What do we learn from high-throughput protein interaction data?

Björn Titz, Matthias Schlesner and Peter Uetz

The biological significance of protein interactions, their method of generation and reliability is briefly reviewed. Protein interaction networks adopt a scale-free topology that explains their error tolerance or vulnerability, depending on whether hubs or peripheral proteins are attacked. Networks also allow the prediction of protein function from their interaction partners and therefore, the formulation of analytical hypotheses. Comparative network analysis predicts interactions for distantly related species based on conserved interactions, even if sequences are only weakly conserved. Finally, the medical relevance of protein interaction analysis is discussed and the necessity for data integration is emphasized.

Keywords: high-throughput screening, interactomics, networks, protein–protein interactions, proteomics

Currently, more than 150 bacterial and approximately 15 eukaryotic genomes have been completely sequenced [101]. These sequencing projects provide us with a wealth of information about these organisms. Theoretically, most gene products of these genomes can be predicted from their sequence. Nevertheless, the biochemical activities and biological roles of many gene products remain unclear. Surprisingly, even in new genome sequences approximately one third of the genes cannot be annotated functionally, either because there is no unambiguous homology or homologous genes lack sufficient annotation.

High-throughput functional analysis appears to be the perfect tool to turn the significant number of uncharacterized open reading frames (ORFs) into biological knowledge. Although high-throughput screening (HTS) usually fails to yield a detailed understanding of a protein’s function, it often provides the first evidence for function and therefore, an in-route to further characterization. Currently established HTS methodology includes expression profiling using DNA microarray technology, systematic knockout studies, high-throughput localization studies and protein–protein interaction mapping approaches [1].

This review focuses on protein–protein interaction mapping (interactomics), mainly by two-hybrid approaches. Three questions will be addressed:

• What can we learn from the interaction data generated for several organisms?
• What other information is needed to derive biological conclusions from these data?
• How can such additional data improve our conclusions?

Biological significance of protein–protein interactions

Protein–protein interactions greatly expand the flexibility of proteins beyond their individual activities. For example, the dimeric transcription factors Myc and Max must associate in order to recognize their DNA-binding motif. The Myc/Max dimer allows regulation by altering the concentration of each protein but also by the expression of competitive inhibitors, such as Mad, which binds to and blocks Max. Such combinatorial regulation also expands evolutionary flexibility since each gene’s encoded binding partner can duplicate. These additional proteins can adopt different specificities and eventually biological roles. For an extensive discussion of...
protein interactions and their biological significance the reader is referred to standard textbooks of molecular biology [35].

**Generation of protein–protein interaction data**

Although a number of methods are available for high-throughput analysis of protein-protein interactions, the most commonly used are the yeast two-hybrid (Y2H) system and a combination of protein-complex purification with subsequent analysis by mass spectrometry (MS) [4–6].

The first genome-wide two-hybrid screen was performed by Bader and coworkers for the study of protein interactions in bacteriophage T7 [4]. The first genome-wide protein-protein interaction studies of a free-living organism have been published by Uetz and coworkers [5] and Ito and coworkers [6]. These studies also used the Y2H system and obtained high confidence maps of approximately 3800 and 2200 unique interactions, respectively.

Protein complex purifications from these organisms have not been carried out successfully on a larger scale, although this may be possible with improved protocols and MS sensitivity. No matter how they are generated, interaction data have been used by both experimentalists and theorists for further analysis. A breakdown of such uses is shown in Figure 1 and discussed below in more detail.

**Reliability of high-throughput data**

Before conclusions from high-throughput interaction data can be drawn, it is necessary to briefly discuss the quality of available data sets.

No method is able to identify all protein-protein interactions. That is, each experimental strategy generates a significant number of false negatives. The sources of this systematic error are poorly understood. Two-hybrid false negatives may be caused by steric effects due to the use of two fusion proteins (two-hybrid) or it may involve weak interactions within complexes that require cooperative effects to be stabilized and therefore may generate a two-hybrid signal [23]. Conversely, a major bottleneck for MS analysis are low abundance proteins and proteins that are only weakly associated with protein complexes and hence tend to get lost during purification. False positives are usually a more serious problem since they result in erroneous data and thus misleading conclusions. In Y2H studies, some bait constructs activate the reporter gene without interacting with a prey and so may generate large numbers of technical false positives. On the other hand, biological false positives represent true interactions that take place in the Y2H system but have no biological relevance [24]. A case in point are interacting proteins that are usually expressed in different cell types.

