Protein interaction networks

Matteo Pellegrini†, David Haynor and Jason M Johnson

The study of protein interactions is playing an ever increasing role in our attempts to understand cells and diseases on a system-wide level. This article reviews several experimental approaches that are currently being used to measure protein–protein, protein–DNA and gene–gene interactions. These techniques have now been scaled up to produce extensive genome-wide data sets that are providing us with a first glimpse of global interaction networks. Complementing these experimental approaches, several computational methodologies to predict protein interactions are also reviewed. Existing databases that serve as repositories for protein interaction information and how such databases are used to analyze high-throughput data from a pathway perspective is also addressed. Finally, current efforts to combine multiple data types to obtain more accurate and comprehensive models of protein interactions are reviewed. It is clear that the evolution of these experimental and computational approaches is rapidly changing our view of biology and promises to provide us with an unprecedented ability to model cells and organisms at a system-wide level.

Expert Rev. Proteomics 1(2), (2004)

CONTENTS

†*Author for correspondence*
Rosetta Inpharmatics LLC,
401 Terry Ave.,
Seattle, WA 98109, USA
Tel.: +1 206 802 6427
Fax: +1 206 802 6411
matteo_pellegrini@merck.com

KEYWORDS:
Bayesian networks, ChIP on chip, phylogenetic profiles, protein interactions, protein interaction databases, networks, Rosetta Stone, synthetic lethality, tandem affinity tag, transcription factor binding sites, two-hybrid

One of the grand challenges for molecular biology is to reconstruct the complete network of protein interactions within cells. This so-called interactome will shed unprecedented light on the inner workings of cellular machinery. Analysis of the network should also permit scientists to select protein targets for therapeutic intervention by understanding the underlying mechanisms of action. Furthermore, interaction maps may reveal how drug–protein interactions lead, through primary or secondary mechanisms, to toxic side effects. Eventually, protein networks may also be used to construct comprehensive dynamic models of molecular interactions within cells, allowing scientists to quantitatively predict the outcome of experiments. However, despite the fact that high-throughput sequencing has facilitated the prediction of proteins coded within a genome, thus providing a list of the interactome's constituents, constructing such powerful models on a genome-wide scale remains a distant goal.

Within cells, proteins interact with other proteins, metabolites and nucleic acids. Our understanding of protein–metabolite interactions is perhaps the most complete, deriving from decades of enzymological studies. Protein–protein interactions have also been measured using a variety of assays, such as immunoprecipitations and the yeast two-hybrid approach. Recently, these techniques have been scaled up to measure interactions on a genome-wide level [1]. High-throughput techniques have also been developed to systematically identify protein complexes using affinity purification techniques followed by mass spectrometry (MS) to sequence proteins [2,3]. Genome-wide protein–DNA interactions may be measured using chromatin immunoprecipitation (ChIP) in conjunction with expression microarrays [4]. Finally, genetic interaction networks, identified by systematically constructed double knockout strains, have been used to provide a different view of functional linkage on a large scale in yeast [5]. Although these approaches, depicted in FIGURE 1, have been systematically applied to the lower eukaryotes, they are only now beginning to be applied on a large scale to measure protein interactions in mammalian cells, where they have greater therapeutic relevance.

 ISSN 1478-9450

Along with experimental approaches to detect protein interactions, computational methods have also been developed. These methods search for pairs of proteins that have coevolved, implying that they are likely to be interacting within the cell. Coevolution may be detected by searching for pairs of proteins that are fused in some organism, found within the same sets of organisms or have similar phylogenetic distances to members of their respective families. Although computationally derived interactions are generally not as reliable as experimentally measured ones, they provide a more complete and accurate understanding of protein interactions in combination with experimental data.

Over the past few years, several protein interaction databases have been developed. These databases are populated by recent high-throughput data and some of the smaller sets of interactions reported in the literature over the past few decades. The latter is an important source of content that significantly enhances the data from high-throughput assays. However, despite advances in natural language processing, it remains challenging to automatically extract information on protein interactions from the literature, and therefore most of these data are still extracted manually. As a result, current protein interaction databases have mined only a small fraction of all the interactions in the literature and cover an even smaller fraction of protein interaction space. Nonetheless, the interactions within these databases often provide clear molecular mechanisms for many important biological processes.

Despite our limited ability to reconstruct a genome-wide protein interaction network for human cells, our attempts to identify smaller interaction modules within the network have proven more successful. Modules consist of conceptually distinct pathways responsible for, among other tasks, the metabolism of small molecules or the transduction of signals across cell membranes. Other modules contain the subunits of protein complexes. Several public databases catalog the extensive knowledge that has accumulated regarding the protein components of these pathways and complexes.

How does the elucidation of protein interaction networks advance drug discovery research? RNA expression microarrays and other high-throughput molecular profiling approaches have become integral to modern drug discovery research; however, the utility of these data to guide drug discovery is sometimes limited by the ability to interpret the biological meaning of the results. Analyzing large data sets from a pathway perspective is one approach that can enhance understanding of the biological mechanisms affected in an experiment. This type of analysis may involve, for example, asking which pathways are perturbed in a disease population with respect to a healthy one, which pathways are activated in adverse responses to a compound, or which pathways distinguish good from bad patient prognoses. Since the number of pathways is significantly smaller than the number of genes, and many pathways are at least partially characterized, this type of analysis can simplify the interpretation of experimental results when many genes are involved.

