on graphs and metric computation	URL links to external source for node and edge properties	Data server for central data storage; List of connected groups
Conclusions	**Pros**	• Flexible and extensible through many existing and user defined plug-ins • Superposition of gene expression and other data	• Direct import and quick visualization from GRID DB • Superposition of different datasets	• General network vis. and analysis tool • Multiple formats for exporting images • Rich set of operations on graphs and metric computation	• Interaction filter based on PSI-MI controlled vocabulary terms • New analyses as plug-ins using the Tulip graph management platform	• Central storage of data on server
	Cons	• Requires substantial preprocessing of data, e.g. special network formats and data attribute lists	• Limited visualization possibilities for external data sets (outside of GRID)	• Single platform • Not specifically designed for molecular interaction networks • Requires much data preprocessing	• Single platform • Limited visualization functionality	• No visualization of protein or interaction attributes (e.g. expression) • Only one filter • Very brief documentation

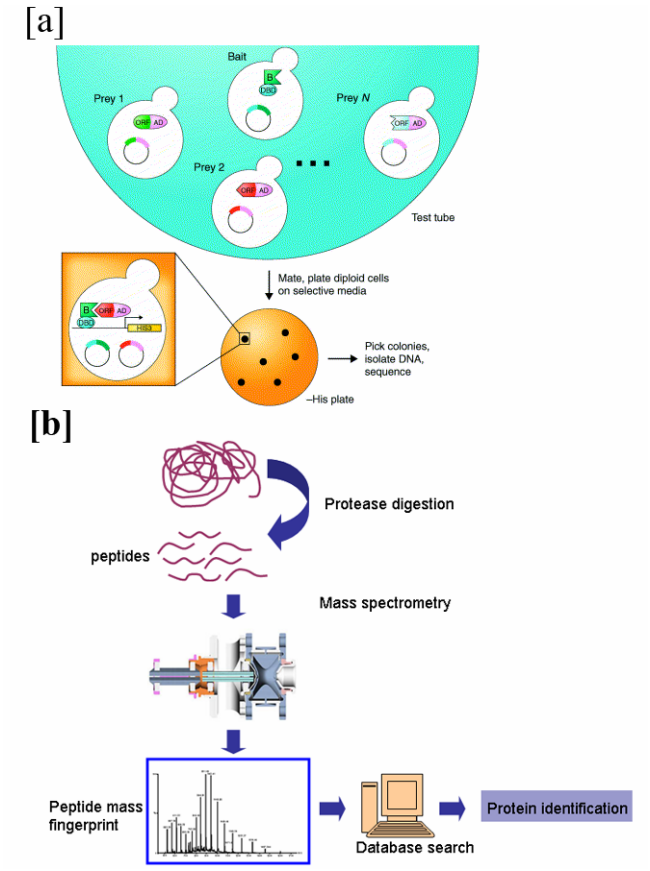

Figure 1: Principles of two high through-put technologies for identifying protein interactions. [a] Yeast two hybrid system. Typical two-hybrid screens use a library of random DNA or cDNA fused to a transcriptional activation domain (AD), expressed in yeast ('preys'; circles denote plasmids). The library clones are mated to a strain of opposite mating type that expresses a protein of interest ('bait', B) as a fusion to a DNA-binding domain (DBD). If bait and prey interact in the resulting diploid cells, they reconstitute a transcription factor, which activates a reporter gene whose expression allows the diploid cell to grow on selective media (here, without histidine). Positive clones have to be picked, their DNA isolated and the encoded plasmids sequenced in order to identify interacting proteins. Reproduced with permission from Uetz 2002. **[b]** Mass spectrometry. Intact proteins are proteolytically digested. The resulting peptide mixture is fractionated and introduced into a mass spectrometer. The mass spectrometer is responsible for separating peptide ions by their mass-to-charge (*m/z*) ratio. The peptides then serve as surrogate markers for the protein sequence. Proteins are identified by searching the resulting mass spectra through sequence databases.

