REVIEW

Getting connected: analysis and principles of biological networks

Xiaowei Zhu,[1,2] Mark Gerstein,[3] and Michael Snyder[1,2,4]

[1]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA;
[2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA;
[3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA

The execution of complex biological processes requires the precise interaction and regulation of thousands of molecules. Systematic approaches to study large numbers of proteins, metabolites, and their modification have revealed complex molecular networks. These biological networks are significantly different from random networks and often exhibit ubiquitous properties in terms of their structure and organization. Analyzing these networks provides novel insights in understanding basic mechanisms controlling normal cellular processes and disease pathologies.

Proper execution of complex biological systems occurs through the intricate coordination of a large number of events and their participating components. Cellular proliferation, differentiation, and environmental interactions each requires the production, assembly, operation, and regulation of many thousands of components, and they do so with remarkable fidelity in the face of many environmental cues and challenges. Understanding how cellular and developmental events occur at a molecular level with such precision has become a major focus for modern molecular biology, and considerable effort has been devoted to determining the regulatory networks that control and mediate complex biological processes.

Until recently dissection of biological networks has occurred through the efforts of individual laboratories working on one or a few components, limiting a thorough understanding of individual biological processes in the context of the entire cellular network. Detailed analysis of specific components and their interacting partners or substrates can be used to assemble high-confidence pathways. For example, analysis of the NF-κB and TGF-β signaling pathways has revealed many components whose functions are reasonably well known for each of these pathways (Mishra et al. 2005; Karin 2006). Nonetheless, in spite of the intensive study of such pathways, new components of these pathways continue to be discovered (Covert et al. 2005; Ma et al. 2006), indicating that our analysis of even the most well-studied pathways is likely to be incomplete.

The advent of high-throughput techniques has allowed the large-scale identification of components (genes, RNAs, and proteins), their expression patterns, and their biochemical and genetic interactions. Although useful for generating large amounts of biological information, the data from such studies are often incomplete and contain errors. Nonetheless, they can provide valuable information about the functions of individual components and unexpected relationships between components and cellular processes. For example, Arg5,6, a well-characterized metabolic enzyme, was identified to have a DNA-binding activity through a proteome microarray screen and was later confirmed to regulate gene expression in vivo (Hall et al. 2004). Thus far a variety of large-scale data sets have been identified and used to assemble different networks. Below we briefly describe the different types of biological networks and general features and principles that result from the analysis of such networks.

Types of biological networks

Interaction data gathered through both individual studies and large-scale screens can be assembled into a network format whose topological structure contains significant biological properties. To date, at least five types of biological networks have been characterized in detail: transcription factor binding, protein–protein interactions, protein phosphorylation, metabolic interactions, and genetic interaction networks (examples of each of these networks and their sizes are presented in Fig. 1 and Table 1). Each of these networks is discussed briefly below.

Transcription factor-binding networks

Transcription factor-binding networks have been assembled in two ways: (1) The analysis of individual components has been used to develop intricate maps in sea urchins and other model organisms (Davidson et al.

[*Keywords*: Network topology; regulatory circuits; scale-free networks]

[4]**Corresponding author.**
E-MAIL michael.snyder@yale.edu; FAX (203) 432-6161.

Article is online at http://www.genesdev.org/cgi/doi/10.1101/gad.1528707.

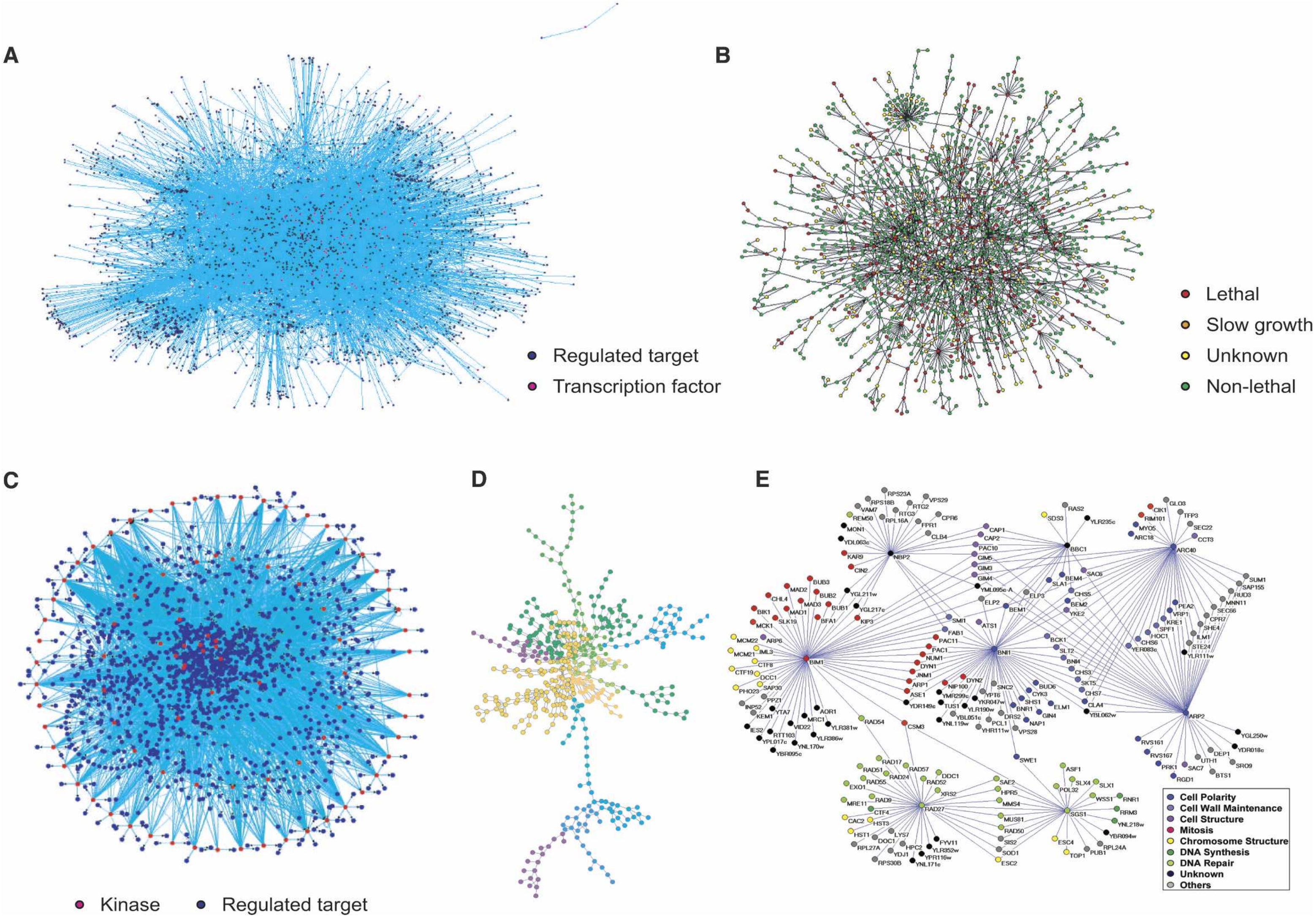


Figure 1. Examples of the five major biological networks. (*A*) A yeast transcription factor-binding network, composed of known transcription factor-binding data collected with large-scale ChIP–chip and small-scale experiments. This figure was generated with the program Pajek (de Nooy et al. 2005). (*B*) A yeast protein–protein interaction network, containing protein–protein interactions identified by yeast two-hybrid and protein complexes identified by affinity purification and mass spectrometry (Barabasi and Bonabeau 2003). (Reprinted by permission from Macmillan Publishers Ltd: *Nature* [Jeong et al. 2001], © 2001.) Nodes are colored according to the mutant phenotype. (*C*) A yeast phosphorylation network comprised primarily of in vitro phosphorylation events identified using protein microarrays (Ptacek et al. 2005). The figure was generated with Osprey 1.2.0. (Breitkreutz et al. 2003). (*D*) An *E. coli* metabolic network with 574 reactions and 473 metabolites colored according to their modules (Reprinted by permission from Macmillan Publications Ltd: *Nature* [Guimera and Nunes Amaral 2005], © 2005). (*E*) A yeast genetic network constructed with synthetic lethal interactions using SGA analysis on eight yeast genes (From Tong et al. 2001; reprinted with permission from AAAS). Nodes are colored according to their YPD cellular roles.

Table 1. *Current status of biological networks*

Type of network	Species	Number of nodes	Number of interactions	Reference
Transcription factor-binding network	*S. cerevisiae*	3528	7419	Yu et al. 2003[a]
		3207	11231	Harbison et al. 2004[b]
Protein–protein interaction	*C. elegans*	2788	4441	Stark et al. 2006
	D. melanogaster	7546	25403	
	Homo sapiens	7509	20979	
	Mus musculus	209	393	
	S. cerevisiae	5325	51773	
Phosphorylation network	*S. cerevisiae*	1325	4200	Ptacek et al. 2005
Metabolic network	*E. coli*	473	574	Guimera and Nunes Amaral 2005
	S. cerevisiae	646	1149	Tong et al. 2004
Genetic network	*S. cerevisiae*	3258	13963	Reguly et al. 2006[c]

[a] Transcriptional factor-binding data collected at rich-media condition.
[b] Transcriptional factor-binding data collected at a variety of growth conditions.
[c] Synthetic lethal interactions among nonessential genes.