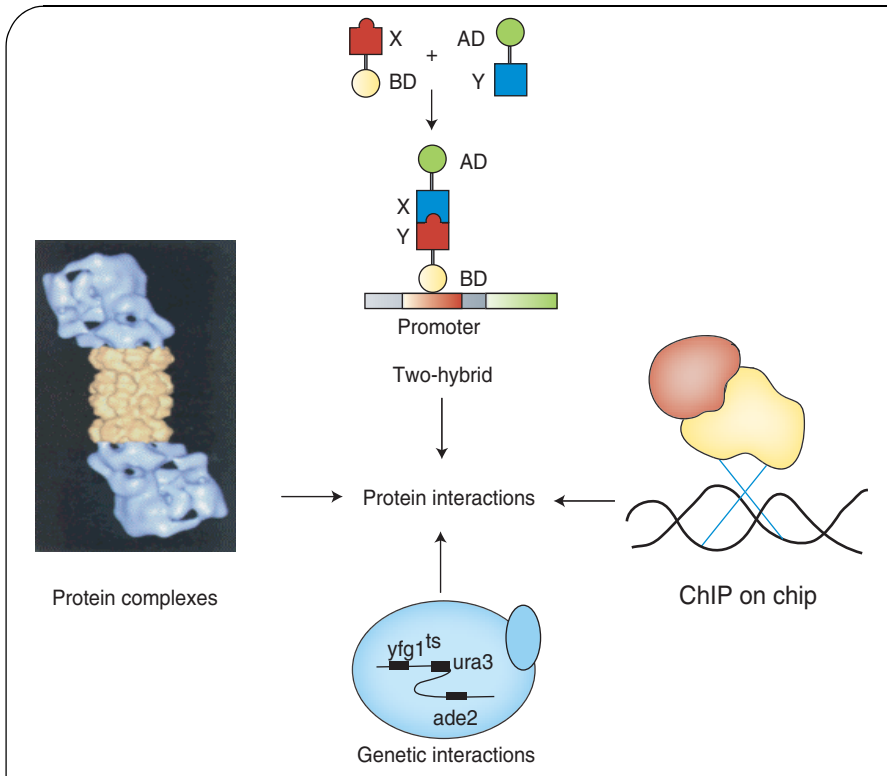

Figure 1. Schematic representation of the four principal experimental methods used to determine protein interactions. The two-hybrid assay to measure protein–protein pair-wise interactions (top, from [102]), the ChIP on Chip assay to measure protein–DNA interactions (right, from [103]), the construction of double knock-out mutants in yeast to measure gene–gene interactions (bottom, from [104], and the tandem affinity purification tag assay to pull down and identify the subunits of protein complexes (left, from [105]).
ChIP: Chromatin immunoprecipitation.

The availability of large data sets that measure protein–protein, protein–DNA and genetic interactions has also created a need for methods to integrate these different data types for a more comprehensive view of protein interactions. For example, recent drug discovery efforts are combining the measurements of expression microarray data with systematically measured genotypes to map the genetic basis of transcriptional regulation and disease [6]. The resulting networks of protein interactions are enhancing efforts to reconstruct the genetic basis of disease and discover novel targets.

Although we are only now beginning to decipher protein interaction networks, the impact that this research is having on drug discovery can already be seen. As we move closer to the goal of quantitatively modeling human cells and cell systems, the emerging protein interaction networks provide the raw materials for early systems modeling efforts.

Experimental measurements of protein interactions

Two techniques are widely used to measure protein–protein interactions in a high-throughput fashion: yeast two-hybrid assays and pull-down experiments on tagged proteins followed by MS to identify interactors. These approaches have been used to develop large-scale protein interaction maps in *Saccharomyces cerevisiae* [1,3], *Helicobacter pylori* [7], *Drosophila melanogaster* [8], and *Caenorhabditis elegans* [9]. Various efforts are also underway to apply these techniques to measure interactions in the human proteome [10]. A more detailed description of these two techniques is presented below.

Other high-throughput techniques have also been developed to measure protein–DNA interactions. The most advanced is the use of ChIP to pull down DNA fragments that are subsequently identified using hybridization microarray chips. This so-called ChIP on Chip approach identifies many of the binding sites of transcription factors, and has been used to map the binding sites of many known yeast transcription factors [4].

Finally, recent efforts to map genetic interaction networks will also be described. These networks identify pairs of genes that buffer each other and therefore only cause a phenotype when deleted in combination. For example, a genetic interaction screen has been performed in yeast to identify genes that are not essential for cell survival when deleted one at a time, but are lethal when deleted in combination [5]. Genetic interactions often occur between genes that participate in the same biological pathway or are physically interacting.

Protein–protein interaction assays

One of the most frequently used techniques to study pair-wise interactions between proteins is the yeast two-hybrid assay. This assay involves the use of GAL4, a transcription factor that in the presence of galactose activates transcription of the *GAL* genes, which encode galactose metabolic proteins. GAL4 contains two domains, an activation domain and a DNA-binding domain. In 1989, Fields and Song suggested that GAL4 hybrids could be used to report protein interactions [11]. The strategy consists of forming two fusions between one protein and the GAL4 activation domain and another protein with the GAL4 DNA-binding domain. If the two proteins interact, then the GAL4 activation domain and the GAL4 DNA-binding domain will be brought into proximity with each other and the reconstituted factor will activate the transcription of a reporter gene, which has been engineered to contain the GAL4 promoter. Thus, detection of the reporter gene in yeast implies that the two proteins have interacted in the assay.

Over the past few years this strategy has been scaled up so that it is now possible to efficiently measure thousands of binding events [1,6–9,12,13]. The most extensive interaction map, measuring the binding of *D. melanogaster* proteins, contains 20,405 interactions among 7048 proteins [8]. The resulting network has regions of high local connectivity, representing interactions between subunits of protein complexes.

Typically, the observation of a single binding event using the yeast two-hybrid assay is not a reliable indication that two proteins interact *in vivo* [14,15]. One reason is that the two interacting proteins are overexpressed in this assay and the observed interaction may not therefore be present in the wild type cells where the concentrations may be significantly lower.

Due to these limitations, various strategies have been developed to identify which of the interactions reported in a two-hybrid screen are likely to be biologically relevant. In order to evaluate whether an interaction is biologically relevant, various supporting information such as annotation, cellular localization and messenger RNA (mRNA) expression levels have been used. The underlying assumption is that true interactions are likely to occur between proteins involved in the same biological process, proteins found in the same cell compartment, and proteins whose mRNA are coexpressed.