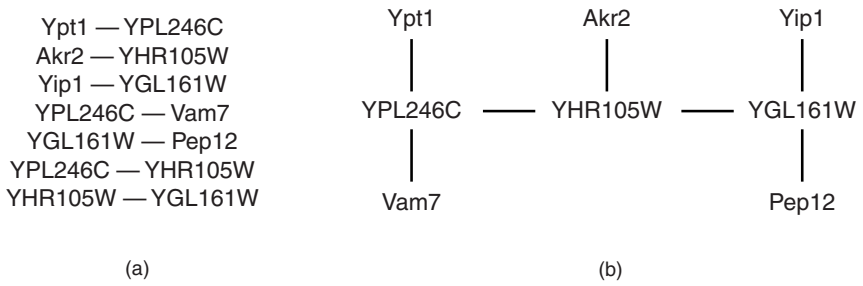

Figure 2: List [a] versus graphical network representation [b] of protein interactions. The two representations differ in *localization* (a protein occurs multiple times in the list but exactly once in the layout); *context* (in the layout, the neighbors of a protein are easily identified and studied; and *mental image* (the network layout allows proteins to be memorized by position) (Eades et al. 1991). In positioning the nodes, secondary information can be employed to guide the layout; for example, proteins can be spatially grouped by localization or function. In this way, a particular arrangement of the proteins can even increase the information content.

