Analyzing Cellular Biochemistry in Terms of Molecular Networks

Yu Xia[1,5], Haiyuan Yu[1,5], Ronald Jansen[2,5], Michael Seringhaus[1], Sarah Baxter[1], Dov Greenbaum[1], Hongyu Zhao[3], Mark Gerstein[1,4,6]

[1] Department of Molecular Biophysics and Biochemistry, P.O. Box 208114, Yale University, New Haven, CT 06520; email: yuxia@csb.yale.edu, haiyuan.yu@yale.edu, michael.seringhaus@yale.edu, sarah.baxter@yale.edu, dov.greenbaum@yale.edu, mark.gerstein@yale.edu

[2] Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 307 East 63rd Street, 2nd floor, New York, NY 10021; email: jansenr@mskcc.org

[3] Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520; email: hongyu.zhao@yale.edu

[4] Department of Computer Science, Yale University, New Haven, CT 06520

[5] These authors contributed equally to this review.

[6] Corresponding author. Phone: 203-432-6105; efax: 360-838-7861

Running Title: Biomolecular network analysis

Key Words: genome-wide high-throughput experiments, protein-protein interaction networks, regulatory networks, integration and prediction, network topology

Abstract

One way to understand cells and circumscribe the function of proteins is through molecular networks. These take a variety of forms including protein-protein interaction networks, regulatory networks linking transcription factors and targets, and metabolic networks of reactions. We first survey experimental techniques for mapping networks (e.g. the yeast two-hybrid screens). We then turn our attention to computational approaches for predicting networks from individual protein features, such as correlating gene expression levels or analyzing sequence co-evolution. All the experimental techniques and individual predictions suffer from noise and systematic biases. These can be overcome to some degree through statistical integration of different experimental datasets and predictive features (e.g. within a Bayesian formalism). Next, we discuss approaches for characterizing the topology of networks, such as finding hubs and analyzing sub-networks in terms of common motifs. Finally, we close with perspectives on how network analysis represents a preliminary step towards systems-biology modeling of cells.

Contents

Introduction

An important idea emerging in post-genomic biology is that the cell can be viewed as a complex network of interacting proteins, nucleic acids, and other biomolecules (1, 2). Similarly complex networks are also used to describe the structure of a number of wide-ranging systems including the Internet, power grids, the ecological food web, and scientific collaborations. Despite the seemingly vast differences among these systems, they all share common features in terms of network topology (3-11). Therefore, networks may provide a framework for describing biology in a universal language understandable to a broad audience.

Many fundamental cellular processes involve interactions among proteins and other biomolecules. Comprehensively identifying these interactions is an important step towards systematically defining protein function (2, 12), as clues about the function of an unknown protein can be obtained by investigating its interaction with other proteins of known function.

A biomolecular interaction network can be viewed as a collection of nodes (representing biomolecules), some of which are connected by links (representing interactions). There are many classes of molecular networks in a cell, each with different types of nodes and links. We list a representative subset below:

(1) Protein-protein physical interaction networks. Here nodes represent proteins, and links represent direct physical contacts between proteins. In addition to direct interaction, two proteins can interact indirectly through other proteins when they belong to the same complex.

(2) Protein-protein genetic interaction networks. In general, two genes are said to interact genetically if a mutation in one gene either suppresses or enhances the phenotype of a mutation in its partner gene (13). Some researchers restrict the term 'genetic interaction' to a pair of so-called synthetic lethal genes, meaning that cell death occurs when this pair of genes is deleted simultaneously, though neither deletion alone is lethal. Synthetic lethal relationships may exist between functionally redundant genes, and therefore can be used to determine the function of unknown genes.

(3) Expression networks. Large-scale microarray experiments probing mRNA expression levels yield vast quantities of data useful for constructing expression networks. In an expression network, genes that are co-expressed are considered connected (14-16). Genes linked in an expression network are not necessarily co-regulated, as unrelated genes can sometimes show correlated expression simply by coincidence. The structure of an expression network can vary greatly across different experiments, and even within the same experiment, networks produced by different clustering algorithms are often distinct.

(4) Regulatory networks. Protein-DNA interactions are an important and common class of interactions. Most DNA-binding proteins are transcription factors that regulate the expression of target genes. A regulatory network consists of transcription factors and their targets, with a specific directionality to the connection between a transcription factor and its target (17, 18). Transcription factors can either up- or down-regulate expression of their target genes.

(5) Metabolic networks. These networks describe the biochemical reactions within different metabolic pathways in the cell. Nodes represent metabolic substrates and products, while links represent metabolic reactions (19).

(6) Signaling networks. These networks represent signal transduction pathways through protein-protein and protein-small molecule interactions (20). Nodes represent proteins or small molecules (21), while links represent signal transduction events.

These biomolecular networks are the focus of this review. We will first discuss how networks can be reconstructed, from a combined experimental and computational perspective. Later, we will discuss how networks can be analyzed to yield biological insight.

Survey of Experimental Techniques

There are several experimental methods for uncovering protein-protein and protein-DNA interactions in biological systems on a large scale. Here we review the most current, powerful and common of these.

Yeast two-hybrid screens

The yeast two-hybrid (Y2H) system (22) has been widely used in protein-protein physical interaction assays. The system uses putative interacting proteins to broker an in vivo reconstitution of the DNA binding domain (DB) and activation domain (AD) of the yeast transcription factor Gal4p. Hybrid proteins are created by fusing the two proteins or domains of interest (generally called 'bait' and 'prey') to the DB and AD regions of Gal4p, respectively. These two hybrid proteins are introduced into yeast, and if transcription of Gal4p-regulated reporter genes is observed, the two proteins of interest are deemed to have formed an interaction – thereby bringing the DB and AD domains of Gal4p together and reconstituting the functional transcriptional activator.

Unlike most biochemical analyses of protein-protein interaction such as co-immunoprecipitation, crosslinking and chromatographic co-fractionation (22), the two-hybrid system does not demand any protein purification, isolation or manipulation – the proteins to be tested are expressed by the yeast cells, and a result is easily seen by in vivo reporter gene assays. The two-hybrid technique is therefore applicable to nearly any pair of putative interacting proteins.

There exist three main approaches for large-scale two-hybrid studies (23). The matrix approach (one versus one) systematically tests pairs of proteins for an interaction phenotype; a positive result can indicate that these particular proteins interact. Array experiments (one versus all) examine the interactions of a single DB fusion protein against a pool of AD fusions; depending on the size of the AD pool, whole-proteome coverage can be achieved against the single DB fusion. Pooling studies (all versus all) involve yeast strains expressing different DB fusions being mass-mated with strains expressing AD hybrids; with such experiments, it is conceptually possible to test every protein in the organism against every other protein.

The first large-scale, systematic search for yeast protein-protein interactions was conducted in 1997 (24). In the year 2000, Uetz et al. published the results (25) of two different large-scale screens on all full-length predicted ORFs. The first approach involved a protein array of roughly 6,000 yeast transformants, each transformant expressing one yeast ORF-AD fusion. 192 yeast proteins were screened against this array. In the second screen, a library of cells was generated and pooled, such that all 6000 AD fusions were present. Nearly all predicted yeast proteins, expressed as DB fusions, were screened against this library and positives were identified by sequencing. Later, Ito et al. (26, 27) reported another systematic identification of yeast interacting protein pairs with a whole-genome level two-hybrid screen. Their comprehensive approach involved cloning all yeast ORFs as both bait and prey, and testing about 4×10^6 mating reactions (roughly 10% of all possible combinations). The researchers pooled constructs such that each pool expressed either 96 DB fusions or 96 AD fusions, and screened all possible combinations of these pools. False positives were controlled by requiring a positive interaction result on at least three independent occasions.

Overlap between the Ito and Uetz screens was low, indicating that both studies, while extensive, sampled only a small subset of yeast protein interactions (28, 29).

It is also possible to use large-scale two-hybrid screens to explore interactions relevant to a specific pathway or biological process. Drees et al. (30) screened 68 Gal4p DB fusions of yeast proteins associated with cell polarity against an array of yeast transformants expressing roughly 90% of predicted yeast ORFs. In addition, large-scale two-hybrid screens are not confined to yeast proteins: Working with proteins involved in vulval development, Walhout et al. (31) conducted large-scale interaction mapping in the nematode *C. elegans*, while Boulton et al. (32) combined protein-protein interaction mapping with phenotypic analysis in *C. elegans* to explore DNA damage response interaction networks.

Comprehensive in vivo pull-down techniques

In vivo pull-down describes a class of techniques that use either a native or modified bait protein to identify and precipitate interacting partners. Most experiments concerned with studying protein-protein interactions through pull-down techniques consist of three parts: bait presentation, affinity purification, and analysis of the recovered complex (33).

Compared with the two-hybrid system, the main advantages to in vivo pull-down techniques are the relative ease of analyzing complete complexes, and the use of native, processed and post-translationally modified protein as a reagent to target potential interactors in its natural environment and at normal abundance levels (34). If a suitable antibody exists to the native protein, endogenous protein can be used. However, since insufficient antibodies exist to attack most unmodified proteins with the requisite specificity and affinity, more general techniques such as tagging are typically used for large-scale assays. Generic tagging involves the addition of a sequence onto the gene of interest, encoding a tag recognized by a convenient antibody. HA-tagging is a common epitope-tagging approach that has been used successfully (35). A recent tagging strategy facilitating recovery of highly pure protein preparations is the tandem affinity purification (TAP) system, consisting of a calmodulin-binding domain and the protein-A Ig-binding domain separated by the TEV protease target sequence (36). Bait protein is recovered with an immunoglobulin-bound solid support, and after washing, released from this support by protease cleavage. Following this initial purification, the recovered sample is passed over a calmodulin column, pending elution with EGTA or other Ca^{2+} chelators. This two-stage purification ensures low background noise and correspondingly high sample purity, but risks losing weak interacting partners or complex components due to the harsh purification procedure.

After the bait/interactor complex is purified, components of this complex can be identified by mass spectrometry (MS). The many recent advances in MS technology (MALDI-TOF, ESI, tandem MS/MS and others) have enabled accuracy to increase while permitting ionization (and therefore, characterization) of larger biomolecules. In general, MS proteomics experiments comprise five stages (33): the first three involve purification (typically culminating in 1D gel electrophoresis), tryptic digestion to generate short peptides, and HPLC separation of the tryptic digest; the final two steps are the tandem mass spectrometry assays. The high accuracy of MS spectra, combined with knowledge of the genomic sequence of the organism in question, permits rapid and accurate identification of the proteins involved in the recovered complex.

Two large-scale projects dealing with the yeast ‘interactome’ were recently completed by Gavin et al. (37) and Ho et al. (38). Gavin et al. purified 589 bait proteins from a library of 1,548 tagged strains, and from these identified 1,440 distinct participant proteins in 232 complexes. Ho et al. purified 725 bait proteins from which 1,578 interacting proteins were identified. Both studies used extensive literature comparisons to characterize the complexes they found, and both reported significant participation by previously unknown or un-annotated genes (35, 37, 38).