For example, Kemmeren and coworkers used expression data to evaluate the fraction of protein interactions pairs from a variety of data sets that are likely to be truly interacting [15]. They discovered that among pairs of proteins with high confidence interactions, 70% were coexpressed. In contrast, for high-throughput protein interaction data sets, the percentage of coexpressed proteins varied from 26 to 56%, implying that interactions measured using high-throughput assays tend to generate a significant number of false positives and that the false-positive rate differs substantially depending on the experimental protocol used. However, by identifying the coexpressed pairs within a large protein–protein interaction screen, one is able to recover the high confidence ones.

Some additional limitations of the two-hybrid approach include the difficulty of detecting interactions involving membrane proteins. To study membrane proteins, one must construct GAL4 fusions with only the extracellular or cytoplasmic domains of membrane proteins, adding an additional level of complexity to the assay and resulting in the under-representation of these domains in a large screen. The protein classes that are least represented in the *D. melanogaster* data are plasma membrane proteins including receptors, ion channels and peptidases [8].

Large-scale two-hybrid screens have been reported for *S. cerevisiae, C. elegans* and *D. melanogaster* but not yet for humans. For drug discovery applications, a map of human interactions will be more valuable than for these model organisms. Nonetheless, interaction maps in model organisms could be useful in human drug discovery. For instance, the *D. melanogaster* map allowed the authors to hypothesize that therapeutic inhibition of calcineurin phosphatases may be an attractive strategy to treat human lymphomas [8].

Experimental measurements of protein complexes

Another approach that can be utilized to map protein interactions is to tag a protein in the cell and then pull down the tagged protein together with other proteins bound to it. The identity of the interacting partners may be revealed using MS. In contrast to the two-hybrid assays, which detect pair-wise interactions between proteins, here an entire protein complex can be identified since many proteins are typically found to interact with the tagged protein.

Two versions of this approach have recently been implemented by groups that set out to map protein interactions in *S. cerevisiae* [2,3]. In the first version, the tandem affinity purification (TAP) tag was used [16]. The tag is coded into the 3′ end of the chosen gene using homologous recombination. This ensures that the tagged protein is expressed at native levels within the cell. The TAP tag has two components that are used for an initial protein purification step, followed by cleavage of the first tag. The second tag is then used for a subsequent purification step. Following affinity purification of the tagged protein, the product is separated using 1D sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). The identity of the proteins in the various bands are then determined using matrix-assisted laser desorption/ionization (MALDI) time-of-flight (TOF) MS.

This methodology has also been applied to map the interactions involved in tumor necrosis factor (TNF)-α signaling in human cells [10]. The authors tagged 32 known and candidate members of the pathway and identified 221 molecular associations. As expected, many associations were found among the subunits of the nuclear factor (NF)-κB and IκB kinase (IKK) complexes. This methodology was able to detect 70% of the known interactions between the components of this pathway. In addition to these, the study revealed additional complexes that were not previously known to be involved in this signaling cascade. The involvement a few of these with the TNF-α signaling pathway was validated using an RNA interference (RNAi) assay; specifically, the suppression of these genes using RNAi modulated the NF-κB-dependent luciferase reporter activity.

In another implementation of these pull-down assays, genes are tagged with the FLAG epitope tag [2]. In contrast to the previous approach, the genes are transiently overexpressed in a plasmid fused to the FLAG epitope tag with heterologous GAL1 or tet promoters. The subsequent steps are similar to those described above: the proteins are affinity purified, passed through an SDS-polyacrylamide gel and then identified using tandem MS fragmentation.

The sensitivity and specificity of these procedures are difficult to determine, but may be approximated by comparing the measured interactions with known ones. For example, in one study, the yeast Munich Information Center for Protein Sequences databse (MIPS) complexes were used as a reference set [17]. Using this set, the TAP tag method generated a coverage (i.e., fraction of a reference set of interactions observed in the data) of 20% and an accuracy (i.e., fraction of observed interactions confirmed by the reference set) of 10% while the FLAG/overexpression approach yielded a coverage of 8% with an accuracy of 2% [18].

For drug discovery applications it is not only important to understand the interaction between subunits of proteins complexes, but also the structure of the complex. A recent study has combined the topological information obtained from pull-down assays with electron microscopy data of the complexes to model their 3D structures [19]. This approach demonstrates how protein–protein interaction networks may eventually contribute to detailed atomic level descriptions of protein complexes, which may be helpful for the development of drugs that interfere with the function of the complexes.

Protein–DNA binding studies

Many proteins act as transcriptional regulators by binding to specific DNA sequences, typically upstream of genes, to affect the rate of transcription. According to the MIPS database [17], 344 transcription factors have been identified within the yeast *S. cerevisiae*. These factors may be identified by virtue of their homology to known transcription factors and to DNA-binding domains. The binding sites of 106 of these transcription factors have recently been mapped [4]. The strategy used to map the binding sites of yeast transcription factors firstly consisted of adding myc epitope tags into the genomic sequence of the C-terminus of each regulator. ChIP was then used to pull down the tagged protein and identify the DNA sequences bound to it by hybridizing these against DNA microarrays. In total the authors were able to identify 3985 high-confidence protein–DNA interactions.

From this set of interactions, the authors were able to reconstruct network motifs and classify them into six types:

- Autoregulatory motifs, in which a transcription factor binds to its own gene
- Multicomponent loops, in which a regulator binds to a gene that affects the transcription of the first regulator
- Feedforward loops, in which two regulators affect a gene, but then one of them also affects the other
- Single input motifs, in which one factor affects multiple genes
- Multi-input motifs, in which multiple factors affect multiple genes
- Motifs where one regulator binds to a gene of another regulator that binds to the gene of another regulator, and so on forming a regulatory chain

This same ChIP on Chip approach has also been applied to study the binding sites of human transcription factors [20–23]. One example of such studies involved the measurement of the binding of three transcription factors known to be important in liver and pancreas: hepatocyte nuclear factor (HNF)-1α, HNF4α, and HNF6 [24]. ChIP assays were used to measure the binding sites of these three factors in the upstream regions of 13,000 human genes. It was found that the three factors bind different promoters in liver cells and pancreatic islet cells, with relatively little overlap. Furthermore, the number of binding sites for each factor varies greatly, from 187 in liver for HNF1α to 910 for HNF4α. The observation that HNF4α binds to a large number of promoters in islet cells provides a mechanistic explanation for recent observations that polymorphisms in *HNF4A* (the gene encoding HNF4α) may increase the risk of Type II diabetes by disrupting the transcriptional program of these cells [25].