2002); and (2) the large-scale identification of transcription factor-binding sites using chromatin immunoprecipitation followed by probing of genomic microarrays (ChIP–chip) or DNA sequencing (ChIP–PET or STAGE) has been used to assemble networks in yeast and other organisms (Horak and Snyder 2002; Kim et al. 2005; Wei et al. 2006).

Thus far a large number of ChIP mapping experiments have been performed in yeast and mammalian cells. The data from ChIP experiments are often of variable quality, particularly in mammalian cells. Most of the initial ChIP–chip experiments used genomic arrays comprised of PCR products that allowed crude mapping of binding sites and often lower-quality results. More recent experiments use oligonucleotide arrays that allow higher-resolution mapping of the binding regions (Cawley et al. 2004; Borneman et al. 2006). The calling of targets is not trivial as there is a considerable range of signals and probability values associated with each target, often leading to arbitrary assignment of thresholds to the data. Nonetheless, interesting networks have been assembled using these data sets.

For yeast, >250 ChIP–chip experiments have been performed using cells incubated in a variety of experimental conditions or treated with different stimuli, and >10,000 interactions have been reported (Horak et al. 2002; Lee et al. 2002; Harbison et al. 2004; Borneman et al. 2006). These have been assembled into a variety of global networks and subnetworks. For mammalian cells, a large number of experiments have also been performed, often by analyzing selected regions of the genome (Martone et al. 2003; Cawley et al. 2004) or promoter regions (Li et al. 2003). For example, the global identification of targets of three factors involved in embryonic stem cell maintenance has suggested pathways important for stem cell self-renewal (Boyer et al. 2005). Similarly, the analysis of targets of three major transcription factors has revealed a transcriptional map of skeletal myogenesis (Blais et al. 2005).

By combining binding data with expression data, the putative effect of binding on transcriptional output (i.e., activation or repression) can often be obtained. For inducible factors, studies with human NF-κB and STAT1 indicate that only a subset (30%–40%) of differentially expressed genes appear to be direct targets of the factor of interest; presumably many differentially expressed genes are regulated by factors other than the one of interest. Likewise, only a small fraction of binding sites appear to be directly modulating nearby gene expression, as many binding sites do not reside near genes whose expression is altered. For example, the majority of NF-κB- and STAT1-binding sites reside near genes whose expression is not altered by the conditions that activate the factor (Martone et al. 2003; Cawley et al. 2004; Hartman et al. 2005). In addition, experiments with yeast have shown that deletion of a transcription factor typically affects only a subset of targets (Gasch et al. 2000). These observations indicate that many binding sites lack biological function, or more likely, are functionally redundant with other regulatory sites or affect gene expression under other conditions. For the case of mammalian systems, they might also operate on genes that reside at distant locations (Carroll et al. 2005).

Protein–protein interaction networks

Protein–protein interaction maps represent the largest and most diverse data sets available to date. The first maps were generated using two-hybrid studies in which interactions of protein partners are accessed in yeast using a transcriptional readout (Uetz et al. 2000; Ito et al. 2001). Large-scale two-hybrid studies have been used to study interactions in other organisms such as *Drosophila*, *Caenorhabditis elegans*, and humans (Giot et al. 2003; Li et al. 2004; Rual et al. 2005). More recently, high-throughput studies using affinity purification followed by identification of associated proteins using mass spectrometry have resulted in large data sets of protein interactions. Two recent studies have described the purification of most proteins present in a eukaryotic cell, and both identified ∼500 protein complexes in yeast (Gavin et al. 2006; Krogan et al. 2006). Considering the coverage of the experiments, these studies suggest there are ~800 protein complexes in yeast. Extrapolation to the human proteome based on gene number predicts an estimate of 3000 human protein complexes.

Interactions studies each have technical concerns associated with them (Goll and Uetz 2006). Two-hybrid studies may reveal interactions that do not normally occur in vivo. Affinity purification, on the other hand, may yield protein contaminants and may not detect interactions in which binding partners are present substoichiometrically in a complex. Comparison between these data sets reveals only partial overlap even for the most comprehensive studies. This is likely due to the incomplete coverage of each study and diverse computational methods or stringencies applied to interpret the raw data sets. Nonetheless, these interaction maps, when integrated together, have revealed global topological and dynamic features of interactome networks that relate to known biological properties (see below).

Protein phosphorylation networks

Studies of yeast and humans have suggested that 30% of cellular proteins are phosphorylated in vivo (Cohen 2000; Ficarro et al. 2002; Manning et al. 2002a); this figure is most likely a large underestimate of the number of phosphorylated residues since comprehensive mapping studies have not been performed. Consistent with the importance of phosphorylation as a regulatory mechanism, eukaryotes devote ~2% of their protein-coding genes to protein kinases, ranging from 122 for yeast to 518 for humans (Zhu et al. 2000; Manning et al. 2002b).

Until recently, protein phosphorylation has generally been mapped on a limited scale. However, newly developed approaches in mass spectrometry have allowed the identification of a large number of phosphorylated residues including those regulated during cell stimuli and developmental responses (Ficarro et al. 2002; Gruhler et al. 2005; Ptacek and Snyder 2006). These approaches usually involve enrichment of phospho-proteins using matrices that bind phospho-modified proteins. For example, one study of the developing forebrain and midbrain tissues of embryonic mice used strong cation exchange columns followed by tandem mass spectrometry to identify >500 serine, threonine, or tyrosine phospho-sites (Ballif et al. 2004). Other studies have used immunoprecipitation to enrich for tyrosine phospho-proteins followed by mass spectrometry; these have led to discovery of novel phospho-tyrosine protein modifications in human T cells (Brill et al. 2004; Tao et al. 2005).

In addition to the identification of phosphorylated residues, two new approaches have shed light on discovering substrates of protein kinases. The use of modified kinases that accept only radiolabeled ATP analogs has revealed many substrates for several yeast kinases including the cyclin-dependent kinases Pho85 and Cdc28 (Dephoure et al. 2005; Loog and Morgan 2005). A second approach used a proteome microarray containing 4400 yeast proteins to detect in vitro substrates for the majority of yeast protein kinases. This study identified ~4200 phosphorylations affecting >1300 substrates (Ptacek et al. 2005). These different studies have identified a large number of phosphorylation events, many of which were validated in vivo. Many of the phosphorylations involved substrates that operate in a known pathway of the kinase; however, several validated substrates function in different cellular processes from those known for the kinase, thereby revealing new functions for the protein kinases.

Metabolic interaction networks

The wealth of biochemical data generated in the past century when combined with genome sequences allows the construction of metabolic networks. The metabolic network usually focuses on the mass flow in basic chemical pathways that generate essential components such as amino acids, sugars, and lipids, and the energy required by the biochemical reactions. As such, these networks typically present both protein and metabolite information. Literature curation and genome annotation have elucidated many complex biochemical pathways (Kanehisa and Goto 2000; Overbeek et al. 2000) from which various metabolic networks have been reconstructed in a wide variety of organisms such as *Escherichia coli* (Reed et al. 2003), *Saccharomyces cerevisiae* (Duarte et al. 2004), and human mitochondria (Vo et al. 2004).

Interactions in metabolic networks are closely related to the gene functions, and therefore have great potential for immediate applications in the interpretation of gene roles. Considerable attention has been focused on the network dynamics using constraint-based analyses such as flux balance analysis (FBA), which assumes the steady state of all metabolites and that the organisms will optimize the metabolite fluxes to maximize biomass production (Segre et al. 2002; Famili et al. 2003; Forster et al. 2003). This approach has led to many successful predictions. For example, an in silico flux model was used to predict the phenotypes of yeast strains containing gene deletion mutations grown under various media conditions and achieved a remarkable 83% accuracy (Duarte et al. 2004). In addition, a flux model on a yeast metabolic network was able to explain enzyme dispensability; that is, how loss-of-function mutations of many yeast enzymes result in viable strains (Papp et al. 2004). This model suggested that the majority of nonessential enzymes are vital for cell growth under certain previously untested conditions, whereas only a small subset are compensated by isoenzymes or parallel pathways. Other successful constraint-based analyses in metabolic networks have also been performed. These include (1) re-engineering micro-organisms with gene deletions for the purpose of manipulating their chemical products (Burgard et al. 2003) and (2) evaluating steady-flux distributions in human mitochondria using constraints related to normal, disease, and dietetic treatment conditions (Thiele et al. 2005). Additional examples of constraint-based analysis can be found in a detailed review (Price et al. 2004). Although many metabolic network studies were developed in micro-organisms and *S. ce-*

revisiae. These studies may also shed light in other organisms since the fundamental network structures may be conserved in evolution. Topological analysis of metabolic networks in 43 organisms covering all three life domains revealed highly similar topological properties, although great diversity exists among individual pathways and components (Jeong et al. 2000).