Several approaches were used to minimize the number of false positives in high-throughput studies. Uetz and coworkers [5] eliminated two-hybrid interactions that could not be reproduced, while Ito and coworkers [6] defined interacting protein pairs found three or more times as the (supposedly reliable) core data set. More elaborate statistical scores were proposed by Rain and coworkers [25] for the Helicobacter interaction map and by Bader and colleagues [26] for yeast and other data sets.

Rain and coworkers screened bait proteins against a genomic fragment prey library and considered overlapping prey fragments as the most reliable. This approach combines reproducibility and identifies the interacting-domain at the same time.
The critical point of any attempt to estimate the number of true and false positives in an HTS interaction study is the choice of the true positive data set against which the new interactions are evaluated. Bader and coworkers used the data set of known protein complexes to derive other parameters that allow the scoring of Y2H data [19]. A similar statistical model was applied to the whole Drosophila data set resulting in a high confidence protein interaction network, which the authors estimated to retain 40% interactions of biological significance [19].

Edwards and coworkers selected known interactions from 3D structures (RNA polymerase II, proteasome and the Ap2/3 complex) and additional complexes from the literature [20]. The crystal structures of complexes approximate the absolute truth regarding stable protein interactions since they reveal all interactions in atomic detail, at least for the proteins that have been co-crystallized. Based on crystal structures, Edwards and coworkers found a false-negative rate of 51–98% for Y2H and 15–50% for yeast two-hybrid experiments, respectively. In this context it is remarkable that conventional low-throughput methods also produce a large fraction of false positives, for example, 61% in a pull-down study of RNA polymerase II [21].

Several studies showed that interacting proteins tend to be coexpressed at the messenger RNA (mRNA) level under various experimental conditions [15,18]. However, while coexpression of the two partners increases the confidence in a protein-protein interaction, it is only an indirect measure of its reliability. While proteins in a complex must be expressed at similar levels in order to maintain their stoichiometric ratios, this is not necessarily true for transient interactions that are often found in Y2H screens.

**Topology of protein interaction networks**

Protein interactions identified on a genome-wide scale are commonly visualized as protein interaction networks. Such networks are graphs with proteins as nodes and interactions as edges (FIGURE 3). Although this representation does not reflect the true nature of protein interactions (which is rather composed of dynamically forming complexes), it serves as a useful mental map and allows for the analysis of certain network properties.

Many biological networks, including protein interaction networks and metabolic networks, have a so-called scale-free topology. Scale-free networks are characterized by a few highly connected nodes (hubs) and many less-well connected peripheral nodes. The distribution of the node degree $k$ follows a power law (FIGURE 3D)

$$P(k) \propto k^{-\gamma}$$

The scale-free nature explains several properties of protein interaction networks. For example, highly connected hubs often appear to have central roles in a network, which would make them vulnerable to attack by mutation or drugs. Jeong and coworkers have shown that the lack of homogeneity of a network results in tolerance to errors [20]. Random mutations in the yeast genome do not appear to affect the overall topology of the network. By contrast, when the most connected proteins are computationally eliminated, the network diameter increases rapidly (i.e., the minimum number of nodes between two arbitrary proteins). Although proteins with five or fewer links constitute approximately 93% of the total number of proteins in the data set of Jeong and coworkers, they found that only approximately 21% of them are essential. By contrast, only some 0.7% of the yeast proteins with known phenotypic profiles had more than 15 links, but a deletion of 62% of these proved lethal.

Experimentally derived interaction networks, such as that shown in FIGURE 3A, can be extremely complex and biological meaning is not immediately obvious in them. However, biological systems are hierarchically organized into functional modules and submodules [23]. For example, cells produce ATP via a set of modules, such as the glycolytic pathway, the Krebs cycle and the protein complexes involved in oxidative phosphorylation. Even if their annotation cannot be used for clustering as shown in FIGURE 3C, several groups have developed algorithms to identify functional clusters (cliques) in protein interaction networks. For example, Spinosa and Money developed an algorithm that was able to recover many previously known protein complexes (e.g., the anaphase-promoting complex) and functional modules (e.g., the yeast phenotypic response pathway) [25]. In addition, new complexes (e.g., a complex of six proteins including the YIP1 Golgi membrane protein) and new members of complexes (e.g., two 40S small ribosomal subunits in the Lsm splicing complex) were identified and thus these methods can provide information about single proteins and their biological context.

The interconnections between different modules can be derived from individual protein interactions and their functional annotation (FIGURE 3E). When all proteins of a certain functional class (or module) are collapsed into one node each, the protein interactions can be used to visualize their relationships.
Protein interaction networks are scale-free networks. In contrast to exponential random networks, in which all proteins (nodes) are regarded as equal, scale-free networks have highly connected proteins which are more likely to interact with new proteins added to the network. Exponential networks are therefore statistically homogenous, whereas scale-free networks have a few highly connected proteins (hubs) and many proteins with few interactions. The signature of scale-free networks is the power law distribution of the node degree (k, number of interacting partners of a protein), $P(k) \sim k^{-\gamma}$, whereas the node degree follows a Poisson distribution in the exponential network model. Reprinted with permission from [23] and [48].