Protein chips

The application of microarray technology to proteomics yielded the protein chip, an advanced in vitro technique for protein functional assays on a large scale. Protein chip technology is directly applicable to protein interaction networks, since the large number of immobilized proteins can be probed with labeled substrate in a single experiment.

Arenkov et al. (39) reported the creation of a polyacrylamide-based protein microchip, containing 0.2nl spots of gel substrate in which proteins were immobilized; this platform allowed electrophoresis to be used to enhance mixing of substrate. MacBeath and Schreiber’s protein chip (40) uses microarray technology and robotics to spot nanoliter volumes of protein onto aldehyde-coated glass slides. The abundance of lysine residues in most proteins, combined with a reactive N-terminal amine, permit proteins to become covalently linked to the slide surface in a number of possible orientations.

Shortly thereafter, Zhu et al. (41) described another type of protein chip, also mounted on a glass slide but comprising a system of 300nl silicone elastomer microwells for physical separation of samples during processing. As with the MacBeath protein arrays, the target protein was covalently linked to the chip, though here the chemical crosslinker GPTS was used. The following year, the same group announced the creation of the first whole-proteome chip (42), a glass slide similar to MacBeath & Schreiber's initial protein chip, but containing over 80% of known yeast ORF gene products attached to nickel-coated slides via 6-His tags. Zhu et al. demonstrated the effectiveness of the proteome chip for protein-protein interaction studies by probing with biotinylated calmodulin in the presence of calcium; calmodulin binding partners were visualized by probing with Cy3-labeled streptavidin. This demonstrated that biotinylated constructs of virtually any protein could be used to probe the proteome chip, thereby visualizing protein-protein interactions. In addition to uncovering several known calmodulin interactors, the researchers found a significant number of novel interaction partners.

Structure determination of biomolecular complexes

An atomic view of physical interactions between biomolecules can be achieved by solving three-dimensional structures of biomolecular complexes, most often accomplished with X-ray crystallography and NMR spectroscopy. In particular, X-ray crystallography is able to produce the most spatially accurate description of biomolecular interactions. Though technically challenging, significant advances have been made in recent years and X-ray crystallography can now be applied to complexes as large as several megadaltons. For a detailed review of various structural determination methods for biomolecular complexes, see (43).

Comparing in vivo and in vitro techniques

The caveats associated with genomic-level data sets stem largely from the experimental techniques used to generate them, and in particular, care should be taken to note whether interaction results originate from in vivo or in vitro studies. A major advantage of in vivo pull-down techniques is that near-native interactions can be probed, provided that tagging and bait expression do not interfere with the replication of endogenous levels of protein activity – proper folding, post-translational modification and the accessibility of biologically relevant binding partners are generally assumed. Still, the abundance of proteins and solutes in the cell means contaminants often co-purify, potentially yielding misleading results. In vivo experiments generally offer little or no direct control over reaction conditions (especially in the case of large-scale studies) while in vitro assays permit exquisite control over ion concentration, temperature, and other factors. The assumption that in vivo assay conditions are biologically meaningful is sometimes inapplicable to interactions probed by the yeast two-hybrid technique, which must occur in the yeast nucleus. In vitro and two-hybrid approaches are unlikely to recover only significant binding partners, and risk false-positive results if interacting proteins localize to different cell compartments, express at different times in the cell cycle, or are otherwise inaccessible to binding under normal conditions. Still, in vitro techniques such as protein chip assays are convenient to record, since results can be visualized for individual putative interacting partners; compare this to the grouped results of many pooling techniques where over- or under-representation in bait/prey pools can influence results, and positives must be identified by sequencing or barcode analysis.

Methods for determining protein-protein genetic interactions

Synthetic lethal screens are used to identify genetic interactions between proteins. Small-scale synthetic lethal screens have been used to identify genes involved in many cellular processes (44-46). Recently, Tong et al. introduced a systematic method to construct large-scale double mutant arrays, termed synthetic genetic array (SGA) analysis, in which double mutants were created by crossing a query mutation to an array of roughly 4700 deletion mutants, and non-viable double-mutant meiotic progeny were identified. SGA analysis has generated a genetic network of 291 interactions among 204 genes (13).

Methods for determining protein-DNA interactions

Protein-DNA interactions can be determined by three core methods:

(1) Gel shift. Compared with protein molecules, DNA molecules are much smaller and therefore have much higher mobility in a polyacrylamide gel. Under favorable conditions, unbound DNA can be distinguished from DNA associated with proteins based on their relative mobility (47, 48). Recently, several enhanced methods, such as capillary electrophoretic mobility shift assay (CEMSA) (49), have been proposed to improve the performance of this approach.

(2) DNA footprinting. A 5' end-labeled, double-stranded target DNA segment is partially degraded by DNase both in the presence and absence of the putative binding protein. Degraded fragments are visualized by electrophoresis and autoradiography. The binding site on the DNA will be

protected by the binding protein from DNAase degradation (48, 50). Compared with gel shift methods, DNA footprinting not only confirms the interaction between the DNA and the binding protein, but can also elucidate the specific binding site of the protein.

(3) In vivo cross-linking and immunoprecipitation. The binding protein is first covalently linked to DNA in situ using any of a variety of common cross-linking reagents; among these, UV and formaldehyde have been widely used. After crosslinking, chromosomal DNA is sheared; the protein is precipitated using a specific antibody, and bound DNA fragments co-precipitate. Reversal of crosslinks releases bound DNA, so fragments can be identified by PCR and electrophoresis (51, 52). This method is also called chromatin immunoprecipitation (ChIP).

Recently, with the advent of microarray technology, novel methods have been introduced to rapidly determine the binding sites of transcription factors on a genome-wide scale (17, 18, 53, 54).

(1) ChIP-chip (Chromatin-Immunoprecipitation and microarray/chip technique). This method combines the ChIP technique with DNA microarray technology. Thousands of DNA fragments purified by the ChIP method are identified simultaneously by microarray experiments (53). Using ChIP-chip, Lee et al. were able to create a yeast regulatory network consisting of 106 transcription factors and 2363 target genes (17).

(2) DamID (DNA Adenine Methyltransferase Identification). The use of cross-linking reagents can produce artifacts in ChIP-chip experiments. To overcome this problem, van Steensel and Henikoff introduced a new technique to map protein-DNA interactions, termed DamID (55, 56). The DNA

binding protein of interest is genetically fused with *Escherichia coli* DNA adenine methyltransferase (Dam). Dam methylates the N^6-position of adenine in the sequence GATC, which occurs on average every 200-300 base pairs in the fly genome. Upon in vivo binding of the protein to its target DNA sites, DNA around the target sites is preferentially methylated by the tethered Dam enzyme. Subsequently, genomic DNA is digested into small fragments by *DpnI*. DNA fragments without methylated GATCs are removed by *DpnII* digestion. The remaining methylated fragments are amplified by selective PCR and quantified by microarray analysis (54-56). Recently, Sun et al. successfully mapped protein-DNA interactions at high resolution along large segments of genomic DNA from *Drosophila melanogaster* using the DamID technique and genomic DNA tiling path microarrays (54).

Conceivably, data generated by these different methods can be used to cross-validate one another, thereby producing more comprehensive information. While each method yields only a subset of the total interactions present, a more complete yeast regulatory network consisting of 180 transcription factors and 3474 target genes has been produced through the synthesis of all available datasets (57).

Databases for biomolecular interactions

Many databases have been created to store the tremendous amount of data required for, and contained in these networks, some of which are summarized in Table 1 (58-67). Some databases are more comprehensive than others; for instance, MIPS contains not only protein-protein physical interaction data, but genetic interaction information as well (60).

Computational Approaches for Predicting Interactions

In addition to experimentally determined interaction datasets, a vast amount of biological information is contained in the ever-growing datasets of protein sequences, structures, functions, expressions, and literature. Here we review computational methods that extract interaction information from these datasets.

Computational approaches for predicting protein-protein interactions

Predicting protein functional relationships based on comparative genomics

Several methods exist to predict functional relationships between pairs of proteins based on their patterns of occurrence, and their location across multiple genomes. The first method identifies protein pairs that are adjacent along the chromosome. Protein pairs are likely to share similar functions if such chromosomal proximity is conserved across multiple genomes (68-70). In addition, conserved gene order can also be used as an indicator for functional interaction (71). These methods are inspired by the experimental observation that functionally related proteins in bacteria tend to cluster along the chromosome to form operons; their applicability in eukaryotes is less clear.

The second method predicts protein functional interaction based on patterns of domain fusion (72, 73). Sometimes two protein domains exist as separate proteins in one genome, but are fused together into a single protein in another genome. In such a case, the domains are likely to be functionally related (74).

The third method analyzes patterns of occurrence of proteins in multiple genomes. For each protein, a phylogenetic profile is constructed that indicates whether or not the protein is present in each genome. From an evolutionary standpoint, protein pairs with similar phylogenetic profiles tend to 'travel together', and are candidates for functional interaction (75-77).

Predicting protein-protein interactions based on detailed sequence and structural analysis

Two methods exploit the hypothesis that interacting proteins tend to co-evolve. In the first method, the co-evolution of interacting protein families is measured by the similarity of phylogenetic trees constructed from multiple sequence alignments of the two protein families (78, 79). When this technique is applied on a genomic scale, phylogenetic trees for all proteins can be constructed. Proteins with similar phylogenetic trees are more likely to interact with one another than those without (80). In the second method, the co-evolutionary signal in multiple sequence alignments is further analyzed in terms of correlated mutations: a protein pair is likely to interact if there is accumulation of correlated mutations between the interacting partners (81).

Certain pairs of sequence motifs and structural families preferentially interact. To identify such pairs, one first classifies known protein interactions in terms of interactions between sequence motifs and structural families (82, 83). Pairs of sequence motifs and structural families that are overrepresented in the interaction dataset can then be identified. A new protein pair is likely to interact if it can be classified into one of these overrepresented sequence motif or structural family pairs.

It is also possible to predict protein-protein interactions from sequence information using machine learning techniques. For example, using a database of known interactions, a support vector machine learning system can be trained to predict interactions based on sequence information and associated physicochemical properties such as charge, hydrophobicity, and surface tension (84).

With progress in structural genomic projects and structure prediction methods, structural models can be built for an increasing fraction of genomic proteins, with varying degrees of accuracy. For two candidate proteins, each equipped with accurate structural models, it is possible to assess the likelihood of interaction in vitro by calculating the lowest free energy for the protein complex. This process – called docking – has proven increasingly successful in structure prediction of protein complexes, as indicated in the CAPRI meetings (85). However, docking is a time-consuming procedure and its accuracy needs further improvement; in its current form, it is not feasible to predict protein interactions on a genomic scale with this technique.

Databases of solved 3D structures for protein complexes provide additional information that can be exploited for predicting protein-protein interactions. The full set of known 3D complexes can be used to search for all complex homologues in yeast (86). In this method, called multimeric threading, sequences of every protein pair are aligned (or threaded) to a 3D complex template to optimize a compatibility scoring function compiled from known 3D complexes. Top protein pairs with the best compatibility scores are likely to interact in a way similar to the 3D complex template.