However, it should be noted that many current ChIP on chip efforts only look for binding sites in regions that are proximal to the gene and may therefore miss important but distant

enhancer binding sites. For instance, Odom and coworkers constructed arrays that spanned only 700 base pairs upstream and 200 base pairs downstream of genes [24]. Using regularly spaced microarray probes spanning the genome [26], genome-wide ChIP on Chip experiments have also been conducted [20]. Cawley and coworkers looked at the binding of three transcription factors, Sp1, cMyc and p53. They performed ChIP on Chip experiments using arrays that contained on average one probe for every 35bp sequence on chromosomes 21 and 22. They identified a total of 353, 756 and 48 high-confidence binding sites, respectively, for the three factors in Jurkat and HCT116 cells.

The main conclusions drawn from this study are that these three transcription factors probably bind 12,000 (Sp1), 25,000 (cMyc) and 1600 (p53) sites within the genome, far more than one would have anticipated before full genome arrays were available. Furthermore, only 22% of these binding sites are located at the traditional 3′ end of genes. The authors also noticed that these factors regulate the transcription of many noncoding RNAs. These RNAs were shown to be differentially expressed when cells were exposed to retinoic acid and are therefore likely to be biologically important. It is clear that these initial full genome protein DNA binding studies are revealing a far more complete and unexpected view of transcription.

Experimental measurements of genetic interactions

Through systematic deletion of all yeast genes, it was discovered that approximately 80% of genes are nonessential (i.e., yeast cells are able to grow and function when any single one of these genes is deleted) [27]. However, in some cases, the combined knockout of two individually nonessential genes is lethal. Genes that are nonessential individually, but essential when both are mutated, buffer each other's function. These types of relationships have been named genetic interactions.

Recently, advances in robotics have permitted the systematic search of genetic interactions in *S. cerevisiae* [5,28]. By crossing strains of yeast with single mutations, it is possible to generate all possible double mutants. By screening these double mutants for viability, it is possible to identify all genetic interactions.

The most comprehensive network of genetic interactions in yeast to date was generated by crossing 132 strains missing a query gene with 4700 strains lacking one of the nonessential genes. The network generated from these crosses contains about 4000 interactions among approximately 1000 genes [28]. In this network the average query gene has 34 genetically interacting partners. This is larger than the average of eight interactions found in yeast two-hybrid screens, but only a subset of genetic interactions are between physically binding proteins.

One of the most significant properties of genetic interactions is that they frequently occur between genes that act within the same biological process. For example, 12% of genetic interactions are between genes that belong to the same Gene Ontology biological process category, an observation that has an associated p value of 1e-322 of occurring by chance. It is also found that genetic interactions often occur between subunits of a protein complex (p = 1e-68), proteins with the same subcellular localization (p = 1e-70) and proteins with the same mutant phenotype (p = 1e-316) [28].

Much of the heritable mortality and morbidity in the world is caused by multilocus diseases. The study of genetic interactions could shed light on these human diseases. For example, in cystic fibrosis it is known that secondary mutations to genes other than the cystic fibrosis transmembrane conductance regulator may aggravate the symptoms of the disease. Although genetic networks have not been measured in human cells, it is possible that RNA interference techniques may be used to suppress the expression of pairs of genes to map these types of interactions.

Computational methods for predicting protein interactions

Along with the experimental techniques to measure protein interactions listed above, several computational approaches for predicting protein interactions have also been developed over the past few years. These approaches rely on the hypothesis that interacting proteins tend to evolve in a constrained fashion, since mutations in one protein may affect its ability to interact and thus affect another protein. Two techniques that one may use to identify protein interactions computationally are described below: the detection of protein fusions and phylogenetic studies that reveal protein interactions.

Protein fusions

Protein sequence alignments are traditionally used to identify homologous proteins. However, it is possible to slightly modify the typical alignment procedure to identify protein fusions [29,30]. In this case, one seeks two nonhomologous proteins that align to different regions of another protein. In other words, these two proteins are essentially fused into a single longer polypeptide chain. The longer protein has been dubbed the Rosetta Stone protein, because it often reveals that the two fused proteins are interacting.

However, this approach may also yield a significant number of spurious fusion events due to the fact that protein sequences contain conserved domains. Some of these domains are repeated hundreds of times throughout a protein sequence database. For example, the zinc finger binding domain is found in hundreds of human proteins. Since these domains vary in sequence, they often appear to be nonhomologous by standard alignment techniques. This may lead to a spurious fusion observation between two proteins that contain different zinc finger domains that align to another protein with multiple zinc fingers.

To screen out these false fusion events, it has recently been suggested that one may use the hypergeometric distribution in EQUATION 1 [31]:

$$P(k|n, m, N) = \frac{\binom{n}{k}\binom{N-n}{m-k}}{\binom{N}{m}}$$

where k represents the number of Rosetta Stone proteins found between two nonhomologous proteins A and B, n the number of homologs of protein A, m the number of homologs of protein B and N the total number of proteins within the database. According to this function, proteins that have many homologs are less likely to be involved in real protein fusion and are more likely to appear fused since they contain a commonly found domain.

Predicting protein interactions using phylogenetic analysis

Over the past few years, the numbers of fully sequenced genomes has grown dramatically to include over 100 organisms. From the analysis of homologs across genomes it is possible to construct phylogenetic profiles [32]. These are binary arrays computed for each protein in a genome and encode whether a homolog of the protein is present in any of the fully sequenced genomes. Proteins with similar phylogenetic profiles are effectively coevolving, since they are often found together in organisms. It has been found that subunits of cellular complexes often coevolve. Therefore, this technique may be used to identify protein–protein interactions [33].