The protein interaction network of yeast reveals different levels of organization. Computer algorithms can deduce molecular modules: protein complexes and pathways directly from the topology of protein interaction networks [26, 49]. Complex protein interaction networks can be collapsed into a meta-network showing the interactions between functional categories. (B) and (D) reprinted with permission from [32].
example, in Figure 3D (top middle) the 68 proteins involved in amino acid metabolism are connected by 23 protein interactions. More importantly, this class of proteins also interacts with proteins involved in protein degradation (arguably to generate amino acids), the cell cycle (which controls almost everything and therefore is highly connected by definition) and, surprisingly, chromatin structure. Unexpected interactions such as the one between amino acid metabolism and chromatin structure point to hitherto unnoticed crosstalk between biological pathways and functions which in this case may be regulatory in nature. The fact that some groups (such as the cell cycle proteins) are highly connected indicates their central regulatory role for most other processes in a cell.

Another method for the detection of complexes in protein interaction networks based on k-centers was used to detect a novel nuclear network in yeast [27,28]. A k-core is a subnetwork of the protein interaction network in which each protein is connected to at least k proteins of this subnetwork. Therefore, this set of proteins forms a highly connected complex in the protein interaction network. The identified nuclear protein interaction network showed a structure corresponding to the known electron microscopic substructure of the nucleolus (fibrillar component, dense fibrillar component and granular component) [29]. This illustrates that the close examination of protein interaction networks can reveal molecular structures, without any prior knowledge of protein functions.

Lessons from single interactions

The ultimate goal of molecular biology is the mechanistic explanation of specific biological phenomena. For such explanations a detailed understanding of single proteins is necessary. Protein interaction data often provide critical information on the molecular behavior of a protein and almost always allow the formulation of some biological hypotheses. The chromosome cohesion proteins illustrate this point [30-32]; a few interactions of the Smc and Scc proteins in yeast and their predicted coiled-coil structure suggested a model that explained their ability to hold chromatin together.

Obviously, the lower reliability of high-throughput interaction data has to be taken into account and hypotheses building should start with the most plausible interaction and then proceed to less likely ones. However, the power of interaction mapping is also based on the fact that it is not dependent on previous knowledge of a certain protein. Therefore, completely unexpected interactions may lead to spectacular new discoveries. For example, interactions between membrane proteins and transcription factors have usually been considered as false positives. However, during the past couple of years it has been shown in a number of cases that such interactions represent novel methods of regulating transcription directly by membrane receptors. Well-studied examples include the retinoid receptor family and the retinoid-related orphan receptor (VDR [33]). Ames et al. showed that the interaction between the VDR and the retinoid-related orphan receptor (VDR [33]) could explain the effects of vitamin D on gene expression.

Such predictions have also been experimentally tested. Kemmen and coworkers verified the predicted interaction of five proteins that had interactions with known proteins that were also overexpressed in a deletion strain of an uncharacterized ORF (YLR270W) shown to interact with a protein required for thiamine resistance (NTH1, neutral thiamine gene) showed sensitivity to heat shock. Ideally, high-throughput interaction data are used by more traditional cell biological studies [34,35]. For example, Toso and coworkers examined the role of Sklp1, a protein that binds to DNA.
A role of Ski8p in meiotic DNA recombination was suggested by the mutational phenotype. However, due to its known role in cytoplasmic RNA degradation (nonpoly[A] and double-stranded RNA), an indirect role of Ski8p was assumed. However, a direct interaction between Ski8p and a protein involved in meiotic recombination, Skp1, in a comprehensive Y2H study led the authors to examine a direct effect of Ski8p on meiotic recombination which was subsequently proven [7].

Evolution of protein interaction networks

It has been suggested that proteins involved in interactions are more conserved than proteins that participate with a smaller number of interaction partners [14]. However, Jordan and coworkers demonstrated that only proteins with the largest number of interactions (the hubs of the protein interaction network) show a slower evolution rate [15]. Thus, the correlation found by Fraser and colleagues may be an artefact caused by a small subset of proteins rather than a general phenomenon [16].