Extracting protein interactions from literature

A number of methods have been developed to extract protein interactions from literature. These methods can be grouped into two categories. Methods in the first category use machine learning techniques to screen the literature for articles containing information about protein interactions (87); selected articles are then curated by hand. Methods in the second category automatically extract protein interaction events from biomedical articles. Techniques used range from statistical analysis of co-occurrence of names of biomolecules (88), to natural language processing (89). For detailed reviews of information extraction methods for molecular biology, see (90, 91).

Annotation transfer of protein interactions

Sequence homology offers an efficient way to map genome-wide interaction datasets between different organisms, based on the concept of 'interolog'. This will be discussed later in the section entitled "Cross-referencing different networks".

Correlation of protein functional genomic features as predictors for protein interactions

In addition to sequence and structural information, functional genomic datasets are also available for certain organisms. Much of this functional genomic information is applicable to the study of protein interactions. Consider each class of functional genomic data as a protein feature; two proteins are therefore more likely to interact if these genomic features are correlated. A list of potential functional genomic features for proteins is given below.

(a) mRNA expression. Interacting proteins tend to have correlated expression profiles (16, 92). Protein abundance can be indirectly and quite crudely measured by the presence or absence of the corresponding mRNA transcripts, though large differences can exist between the mRNA and protein abundance (93). Still, several studies have reported a significant correlation of mRNA transcript levels among proteins that interact (92, 94, 95). This correlation is more prominent for proteins in permanent complexes, and less noticeable for those participating in transient complexes (92).

(b) The phenotype of knockout mutants (96, 97) can serve as another potential indicator, suggesting whether two proteins are subunits of the same complex. The genetic deletion of different subunits of the same complex may disturb the function of a complex in the same way, thus producing a similar phenotype. Synthetic lethal interactions are generally enriched in genes that encode members of the same complex (13). More generally, if proteins function in related cellular processes, they have an increased chance of being in the same complex.

(c) To form an interaction, proteins must localize to the same subcellular compartment at the same time. Co-localization thus serves as a useful predictor for protein interaction. A large amount of protein subcellular localization data is available for yeast (98).

Circumstantial evidence, such as the indicators given above, is rarely strong enough to directly predict protein-protein interactions. However, when these datasets are properly combined, quite reliable predictions can result.

Integration of protein-protein interaction datasets

We have seen that protein-protein interaction datasets come from a variety of different experimental and computational sources. To gain a comprehensive understanding of the 'interactome', we must integrate these disparate interaction datasets. There are two key reasons for integrating multiple protein-protein interaction datasets. First, different interaction datasets cover different subsets of the proteome, so it is reasonable to consider their union. Second, the degree of confidence in a protein-protein interaction depends upon how much evidence supports it (99-102). Usually, when multiple, distinct data sources all contribute evidence for a predicted interaction, we gain increased confidence in the validity of our prediction. It is important to note that different experimental methods carry with them different systematic errors – errors that cannot be corrected by repetition.

Integration of multiple datasets of physical protein-protein interactions: RNA polymerase II

The value of integrating multiple datasets of physical protein-protein interactions was demonstrated in a recent study by Edwards et al., who compared the crystal structure of RNA polymerase II with protein-protein interaction experiments on the same set of proteins (29). The protein-protein interaction experiments – including cross-linking, pull-down and 'far western' blotting studies –

were carried out while this structure was still unknown (29, 103-107). The subsequent publication of the crystal structure allowed a retrospective assessment of the success of these experiments.

The comparison showed that the individual protein-protein interaction experiments tended to measure subsets of the potential interactions in the RNA polymerase II structure. Furthermore, individual experiments missed many interactions present in the true structure (‘false negatives’) among the protein pairs that were tested, and found spurious protein-protein interactions absent from the true structure (‘false positives’). The best pull-down experiment was inconsistent with the crystal structure for 23% of the protein pairs, while some experiments were incorrect nearly 50% of the time.

To reduce these error rates, different datasets can be combined. The simplest rules for integration of multiple datasets are the AND- and OR-rules. The AND-rule predicts a positive interaction only when all datasets agree (intersection), while the OR-rule predicts an interaction when at least one dataset gives a positive result (union). The AND-rule tends to give more accurate results, but offers low coverage because few cases exist where all available datasets agree. The OR-rule tends to yield maximum sensitivity (that is, the discovery of the highest number of true positives), but simultaneously produces the highest number of false positives.

An intuitive method of combining the datasets is a majority voting procedure (Figure 1), in which the different experimental results contribute an additive positive or negative vote towards the final result. If the majority of datasets detect an interaction between a protein pair, the pair is predicted to interact, whereas the pair is considered non-interacting if the majority of datasets do not measure an interaction. A major caveat of this procedure is that each dataset implicitly carries the same weight, despite the fact that some datasets contain more reliable results, and other datasets may be redundant. In fact, in the RNA polymerase II example, the prediction by the voting procedure offers virtually no improvement in accuracy compared with the results of the individual interaction experiments (Figure 1). Altogether, the voting procedure has higher coverage than the individual experiments, a trivial result of the integration.

Machine-learning methods provide more sophisticated data integration procedures that take into account data reliability and redundancy, often leading to better results in both coverage and accuracy. An effective method is the Bayesian network, in particular the naïve Bayesian network in its simplest form. Bayesian networks have previously been applied successfully in computational biology research, ranging from the prediction of subcellular localization of proteins (108) to the combination of different gene prediction algorithms (109, 110).

The Bayesian network combines different interaction datasets in a probabilistic manner, assigning a probability to the prediction result rather than just a binary classification. Each individual dataset is essentially weighted by its accuracy and redundancy. The naïve Bayesian network yields optimal results when the different datasets contain uncorrelated evidence; but even when this condition is not met, the results are often useful. In the RNA polymerase II example, naïve Bayesian network integration leads to an increase in accuracy ranging from 5 to 26% compared to the individual experiments (Figure 1). Details on using Bayesian networks for integrating interaction datasets can be found in the Appendix.

Integration of genome-scale protein-protein interaction data

Similar data integration methods can be used on a genomic scale. This is important because several studies have demonstrated that a large number of false positives occur in the results of individual interaction experiments carried out in a high-throughput manner and on a large scale, calling into question the general validity of such experiments. A fair estimate might be that the number of false positives in high-throughput studies is on the same order of magnitude as the actual number of true positive interactions; this reflects the fact that the number of interacting proteins in any cell is perhaps several orders of magnitude smaller than the number of all possible combinations between the proteins in the entire proteome. Screening for protein-protein interactions in the proteome is therefore equivalent to using a diagnostic test for screening for people with a rare disease in the general population: an experiment with a small false positive rate would still yield a high, absolute number of false positives simply because the pool of tested candidates is so large. Thus, a natural strategy to overcome this problem is the combination of multiple interaction data sources and other genomic data.

De novo prediction of protein complexes

Jansen et al. (111) recently showed how protein complexes can be predicted de novo with high confidence when multiple genomic datasets are integrated. In this study, the MIPS complexes catalog was used as a sample of well-characterized protein complexes (determined from more reliable small-scale interaction studies), and a list of negative examples (non-interacting protein pairs) was constructed from proteins that were observed to have different subcellular localizations

(60, 98). While such a list of negatives may be imperfect, they are expected to be strongly enriched in non-interacting protein pairs when compared to randomly chosen proteins. These datasets (‘gold standards’) serve as a reference for observing whether the prediction results are correct (‘testing’), and for determining the parameters of possible integration methods (‘training’).

It is possible to quantify how the different values in the individual genomic features fare in predicting whether two proteins are members of the same complex (Figure S1; follow the Supplemental Material link in the online version of this chapter or at http://www.annualreviews.org/. More details can be found at http://www.genecensus.org/intint/). These different genomic features can then be combined using naïve Bayesian networks (analogous to the method employed in the aforementioned example of RNA polymerase II). Cross-validation with the reference datasets shows that the predictions are highly enriched in positive protein pairs (interacting proteins) rather than negative protein pairs (negatives). Figure 2 shows an example of the de novo prediction results: a set of rRNA processing proteins were predicted to be present in the same complex, and subsequently validated with TAP-tagging experiments. Figure 2 also shows the value of integrating multiple datasets: the confidence with which proteins can be predicted to be in the same complex (here measured in terms of the “likelihood ratio”) is low in the individual datasets, but high in the combined data.

To conclude, the integration of multiple interaction data sources – or data providing circumstantial evidence about protein-protein interactions – can lead to reliable predictions of protein-protein interactions, even if the individual datasets are related to these interactions only in a statistical sense and contain many false positives. If performed correctly, integration of multiple interaction datasets should yield an error rate lower than the component datasets. Machine-learning methods, such as Bayesian networks, have advantages over more simple-minded integration procedures.

We have seen how Bayesian networks can be used as a means to integrate multiple data sources. But in addition to integrating and correlating sets of data, Bayesian networks can also be used to model the regulatory relationships between individual proteins. In the former case, the Bayesian network is used primarily as a tool for integration and classification, while the latter application aims at modeling the interdependency of gene and protein activities, as we will discuss below.

Reconstructing biological pathways and regulatory networks from quantitative measurements

A large amount of data has been produced by quantitatively monitoring the concentrations of biomolecules in a cell, such as mRNA expression levels. Many computational methods have been developed to reconstruct biological pathways and networks from these quantitative measurements, including correlation metric construction (112), Boolean networks (113-115), and Bayesian networks (116, 117). Here we discuss Boolean networks and Bayesian networks in detail.

A Boolean network is a system of interconnected binary elements, defined by a set of nodes and a group of Boolean functions. Each node exists in one of two states; this is applicable to any binary condition, for example on/off or active/inactive. In general, these two states are assigned numerical values of 1 and 0. A Boolean operation is a function taking input from a set of binary variables, and producing output to a single binary variable. Boolean networks can be used to describe the dynamics of a biological system, in that all nodes are updated synchronously, moving the system into its next state. Because the number of all possible states of the system is limited and the transition rules are defined deterministically and do not depend on time, the system either reaches a cycle or converges to an attractor. The attractor can be a steady state or a limit cycle. Attractors can be regarded as the 'target area' of the organism, for instance, cell types following differentiation and development. Although Boolean networks have been considered and developed as an approximation model for biological networks, they are inherently deterministic, and thus do not reflect the inherent randomness that is an integral part of biology. Probabilistic Boolean networks incorporate stochastic variations (115), but the identification of models and the estimation of model parameters under these generalized Boolean networks can pose both theoretical and computational challenges. Another serious limitation of the Boolean network is that all possible variables must be assigned to binary states, while most biological activities exhibit continuous measurements. Most recent studies have focused more on the properties of Boolean networks, so the usefulness of Boolean networks as a general modeling and computational tool for biological pathways has yet to be demonstrated.

Recently, there has been enormous interest in modeling gene expression data with Bayesian networks (see for example (118)). Due to the stochastic nature of biological processes and various measurement errors, the Bayesian network has won support as a suitable technique with which to study gene expression data. Simply put, a Bayesian network is a graphical representation of a joint probability distribution. It consists of two parts: B_s and B_p, where B_s is a directed acyclic graph (DAG, meaning a directed graph where no path starts and ends at the same node), and B_p is a set of local joint probability distributions describing statistical associations. Causal inferences can be made from these associations, by statistically testing the associations between variables, or using a certain measure to score all possible structures and searching for those with high scores. In general, the scoring method is better and more intuitive, and much research has focused on this issue (89, 119).