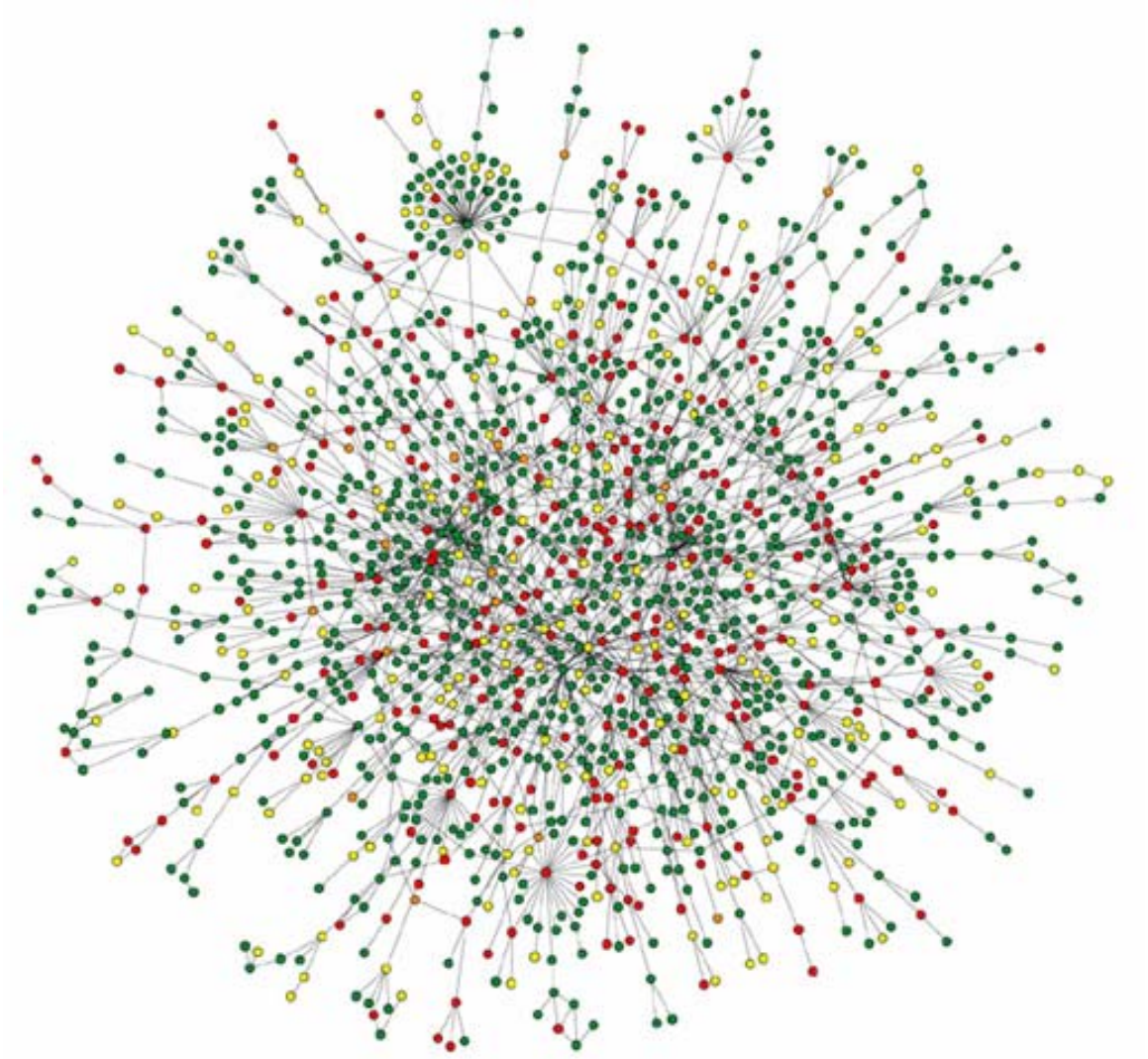

Figure 3: A map of protein-protein interactions in *Saccharomyces cerevisiae* based on a early systematic yeast two-hybrid experiment (Uetz et al. 2000), illustrates that a few highly connected nodes hold the network together. The color of a node indicates the phenotypic effect of removing the corresponding protein (red = lethal, green = non-lethal, orange = slow growth, yellow = unknown). Reproduced with permission from Barabasi and Oltvai, 2004.