However, it is difficult to discern subtle differences in the evolution of paralogs using phylogenetic profiles. Paralogs are homologous proteins that have emerged by duplication within a species and would necessarily have very similar phylogenetic profiles. In order to study the subtle differences in interactions and function between paralogs, it is necessary to describe their evolution more completely using standard phylogenetic distance estimation techniques.

In order to estimate the evolutionary distance within a group of homologous proteins, one must first construct a multiple sequence alignment. This may be accomplished using the ClustalW program [34]. Once the multiple alignment has been built it is possible to estimate the evolutionary distance between any two sequences using the alignment score. It is then possible to deduce the different interacting partners of paralogs by comparing the distance matrices of two protein families that are known to contain interacting pairs.

One might imagine that if two proteins interact, the evolution of one might be correlated with the other. For instance, mutations that occur on a ligand might be compensated by mutations to its receptor in order to maintain the ligand–receptor binding affinity. This phenomenon has in fact been demonstrated in the case of chemokines and their associated receptors [35,36]. By correctly aligning the distance matrices of ligands and ligand receptors, it is possible to partially reconstruct which ligand is likely to bind which receptor [37,38].

Databases of protein interactions

Several databases that have been developed to store protein interactions have been reported in the literature. Some of these are general repositories of varied types of interactions such as the Biomolecular Interaction Database (BIND) [39]. Other databases, such as the Database of Interacting Proteins (DIP) [40], IntAct [41], and the Molecular Interaction Database (MINT) [42], report only protein–protein interactions. The contents of DIP, a representative protein–protein interaction database, are listed in TABLE I. Although for certain organisms (e.g., *S. cerevisiae*) the map of protein interactions is quite extensive, it is still very sparse for humans.

Table 1. Contents of the Database of Interacting Proteins in 2004 [40].

Organism	Proteins	Interactions	Experiments
Drosophila melanogaster	7052	20,988	21,012
Saccharomycs cerevisiae	4749	15,642	19,116
Caenorhabditis elegans	2638	4030	4075
Helicobacter pylori	710	1425	1425
Homo sapiens	896	1371	1989
Escherichia coli	421	516	971
Mus musculus	193	284	383
Rattus norvegicus	84	107	154

There are a several factors that limit the coverage of these databases. The primary factor is that the experiments reported in the literature measure only a fraction of all biologically relevant protein interactions. Traditionally, these interactions were measured using low-throughput techniques that permitted the measurement of only one or a few interactions at a time. Although these techniques are very accurate, they cover only a tiny fraction of the expected interactome. As previously discussed, high-throughput protein–protein interaction assays are now being performed on a variety of organisms. Although these have dramatically expanded our coverage of protein interaction space, they are still measuring only a small fraction of biologically relevant interactions. Furthermore, as discussed above, the approaches generate high false-positive and -negative rates, in part since they are often measured under conditions that are different from those *in vivo*.

The other factor that complicates the collection of protein interactions from the literature is that there are no reliable automated procedures for accomplishing this. Protein interaction data are laboriously extracted from the literature by scientists one article at a time. This manual approach has to date captured reported interactions in only a small fraction of the 12 million abstracts in PubMed [101]. Furthermore, usually only the abstracts of the articles are searched for information and not the full text.

However, several techniques have been developed to aid scientists in selecting which papers are likely to report a protein–protein interaction. One such approach measures the frequency of keywords that discriminate papers that report protein interactions from those that do not [43]. Using Bayesian statistics, they

estimate the likelihood that a paper reports a protein interaction based on the presence of the hundred or so discriminating words in the abstract. By prioritizing papers according to these criteria, the input of protein interactions into DIP has been rendered more efficient. A similar approach named PreBIND [44] is used to select papers that are likely to contain protein interaction data to be inserted into the BIND database.

A more automated but less precise approach for identifying protein interactions in text is to look for the co-occurrence of protein names. A pair of proteins that often appears together in abstracts is likely to be associated in the same biological process. A fraction of these might even be physically interacting. It is important yet difficult in this type of analysis to account for the multiple synonyms that are given to each protein when cataloguing the coincidence of proteins in abstracts. The results of one of these analyses is reported in the PubGene database [45].

In the near future, the extraction of protein interaction data from the literature should be facilitated by the introduction of a standard data model for interactions. Recently, a model has been proposed by the Proteomics Standards Initiative, which is a working group of the Human Proteome Organization [46]. This standard has been accepted by most of the databases described above, and should ultimately be required for the publication of protein interaction data in the literature.

Databases of protein pathways

Although it is important to reconstruct the network of pair-wise interactions between proteins, it is often simpler but still useful to classify proteins into sets that participate in a common biological process. These sets correspond to partially distinct biological modules that perform a specific function [47]. These could be proteins that are the subunits of a protein complex, components of a signal transduction pathway or enzymes that act on related metabolites and comprise a biochemical pathway. Although it is clearly useful to know how proteins within these sets are interacting among themselves, for certain types of analyses described below, this level of detail is not necessary.

One of the most widely used classification schemes of this type is that developed by the Gene Ontology Consortium [48]. As a part of this effort, three ontologies have been developed to group proteins according to their biological process, molecular function and cellular component. Each ontology represents relationships between terms through a directed acyclic graph. These graphs allows scientists to describe complex relationships between terms that are increasingly more detailed and are associated with decreasing numbers of proteins. An example of a simple ontology is shown in BOX 1 for the histidine biosynthesis pathway, where, for instance, one chain of terms from the most general to the most specific is the following: physiological process, metabolism, amine metabolism, amino acid metabolism, histidine family amino acid metabolism, and histidine metabolism.

Another valuable resource for information on protein pathways is the Kyoto encyclopedia of genes and genomes [49]. This database contains dozens of pathways spanning a broad range of biological processes: metabolism, genetic information processes and human diseases. For each pathway, the enzymes involved are presented and mapped to proteins in a specific organism. The pair-wise relationships between proteins and metabolites are also contained in the pathway descriptions.