Dynamic Bayesian networks (DBN) represent a generalization of Bayesian networks and Markov Chains. With DBN modeling, we can model the stochastic evolution of a set of random variables over time (120). Bayesian networks have been used to model gene expression data at various scales. Some studies have modeled roughly 800 yeast cell-cycle genes (116). Other groups have focused on a more limited number of genes. For example, the yeast pheromone response pathway (~32 genes) was recently studied (117). A detailed analysis of just three genes involved in the yeast galactose pathway was reported (118). Although the application of both Bayesian networks and DBN to modeling gene expression has been discussed, their usefulness remains to be shown and analyses of more well-understood genetic pathways are needed.

There are some limitations to current BN and DBN approaches. From a statistical perspective, expression levels must be discretized, undoubtedly leading to loss of information. Although we can simplify the computation (as well as obtain a stable result) through such discretization, we need to explore alternative ways to discretize data and, more importantly, find reliable approaches to analyze continuous data.

Two major limitations exist to using Bayesian networks to model biological pathways. First, all observations are assumed to stem from the same distribution, which clearly cannot model the dynamics of biological systems as well as responses to environmental perturbations. Second, there is the identifiability problem, in that many distinct DAGs may result in the same joint probability distributions. Although the DBN may partially address these problems, the computational and theoretical implications of extension to more general models require further investigation. Although it has been reported in the literature (117) that the Bayesian network methodology was able to correctly identify the true biological model from two competing hypotheses, it became clear that this particular analysis was driven by two outlying observations from a total of 55 observations (H. Zhao and B. Wu, unpublished results). The Bayesian networks also failed to detect the galactose pathway from genomics data reported in (121). Furthermore, when a DBN was applied to time-course data in Drosophila (122), it failed to identify the correct transcriptional regulatory network among three genes showing expression patterns clearly consistent with known biology (H. Zhao and B. Wu, unpublished results). A closer inspection of the cause of DBN failure showed that the stationarity assumption underlying this approach may be too strong and inappropriate. Our experience with Bayesian networks and DBN suggests that a considerable amount of work needs to be done to improve current methods before meaningful results can be reliably extracted from genomic data.

Clearly, better statistical methods are needed to reconstruct biological pathways from quantitative measurements. In addition, improvements along other directions are possible. First, additional quantitative measurements performed on a systematically perturbed network can help define the network architecture with increasing accuracy (121, 123). Second, cross-species comparison can help reveal the conserved core network. Evolutionarily conserved co-expression implies selective advantage, and therefore, a functional relationship (124). Third, the aforementioned analyses need to be combined with other types of information, such as shared functional classification (125), shared promoter motifs (126), protein-protein interaction data (127) and protein-DNA binding data (128). In the end, all these diverse genomic datasets need to be integrated in a proper way for an accurate reconstruction of biomolecular networks.

Approaches for Analyzing Large Networks of Interactions

Once molecular networks have been reconstructed, we can then proceed to compare and contrast them in terms of global and local topology, and relate structural properties of networks to protein properties, such as function or essentiality. These topics, generally termed network analysis, are reviewed here.

Network topology

The classical random network theory, introduced by Erdös and Rényi (10, 129), has been generally used to model complex networks. This model assumes that each node in a network is connected to another node randomly with probability p and the degrees of the nodes follow a Poisson distribution, which has a strong peak at the average degree, K. Most random networks are highly homogeneous, in that most nodes have the same number of links (degree), $k_i \approx K$, where k_i is the degree of the ith node. The chance of having nodes with k links falls off exponentially (i.e. $P(k) \approx e^{-k}$) for large k.

To explain the heterogeneous nature of complex networks, Barabási and colleagues recently proposed a “scale-free” model in which the degree distribution in many large networks follows a power-law distribution ($P(k) \approx k^{-r}$) (5). The most remarkable point about this distribution is that most of the nodes within these networks have very few links, but a few (the hubs) are exceptionally highly connected. Concurrently, Watts & Strogatz found that many networks also have a “small-world" property (3), meaning they are defined as being both highly clustered and containing small characteristic path lengths.

The relevance of such structures is apparent in multiple disciplines. A recent practical example is the North American power grid structure. While haphazardly constructed, the grid has evolved into a network that is defined by a power law; most nodes in the grid are linked to few other nodes, while some hubs are highly connected to many other nodes. The power law adds a level of robustness to the network, meaning many individual nodes can fail without destroying the whole grid. Conversely, when several hubs, or too many nodes (130) fail, the entire network will collapse and a large regional blackout will result. This example highlights two important characteristics of networks that are pertinent to the protein interaction network. Firstly, networks often evolve naturally into power law networks – hubs evolve naturally from nodes due to inherent characteristics of the original node, i.e. its importance, fitness, or relative age within the network (10). Second, power law networks are robust: a network defined by a power law has an inherent design that makes it less susceptible to random destabilizing events. The small world concept is important in social networks, connecting multiple and otherwise unassociated cliques, causing the networks to have a higher than otherwise logically suspected degree of clustering (the so-called 'six degrees of separation'). Similarly, protein networks are often found to adhere to the small world property, primarily due to the interconnectivity of a select group of nodes.

Topological analysis of these networks provides quantitative insight into their basic organization. Generally, there are four topological parameters in network analysis (Figure 3) (3-5, 7-11, 19):

(1) Average degree (K). The degree of a node is the number of links that this node has with other nodes. The average degree of the whole network is the average of the degree of all its individual nodes.

(2) Clustering coefficient (C), defined as the ratio of the number of existing links between a node's neighbors and the maximum possible number of links between them. The clustering coefficient of the network is the average of the individual coefficients. This statistic can be used to determine the completeness of the network.

(3) Characteristic path length (L). The graph theoretical distance between two nodes is the minimum number of edges that is necessary to traverse from one node to the other. The characteristic path length of a network is the average of these minimum distances: it gives a measure of how close nodes are connected within the network.

(4) The diameter (D) of a network is the longest graph theoretical distance between any two nodes in the graph.

Networks can be divided into two broad categories: directed and undirected. Physical interaction, genetic interaction and expression networks are *undirected*, meaning no directionality or causality is implied in the interactions. Stated differently, "node A is linked to node B" is equivalent to "node B is linked to node A". These undirected networks should be sharply distinguished from other biological networks, such as regulatory networks, metabolic networks, and signaling networks; these *directed* networks do imply directionality in their linkages. A node in the directed network may have an in-degree and an out-degree (see Figure 3C), which are completely independent. The in-degree of a node is the number of edges pointing toward this node, while its out-degree is the number of edges pointing out of this node. The clustering coefficient cannot be calculated for directed networks (10).

Sub-structures within networks

Complex networks, such as protein-protein physical interaction networks (herein referred to simply as interaction networks) and regulatory networks, usually contain biologically meaningful sub-structures. Within interaction networks, protein complexes will theoretically appear as a clique, a fully connected sub-graph. However, because of the limitation of the interaction identification techniques, some of the links within the same complex may be missing. Therefore, in reality, most complexes are quasi-cliques within the interaction networks (131).

Compared with yeast two-hybrid methods, in vivo pull-down methods detect protein complexes, rather than binary protein-protein interactions. In order to break down the complexes into binary interaction pairs, Bader and Hogue proposed two models: spoke and matrix (132). The ‘spoke’ model assumes that only the bait proteins directly interact with each component of the complex. The ‘matrix’ model assumes that each component interacts with all other components in the same complex. An important assumption in extrapolating gene function is “guilt by association”, i.e. two interacting proteins should share the same function. In the paper, the authors were able to show that interacting pairs produced by the spoke model are more likely to share common functions than those produced by the matrix model. In addition, Bader and Hogue proposed a new method to determine protein complexes within interaction networks, termed "*k*-core." A *k*-core is a graph of minimal degree *k*. To date, high-throughput interaction identification methods, such as yeast two-hybrid methods and in vivo pull-down methods, all have high false-positive rates. However, within a k-core, links between proteins are strengthened by one another as a joint probability, which largely increases the accuracy of the predicted interactions. Using *k*-core method, Bader and Hogue were able to identify many well-known complexes, as well as some novel but reasonable ones, within the yeast interaction networks (132).

As discussed above, regulatory networks are different from interaction networks in that regulatory networks are directed networks. There are six basic network motifs within regulatory networks (Figure 4) (17, 133, 134): 1) Single input motif, in which a set of targets are regulated by only one transcription factor. 2) Multi-input motif, where a set of targets are regulated by more than one regulator, and these are the only regulators for these targets. 3) Feed-forward loop, where one transcription factor (TF1) regulates another transcription factor (TF2), and both factors regulate their targets together. 4) Auto-regulation, in which one transcription factor regulates itself. 5) Multi-component loop, where one factor regulates a second factor, which in turn regulates the first one. 6) Regulator chain, in which for several regulators, one regulates another in a chain fashion.

Application of topological analysis

The ultimate goal of functional genomics is to determine the function of every gene product in fully sequenced genomes. Different prediction schemes have been proposed, such as the concept that co-expressed genes share similar functions, or that interacting proteins have the same functions.

Given the relative ease with which large-scale protein-protein interaction datasets can now be produced, functional genomics relies on interaction data to determine an unclassified protein's function based on its interacting partners. Traditionally, pairs of interacting proteins have been thought to share similar functions (25, 26); but because proteins normally interact with more than one partner, and the interacting partners for the same protein do not generally share the same functions, this idea is clearly problematic. A better method, known as the 'majority rule' method (135, 136), assigns an unknown protein to the functional class to which the majority of its partners belong. Obviously, the method is still inefficient: because only a small portion of the genes in fully-sequenced genomes have functional annotations, the functional assignment of an unknown protein will affect the assignment of its interaction partners, which will in turn affect the assignment of this protein itself. This circular reasoning serves to amplify possible errors in function prediction. Therefore, in order to efficiently predict protein functions based on interaction networks, the global topological structures of the networks have to be taken into consideration. Here, two such methods will be discussed in detail.

Bu et al. introduced a method to determine quasi-cliques within interaction networks using spectral analysis (131). These quasi-cliques were proven to be biologically relevant functional groups, which is similar to the concept of "protein complexes" as discussed above. In order to perform spectral analysis, an interaction network is represented by an $N \times N$ adjacency matrix. N is the total number of proteins within the network. The adjacency matrix is defined as $\boldsymbol{A} = (a_{ij})$, where $a_{ij} = 1$ if protein i interacts with protein j, and $a_{ij} = 0$ if not. For each eigenvector of the matrix with a positive eigenvalue, the proteins corresponding to absolutely larger components tend to form a quasi-clique, meaning every two of them tend to interact with each other. The quasi-cliques were defined based on the following criteria: 1) each quasi-clique must contain at least 10 proteins, 2) the proteins were sorted by their absolute weight value in an eigenvector, and the top 10% were selected, and 3) each protein in a quasi-clique must interact with at least 20% of that clique's members. The clustering coefficient of each quasi-clique was tuned for a high degree of interconnectivity. Within yeast interaction networks, 48 quasi-cliques were successfully identified. The proteins of each quasi-clique indeed tend to share common functions based on MIPS functional classification (131).