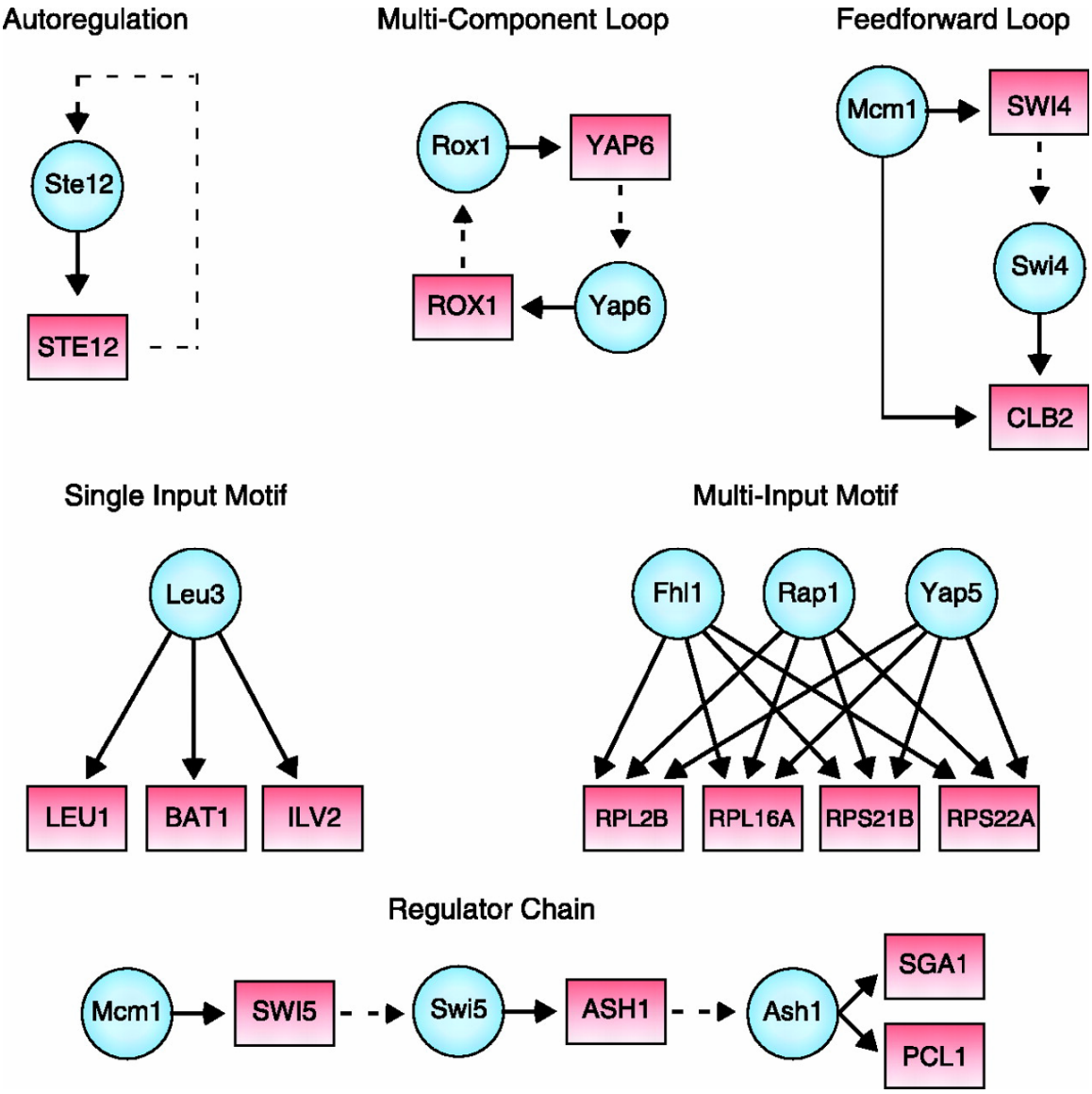

Figure 4. Examples of network motifs in the yeast regulatory network. Regulators are represented by blue circles; gene promoters are represented by red rectangles. Binding of a regulator to a promoter is indicated by a solid arrow. Genes encoding regulators are linked to their respective regulators by dashed arrows. For example, in the autoregulation motif, the Ste12 protein binds to the *STE12* gene, which is transcribed and translated into Ste12 protein. Reproduced with permission from Lee et al. 2002.