Box 1. Ontology for the histidine biosynthesis pathway.

$Gene_Ontology	GO:0003673
<biological_process	GO:0008150
% physiological process	GO:0007582
% metabolism	GO:0008152
% amine metabolism	GO:0009308
% amino acid metabolism	GO:0006520
% carboxylic acid metabolism	GO:0019752
% amino acid and derivative metabolism	GO:0006519
% histidine family amino acid metabolism	GO:0009075
% histidine metabolism	GO:0006547
% amino acid and derivative metabolism	GO:0006519
% amino acid metabolism	GO:0006520
% carboxylic acid metabolism	GO:0019752
% amine metabolism	GO:0009308
% histidine family amino acid metabolism	GO:0009075
% histidine metabolism	GO:0006547
% organic acid metabolism	GO:0006082
% carboxylic acid metabolism	GO:0019752
% amino acid metabolism	GO:0006520
% amino acid and derivative metabolism	GO:0006519
% amine metabolism	GO:0009308
% histidine family amino acid metabolism	GO:0009075
% histidine metabolism	GO:0006547
<cellular_component	GO:0005575
<molecular_function	GO:0003674

Pathway enrichment analysis

Many high-throughput technologies used in drug discovery research generate large lists of genes or proteins that must be further analyzed in order to generate hypotheses from the experiments. For example, the list of significantly perturbed genes in an expression microarray experiment measuring the

response to a drug in a rat may be analyzed to understand if pathways associated with known toxicities have been perturbed. In general, lists of proteins or genes could come from pull-down assays, expression microarray experiments or from MS proteomics studies.

One approach for annotating lists of genes is to compare them with previously annotated gene sets from pathway databases. In other words, if a list of genes upregulated in an expression microarray experiment overlaps significantly with the Gene Ontology histidine biosynthesis genes, one would conclude that the experimental condition used has stimulated the activity of this pathway. An example of an application that performs this type of analysis is the Expression Analysis Systematic Explorer (EASE) [50]. Among other capabilities, this tool measures the overlap between and initial list of genes with Gene Ontology biological process categories. The significance of the overlap is calculated using a hypergeometric probability distribution, to estimate the probability of finding the observed overlap by chance. As an example, the authors automatically computed the Gene Ontology terms associated with a gene expression study by Kayo and coworkers on the influence of aging and caloric restriction to the transcriptional profile of skeletal muscle in rhesus monkeys [51]. They find that the terms computed with EASE (mitochondrion and electron transport) matched the terms Kayo and coworkers had found through a literature search. However, in contrast to the approximately 200 h required for the literature search, EASE was able to perform the analysis automatically in just a few minutes.

Since one can associate a given gene list from an expression experiment with known pathways, it is also possible to identify pathways that are differentially expressed between two patient populations (e.g., patients with and without a specific disease). This type of analysis was recently performed on data collected from healthy and diabetic patients [52]. The approach used was termed Gene Set Enrichment Analysis, and attempted to identify the pathway that contained the most differentially expressed genes between the two populations. The analysis identified the oxidative phosphorylation pathway as the most differentially expressed. Further analysis demonstrated that the transcription factor PPAR-γ coactivator 1α, mutations in which correlate with diabetes, is a regulator of this pathway.

Reconstructing & predicting networks from multiple data sources

As discussed above, multiple high-throughput experimental techniques have been developed to study protein interactions. As a result, integration of these data has emerged as an important field of research in its own right. It seems clear that by combining disparate data sets, a more accurate and comprehensive view of protein interactions will emerge.

A simple Bayesian approach to combine multiple data types to reconstruct yeast protein–protein interaction networks has been described by Jansen and coworkers [53]. They accumulated four types of data that could be analyzed to reconstruct interacting proteins: yeast two-hybrid screens, mRNA expression arrays, Gene Ontology terms, pathway annotation from the MIPS database [17], and gene knockout phenotypes. They established a set of true positive interactions (proteins in the same MIPS complex) along with a set of true negative ones (proteins localized in different cellular compartments). For each method, it was then possible to compute the ratio of the probability that the observed relationship between two proteins was found between true positive and false-positive pairs. These ratios for each data type are then multiplied to yield the final likelihood ratio that two proteins are interacting:

$$O_{post} = \left(\prod_{i=1}^{4} \frac{P(f_i|pos)}{P(f_i|neg)} \right) \frac{P(pos)}{P(neg)}$$

The authors demonstrate that the interactions predicted by this combined approach are of higher accuracy and coverage than those generated by any individual method. For example, interactions with a posterior likelihood threshold greater than ten generated true positive to false-positive ratios of 1 for the combined data and less than 0.1 for the individual data sets [53]. These types of data integration models will likely be species specific. The methodology presented above to integrate disparate data types yields improved predictions of which proteins are interacting but does not allow one to generate hypotheses about the directionality of these interactions. For example, if we inhibit the expression of a gene using RNAi, we would like to know which other genes will change expression levels. In order to generate such predictions, one needs to construct a network with directed edges that imply that changes to one gene cause changes in the other.

An example of the use of combined data to infer directed interaction networks is that presented by Zhu and coworkers [6]. Their approach exploits the collection of high-throughput genotyping data in combination with expression profiling of tissues in inbred mouse strains [54]. Combining these data allows them to infer correlations between genotypes and gene expression levels and thus to generate expression quantitative trait loci (eQTL). These eQTL indicate polymorphisms that affect the expression of genes. By observing how polymorphisms differentially affect coexpressed genes, the authors infer which of the two genes is likely to be causing transcriptional changes in the other. These causality relationships are then used as prior information to bias the construction of Bayesian networks built from expression and genotype data from crosses of inbred mouse strains. Zhu and coworkers are able to demonstrate the predictive power of their network by considering the subnetwork downstream of 11β hydroxysteroid dehydrogenase (HSD11). Previous studies demonstrated that an inhibitor of HSD11 caused significant changes in the expression of 176 genes. Of the 33 genes downstream of HSD11 in these networks, 16 overlapped the set of 176. The probability of this overlap occurring by chance is only 1.1e(-5), demonstrating that the network can be used to predict the outcome of drug perturbation experiments.