Concurrently, Vazquez et al. proposed a global optimization method to predict functions of unknown proteins within the interaction networks (137). Simply put, the global optimization method computes a score for any particular configuration of the functional assignment for the whole protein interaction network. The score is lower if fewer interacting pairs are assigned to distinct functional classes. The method aims to find the configuration of functional assignment with the lowest possible score through global optimization. Since there can be more than one optimal solution for this kind of problem, a 'simulated annealing' technique was introduced to determine the optimal configurations, and the most frequent functional assignment for a certain protein in all optimal solutions was assigned as its function. In order to evaluate the success rate, a fraction of classified proteins were considered as unclassified in the input data; for proteins with more than one interacting partner, the performance of the global optimization method is much better than that of the majority rule method (137).

Cross-referencing different networks

So far, we have discussed many distinct types of networks. The fact that such networks exist in all species means that the total number of different networks is far larger. It is impossible to investigate every network of every organism in equal detail; however, in model organisms, particularly *Saccharomyces cerevisiae*, a vast amount of data has been accumulated for all these types of networks. Mapping the networks in model organisms to other species by homology provides insight into how to exploit the usefulness (and prevent the potential pitfalls) of annotating unknown genes in other less characterized species. To this end, Walhout et al. introduced the concept of 'interolog' to transfer interaction networks from one species to another (31). Interologs are defined as orthologous pairs of interacting proteins in different organisms. Thus if interacting proteins X and Y in one organism have interacting orthologs X' and Y' in another species, the X-Y and X'-Y' interactions are called interologs (31). Subsequently, based on 216 worm protein pairs and 72 yeast protein pairs, Matthews et al. experimentally estimated the accuracy of this 'interolog' method to be in the range of 16% - 31% for two species that are evolutionarily distant by about 900 million years (138).

Cross-species comparison of interaction networks tells us how these networks evolve. Similarly, comparison of different networks within the same organism often sheds light on the basic organization principles of the cell. Yu et al. analyzed the regulatory and expression networks for *Saccharomyces cerevisiae* and were able to show that co-regulated genes are generally co-expressed and the correlation in expression profiles is highest for genes targeted by multiple transcription factors. Furthermore, co-regulated gene pairs tend to share cellular functions, and there are subdivisions within individual network motifs that separate the regulation of genes of distinct functions. The expression profiles of transcription factors and their target genes display more complex relationships than simple correlation, with the regulatory response of target genes often being delayed (57).

Interaction Networks and Systems Biology

Mapping and understanding molecular interaction networks represent the first steps towards modeling how a cell actually operates in time and space. As a result of genome-wide high-throughput experiments, we now have a complete 'parts list' of functional elements for many genomes, and soon we will also have a complete catalog of how these functional elements interact with and regulate each other. The grand challenge, then, will be to put all the pieces back together to create predictive models of cellular behavior. Such is the goal of systems biology, an emerging field that quantitatively measures and models the behavior of a cell from a systems perspective, as a result of collective spatial-temporal dynamics of its interacting components (139, 140). Here we briefly review some of the challenges and initial successes of using interaction networks to model cellular behavior.

The first challenge is to create a three-dimensional view of molecular interaction networks in a cell. This is important because biomolecules are 3D objects; they function and interact through spatially precise atomic interactions in crowded microenvironments. The second challenge is to capture the dynamic and context-dependent nature of interaction networks. In addition to mapping out all possible interactions, it is also important to know under which conditions (cellular state, environment type, protein modification type, et al.) each interaction is present in a cell. Finally, the third challenge is to quantitatively measure interaction networks. Interaction networks tell us whether or not two molecules interact. In order to model cellular behavior, it is also important to know how strongly and how quickly they interact. Such requirements call for continuing improvements in quantitative high-throughput methods.

Despite these considerable challenges, a number of recent modeling successes indicate a promising future for systems biology. For well-characterized interaction networks with a small number of genes and proteins, it is possible to build a detailed kinetic model of the system, and simulations are generally in good agreement with experiments. Such systems include, among others, bacteriophages (141), bacteria chemotaxis (142), circadian clocks (143), and signaling pathways (144). Softwares are available that could potentially scale up these detailed simulations to the level of an entire cell (145, 146), but the predictive power of the whole-cell simulations is limited by the fact that the vast majority of underlying kinetic parameters remain unknown.

Alternatively, it is possible to model the behavior of interaction networks in a coarse manner without knowing detailed kinetic parameters. Such modeling can be applied on a genomic scale or on less well-characterized systems, since it requires only a few parameters beyond the topology and stoichiometry of the network. A prime example is the flux balance analysis of metabolic networks (147), where the steady-state behavior of the entire network can be modeled reasonably well. Similarly, much of the logic and dynamics of a bacteria cell-cycle regulatory network can be understood and possibly modeled without a full set of kinetic parameters (148).

These modeling efforts are providing us with an increasing number of insights into the design principles of biomolecular networks. For example, biomolecular networks can be grouped into modules (1). Functional elements within a module interact strongly with each other and carry out a common function in a concerted fashion. Biomolecular networks are resilient towards common external and internal perturbations (149). Furthermore, noise is an integral part of the functioning of biomolecular networks (150).

In summary, molecular interaction networks are at the core of current functional genomic research. These networks represent an appealing framework upon which different genomic data can be integrated, and analysis of these networks has yielded the first clues about their organizational and design principles. Furthermore, these networks lay the foundation for systems biology analysis of the cell. With combined experimental, computational and theoretical efforts, a complete mapping of all interaction networks, and ultimately a rational understanding of cellular behavior, will become a reality.

Acknowledgments

We thank Anuj Kumar for reading the experimental section of the draft and reference selection. R. Jansen is supported by the Seraph Foundation. H. Zhao is supported by NSF grant 0241160. This work is supported by a grant from NIH/NIGMS (GM054160-07).

Figure Legends

Figure 1: Comparison of the crystal structure of RNA polymerase II and the protein-protein interaction experiments. The RNA polymerase II structure consists of 10 protein subunits, allowing for 45 different pairings between these proteins (shown in the first two columns). The third column shows which of the protein pairs are in physical contact (with a contact interface area ≥ 800 Å^2, shown by the gray squares). The following columns show the results from three far western, three pulldown and one cross-linking experiment. The experimental results are indicated as either positive (+), when an interaction was found, or negative (-) when no interaction was detected. The results are either colored green (true), when they agreed with the crystal structure contacts, or red (false), when they disagreed. Blank fields in the table correspond to protein pairs that were not tested in the experiments. The two columns on the right show the results of integrating the seven experiments into one prediction of interactions. The column "Voting" shows the difference between positive and negative experimental results for each protein pair. The column "Bayes" shows the posterior odds of having a real protein-protein interaction, based on the experimental data, calculated with a naïve Bayesian network. The row "Coverage" shows how many protein pairs were measured in each experiment or how many protein pairs are covered by the voting and the Bayesian network procedure. These can be divided into TP (true positive), FN (false negative), TN (true negative), and FP (false positive). The bottom three rows show the accuracy of individual experiments, defined as (TP + TN)/Coverage. The rows "Voting" and "Bayes" show the accuracy of the voting procedure and the naïve Bayesian network for the same subset of protein pairs that were measured in the individual experiments. The graph at the bottom

shows the "Accuracy increase", the difference between the accuracies of the integration methods and the individual experiments. Gray bars represent Bayes, and black bars represent Voting.

Figure 2: (**Top**) Graphical representation of a complex around the protein Nsr1 that was predicted by combining genomic features such as essentiality, expression, and function. All proteins shown have posterior odds of being in the same complex greater than 1 (assuming prior odds of 1/600). Some of the protein pairs were experimentally verified by TAP-tagging, whereas other protein pairs were shown to be interacting by previous proteomics studies (25, 27, 37, 38). (**Bottom**) A histogram of the distribution of likelihood ratios among the individual experiments and the combined data shows the value of integrating multiple data sources: the combined data contains a much larger number of protein pairs with high likelihood ratios.

Figure 3: Schematic illustration of different topological parameters within undirected and directed networks. (**A**) In an undirected network, the diameter of the essential protein network (shown as the red line) is the maximum distance between any two essential proteins. The path can go through non-essential proteins, but has to start and end at essential ones. The diameter of the non-essential protein network can be similarly defined. (**B**) Protein B has 4 interaction partners, among which there are only 2 connections, whereas there could potentially be 6 (shown as the dotted lines). Therefore, the clustering coefficient for B is 1/3. Protein A interacts with 5 different proteins, which belong to 3 different complexes. Therefore, the complex degree of A is 3. In (**A**) and (**B**), black circles: essential proteins; open circles: non-essential proteins; gray circles: marginally essential proteins. The size of the circle represents the degree of the node. (**C**) The out-degree of

transcription factor A is 3 and its in-degree is 2. Genes without transcription factor activities (shown as rectangles) only have in-degrees.

Figure 4: Depiction of the six basic regulatory motifs. Circles: transcription factors, rectangles: targets. Detailed descriptions are given in the section entitled "Sub-structures within networks".

Tables

Table 1: Summary of the databases for biomolecular interactions

Name	*URL*	Type of Networks	References
DIP	http://dip.doe-mbi.ucla.edu/	Physical	(58)
BIND	http://www.bind.ca/	Physical	(59)
HPRD	http://www.hprd.org/	Physical	(67)
MIPS	http://mips.gsf.de/	Physical Genetic	(60)
YPD	http://www.incyte.com/sequence/proteome/index.shtml	Physical Genetic Regulatory	(61)
TRANSFAC	http://transfac.gbf.de/TRANSFAC/	Regulatory	(62)
RegulonDB	http://www.cifn.unam.mx/Computational_Genomics/regulondb/	Regulatory	(63)
KEGG	http://www.kegg.com/	Metabolic	(64)
MetaCyc	http://metacyc.org/	Metabolic	(65)
AFCS	http://www.cellularsignaling.org/	Signaling	(66)

Appendix

Details on Using Bayesian Networks for Integrating Interaction Datasets

Given multiple experimental results e_i (from N different experiments, with $i = 1 \ldots N$), the posterior odds of a protein-protein interaction can be computed as follows with a naïve Bayesian network:

$$O_{post} = \prod_{i=1}^{N} L_i(e_i) O_{prior} \quad (1)$$

Here, O_{post} is defined as:

$$O_{post} = \frac{P(I=+\,|\,e_1,e_2 \ldots e_N)}{P(I=-\,|\,e_1,e_2 \ldots e_N)} = \frac{P(I=+\,|\,e_1,e_2 \ldots e_N)}{1-P(I=+\,|\,e_1,e_2 \ldots e_N)} \quad (2)$$

while O_{prior} is:

$$O_{prior} = \frac{P(I=+)}{P(I=-)} = \frac{P(I=+)}{1-P(I=+)} \qquad (3)$$

Thus the posterior odds describe the odds of having a protein-protein interaction ($I = +$) given that we have the information from the N experiments, whereas the prior odds are related to the chance of randomly finding a protein-protein interaction when no experimental data is known. If $O_{post} > 1$, the chances of having an interaction are higher than having no interaction. For the RNA polymerase II example given in the main text, the prior odds were set to $13/(45 - 13) \approx 0.41$, i.e., the ratio of protein pairs observed to be in contact in the crystal structure of RNA polymerase II divided by the remaining protein pairs, but they could also be determined by counting the number of protein-protein interactions in comparable protein structures.