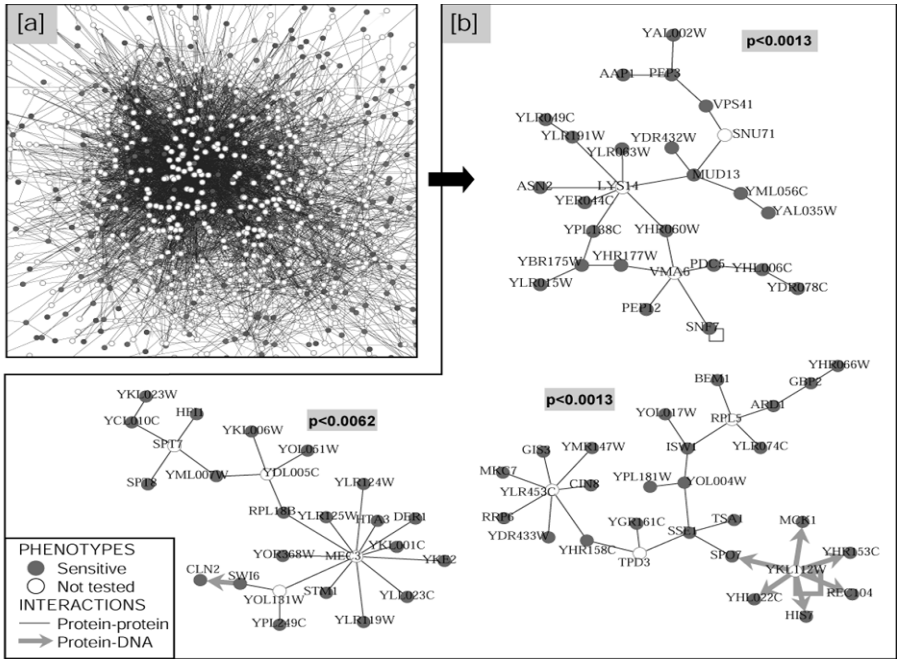

Figure 5: Screening damage phenotypes vs. the interaction network [a] A protein interaction network was integrated with 1,615 yeast deletion phenotypes gathered in response to MMS. **[b]** A search of the network found protein complexes containing significant numbers of MMS-essential proteins. Three of four identified regions are shown. Dark gray nodes represent MMS-essential proteins; white nodes were untested. Reproduced with permission from Begley et al. 2002.

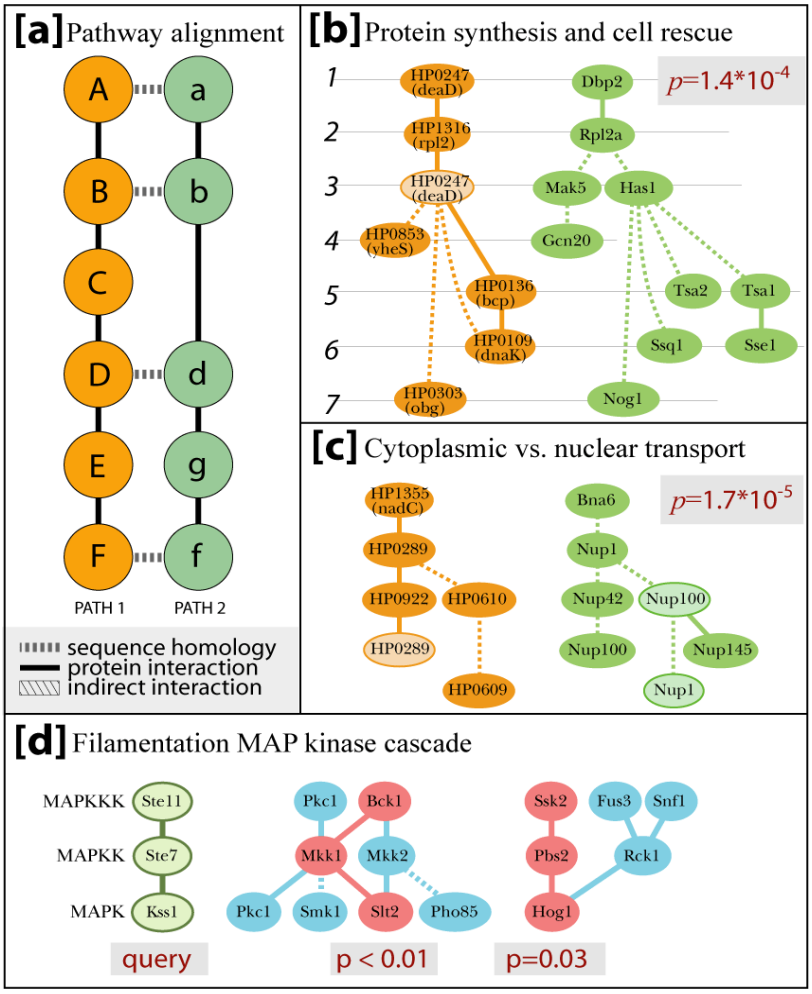

Figure 6: PathBLAST network alignment across species. [a] A model pathway alignment between two protein networks, where interactions in a pathway appear vertically and horizontal dotted lines link proteins with significant sequence similarity. Insertions (e.g., protein C) or mismatches (e.g.. proteins E and g) in the alignment are permitted but penalized. Panels **[b-c]** show aligned regions from the networks of *H. pylori* (orange; left) vs. *S. cerevisiae* (green; right). Bacterial/yeast protein pairs with significant sequence similarity are placed on the same row (e.g., deaD and Dbp2 in row 1 of [b]). **[d]** Querying the yeast network with a specific MAP kinase pathway involved in the yeast filamentation response. In panels [b-d], solid links indicate direct protein interactions, whereas dotted links indicate a single protein insertion (additional protein in one of the compared network).