Conclusions

Protein networks present a detailed view of molecular interactions and the molecular basis of biological processes. Although the false-positive and -negative rates for networks generated from high-throughput methods are currently high, new experimental techniques and new methods for integrating multiple interacting data types will allow these networks to become powerful predictive tools. In order to use this knowledge to impact human disease research, we must understand how genetic or environmental perturbations to these networks generate disease phenotypes. Subsequently, we must uncover how perturbations to individual components of the network using drugs can reduce or eliminate the disease phenotype.

The study of genetics has been one of the primary tools used to infer how changes to individual genes, and ultimately protein networks, causes disease. More recently, high-throughput RNAi assays have permitted the mapping of perturbations in gene expression to phenotypic measurements. For example, a recent genome-wide RNAi screen of *D. melanogaster* embryonic hemocyte (blood cell) lines revealed 438 genes that showed strong growth phenotypes [55]. Among these, 50 had homology to genes linked to disease in humans, suggesting that the study of model organisms may be valuable to study human diseases.

The integration of information from large-scale protein interaction networks with phenotypic screens could fundamentally change our drug discovery process. It is possible that such techniques could in the future generate a large number of new therapeutic intervention points leading to novel treatments of diseases.

Expert opinion

Comprehensive surveys of protein interaction maps are now accessible through high-throughput technologies. The first surveys looked at protein–protein interactions while more recently protein–DNA and gene–gene interaction maps have also been measured. Although most of the initial work was performed in *S. cerevisiae*, we are now seeing the first interaction maps for more complex organisms such as *D. melanogaster*. In parallel to this experimental work, computational methods have also been developed to predict protein interactions from protein sequence data. A large emphasis of computational work is now being devoted to the integration of all these data to arrive at the most accurate and comprehensive networks possible. Nonetheless, even the most sophisticated efforts are still only beginning to glimpse protein interactions in detail. Although these networks are providing us with global views of cellular organization, they are not yet developed enough to systematically predict the outcomes of perturbation experiments.

Five-year view

The major milestones that will likely be met during the next 5 years will be the application of high-throughput technologies to measure interactions in human cells. We will most likely have measured genome-wide networks of protein–protein, protein–DNA, protein–metabolite and gene–gene interactions in human cells within 5 years. This data, when coupled with more sophisticated network modeling approaches than those used today, will generate predictive networks. Predictive networks will be used to model the effect of perturbations on the cell. These models will likely be able, with a reasonable degree of accuracy, predict the outcome of RNA interference experiments or the effect of adding drugs to cells. If so, they will prove to be an indispensable tool for the development of novel therapeutics.

Key issues

- High-throughput experimental techniques have been developed to measure protein–protein, protein–DNA and gene–gene interactions.
- Computational methods allow us to predict protein interactions from the study of protein fusions and coevolution.
- Several repositories of protein interaction data now exist, for example, the Database of Interacting Proteins and the Biomolecular Interaction Database.
- Gene set enrichment analysis may be used to identify biochemical pathways that are active within a data set.
- Multiple data types may be combined using Bayesian networks to generate more accurate and comprehensive views of protein interactions than those provided by a single data type.
- Our understanding of global protein interaction networks is advancing at a rapid pace and promises to revolutionize our understanding of system-wide molecular biology.

References

Papers of special note have been highlighted as:

- • of interest
- •• of considerable interest

1 Uetz P, Giot L, Cagney G *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403(6770), 623–627 (2000)

2 Ho Y, Gruhler A, Heilbut A *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415(6868), 180–183 (2002).

3 Gavin AC, Bosche M, Krause R *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415(6868), 141–147 (2002).

4 Lee TI, Rinaldi NJ, Robert F *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298(5594), 799–804 (2002).

5 Tong AH, Evangelista M, Parsons AB *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294(5550), 2364–2368 (2001).

6 Zhu J, Lum PY, Lamb J *et al*. An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res.* 105(2–4), 363–374 (2004).

•• **Demonstrates that causal networks that predict the outcome of drug perturbations may be constructed from expression and expression quantitative trait loci data.**

7 Rain JC, Selig L, De Reuse H *et al.* The protein–protein interaction map of *Helicobacter pylori. Nature* 409(6817), 211–215 (2001).

8 Giot L, Bader JS, Brouwer C *et al.* A protein interaction map of *Drosophila melanogaster. Science* 302(5651), 1727–1736 (2003).

9 Li S, Armstrong CM, Bertin N *et al.* A map of the interactome network of the metazoan *C. elegans. Science* 303(5657), 540–543 (2004).

10 Bouwmeester T, Bauch A, Ruffner H *et al.* A physical and functional map of the human TNF-α/NF-κB signal transduction pathway. *Nature Cell. Biol.* 6(2), 97–105 (2004).

• **First application of high-throughput protein interaction assays to human cells.**

11 Fields S, Song O. A novel genetic system to detect protein–protein interactions. *Nature* 340(6230), 245–246 (1989).

12 Ito T, Tashiro K, Muta S *et al.* Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA* 97(3), 1143–1147 (2000).

13 Ito T, Tashiro K, Kuhara T. Systematic analysis of *Saccharomyces cerevisiae* genome: gene network and protein-protein interaction network. *Tanpakushitsu Kakusan Koso.* 46(Suppl. 16) 2407–2413 (2001).

14 Deane CM, Salwinski L, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment of the reliability of high-throughput observations. *Mol. Cell. Proteomics* 1(5), 349–356 (2002).

15 Kemmeren P, van Berkum NL, Vilo J *et al.* Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell.* 9(5), 1133–1143 (2002).

16 Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M, Seraphin B. A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* 17(10), 1030–1032 (1999).

17 Mewes HW, Amid C, Arnold R *et al.* MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.* 32, D41–D44 (2004).

18 von Mering C, Krause R, Snel B *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 417(6887), 399–403 (2002).