$L_i(e_i)$ describes the "likelihood ratio" of the experimental result e_i, and can be computed from the table in Figure 1 as follows:

$$L_i(e_i=+1)=\frac{TP_i}{TP_i+FN_i}\frac{FP_i+TN_i}{FP_i} \qquad (4)$$

and

$$L_i(e_i=-1)=\frac{FN_i}{TP_i+FN_i}\frac{FP_i+TN_i}{TN_i} \qquad (5)$$

where the subscript i refers to a particular column in the table. (We assume here for simplicity that an experiment either has a positive or a negative result, i.e., $e_i = \pm 1$). For a perfect experiment with no errors, one would observe $FP_i \to 0$ and $FN_i \to 0$, such that $L(e_i = +1) \to \infty$ and $L(e_i = -1) \to 0$.

The naïve Bayes procedure can be intuitively understood by comparing it to the voting procedure. In the voting procedure the experimental results are simply added up:

$$s=\sum_{i=1}^{N} e_i \qquad (6)$$

Then, when $s > 0$, we consider the protein pair to be interacting (and non-interacting otherwise). Note that all experiments are weighted equally. In contrast, the naïve Bayes procedure weights each experiment differently based on the likelihood ratio values. This analogy to a weighted voting procedure can be seen if we take the logarithm of equation (1):

$$\log\left(O_{post}\right) = \sum_{i=1}^{N} \log\left(L_i(e_i)\right) + \log\left(O_{prior}\right) \qquad (7)$$

Here, a protein pair is considered to be interacting if $\log(O_{post}) > 0$, while the term $\log(L_i(e_i))$ corresponds to the weight of experiment e_i. The difference with the voting procedure is the inclusion of the term $\log(O_{post})$, which represents the chance of randomly finding a protein-protein interaction without experimental information.

References

1. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. 1999. *Nature* 402: C47-52

2. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. 2000. *Nature* 405: 823-6

3. Watts DJ, Strogatz SH. 1998. *Nature* 393: 440-2

4. Albert R, Jeong H, Barabasi AL. 1999. *Nature* 401: 130-1

5. Barabasi AL, Albert R. 1999. *Science* 286: 509-12

6. Huberman BA, Adamic LA. 1999. *Nature* 401: 131

7. Albert R, Jeong H, Barabasi AL. 2000. *Nature* 406: 378-82

8. Amaral LA, Scala A, Barthelemy M, Stanley HE. 2000. *Proc. Natl. Acad. Sci. USA* 97: 11149-52

9. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. *Nature* 411: 41-2

10. Albert R, Barabasi AL. 2002. *Rev. Mod. Phys.* 74: 47-97

11. Girvan M, Newman ME. 2002. *Proc. Natl. Acad. Sci. USA* 99: 7821-6

12. Lan N, Montelione GT, Gerstein M. 2003. *Curr. Opin. Chem. Biol.* 7: 44-54

13. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, et al. 2001. *Science* 294: 2364-8

14. Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. *Proc. Natl. Acad. Sci. USA* 95: 14863-8

15. Altman RB, Raychaudhuri S. 2001. *Curr. Opin. Struct. Biol.* 11: 340-7

16. Qian J, Dolled-Filhart M, Lin J, Yu H, Gerstein M. 2001. *J. Mol. Biol.* 314: 1053-66

17. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al. 2002. *Science* 298: 799-804

18. Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, et al. 2002. *Genes Dev.* 16: 3017-33

19. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. 2000. *Nature* 407: 651-4

20. Pawson T, Scott JD. 1997. *Science* 278: 2075-80

21. Sambrano GR, Chandy G, Choi S, Decamp D, Hsueh R, et al. 2002. *Nature* 420: 708-10

22. Fields S, Song O. 1989. *Nature* 340: 245-6

23. Walhout AJ, Vidal M. 2001. *Nat. Rev. Mol. Cell Biol.* 2: 55-62

24. Fromont-Racine M, Rain JC, Legrain P. 1997. *Nat. Genet.* 16: 277-82

25. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. 2000. *Nature* 403: 623-7

26. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, et al. 2000. *Proc. Natl. Acad. Sci. USA* 97: 1143-7

27. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. 2001. *Proc. Natl. Acad. Sci. USA* 98: 4569-74

28. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, et al. 2002. *Nature* 417: 399-403

29. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M. 2002. *Trends Genet.* 18: 529-36

30. Drees BL, Sundin B, Brazeau E, Caviston JP, Chen GC, et al. 2001. *J. Cell Biol.* 154: 549-71

31. Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, et al. 2000. *Science* 287: 116-22

32. Boulton SJ, Gartner A, Reboul J, Vaglio P, Dyson N, et al. 2002. *Science* 295: 127-31

33. Aebersold R, Mann M. 2003. *Nature* 422: 198-207

34. Ashman K, Moran MF, Sicheri F, Pawson T, Tyers M. 2001. *Sci. STKE* 2001: PE33

35. Kumar A, Snyder M. 2002. *Nature* 415: 123-4

36. Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B. 1999. *Nat. Biotechnol.* 17: 1030-2

37. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. 2002. *Nature* 415: 141-7

38. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. 2002. *Nature* 415: 180-3

39. Arenkov P, Kukhtin A, Gemmell A, Voloshchuk S, Chupeeva V, Mirzabekov A. 2000. *Anal. Biochem.* 278: 123-31

40. MacBeath G, Schreiber SL. 2000. *Science* 289: 1760-3

41. Zhu H, Klemic JF, Chang S, Bertone P, Casamayor A, et al. 2000. *Nat. Genet.* 26: 283-9

42. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, et al. 2001. *Science* 293: 2101-5

43. Sali A, Glaeser R, Earnest T, Baumeister W. 2003. *Nature* 422: 216-25

44. Bender A, Pringle JR. 1991. *Mol. Cell. Biol.* 11: 1295-305

45. Wang T, Bretscher A. 1997. *Genetics* 147: 1595-607

46. Mullen JR, Kaliraman V, Ibrahim SS, Brill SJ. 2001. *Genetics* 157: 103-18

47. Garner MM, Revzin A. 1981. *Nucleic Acids Res.* 9: 3047-60

48. Seguin C, Hamer DH. 1987. *Science* 235: 1383-7

49. Fraga MF, Uriol E, Borja Diego L, Berdasco M, Esteller M, et al. 2002. *Electrophoresis* 23: 1677-81

50. Galas DJ, Schmitz A. 1978. *Nucleic Acids Res.* 5: 3157-70

51. Kuo MH, Allis CD. 1999. *Methods* 19: 425-33

52. Simpson RT. 1999. *Curr. Opin. Genet. Dev.* 9: 225-9

53. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. 2001. *Nature* 409: 533-8

54. Sun LV, Chen L, Greil F, Negre N, Li TR, et al. 2003. *Proc. Natl. Acad. Sci. USA* 100: 9428-33

55. van Steensel B, Henikoff S. 2000. *Nat. Biotechnol.* 18: 424-8

56. van Steensel B, Delrow J, Henikoff S. 2001. *Nat. Genet.* 27: 304-8

57. Yu H, Luscombe NM, Qian J, Gerstein M. 2003. *Trends Genet.* 19: 422-7

58. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. 2002. *Nucleic Acids Res.* 30: 303-5

59. Bader GD, Betel D, Hogue CW. 2003. *Nucleic Acids Res.* 31: 248-50

60. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, et al. 2002. *Nucleic Acids Res.* 30: 31-4

61. Csank C, Costanzo MC, Hirschman J, Hodges P, Kranz JE, et al. 2002. *Methods Enzymol.* 350: 347-73

62. Wingender E, Chen X, Fricke E, Geffers R, Hehl R, et al. 2001. *Nucleic Acids Res.* 29: 281-3

63. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, et al. 2001. *Nucleic Acids Res.* 29: 72-4

64. Kanehisa M, Goto S. 2000. *Nucleic Acids Res.* 28: 27-30

65. Karp PD, Riley M, Paley SM, Pellegrini-Toole A. 2002. *Nucleic Acids Res.* 30: 59-61

66. Gilman AG, Simon MI, Bourne HR, Harris BA, Long R, et al. 2002. *Nature* 420: 703-6

67. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, et al. 2003. *Genome Res.* 13: 2363-71

68. Tamames J, Casari G, Ouzounis C, Valencia A. 1997. *J. Mol. Evol.* 44: 66-73

69. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. *Proc. Natl. Acad. Sci. USA* 96: 2896-901

70. Yanai I, Mellor JC, DeLisi C. 2002. *Trends Genet.* 18: 176-9

71. Dandekar T, Snel B, Huynen M, Bork P. 1998. *Trends Biochem. Sci.* 23: 324-8

72. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. 1999. *Science* 285: 751-3

73. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. 1999. *Nature* 402: 86-90

74. Yanai I, Derti A, DeLisi C. 2001. *Proc. Natl. Acad. Sci. USA* 98: 7940-5

75. Tatusov RL, Koonin EV, Lipman DJ. 1997. *Science* 278: 631-7

76. Gaasterland T, Ragan MA. 1998. *Microb. Comp. Genomics* 3: 199-217

77. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. *Proc. Natl. Acad. Sci. USA* 96: 4285-8

78. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. *J. Mol. Biol.* 299: 283-93

79. Goh CS, Cohen FE. 2002. *J. Mol. Biol.* 324: 177-92

80. Pazos F, Valencia A. 2001. *Protein Eng.* 14: 609-14

81. Pazos F, Valencia A. 2002. *Proteins* 47: 219-27

82. Sprinzak E, Margalit H. 2001. *J. Mol. Biol.* 311: 681-92

83. Park J, Lappe M, Teichmann SA. 2001. *J. Mol. Biol.* 307: 929-38

84. Bock JR, Gough DA. 2001. *Bioinformatics* 17: 455-60

85. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, et al. 2003. *Proteins* 52: 2-9

86. Lu L, Arakaki AK, Lu H, Skolnick J. 2003. *Genome Res.* 13: 1146-54

87. Marcotte EM, Xenarios I, Eisenberg D. 2001. *Bioinformatics* 17: 359-63

88. Stapley BJ, Benoit G. 2000. *Pac. Symp. Biocomput.*: 529-40

89. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. 2001. *Bioinformatics* 17 Suppl 1: S74-82