Genetic and small molecule interaction networks

Combining mutations in two different genes can either synergistically reduce or enhance the growth or fitness of an organism, relative to organisms containing individual mutations. One of the most common interactions analyzed is "synthetic lethality" in which mutations that do not individually cause loss of viability are lethal when combined (Bender and Pringle 1991; Costigan et al. 1992). For many—if not most—species, the majority of genes are not lethal when mutated individually; this is likely because of either genetic redundancy or because the affected genes normally enhance the fitness of the organism rather than are essential for its viability. When mutations are combined in the same strain to produce a phenotype stronger than that caused by an individual mutation, the mutated genes are often thought to reside in parallel redundant pathways, although other interpretations are possible. Regardless of the reason, the ability to combine mutations to produce strong phenotypes provides the opportunity to carry out synthetic lethal analysis on a large scale that provides a wealth of useful information.

Large-scale synthetic lethal screens have been performed in *S. cerevisiae* in which deletion mutations in only 1100 protein-coding genes (of ~6000 total) prevent growth in standard rich medium (Winzeler et al. 1999; Giaever et al. 2002). Genetic interaction screens using either plate (SGA) or microarray readouts (dSLAM) with yeast strains containing mutations in nonessential genes have been used to systematically uncover synthetic lethal interactions (Tong et al. 2001, 2004; Pan et al. 2004). One recent study that combined genetic interactions from high-throughput methods and a literature curation of 53,117 publications in PubMed produced an *S. cerevisiae* genetic network containing 3258 genes and 13,963 interactions; this network revealed a significant overlap with protein–protein interactions (Reguly et al. 2006). For essential genes, strains containing conditional mutations such as those that confer a temperature-sensitive growth defect or with the gene under the control of a tetracycline titratable promoter can be analyzed under conditions that reduce, but do not eliminate, the activity of the gene product (Davierwala et al. 2005). Analysis of these interactions has also revealed functional relationships between genes and a high correlation with other properties, such as mutant phenotypes and cellular localization, thus helping to assign biological roles for unknown genes and infer novel functions to annotated genes.

In addition to synthetic lethal screens, other types of genetic interactions can be measured. These include combining mutations that disrupt inhibitory interactions and thus enhance growth. In fact, interactions that when combined either enhance or reduce growth have been investigated to generate a detailed genetic interaction map, E-MAPs (for epistatic miniarray profiles), for genes involved in the yeast early secretory pathway (Schuldiner et al. 2005). Another type of genetic interaction is a synthetic dosage lethal screen in which overexpressed genes are introduced into a mutant strain background; synthetic dosage lethality can provide additional, and often nonoverlapping interaction data to those found by combining inactivating mutations (Measday et al. 2005). For example, overexpression of genes that inhibit growth in a mutant strain background has been used to screen for genes that would negatively regulate protein kinase substrates (Sopko et al. 2006). Finally, a conceptually similar approach to synthetic lethality is to screen for mutant strains that are hypersensitive to inhibitory small molecules. Thus far, screens have been performed between inhibitory chemical compounds and deletion mutants of all yeast nonessential genes or strains heterozygous for mutations in essential genes (Giaever et al. 2004; Parsons et al. 2004). Such chemical genetic interactions, when integrated with genetic interactions, often suggest pathways targeted by the drugs as well as potential direct drug targets. Thus, this approach offers a powerful tool in deciphering the mechanisms of action of drugs as well defining suitable biological pathways that can be targeted for inhibition.

Other biological networks

The global behavior of gene interactions can also be investigated by networks connecting genes and/or proteins sharing certain properties. A coexpression network, in which genes are connected if their transcripts are coregulated, was assembled in *S. cerevisiae* and contains 4077 genes connected by 65,430 interactions (Stuart et al. 2003; van Noort et al. 2004). Proteins that share other properties, such as biological processes (Tari et al. 2005) and mutant phenotypes (Gunsalus et al. 2005; Ohya et al. 2005), can also be linked with each other and assembled into networks. The coexpression and homolog networks differ from the other networks described above in that the interactions are based on similarities not related to gene function. Nonetheless, they can still be investigated with similar approaches and often exhibit comparable network topology. Moreover, these networks also share the "guilt by association" property with the five biological networks: Highly connected proteins are likely to be functionally related. Therefore studies on these networks may also discover novel protein roles and help to decipher the complex cellular networks, especially when integrated with other biological networks.

Global topology

Interactions are often assembled into network maps comprised of proteins (or genes) termed vertices or nodes and connections between them defined as edges (in undirected networks) or arcs (in directed networks). The directionality of a network is dependent on the characteristics of the biological data. Protein–protein and genetic interactions are usually represented with an undirected network, whereas transcription factor binding, phosphorylation, and metabolic networks have directionality built into their interactions. One feature of nearly all of the interaction studies is that the strength of interactions can vary considerably. For example, the dissociation constant values observed in a biological system can vary by >10 orders of magnitude (Wallis et al. 1995). Such quantitative information, however, is rarely used in most network analyses, and interactions are usually reported as binary measurements. Future studies are likely to overcome these limitations as more accurate measurements are obtained, and weighted values can be assigned to network connections as indicators of the interaction strength.

Network topology plays a vital role in understanding network architecture and performance. Several of the most important and commonly used topological features include degree, clustering coefficient, shortest path length, and betweenness (Fig. 2). Detailed descriptions of each these statistics are listed as follows: (1) Degree: The number of links connected to one vertex is defined as its degree. In directed networks, the number of arcs that end at the node is termed as "in-degree," and the number of arcs that start from the node is termed as "out-degree." A node with high degree is better connected in the network and therefore may play a more important role in maintaining the network structure. (2) Distance: The shortest path length between two vertices is defined as their distance. In an interaction network, the maximum distance between any two nodes is termed as the graph diameter. The average distance and diameter of a network measure the approximate distance between vertices in a network. A network with a small diameter is often termed as a "small world" network (Milgram 1967), in which any two nodes can be connected with relatively short paths. Many real world networks such as metabolic networks have a small world architecture (Watts and Strogatz 1998), which may serve to minimize transition times between metabolic states (Wagner and Fell 2001). (3) Clustering coefficient: The clustering coefficient of one vertex can be calculated as the number of links between the vertices within its neighborhood divided by the number of links that are possible between them. A high clustering coefficient for a network is another indicator of a small world. (4) Betweenness: Betweenness is the fraction of the shortest paths between all pairs of vertices that pass through one vertex or link. Betweenness estimates the traffic load through one node or link assuming that the information flows over a network primarily following the shortest available paths.

Assembly of interactions into networks reveals that current versions of biological networks are not randomly organized but rather have a "scale-free" format containing hubs with many connections and a large number of nodes that have one or a small number of connections (Fig. 3; Barabasi and Oltvai 2004). This organization was originally discovered in World Wide Web interactions and later found to exist in four of the types of networks described above: protein–protein interactions, transcription factor binding, metabolic, and genetic data sets (Barabasi and Albert 1999; Jeong et al. 2000, 2001; Guelzim et al. 2002; Tong et al. 2004). Below we demonstrate that this is also the case for the phosphorylation network as well. Compared with a bell-shaped degree distribution in random networks, scale-free networks have a typical "power law" distribution, $P(k) \propto k^{-\gamma}$, in which k is the degree and $P(k)$ is the probability that a randomly selected node has a degree k. This results in a "fat-tailed" distribution in which there are vertices with

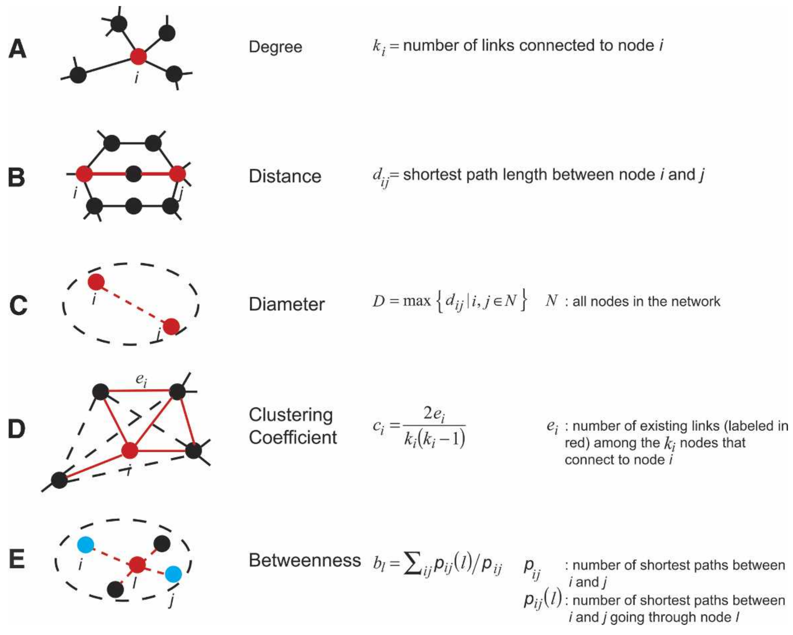

A
i
Degree
k_i = number of links connected to node i
B
i
j
Distance
d_{ij}= shortest path length between node i and j
C
i
j
Diameter
$D = \max\{d_{ij} | i, j \in N\}$
N : all nodes in the network
D
e_i
i
Clustering Coefficient
$c_i = \frac{2e_i}{k_i(k_i-1)}$
e_i : number of existing links (labeled in red) among the k_i nodes that connect to node i
E
i
l
j
Betweenness
$b_l = \sum_{ij} p_{ij}(l)/p_{ij}$
p_{ij} : number of shortest paths between i and j
$p_{ij}(l)$: number of shortest paths between i and j going through node l

Figure 2. Topological parameters. Five commonly used topological parameters are illustrated in both graphs and formulae. (*A*) Degree measures the number of connections one node has. (*B*) Distance is the length of the shortest path between two nodes. (*C*) Diameter is the maximum distance between any two nodes in a network. (*D*) Clustering coefficient measures the percentage of existing links among the neighborhood of one node. (*E*) Betweenness is the fraction of those shortest paths between all pairs of vertices that pass through one vertex or link. All graphs and formulae are based on an undirected network.