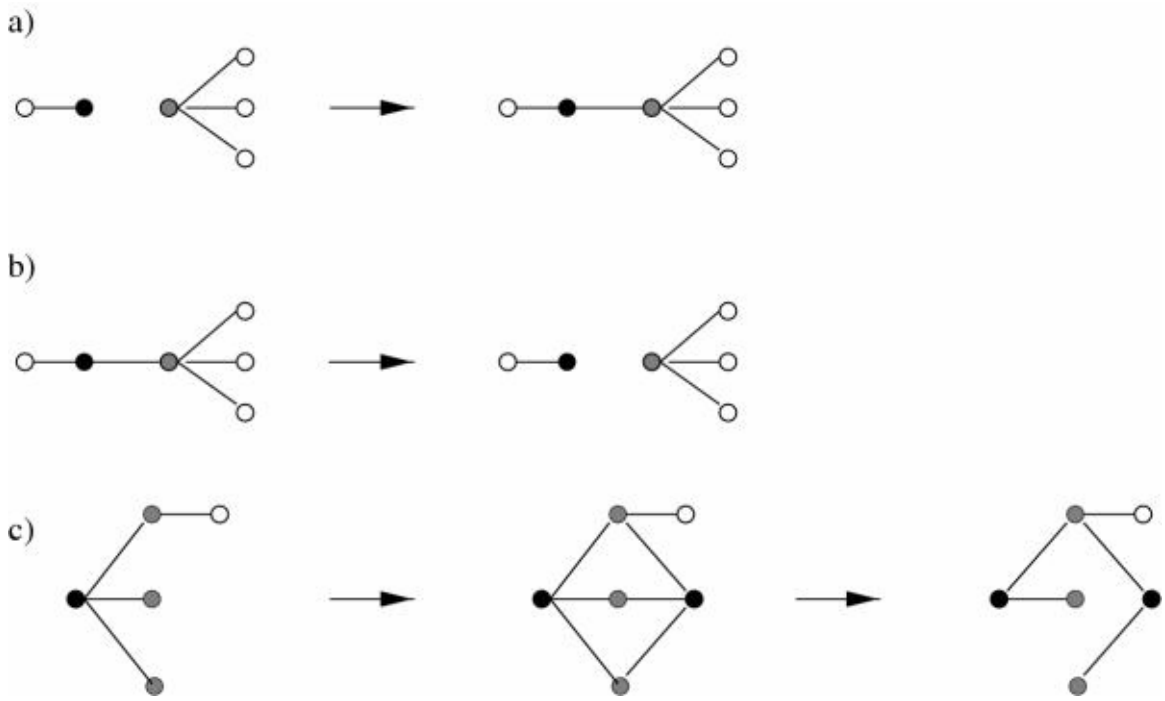

Figure 7: Evolutionary processes shaping protein interaction networks. The progression of time is symbolized by arrows. **[a] Link attachment** and **[b] link detachment** occur through point mutations in the gene encoding an existing protein. These processes affect the connectivities of the protein whose coding sequence undergoes mutation (shown in black) and of one of its binding partners (shown in gray). Empirical data shows that attachment occurs preferentially towards partners of high connectivity. **[a]** and **[b]** are collectively termed link dynamics. **[c] Gene duplication** usually produces a pair of nodes (shown in black) with initially identical binding partners (shown in gray). Empirical data suggests duplications occur at a much lower rate than link dynamics and that redundant links are lost subsequently (often in an asymmetric fashion), which affects the connectivities of the duplicate pair and of all its binding partners. Reproduced with permission from Berg et al. 2004.