90. Hirschman L, Park JC, Tsujii J, Wong L, Wu CH. 2002. *Bioinformatics* 18: 1553-61

91. Blaschke C, Hirschman L, Valencia A. 2002. *Brief. Bioinform.* 3: 154-65

92. Jansen R, Greenbaum D, Gerstein M. 2002. *Genome Res.* 12: 37-46

93. Greenbaum D, Colangelo C, Williams K, Gerstein M. 2003. *Genome Biol.* 4: 117

94. Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, et al. 2002. *Mol. Cell* 9: 1133-43

95. Ge H, Liu Z, Church GM, Vidal M. 2001. *Nat. Genet.* 29: 482-6

96. Ross-Macdonald P, Coelho PS, Roemer T, Agarwal S, Kumar A, et al. 1999. *Nature* 402: 413-8

97. Giaever G, Chu AM, Ni L, Connelly C, Riles L, et al. 2002. *Nature* 418: 387-91

98. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, et al. 2002. *Genes Dev.* 16: 707-19

99. Gerstein M, Lan N, Jansen R. 2002. *Science* 295: 284-7

100. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, et al. 2002. *Science* 295: 321-4

101. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. 1999. *Nature* 402: 83-6

102. Jansen R, Lan N, Qian J, Gerstein M. 2002. *J. Struct. Funct. Genomics* 2: 71-81

103. Acker J, de Graaff M, Cheynel I, Khazak V, Kedinger C, Vigneron M. 1997. *J. Biol. Chem.* 272: 16815-21

104. Kimura M, Ishihama A. 2000. *Nucleic Acids Res.* 28: 952-9

105. Ulmasov T, Larkin RM, Guilfoyle TJ. 1996. *J. Biol. Chem.* 271: 5085-94

106. Miyao T, Yasui K, Sakurai H, Yamagishi M, Ishihama A. 1996. *Genes Cells* 1: 843-54

107. Ishiguro A, Kimura M, Yasui K, Iwata A, Ueda S, Ishihama A. 1998. *J. Mol. Biol.* 279: 703-12

108. Drawid A, Gerstein M. 2000. *J. Mol. Biol.* 301: 1059-75

109. Pavlovic V, Garg A, Kasif S. 2002. *Bioinformatics* 18: 19-27

110. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. 2003. *Proc. Natl. Acad. Sci. USA* 100: 8348-53

111. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, et al. 2003. *Science* 302: 449-53

112. Arkin A, Shen P, Ross J. 1997. *Science* 277: 1275-79

113. Liang S, Fuhrman S, Somogyi R. 1998. *Pac. Symp. Biocomput.*: 18-29

114. Akutsu T, Miyano S, Kuhara S. 2000. *Bioinformatics* 16: 727-34

115. Shmulevich I, Dougherty ER, Kim S, Zhang W. 2002. *Bioinformatics* 18: 261-74

116. Friedman N, Linial M, Nachman I, Pe'er D. 2000. *J. Comput. Biol.* 7: 601-20

117. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. 2002. *Pac. Symp. Biocomput.*: 437-49

118. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. 2001. *Pac. Symp. Biocomput.*: 422-33

119. Heckerman D, Geiger D, Chickering DM. 1995. *Machine Learning* 20: 197-243

120. Dean T, Kanazawa K. 1988. *Proc. AAAI*: 524-9

121. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, et al. 2001. *Science* 292: 929-34

122. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, et al. 2002. *Science* 297: 2270-5

123. Gardner TS, di Bernardo D, Lorenz D, Collins JJ. 2003. *Science* 301: 102-5

124. Stuart JM, Segal E, Koller D, Kim SK. 2003. *Science* 302: 249-55

125. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. 2002. *Nat. Genet.* 31: 370-7

126. Pilpel Y, Sudarsanam P, Church GM. 2001. *Nat. Genet.* 29: 153-9

127. Segal E, Wang H, Koller D. 2003. *Bioinformatics* 19 Suppl 1: I264-I72

128. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, et al. 2003. *Nat. Biotechnol.* 21: In press

129. Erdos P, Renyi A. 1959. *Publ. Math. (Debrecen)* 6: 290-7

130. Cohen R, Erez K, ben-Avraham D, Havlin S. 2000. *Phys. Rev. Lett.* 85: 4626-8

131. Bu D, Zhao Y, Cai L, Xue H, Zhu X, et al. 2003. *Nucleic Acids Res.* 31: 2443-50

132. Bader GD, Hogue CW. 2002. *Nat. Biotechnol.* 20: 991-7

133. Shen-Orr SS, Milo R, Mangan S, Alon U. 2002. *Nat. Genet.* 31: 64-8

134. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. 2002. *Science* 298: 824-7

135. Schwikowski B, Uetz P, Fields S. 2000. *Nat. Biotechnol.* 18: 1257-61

136. Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T. 2001. *Yeast* 18: 523-31

137. Vazquez A, Flammini A, Maritan A, Vespignani A. 2003. *Nat. Biotechnol.* 21: 697-700

138. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, et al. 2001. *Genome Res.* 11: 2120-6

139. Ideker T, Galitski T, Hood L. 2001. *Annu. Rev. Genomics Hum. Genet.* 2: 343-72

140. Kitano H. 2002. *Science* 295: 1662-4

141. Arkin A, Ross J, McAdams HH. 1998. *Genetics* 149: 1633-48

142. Barkai N, Leibler S. 1997. *Nature* 387: 913-7

143. Barkai N, Leibler S. 2000. *Nature* 403: 267-8

144. Bhalla US, Iyengar R. 1999. *Science* 283: 381-7

145. Takahashi K, Ishikawa N, Sadamoto Y, Sasamoto H, Ohta S, et al. 2003. *Bioinformatics* 19: 1727-9

146. Loew LM, Schaff JC. 2001. *Trends Biotechnol.* 19: 401-6

147. Edwards JS, Ibarra RU, Palsson BO. 2001. *Nat. Biotechnol.* 19: 125-30

148. McAdams HH, Shapiro L. 2003. *Science* 301: 1874-7

149. Alon U, Surette MG, Barkai N, Leibler S. 1999. *Nature* 397: 168-71

150. McAdams HH, Arkin A. 1999. *Trends Genet.* 15: 65-9

Supplemental Material

Supplemental Figure Legends

Figure S1: Combining genomic features using Bayesian networks to predict yeast protein-protein interactions. The first column describes the genomic feature. Protein pairs in the essentiality data can take on three discrete values (EE, both essential; NN, both non-essential; and NE, one essential and one not). The values for mRNA expression correlations range on a continuous scale between –1.0 and +1.0. Functional similarity counts are integers between 1 and ~18 million. We binned the mRNA expression correlation values into 19 intervals and the functional similarity counts into 5 intervals. The second column gives the number of protein pairs with a particular feature value (i.e., 'EE') drawn from the whole yeast interactome (~18M pairs). Columns "pos" and "neg" give the overlap of these pairs with the gold-standard positives and the gold-standard negatives. The last three columns on the right give the conditional probabilities of the feature values – *P*(*feature value*|*pos*) and *P*(*feature value*|*neg*) – and the likelihood ratio *L*, the ratio of these two conditional probabilities.

The column "sum(pos)" shows how many gold-standard positives are among the protein pairs with likelihood ratio greater than or equal to *L*, which can be computed by summing up the values in the column "pos" to the left. The column "sum(neg)" shows the number of gold-standards negatives among the protein pairs with likelihood ratio greater than or equal to *L*. Finally, "sum(pos)/sum(neg)" is a measure of how well each feature predicts protein-protein interactions (given a certain likelihood ratio cutoff).

The likelihood ratios of the individual features can be combined using a Naïve Bayesian network, as explained in Equation (1) in the Appendix. The prior odds were set to 1/600, which corresponds to a very conservative estimate that there are at most 30,000 pairs of proteins in the same complex among the 18 million protein pairs in yeast.

Subunit A	Subunit B	Contact	Far western			Pulldown			Cross-linking	Voting	Bayes
			Miyayo et al.	Ishiguro et al.	Yasui et al.	Acker et al. (A)	Acker et al. (B)	Kimura et al.	Ishiguro et al.		
1	2	■							+	1	0.6
1	3		+			+	+	+	+	5	1.3
1	5	■	+			+	+		+	4	1.5
1	6	■		+		-	+		+	2	2.7
1	8	■		+		+	+		+	4	5.1
1	9	■				-	-			-2	0.1
1	10			+		+	+		-	2	0.0
1	11	■		+		-	-		+	0	0.6
1	12			+					+	2	1.9
2	3	■	+			+	+	+	+	5	1.3
2	5		+			+	+		+	4	1.5
2	6			+		-	-		+	0	0.6
2	8			+		+	+		-	2	0.0
2	9	■				-	-			-2	0.1
2	10	■		+		+	+		+	4	5.1
2	11			+		-	-		+	0	0.6
2	12	■		+					+	2	1.9
3	5		+		+	+	+	+		5	0.9
3	6		-	-	-	+	+	-		-2	0.4
3	8		-	-	-	+	+	+		0	0.3
3	9					-	-	-		3	0.2
3	10	■	-	-	-	+	+	-	+	-1	0.6
3	11	■	-	+	+	+	+	+	+	5	4.2
3	12	■	-	-	-			-		-4	0.2
5	6		-	-		+	-		+	1	0.1
5	8		+	-		+	+		-	1	0.0
5	9					-	-			-2	0.1
5	10		-	-		+	+			0	0.3
5	11		-	-		-	-			-4	0.0
5	12		-	-						-2	0.1
6	8			-		-	-		+	-3	0.1
6	9					-	-			-2	0.1
6	10			-		-	-			-3	0.0
6	11			-		-	-			-3	0.0
6	12			-						-1	0.1
8	9					-	-			-2	0.1
8	10			-		+	-			-1	0.1
8	11			-		-	-			-3	0.0
8	12			-						-1	0.1
9	10					-	-			-2	0.1
9	11					-	-			-2	0.1
9	12										0.4
10	11			-		-	-			-3	0.0
10	12			-						-1	0.1
11	12			-						-1	0.1

	TP	2	6	1	6	7	2	10	8	7
	FN	3	2	2	4	3	2	0	5	6
	TN	6	17	2	14	16	2	3	22	29
	FP	4	5	1	11	9	3	7	9	3
Coverage		15	30	6	35	35	9	20	44	45
Accu-racy [%]	**Individual**	53	77	50	57	66	44	65	68	80
	Voting	53	73	50	66	66	44	55	68	NA
	Bayes	**73**	**87**	**67**	**83**	**83**	**67**	**70**	**80**	**80**