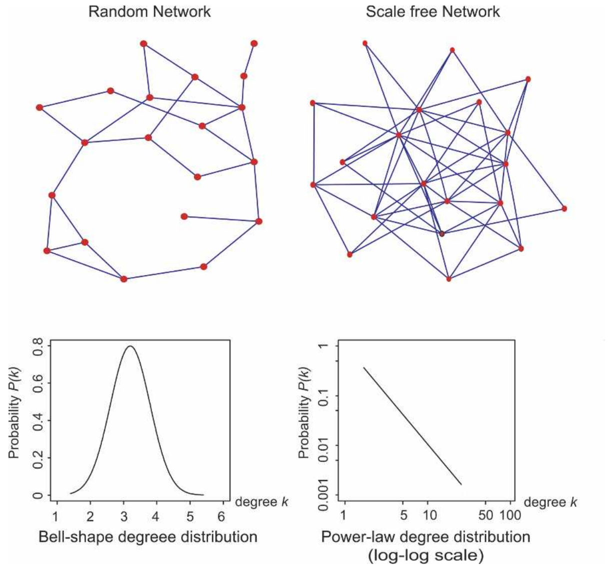


Figure 3. Topological comparison between a random network and a scale-free network. Degree distribution in random networks is bell-shaped. The scale-free network has more high-degree nodes and a power-law degree distribution, which leads to a straight line when plotting the total number of nodes with a particular degree versus that degree in log-log scales.

high degrees termed "hubs." The advantage of this type of organization is that the system is more robust; random loss of individual nonhub vertices is less disruptive in a scale-free network than a random network.

Hub components in a scale-free network are extremely important and therefore usually play essential roles in biological systems. In the yeast protein–protein interaction networks, hubs are more likely to be essential and conserved relative to nonhub proteins (Jeong et al. 2001; Barabasi and Oltvai 2004). Presumably much of the regulation in a network occurs and is mediated through such proteins. Likewise, key components whose activation is sufficient to induce a cellular process (master regulator genes) have been shown to be regulated by many other components and are thus target hubs; these often lie downstream in the process (Weintraub et al. 1989; Borneman et al. 2006). Not all components within a regulatory pathway serve as master regulators, probably because noise introduced into the system may inappropriately activate the process at undesired times. Presumably, components that lie within a network are buffered through both positive and negative regulatory contacts that prevent them from directly activating a biological process. The location of master regulators at the bottom of a highly connected network would allow maximum information input to be interpreted through upstream components and relayed into a final decision output; thus master regulators often represent important regulatory nodes in biological networks. For example, Twist, a master regulator controlling gene expression in embryonic morphogenesis, is responsible for tumor invasion and metastasis (Yang et al. 2004).

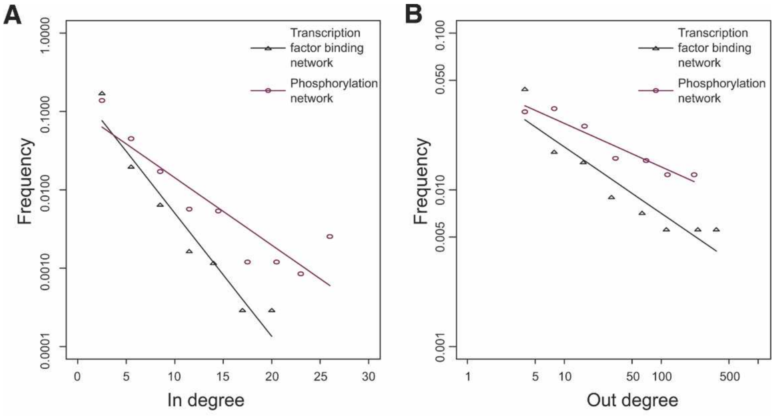
Further analysis of the transcription factor network has also revealed an additional novel aspect of regulatory network hierarchy. When the binding targets of *E. coli* and *S. cerevisiae* transcription factors are analyzed with respect to binding to other transcription factors, a pyramid-shaped hierarchical organization can be assembled with a few key regulators at the top to which few other factors bind and most transcription factors on the bottom as the functional units for specific pathways (Yu and Gerstein 2006). Similar to the middle managers in social networks such as governmental hierarchies, transcription factors in the middle layers often regulate more targets and have higher betweenness, indicating that they may function as bottlenecks in the hierarchy. With more interaction data gathered in the future, such hierarchical structures can also be investigated in other directed networks such as metabolic networks and phosphorylation networks.

Similarities between the transcription and phosphorylation networks

Transcriptional control and post-translational regulation with kinase phosphorylation are two major methods eukaryotes use for gene regulation; each controls a large number of targets. In yeast, humans, and many other organisms, the number of these two types of regulators is within twofold; there are ~250 transcription factors and 122 protein kinases in yeast (Zhu et al. 2000; Harbison et al. 2004) and ~1300 transcription factors and 518 protein kinases in humans. As shown in Figure 4, we have performed a detailed comparison of the network topologies of the yeast transcription factor-binding network and phosphorylation network under rich-nutrient conditions. These networks contain a remarkable number of similarities. First, the two networks share similar degree distributions: exponential in-degree distributions (Fig. 4A) and power law out-degree distributions (Fig. 4B). Second, many topological parameters are comparable between the two networks; however, the phosphorylation network is denser than the transcription factor-binding network and contains more nodes with large in- and out-degrees. Finally, the current phosphorylation network is smaller than the transcription factor-binding network. Both networks are built on incomplete data sets and may contain errors. The yeast phosphorylation data, in particular, are primarily collected from one large-scale study covering only two-thirds of all the yeast kinases. The transcription factor-binding network has more experimental sources and therefore a larger coverage. Since diameter is positively correlated with the network size, and limited sampling of a network often lowers the average clustering coefficient (Friedel and Zimmer 2006), the difference in the network size may explain why the transcription factor-binding network has a larger diameter and a higher clustering coefficient.

Network modules

Although initial studies have characterized the global topological structure of biological networks, recently



C

Network	# Vertices	# Arcs	Average in degree	In degree distribution	Average out degree	Out degree distribution	Average clustering coefficient	Diameter
Transcriptional network	3528	7419	2.13	$p(k) = 0.189e^{-0.362k}$	41.2	$p(k) = 0.050k^{-0.424}$	0.028	14
Phosphorylation network	1238	3671	3.11	$p(k) = 0.104e^{-0.198k}$	45.9	$p(k) = 0.050k^{-0.277}$	0.022	7

Figure 4. The yeast phosphorylation network resembles the transcription factor-binding network in their topological structures. (*A*) The in-degree and out-degree distributions were plotted after the nodes were binned to several degree intervals. Both networks have power-law in-degree distributions and exponential out-degree distributions. (*B*) Many topological parameters are comparable between the two networks, except that the transcriptional network is larger and the phosphorylation network is denser.

much attention has been paid to the local units of the networks. Large subgraph units, assembled by groups of densely associated proteins and connected to each other with loose links, are defined as network modules (Girvan and Newman 2002; Rives and Galitski 2003; Newman 2006). Such community-like network modules have been uncovered in many types of social networks as well as biological networks, in which they often function as essential components of the network. For example, one study of protein interactions in a transcriptional network indicates that different types of transcriptional regulators such as transcription factors, nuclear transporters, and nucleosome remodeling proteins prefer to form modules within each class, and the modules are jointed with sparse connections (Tsankov et al. 2006). The modules often contain proteins of unknown function, and therefore may shed light on protein function predictions. Furthermore, two classes of proteins are revealed by studies of modular structures. "Module organizer" proteins are highly connected to other proteins within modules and are essential to the module functions. "Module connector" proteins link different modules together and are vital for intermodule communications (Rives and Galitski 2003).

Many methods have been developed to identify possible network modules. A traditional method, hierarchical clustering, assigns a weight value to the distance between any two nodes in a network, and then gathers nodes with similar weight vectors together into strongly connected cores (Rives and Galitski 2003). Instead of detecting cores of modules in hierarchical clustering, the Girvan-Newman algorithm focuses on defining the boundaries of modules by searching for edges with high betweenness and therefore those that are more likely to link different modules (Girvan and Newman 2002). Other algorithms have been introduced recently and may demonstrate improvement in module identification (Guimera and Nunes Amaral 2005; Adamcsek et al. 2006; Newman 2006). One concern, however, is that network modules are often dependent on the methods and parameters used in the initial data partitioning, and in general it is difficult to tell which method is better (Barabasi and Oltvai 2004). Furthermore, inaccurate and incomplete data of the interaction networks may also lead to biased module predictions. Nonetheless, networks modules are still ubiquitous structures in most biological networks and may help one to better understand the interplay between network structure and function.