19 Aloy P, Ciccarelli FD, Leutwein C *et al.* A complex prediction: three-dimensional model of the yeast exosome. *EMBO Rep.* 3(7), 628–635 (2002).

20 Cawley S, Bekiranov S, Ng HH *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116(4), 499–509 (2004).

•• **Analysis of transcription factor binding using whole genome chips of two chromosomes.**

21 Li Z, Van Calcar S, Qu C, Cavenee WK, Zhang MQ, Ren B. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl Acad. Sci. USA* 100(14), 8164–8169 (2003).

22 Ren B, Cam H, Takahashi Y *et al.* E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev.* 16(2), 245–256 (2002).

23 Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev.* 16(2), 235–244 (2002).

24 Odom DT, Zizlsperger N, Gordon DB *et al.* Control of pancreas and liver gene expression by HNF transcription factors. *Science* 303(5662), 1378–1381 (2004).

25 Barroso I, Luan J, Middelberg RP *et al.* Candidate gene association study in Type 2 diabetes indicates a role for genes involved in β-cell function as well as insulin action. *PLoS Biol.* 1(1), E20 (2003).

26 Kapranov P, Cawley SE, Drenkow J *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296(5569), 916–919 (2002).

27 Winzeler EA, Shoemaker DD, Astromoff A *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285(5429) 901–906 (1999).

28 Tong AH, Lesage G, Bader GD *et al.* Global mapping of the yeast genetic interaction network. *Science* 303(5659), 808–813 (2004).

•• **Large-scale map of gene–gene interactions in yeast.**

29 Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein–protein interactions from genome sequences. *Science* 285(5428), 751–753 (1999).

30 Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402(6757), 86–90 (1999).

31 Marcotte CJ , Marcotte EM. Predicting functional linkages from gene fusions with confidence. *Appl. Bioinformatics* 1(2), 93–100 (2002).

32 Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* 96(8), 4285–4288 (1999).

33 Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.* 5(5), R35 (2004).

34 Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22(22), 4673–4680 (1994).

35 Goh CS, Cohen FE. Coevolutionary analysis reveals insights into protein–protein interactions. *J. Mol. Biol.* 324(1), 177–192 (2002).

36 Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. Coevolution of proteins with their interaction partners. *J. Mol. Biol.* 299(2), 283–293 (2000).

37 Ramani AK, Marcotte EM. Exploiting the coevolution of interacting proteins to discover interaction specificity. *J. Mol. Biol.* 327(1), 273–284 (2003).

38 Gertz J, Elfond G, Shustrova A *et al.* Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics* 19(16), 2039–2045 (2003).

39 Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 31(1), 248–250 (2003).

40 Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The database of interacting proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451 (2004).

41 Hermjakob H, Montecchi-Palazzi L, Lewington C *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res.* 32, D452–D425 (2004).

42 Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G. MINT: a Molecular INTeraction database. *FEBS Lett.* 513(1), 135–140 (2002).

43 Marcotte EM, Xenarios I, Eisenberg D. Mining literature for protein–protein interactions. *Bioinformatics* 17(4), 359–363 (2001).

44 Donaldson I, Martin J, de Bruijn B *et al.* PreBIND and textomy-mining the biomedical literature for protein–protein interactions using a support vector machine. *BMC Bioinformatics* 4(1), 11 (2003).

45 Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.* 28(1), 21–28 (2001).

46 Hermjakob H, Montecchi-Palazzi L, Bader G *et al.* The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. *Nature Biotechnol.* 22(2), 177–183 (2004).

47 Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature* 402(Suppl. 6761), C47–C52 (1999).

48 Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genet.* 25(1), 25–29 (2000).

49 Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280 (2004).

50 Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. *Genome Biol.* 4(10), R70 (2003).

51 Kayo T, Allison DB, Weindruch R, Prolla TA. Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys. *Proc. Natl Acad. Sci. USA* 98(9), 5093–5098 (2001).

52 Mootha VK, Lindgren CM, Eriksson KF *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are co-ordinately downregulated in human diabetes. *Nature Genet.* 34(3), 267–273 (2003).

•• **Interesting analysis of pathway activation in diabetic patients.**

53 Jansen R, Yu H, Greenbaum D *et al.* A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* 302(5644), 449–453 (2003).

•• **Computational method to combine varied data types to arrive at protein interaction predictions.**

54 Schadt EE, Monks SA, Drake TA *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422(6929), 297–302 (2003).

•• **Simultaneous measurement of expression and genotypes reveals the genetics of transcription.**

55 Boutros M, Kiger AA, Armknecht S *et al.* Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science* 303(5659), 832–835 (2004).

Website

101 PubMed
www.ncbi.nih.gov
(Viewed July 2004)

102 Two-hybrid assay
www.bioteach.ubc.ca/MolecularBiology/AYeastTwoHybridAssay/yeast%20two-hybrid%20transcription.gif
(Viewed July 2004)

103 Figure three ChIP on chip
http://mcardle.oncology.wisc.edu/farnham/images/chips.jpg
(Viewed July 2004)

104 Figure three genetic interactions
www.nature.com/nrg/journal/v2/n9/images/nrg0901_659a_f5.gif
(Viewed July 2004)

105 Figure three chemical structure
http://myhome.hanafos.com/~s9euno/fig3/proteasome.gif
(Viewed July 2004)

Affiliations

- *Matteo Pellegrini*
Rosetta Inpharmatics LLC, 401 Terry Ave, Seattle, WA 98109, USA
Tel.: +1 206 802 6427
Fax: +1 206 802 6411
matteo_pellegrini@merck.com
matteo90024@yahoo.com
- *David Haynor*
Rosetta Inpharmatics LLC, 401 Terry Ave, Seattle, WA 98109, USA
Tel.: +1 206 802 6460
Fax: +1 206 802 6411
david_haynor@merck.com
- *Jason M Johnson*
Rosetta Inpharmatics LLC, 401 Terry Ave, Seattle, WA 98109, USA
Tel.: +1 206 802 6449
Fax: +1 206 802 6411
Jason_johnson@merck.com