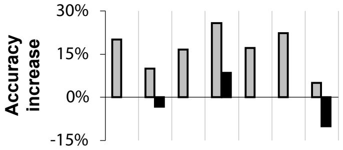
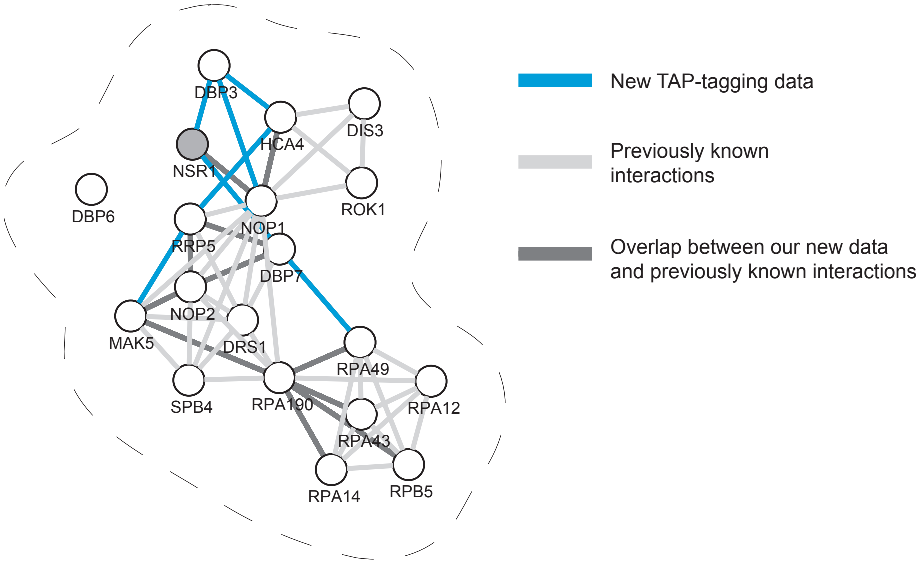


Figure 1



Likelihood ratio	# protein pairs				
	Combined	Essentiality	mRNA expression	MIPS functional	GO biological process similarity
0 - 10	24338971	7554654	15953271	5874302	3125819
10 - 100	221194	0	136353	287503	20467
100 - 1000	25527	0	617	0	0
>= 1000	348	0	0	0	0

Figure 2

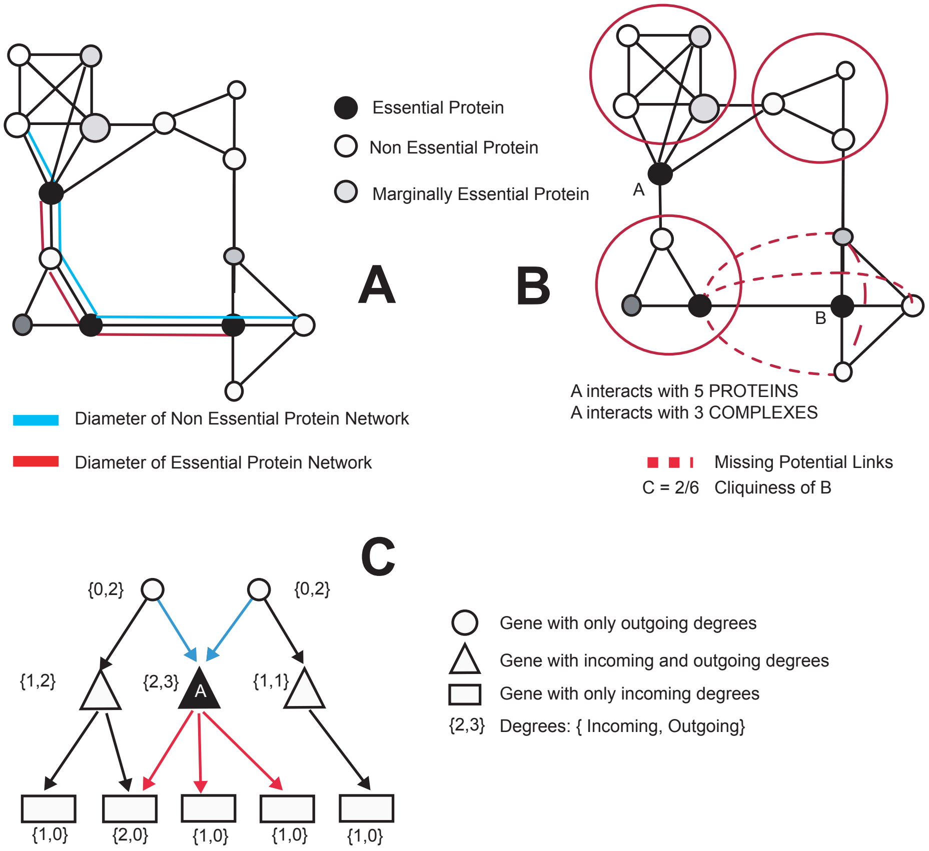
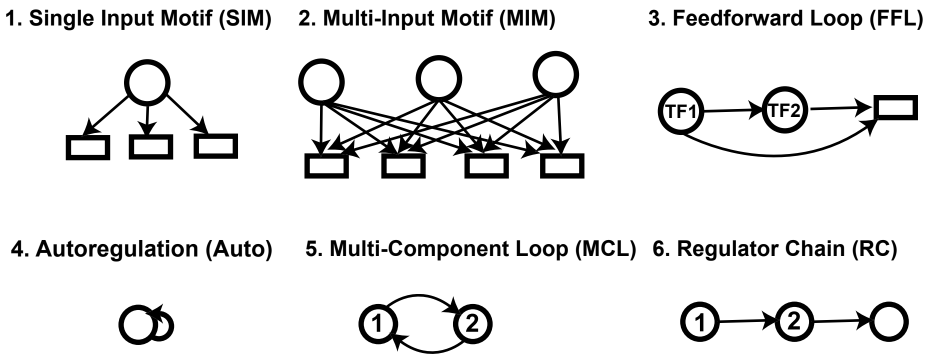

Essential Protein
Non Essential Protein
Marginally Essential Protein
A
B
Diameter of Non Essential Protein Network
Diameter of Essential Protein Network
A
B
A interacts with 5 PROTEINS
A interacts with 3 COMPLEXES
Missing Potential Links
C = 2/6 Cliquiness of B
C
{0,2}
{0,2}
{1,2}
{2,3}
A
{1,1}
{1,0}
{2,0}
{1,0}
{1,0}
{1,0}
Gene with only outgoing degrees
Gene with incoming and outgoing degrees
Gene with only incoming degrees
{2,3} Degrees: { Incoming, Outgoing}

Figure 3



1. Single Input Motif (SIM)

2. Multi-Input Motif (MIM)

3. Feedforward Loop (FFL)

4. Autoregulation (Auto)

5. Multi-Component Loop (MCL)

6. Regulator Chain (RC)

Figure 4

Essentiality		# protein pairs	Gold-standard overlap					P(Ess|pos)	P(Ess|neg)	L
			pos	neg	sum(pos)	sum(neg)	sum(pos)/ sum(neg)			
Values	EE	384,126	1,114	81,924	1,114	81,924	0.014	5.18E-01	1.43E-01	3.6
	NE	2,767,812	624	285,487	1,738	367,411	0.005	2.90E-01	4.98E-01	0.6
	NN	4,978,590	412	206,313	2,150	573,724	0.004	1.92E-01	3.60E-01	0.5
Sum		8,130,528	2,150	573,724	-	-	-	1.00E+00	1.00E+00	1.0

Expression correlation		# protein pairs	Gold standard overlap					P(exp|pos)	P(exp|neg)	L
			pos	neg	sum(pos)	sum(neg)	sum(pos)/ sum(neg)			
Values	0.9	678	16	45	16	45	0.36	2.10E-03	1.68E-05	124.9
	0.8	4,827	137	563	153	608	0.25	1.80E-02	2.10E-04	85.5
	0.7	17,626	530	2,117	683	2,725	0.25	6.96E-02	7.91E-04	88.0
	0.6	42,815	1,073	5,597	1,756	8,322	0.21	1.41E-01	2.09E-03	67.4
	0.5	96,650	1,089	14,459	2,845	22,781	0.12	1.43E-01	5.40E-03	26.5
	0.4	225,712	993	35,350	3,838	58,131	0.07	1.30E-01	1.32E-02	9.9
	0.3	529,268	1,028	83,483	4,866	141,614	0.03	1.35E-01	3.12E-02	4.3
	0.2	1,200,331	870	183,356	5,736	324,970	0.02	1.14E-01	6.85E-02	1.7
	0.1	2,575,103	739	368,469	6,475	693,439	0.01	9.71E-02	1.38E-01	0.7
	0	9,363,627	894	1,244,477	7,369	1,937,916	0.00	1.17E-01	4.65E-01	0.3
	-0.1	2,753,735	164	408,562	7,533	2,346,478	0.00	2.15E-02	1.53E-01	0.1
	-0.2	1,241,907	63	203,663	7,596	2,550,141	0.00	8.27E-03	7.61E-02	0.1
	-0.3	484,524	13	84,957	7,609	2,635,098	0.00	1.71E-03	3.18E-02	0.1
	-0.4	160,234	3	28,870	7,612	2,663,968	0.00	3.94E-04	1.08E-02	0.0
	-0.5	48,852	2	8,091	7,614	2,672,059	0.00	2.63E-04	3.02E-03	0.1
	-0.6	17,423	-	2,134	7,614	2,674,193	0.00	0.00E+00	7.98E-04	0.0
	-0.7	7,602	-	807	7,614	2,675,000	0.00	0.00E+00	3.02E-04	0.0
	-0.8	2,147	-	261	7,614	2,675,261	0.00	0.00E+00	9.76E-05	0.0
	-0.9	67	-	12	7,614	2,675,273	0.00	0.00E+00	4.49E-06	0.0
Sum		18,773,128	7,614	2,675,273	-	-	-	1.00E+00	1.00E+00	1.0

MIPS function similarity		# protein pairs	Gold standard overlap					P(MIPS|pos)	P(MIPS|neg)	L
			pos	neg	sum(pos)	sum(neg)	sum(pos)/ sum(neg)			
Values	1 -- 9	6,584	171	1,094	171	1,094	0.16	2.12E-02	8.33E-04	25.5
	10 -- 99	25,823	584	4,229	755	5,323	0.14	7.25E-02	3.22E-03	22.5
	100 -- 1000	88,548	688	13,011	1,443	18,334	0.08	8.55E-02	9.91E-03	8.6
	1000 -- 10000	255,096	6,146	47,126	7,589	65,460	0.12	7.63E-01	3.59E-02	21.3
	10000 -- Inf	5,785,754	462	1,248,119	8,051	1,313,579	0.01	5.74E-02	9.50E-01	0.1
Sum		6,161,805	8,051	1,313,579	-	-	-	1.00E+00	1.00E+00	1.0

GO biological process similarity		# protein pairs	Gold standard overlap					P(GO|pos)	P(GO|neg)	L
			pos	neg	sum(pos)	sum(neg)	sum(pos)/ sum(neg)			
Values	1 -- 9	4,789	88	819	88	819	0.11	1.17E-02	1.27E-03	9.2
	10 -- 99	20,467	555	3,315	643	4,134	0.16	7.38E-02	5.14E-03	14.4
	100 -- 1000	58,738	523	10,232	1,166	14,366	0.08	6.95E-02	1.59E-02	4.4
	1000 -- 10000	152,850	1,003	28,225	2,169	42,591	0.05	1.33E-01	4.38E-02	3.0
	10000 -- Inf	2,909,442	5,351	602,434	7,520	645,025	0.01	7.12E-01	9.34E-01	0.8
Sum		3,146,286	7,520	645,025	-	-	-	1.00E+00	1.00E+00	1.0

Supplemental Figure S1