Network motifs

The availability of large interaction data sets allows the identification of much smaller common patterns or motifs within large networks that are used with significantly higher frequencies relative to randomized networks. Analysis of transcription factor-binding data in *E. coli* has revealed three different types of motifs: feed-forward loops (FFL), single input modules (SIM), and dense overlapping regulons (DOR) (Shen-Orr et al. 2002). FFL and DOR are also found to be significantly enriched in yeast transcriptional networks (Milo et al. 2002). It is possible that many, and perhaps all, single input motifs in eukaryotes are the result of incomplete data and that most genes probably contain multiple inputs.

We applied a tool, mfinder (Milo et al. 2002), to identify enriched three-element and four-element motifs in an updated yeast transcription factor-binding network and the yeast phosphorylation network. Both data sets were generated in yeast cells grown in rich media conditions. Among all possible three-element motifs, the FFL was found to be well overrepresented in transcriptional networks (Fig. 5). Coherent FFL, in which both transcription factors have the same regulation effects (induction or repression) on the target, may suggest a functional design for gene transcription regulation. Studies have shown that coherent FFLs can control downstream processes in a fashion that is resistant to transient noise, since targets in FFL can only be effectively regulated

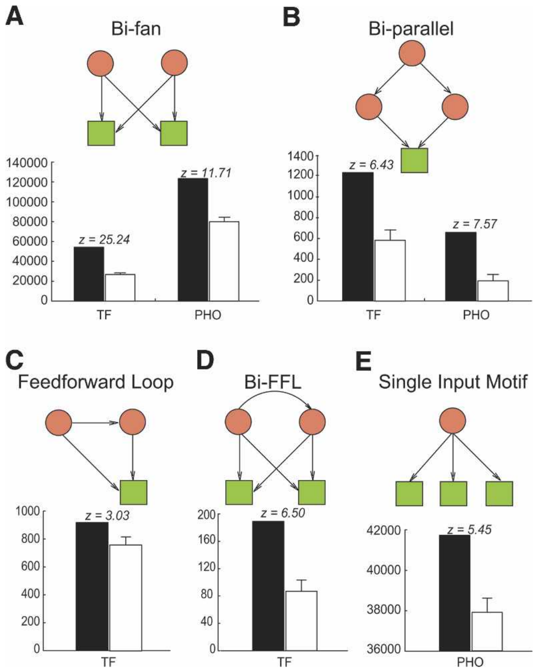


Figure 5. All three-unit and four-unit motifs enriched in the yeast transcriptional factor-binding (TF) network and phosphorylation (PHO) network. The units are colored red as regulators and green as targets. The significance of enrichment is calculated by comparing motif numbers in the transcription factor or phosphorylation networks (solid bars) with the numbers from randomized networks (hollow bars) and indicated by the z-scores. (*A*) Bi-fan motifs, in which two regulators bind common targets, are enriched in both the transcription factor network and phosphorylation network. (*B*) Bi-parallel motifs, in which one regulator controls two other regulators that further regulate one target gene, are enriched in both the transcription factor network and phosphorylation network. (*C*) FFLs, in which one regulator controls another regulator and both of them bind a common target, are enriched in the transcription factor network only. (*D*) Bi-FFL motifs, in which one regulator controls another regulator and both of them bind two common targets, are enriched in the transcription factor network only. (*E*) Single input motifs, in which one regulator binds to multiple targets, are enriched in the phosphorylation network only.

through persistent signals (Shen-Orr et al. 2002). A FFL motif can be easily extended to a four-element motif, "bi-FFL," in which the two regulators collectively control two targets. Bi-FFL motifs are also significantly enriched in yeast transcription factor-binding networks.

Thus far, FFLs are not enriched in the current yeast phosphorylation network. This may be due to the approach used to prepare the network that tends to underestimate the phosphorylation events between kinases, and additional data may be required to properly evaluate this network. However, it is possible that the lack of FFL in phosphorylation networks relative to transcriptional networks also reflects the biology of these networks. Phosphorylation networks are often activated by transient signals that lead to extremely rapid responses on the order of a few minutes. In contrast, transcriptional networks are slower and take longer to reach steady state.

Two four-element motifs were enriched in both the yeast transcriptional network and the phosphorylation network (Fig. 5). A simple version of the DOR motif, the "bi-fan motif," in which two regulators bind common targets, may suggest a way to use a limited number of regulators to precisely control a large number of targets under several different conditions. Moreover, the cooperation of transcription factors to regulate targets can also compensate for the degeneracy and low affinity of single transcription factor-binding sites (Pilpel et al. 2001). The other enriched four-element motif, the "bi-parallel motif," comprises a regulator controlling two other regulators that further regulate one target gene. Bi-parallel motifs are found in both transcriptional and phosphorylation networks and indicate redundancy. In addition to the two four-element motifs shared by both networks, the single input motif (SIM) was found to be overrepresented only in the yeast phosphorylation network. This likely reflects the lack of phosphorylation data currently available.

Network integration

Integration of different experimental resources is used in several different ways: (1) to improve the accuracy of interactions, (2) to identify composite motifs, and (3) to make functional predictions. Integration of similar data sets generated with different methods provides a crucial way to improve data quality and recover missing data. To remove erroneous interactions in the yeast protein–protein interaction network, a "filtered yeast interactome" (FYI) was constructed with high-confidence interactions observed in at least two experimental sources (Han et al. 2004). Studies on *C. elegans* early embryogenesis genes led to an integrative network containing three types of heterogeneous data: protein–protein interaction, expression profiling similarity, and phenotypic profiling similarity (Gunsalus et al. 2005). Further functional analysis demonstrated that gene pairs connected by interactions from multiple sources are more likely from the same GO functional categories, indicating improved accuracy through data integration. In the transcriptional network, integration with the gene expression data set has also proven to be useful to improve the data quality and reveal novel *cis*-regulatory modules (Bar-Joseph et al. 2003).

Recent bioinformatics software platforms enable users to query and integrate very different types of interaction data to learn new information (Breitkreutz et al. 2003; Shannon et al. 2003; Stark et al. 2006). Instead of searching for overlapping interactions, integration of very different types of interaction data can also be performed to reveal composite motifs that contain multiple types of interactions and elements as basic units. An integration of transcription factor binding, protein–protein interactions, and phosphorylation data from yeast has revealed a mega-network of >60,000 interactions (Fig. 6A). Investigations in this mega-network revealed seven three-element kinase-centered composite motifs (Fig. 6B), of which five (motifs 1–5) were shown to be overrepre-

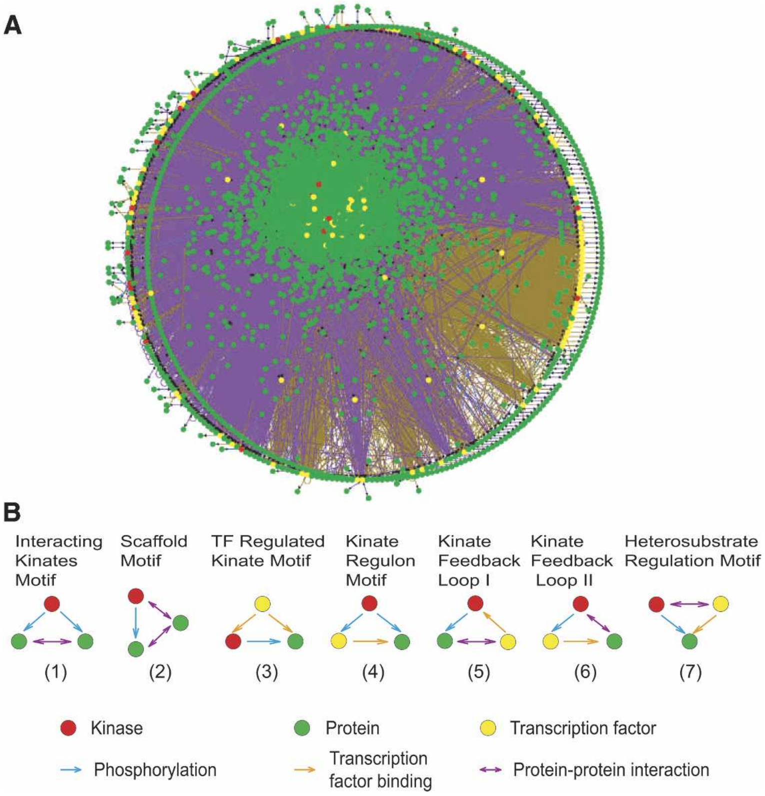


Figure 6. Network integration: mega-network and composite motifs. (*A*) Three types of interactions—phosphorylation (blue), transcription factor binding (yellow) and protein-protein (magenta)—are combined into a mega-network. (*B*) Seven three-element kinase-centered composite motifs are listed. (1) Interacting kinate motif in which one kinase phosphorylates two interacting substrates. (2) Scaffold motif in which one protein interacts with both a kinase and its substrate. (3) Transcription factor-regulated kinate motif in which one transcription factor (TF) regulates the expression of both a kinase and its substrate. (4) Kinate regulon motif in which one kinase phosphorylates both a transcription factor and the target bound by the transcription factor. (5) Kinate feedback loop I motif in which a kinase phosphorylates a protein that interacts with a transcription factor that regulates the expression of that kinase. (6) Kinate feedback loop II in which a kinase phosphorylates a transcription factor whose target physically interacts with the kinase. (7) Heterosubstrate regulation motif with an interacting kinase and transcription factor regulating one target together. Motifs (1) to (5) were found to be enriched in the yeast integrated network.

sented (Ptacek et al. 2005). These composite motifs involve at least one kinase–substrate interaction pair (referred to as "kinates") and one other type of interaction (protein–protein interaction or transcription factor binding). Thus, network integration combines various data sources together and therefore can assist in uncovering proteins that are important in multiple types of interactions and provide a more comprehensive view on their cellular functions. Moreover, this network can be combined with other networks such as biochemical and gene interaction data to reveal a more comprehensive view of regulation in yeast.

In addition to mapping gene roles in a multirelationship network, integration of a variety of relevant genomic data can directly help to predict gene functions and functional relationships such as protein–protein interactions (Jansen et al. 2003; Troyanskaya et al. 2003). Compared with simple combinations of nonweighted interactions, a probabilistic approach integrating confidence-weighted data sources may be superior in modeling real biological data, considering its complicated and heterogeneous nature. For example, certain interaction data sets may be less error-prone and more reliable than others, and moreover, depending on the purposes of the data integration, certain data sources may be more informative and/or relevant than others. In order to overcome the challenge of data source heterogeneity, several early studies developed Bayesian network approaches to incorporate various data sources such as protein–protein interactions, gene expression profiles, and protein localization data for the purpose of predicting protein functions (Troyanskaya et al. 2003) and protein–protein interactions (Jansen et al. 2003). In both studies, the statistical reliabilities of different related genomic data were calculated when comparing with "gold-standard" samples consisting of known positives and negatives, and then extrapolated proteome-wide for novel predictions. Later studies also applied probabilistic models for the discovery of unknown components in pathway-specific protein complexes (Myers et al. 2005). Overall such probabilistic models have proven to be valuable in integrating heterogeneous genomic data and demonstrated a substantial improvement in prediction accuracy.

Network dynamics

Biological networks exhibit complex dynamic behavior, thereby enabling cells to react to various conditions or cell states such as cell cycle progression. Unfortunately, most large-scale data sets do not contain this information; static interactions are often identified from cells exposed to a single condition or at a single time point, often under nonnative conditions (e.g., two-hybrid). Only recently have approaches emerged that attempt to analyze the dynamics of complex biological networks. More interaction data sets have been collected in specific cellular conditions, and more importantly, integration with gene expression profile under various conditions has proven to be very helpful in network dynamics studies.

In protein–protein interaction networks, proteins may vary their partners according to time and location. By integrating gene expression data with a high-quality yeast protein–protein interaction data set, Han et al. (2004) studied the network dynamics in protein–protein interaction networks and revealed two types of hubs: "party hubs" and "date hubs." Party hubs interact with all their partners simultaneously—that is, at the same time and spatial locations—and are more likely to function within the same cellular processes. Date hubs, on the other hand, vary their connections to other proteins at different times and locations and therefore link various biological processes. When considering the modular designs of networks, in silico deletions of these hubs implied that party hubs are more likely to be the module organizers and date hubs to be the module connectors.

The dynamics of the transcriptional network in yeast has been examined on a genomic scale by integrating gene expression data for five cellular conditions with known transcriptional regulatory relationships (Luscombe et al. 2004). A trace-back algorithm was applied to uncover subnetworks that are active under specific conditions. Luscombe et al. (2004) found that these subnetworks exhibit vastly different topologies on both a local and a global level and uncovered two separate groups of cellular states. In so-called exogenous states (e.g., stress response), the network has a shorter diameter and large hubs that should allow cells to respond quickly to external conditions. In endogenous states (e.g., cell cycle), loops and highly intricate connections are more prevalent, indicating a multistage internal program. Different sets of transcription factors become key regulatory hubs at different times, portraying a network that shifts its weight between different foci to bring about distinct cellular states.

Network evolution

Various models have been proposed to explain the development of the scale-free topology of the protein–protein interaction network during evolution. A "network growth" model assumes that nodes with fixed degree are constantly added to the network. The probability that a newly added node interacts with an existing node is proportional to its degree, which leads to a so-called preferential attachment model in which rich nodes get richer during evolution and finally form a scale-free network (Barabasi and Albert 1999). In biological networks, the addition of nodes is due to gene duplication. This model was supported by the fact that older nodes (proteins having orthologs in evolutionarily distant organisms) tend to have higher degrees than newer nodes (proteins having orthologs in evolutionarily close organisms) (Eisenberg and Levanon 2003). However, examination of duplicated genes shows that they will quickly diverge in their connections and thereby rapidly specialize their interacting partners. Thus, most paralogs do not share the same partners. These contradictions lead to a "link dynamics" model that explains the network evolution through interaction loss and preferential interaction gain (Wagner 2001, 2003).

In general, core components of a network tend to be conserved, whereas components at the periphery or false interactions are not. In transcription factor-binding networks, this concept has been applied to identify functional regulatory elements that are conserved in several yeast species (Cliften et al. 2003; Kellis et al. 2003). Studies also have shown that interactions in one organism can be mapped to another organism if both partners are highly conserved (Yu et al. 2004). Conserved protein–protein interaction pairs are termed as interologs (Walhout et al. 2000), whereas conserved transcriptional binding interaction pairs are termed as regulogs (Yu et al. 2004). New interactions in novel organisms can then be discovered through mapping interologs or regulogs.

Although conservation of network components and connections is extremely valuable for mapping conserved interactions and common features among organisms, it is likely that many regulatory interactions are not conserved. Mapping of Ste12- and Tec1-binding sites in closely related yeast *S. cerevisiae*, *Saccharomyces mikatae*, and *Saccharomyces bayanus* reveals extensive divergence in binding sites in these different yeasts (A. Borneman and M. Snyder, unpubl.). These changes likely lead to species diversity and the ability of organisms to occupy distinct ecological niches.

Networks and human disease

Disruption of network architecture is expected to relate to human diseases. One advantage of scale-free networks is robustness—loss of individual components usually maintains overall network topology. This organization in general should make a system relatively immune to defects that target individual components. Loss of multiple components as occurs in many forms of cancer is required for network breakdown. This architecture may explain, in part, the observation that multiple mutations are often required for the onset of cancer (Knudson 1971). Nonetheless, some regions of networks should be more vulnerable to disruptions than others. Loss-of-activity mutations that affect hubs are more likely to cause a defect than those that affect the periphery. In addition, we expect that activating mutations in master regulators (target hubs) are more likely to cause apparent defects in cellular and developmental processes than those that occur elsewhere in the network. Thus, identifying such hubs may suggest possible drug targets for reconstructing the network and therefore curing disease.

Identification of functional roles of unknown pathogenic genes can also shed light on discovering disease pathogenic mechanisms. Proteins connected tightly in biological networks often work in similar processes. Hence, functional annotations of interacting partners may indicate potential roles of unannotated disease-related genes and help us to better understand the pathological mechanisms of the disease. Lim et al. (2006) constructed an interactome map focusing on proteins responsive to human inherited ataxias and purkinje cell degeneration with a yeast two-hybrid screen. The majority of known ataxia-causing proteins were connected with short paths, suggesting that other components in the network might contain candidates responsive to other related inherited ataxias with unknown causative genes. Furthermore, the hubs of this network had crucial roles for disease development in animal models, implying a relationship between the disease and the biological processes in which they are involved: RNA binding or splicing. Such systematic studies can easily be applied to other diseases and organisms and will help to identify crucial components for the disease pathology.

Challenges and future directions

Current studies often draw conclusions for complete interaction networks from limited and possibly erroneous samples of the actual biological networks. The yeast two-hybrid protein–protein interaction network, for example, shows a typical scale-free structure and is often used to infer that the complete yeast protein–protein interaction network has the same properties. Recent studies, however, indicate that the scale-free topology might be generated through the experimental designs, which resulted in a biased sample of the complete data set (Han et al. 2005). Further analyses by Friedel and Zimmer tested the clustering coefficient among several possible topologies, and suggested that the scale-free topology was still most likely to be the organization of the complete protein–protein network, although possibilities of other topologies could still not be completely excluded (Friedel and Zimmer 2006). Moreover, when investigating a more complete protein–protein interaction network, Batada et al. claimed that party hubs and date hubs, which originated from a smaller interaction data set (Han et al. 2004), could no longer be differentiated from each other (Batada et al. 2006). Such debates suggest that our current view of biological networks may still be biased, and more interaction data are needed to better represent the real networks.

The ability to collect large data sets has only just begun. In the future, it should be possible to construct more complete and accurate networks, for example, by identifying the targets of all relevant transcription factors and determining the protein–protein interaction networks of humans and many other organisms. Considerable effort will be required to find the post-translational modifications and factors that control the activity and stability of each protein in different cell states. Finally, large-scale efforts to map post-transcriptional regulation such as miRNAs need to be initiated. All of these interactions and modifications must be accomplished in the appropriate cell state and the dynamics of the process followed. The integration of all interactions/modifications along with their dynamics will reveal the ultimate description of how complex biological processes such as cell proliferation and development occur and can be controlled.

Acknowledgments

We thank L. Nunes Amaral, C. Boone, and A. Barabasi for the permission to use several of their published figures in Figure 1, and T. Gianoulis, A. Sboner, and J. Wu for comments on the manuscripts. This research was supported by grants from NIH.

References

Adamcsek, B., Palla, G., Farkas, I.J., Derenyi, I., and Vicsek, T. 2006. CFinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics* **22:** 1021–1023.

Ballif, B.A., Villen, J., Beausoleil, S.A., Schwartz, D., and Gygi, S.P. 2004. Phosphoproteomic analysis of the developing mouse brain. *Mol. Cell. Proteomics* **3:** 1093–1101.

Barabasi, A.L. and Albert, R. 1999. Emergence of scaling in random networks. *Science* **286:** 509–512.

Barabasi, A.L. and Bonabeau, E. 2003. Scale-free networks. *Sci. Am.* **288:** 60–69.

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

Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R.A., et al. 2003. Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.* **21:** 1337–1342.

Batada, N.N., Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.J., Hurst, L.D., and Tyers, M. 2006. Stratus not altocumulus: A new view of the yeast protein interaction network. *PLoS Biol.* **4:** e317.

Bender, A. and Pringle, J.R. 1991. Use of a screen for synthetic lethal and multicopy suppressee mutants to identify two new genes involved in morphogenesis in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **11:** 1295–1305.

Blais, A., Tsikitis, M., Acosta-Alvear, D., Sharan, R., Kluger, Y., and Dynlacht, B.D. 2005. An initial blueprint for myogenic differentiation. *Genes & Dev.* **19:** 553–569.

Borneman, A.R., Leigh-Bell, J.A., Yu, H., Bertone, P., Gerstein, M., and Snyder, M. 2006. Target hub proteins serve as master regulators of development in yeast. *Genes & Dev.* **20:** 435–448.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122:** 947–956.

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

Brill, L.M., Salomon, A.R., Ficarro, S.B., Mukherji, M., Stettler-Gill, M., and Peters, E.C. 2004. Robust phosphoproteomic profiling of tyrosine phosphorylation sites from human T cells using immobilized metal affinity chromatography and tandem mass spectrometry. *Anal. Chem.* **76:** 2763–2772.

Burgard, A.P., Pharkya, P., and Maranas, C.D. 2003. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84:** 647–657.

Carroll, J.S., Liu, X.S., Brodsky, A.S., Li, W., Meyer, C.A., Szary, A.J., Eeckhoute, J., Shao, W., Hestermann, E.V., Geistlinger, T.R., et al. 2005. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* **122:** 33–43.

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116:** 499–509.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B.A., and Johnston, M.

2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* **301:** 71–76.

Cohen, P. 2000. The regulation of protein function by multisite phosphorylation—A 25 year update. *Trends Biochem. Sci.* **25:** 596–601.

Costigan, C., Gehrung, S., and Snyder, M. 1992. A synthetic lethal screen identifies SLK1, a novel protein kinase homolog implicated in yeast cell morphogenesis and cell growth. *Mol. Cell. Biol.* **12:** 1162–1178.

Covert, M.W., Leung, T.H., Gaston, J.E., and Baltimore, D. 2005. Achieving stability of lipopolysaccharide-induced NF-κB activation. *Science* **309:** 1854–1857.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., et al. 2002. A genomic regulatory network for development. *Science* **295:** 1669–1678.

Davierwala, A.P., Haynes, J., Li, Z., Brost, R.L., Robinson, M.D., Yu, L., Mnaimneh, S., Ding, H., Zhu, H., Chen, Y., et al. 2005. The synthetic genetic interaction spectrum of essential genes. *Nat. Genet.* **37:** 1147–1152.

de Nooy, W., Mrvar, A., and Batagelj, V. 2005. *Exploratory social network analysis with Pajek*. Cambridge University Press, New York.

Dephoure, N., Howson, R.W., Blethrow, J.D., Shokat, K.M., and O'Shea, E.K. 2005. Combining chemical genetics and proteomics to identify protein kinase substrates. *Proc. Natl. Acad. Sci.* **102:** 17940–17945.

Duarte, N.C., Herrgard, M.J., and Palsson, B.O. 2004. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* **14:** 1298–1309.

Eisenberg, E. and Levanon, E.Y. 2003. Preferential attachment in the protein network evolution. *Phys. Rev. Lett.* **91:** 138701.

Famili, I., Forster, J., Nielsen, J., and Palsson, B.O. 2003. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc. Natl. Acad. Sci.* **100:** 13134–13139.

Ficarro, S.B., McCleland, M.L., Stukenberg, P.T., Burke, D.J., Ross, M.M., Shabanowitz, J., Hunt, D.F., and White, F.M. 2002. Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **20:** 301–305.

Forster, J., Famili, I., Palsson, B.O., and Nielsen, J. 2003. Large-scale evaluation of in silico gene deletions in *Saccharomyces cerevisiae*. *OMICS* **7:** 193–202.

Friedel, C.C. and Zimmer, R. 2006. Toward the complete interactome. *Nat. Biotechnol.* **24:** 614–615; author reply 615.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11:** 4241–4257.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440:** 631–636.

Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418:** 387–391.

Giaever, G., Flaherty, P., Kumm, J., Proctor, M., Nislow, C., Jaramillo, D.F., Chu, A.M., Jordan, M.I., Arkin, A.P., and Davis, R.W. 2004. Chemogenomic profiling: Identifying the functional interactions of small molecules in yeast. *Proc. Natl. Acad. Sci.* **101:** 793–798.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., et al. 2003. A protein interaction map of *Drosophila melanogaster*. *Science* **302:** 1727–1736.

Girvan, M. and Newman, M.E. 2002. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99:** 7821–7826.

Goll, J. and Uetz, P. 2006. The elusive yeast interactome. *Genome Biol.* **7:** 223.

Gruhler, A., Olsen, J.V., Mohammed, S., Mortensen, P., Faergeman, N.J., Mann, M., and Jensen, O.N. 2005. Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell. Proteomics* **4:** 310–327.

Guelzim, N., Bottani, S., Bourgine, P., and Kepes, F. 2002. Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.* **31:** 60–63.

Guimera, R. and Nunes Amaral, L.A. 2005. Functional cartography of complex metabolic networks. *Nature* **433:** 895–900.

Gunsalus, K.C., Ge, H., Schetter, A.J., Goldberg, D.S., Han, J.D., Hao, T., Berriz, G.F., Bertin, N., Huang, J., Chuang, L.S., et al. 2005. Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* **436:** 861–865.

Hall, D.A., Zhu, H., Zhu, X., Royce, T., Gerstein, M., and Snyder, M. 2004. Regulation of gene expression by a metabolic enzyme. *Science* **306:** 482–484.

Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P., et al. 2004. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430:** 88–93.

Han, J.D., Dupuy, D., Bertin, N., Cusick, M.E., and Vidal, M. 2005. Effect of sampling on topology predictions of protein–protein interaction networks. *Nat. Biotechnol.* **23:** 839–844.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431:** 99–104.

Hartman, S.E., Bertone, P., Nath, A.K., Royce, T.E., Gerstein, M., Weissman, S., and Snyder, M. 2005. Global changes in STAT target selection and transcription regulation upon interferon treatments. *Genes & Dev.* **19:** 2953–2968.

Horak, C.E. and Snyder, M. 2002. ChIP–chip: A genomic approach for identifying transcription factor binding sites. *Methods Enzymol.* **350:** 469–483.

Horak, C.E., Luscombe, N.M., Qian, J., Bertone, P., Piccirrillo, S., Gerstein, M., and Snyder, M. 2002. Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cere-*
visiae. *Genes & Dev.* **16:** 3017–3033.

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

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. 2003. A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* **302:** 449–453.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabasi, A.L. 2000. The large-scale organization of metabolic networks. *Nature* **407:** 651–654.

Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. 2001. Lethality and centrality in protein networks. *Nature* **411:** 41–42.

Kanehisa, M. and Goto, S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28:** 27–30.

Karin, M. 2006. Nuclear factor-κB in cancer development and progression. *Nature* **441:** 431–436.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.

Kim, J., Bhinge, A.A., Morgan, X.C., and Iyer, V.R. 2005. Mapping DNA–protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nat. Methods* **2:** 47–53.

Knudson Jr., A.G. 1971. Mutation and cancer: Statistical study of retinoblastoma. *Proc. Natl. Acad. Sci.* **68:** 820–823.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., et al. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440:** 637–643.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298:** 799–804.

Li, Z., Van Calcar, S., Qu, C., Cavenee, W.K., Zhang, M.Q., and Ren, B. 2003. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl. Acad. Sci.* **100:** 8164–8169.

Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., et al. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science* **303:** 540–543.

Lim, J., Hao, T., Shaw, C., Patel, A.J., Szabo, G., Rual, J.F., Fisk, C.J., Li, N., Smolyar, A., Hill, D.E., et al. 2006. A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell* **125:** 801–814.

Loog, M. and Morgan, D.O. 2005. Cyclin specificity in the phosphorylation of cyclin-dependent kinase substrates. *Nature* **434:** 104–108.

Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., and Gerstein, M. 2004. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431:** 308–312.

Ma, J., Wang, Q., Fei, T., Han, J.D., and Chen, Y.G. 2006. MCP-1 mediates TGF-β-induced angiogenesis by stimulating vascular smooth muscle cell migration. *Blood* **109:** 987–994.

Manning, G., Plowman, G.D., Hunter, T., and Sudarsanam, S. 2002a. Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci.* **27:** 514–520.

Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. 2002b. The protein kinase complement of the human genome. *Science* **298:** 1912–1934.

Martone, R., Euskirchen, G., Bertone, P., Hartman, S., Royce, T.E., Luscombe, N.M., Rinn, J.L., Nelson, F.K., Miller, P., Gerstein, M., et al. 2003. Distribution of NF-κB-binding sites across human chromosome 22. *Proc. Natl. Acad. Sci.* **100:** 12247–12252.

Measday, V., Baetz, K., Guzzo, J., Yuen, K., Kwok, T., Sheikh, B., Ding, H., Ueta, R., Hoac, T., Cheng, B., et al. 2005. Systematic yeast synthetic lethal and synthetic dosage lethal screens identify genes required for chromosome segregation. *Proc. Natl. Acad. Sci.* **102:** 13956–13961.

Milgram, S. 1967. The small-world problem. *Psychol. Today* **1:** 61–67.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. 2002. Network motifs: Simple building blocks of complex networks. *Science* **298:** 824–827.

Mishra, L., Derynck, R., and Mishra, B. 2005. Transforming growth factor-β signaling in stem cells and cancer. *Science* **310:** 68–71.

Myers, C.L., Chen, X., and Troyanskaya, O.G. 2005. Visualization-based discovery and analysis of genomic aberrations in microarray data. *BMC Bioinformatics* **6:** 146.

Newman, M.E. 2006. Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* **103:** 8577–8582.

Ohya, Y., Sese, J., Yukawa, M., Sano, F., Nakatani, Y., Saito, T.L., Saka, A., Fukuda, T., Ishihara, S., Oka, S., et al. 2005. High-dimensional and large-scale phenotyping of yeast mutants. *Proc. Natl. Acad. Sci.* **102:** 19015–19020.

Overbeek, R., Larsen, N., Pusch, G.D., D'Souza, M., Selkov Jr., E., Kyrpides, N., Fonstein, M., Maltsev, N., and Selkov, E. 2000. WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28:** 123–125.

Pan, X., Yuan, D.S., Xiang, D., Wang, X., Sookhai-Mahadeo, S., Bader, J.S., Hieter, P., Spencer, F., and Boeke, J.D. 2004. A robust toolkit for functional profiling of the yeast genome. *Mol. Cell* **16:** 487–496.

Papp, B., Pal, C., and Hurst, L.D. 2004. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429:** 661–664.

Parsons, A.B., Brost, R.L., Ding, H., Li, Z., Zhang, C., Sheikh, B., Brown, G.W., Kane, P.M., Hughes, T.R., and Boone, C. 2004. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nat. Biotechnol.* **22:** 62–69.

Pilpel, Y., Sudarsanam, P., and Church, G.M. 2001. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* **29:** 153–159.

Price, N.D., Reed, J.L., and Palsson, B.O. 2004. Genome-scale models of microbial cells: Evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2:** 886–897.

Ptacek, J. and Snyder, M. 2006. Charging it up: Global analysis of protein phosphorylation. *Trends Genet.* **22:** 545–554.

Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R., et al. 2005. Global analysis of protein phosphorylation in yeast. *Nature* **438:** 679–684.

Reed, J.L., Vo, T.D., Schilling, C.H., and Palsson, B.O. 2003. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* **4:** R54.

Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.J., Hon, G.C., Myers, C.L., Parsons, A., Friesen, H., Oughtred, R., Tong, A., et al. 2006. Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.* **5:** 11.

Rives, A.W. and Galitski, T. 2003. Modular organization of cellular networks. *Proc. Natl. Acad. Sci.* **100:** 1128–1133.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., et al. 2005. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437:** 1173–1178.

Schuldiner, M., Collins, S.R., Thompson, N.J., Denic, V., Bhamidipati, A., Punna, T., Ihmels, J., Andrews, B., Boone, C., Greenblatt, J.F., et al. 2005. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123:** 507–519.

Segre, D., Vitkup, D., and Church, G.M. 2002. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci.* **99:** 15112–15117.

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

Shen-Orr, S.S., Milo, R., Mangan, S., and Alon, U. 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31:** 64–68.

Sopko, R., Huang, D., Preston, N., Chua, G., Papp, B., Kafadar, K., Snyder, M., Oliver, S.G., Cyert, M., Hughes, T.R., et al. 2006. Mapping pathways and phenotypes by systematic gene overexpression. *Mol. Cell* **21:** 319–330.

Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. 2006. BioGRID: A general repository for interaction datasets. *Nucleic Acids Res.* **34:** D535–D539.

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

Tao, W.A., Wollscheid, B., O'Brien, R., Eng, J.K., Li, X.J., Bodenmiller, B., Watts, J.D., Hood, L., and Aebersold, R. 2005. Quantitative phosphoproteome analysis using a dendrimer conjugation chemistry and tandem mass spectrometry. *Nat. Methods* **2:** 591–598.

Tari, L., Baral, C., and Dasgupta, P. 2005. Understanding the global properties of functionally-related gene networks using the Gene Ontology. *Pac. Symp. Biocomput.* **2005:** 209–220.

Thiele, I., Price, N.D., Vo, T.D., and Palsson, B.O. 2005. Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet. *J. Biol. Chem.* **28:** 11683–11695.

Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H., et al. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294:** 2364–2368.

Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., et al. 2004. Global mapping of the yeast genetic interaction network. *Science* **303:** 808–813.

Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., and Botstein, D. 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). *Proc. Natl. Acad. Sci.* **100:** 8348–8353.

Tsankov, A.M., Brown, C.R., Yu, M.C., Win, M.Z., Silver, P.A., and Casolari, J.M. 2006. Communication between levels of transcriptional control improves robustness and adaptivity. *Mol. Syst. Biol.* **2:** 65.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. 2000. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403:** 623–627.

van Noort, V., Snel, B., and Huynen, M.A. 2004. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.* **5:** 280–284.

Vo, T.D., Greenberg, H.J., and Palsson, B.O. 2004. Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *J. Biol. Chem.* **279:** 39532–39540.

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

Wagner, A. 2003. How the global structure of protein interaction networks evolves. *Proc. Biol. Sci.* **270:** 457–466.

Wagner, A. and Fell, D.A. 2001. The small world inside large metabolic networks. *Proc. Biol. Sci.* **268:** 1803–1810.

Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N., and Vidal, M. 2000. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287:** 116–122.

Wallis, R., Leung, K.Y., Pommer, A.J., Videler, H., Moore, G.R., James, R., and Kleanthous, C. 1995. Protein–protein interactions in colicin E9 DNase-immunity protein complexes. 2. Cognate and noncognate interactions that span the millimolar to femtomolar affinity range. *Biochemistry* **34:** 13751–13759.

Watts, D.J. and Strogatz, S.H. 1998. Collective dynamics of 'small-world' networks. *Nature* **393:** 440–442.

Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z., et al. 2006. A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124:** 207–219.

Weintraub, H., Tapscott, S.J., Davis, R.L., Thayer, M.J., Adam, M.A., Lassar, A.B., and Miller, A.D. 1989. Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. *Proc. Natl. Acad. Sci.* **86:** 5434–5438.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285:** 901–906.

Yang, J., Mani, S.A., Donaher, J.L., Ramaswamy, S., Itzykson, R.A., Come, C., Savagner, P., Gitelman, I., Richardson, A., and Weinberg, R.A. 2004. Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell* **117:** 927–939.

Yu, H. and Gerstein, M. 2006. Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl. Acad. Sci.* **103:** 14724–14731.

Yu, H., Luscombe, N.M., Qian, J., and Gerstein, M. 2003. Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet.* **19:** 422–427.

Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. 2004. Annotation transfer between genomes: Protein–protein interologs and protein–DNA regulogs. *Genome Res.* **14:** 1107–1118.

Zhu, H., Klemic, J.F., Chang, S., Bertone, P., Casamayor, A., Klemic, K.G., Smith, D., Gerstein, M., Reed, M.A., and Snyder, M. 2000. Analysis of yeast protein kinases using protein chips. *Nat. Genet.* **26:** 283–289.

Getting connected: analysis and principles of biological networks

Xiaowei Zhu, Mark Gerstein and Michael Snyder

Genes Dev. 2007 21: 1010-1024
Access the most recent version at doi:10.1101/gad.1528707