Comparative interactomics: predicting homologous interactions

Proteins evolve and so do their interactions. If interacting proteins have a weak homology to another pair of interacting proteins, the interaction will support both their functional and evolutionary homology [38,102]. In order to detect such homologous interactions and pathways, Kelley and coworkers [38] developed the program PathBlast, which aligns two protein–protein interaction networks combining interaction topology and sequence similarity [102]. Using this approach, it was possible to show that the protein–protein interaction networks of yeast and Helicobacter pylori harbor a significant number of evolutionarily conserved pathways. A spectacular example among the conserved subnetworks is a group of proteins involved in basilar membrane transport and nuclear–cytoplasmic transport in yeast. This finding indicates that nuclear–cytoplasmic transport may have originated from a homologous system in basilar plasma membranes.

Pathway comparison cannot only uncover conserved pathways but can also identify additional components that have been found in one organism but not in another. For example, the interaction network of proteins shown in Figure 5A has different interaction partners. This information can be exploited to predict unknown interaction partners based on homology in another model. Such predictions are particularly supported by protein complexes that tend to be well conserved, especially if they usually require several conserved subunits for stability.

Integrating protein interaction data with other HTS data

Obviously, high-throughput data are not sufficient to explain complex biological processes. However, it has been demonstrated that the combination of several data sets can contribute significantly to the understanding of certain processes [15]. In addition, high-throughput approaches can also be used to improve data quality and therefore, their predictive power. For example, it has been shown that the intersection of high-throughput interaction data sets contains more interactions from the same MIPS complex than single data sets [36].

A major drawback of this method is that all high-throughput data sets are far from being comprehensive, which results in a very small intersection between different data sources (e.g., 133 common interactions between Uetz and Inio’s core data sets) [20]. Therefore, a very limited number of interactions are marked as reliable using this method. A more elaborate approach is the use of a Bayesian network, which allows for the probabilistic combination of multiple data sets. It has been shown that the fraction of false positives and false negatives can be reduced using this method [18]. This approach has also been used in a comprehensive study by Jensen and coworkers [40], in which the high-throughput interaction data sets for the yeast proteome (Y2H and in vivo pull-downs) were combined with genomic features only weakly associated with an interaction (e.g., expression of two proteins) to generate a more reliable interaction data set.

Can a combination of high-throughput data replace traditional experiments?

As has been seen, HTS data are often of lower quality than individually obtained data. On the other hand, HTS data are often better controlled internally since they have been collected under standard conditions. What if all kinds of data were collected under such standardized conditions and were subsequently combined? For example, why are intracellular transport processes not studied by traditional experiments?

Pathway comparison cannot only uncover conserved pathways but can also identify additional components that have been found in one organism but not in another. For example, the interaction network of proteins shown in Figure 5A has different interaction partners. This information can be exploited to predict unknown interaction partners based on homology in another model. Such predictions are particularly supported by protein complexes that tend to be well conserved, especially if they usually require several conserved subunits for stability.

Integrating protein interaction data with other HTS data

Obviously, high-throughput data are not sufficient to explain complex biological processes. However, it has been demonstrated that the combination of several data sets can contribute significantly to the understanding of certain processes [15]. In addition, high-throughput approaches can also be used to improve data quality and therefore, their predictive power. For example, it has been shown that the intersection of high-throughput interaction data sets contains more interactions from the same MIPS complex than single data sets [36].

A major drawback of this method is that all high-throughput data sets are far from being comprehensive, which results in a very small intersection between different data sources (e.g., 133 common interactions between Uetz and Inio’s core data sets) [20]. Therefore, a very limited number of interactions are marked as reliable using this method. A more elaborate approach is the use of a Bayesian network, which allows for the probabilistic combination of multiple data sets. It has been shown that the fraction of false positives and false negatives can be reduced using this method [18]. This approach has also been used in a comprehensive study by Jensen and coworkers [40], in which the high-throughput interaction data sets for the yeast proteome (Y2H and in vivo pull-downs) were combined with genomic features only weakly associated with an interaction (e.g., expression of two proteins) to generate a more reliable interaction data set.

Can a combination of high-throughput data replace traditional experiments?

As has been seen, HTS data are often of lower quality than individually obtained data. On the other hand, HTS data are often better controlled internally since they have been collected under standard conditions. What if all kinds of data were collected under such standardized conditions and were subsequently combined? For example, why are intracellular transport processes not studied by traditional experiments?

Pathway comparison cannot only uncover conserved pathways but can also identify additional components that have been found in one organism but not in another. For example, the interaction network of proteins shown in Figure 5A has different interaction partners. This information can be exploited to predict unknown interaction partners based on homology in another model. Such predictions are particularly supported by protein complexes that tend to be well conserved, especially if they usually require several conserved subunits for stability.
export, which mainly affects Golgi-to-vacuole transport. More subtle effects of YHR105W on protein transport must now be studied, as it is entirely possible that the interaction has a modulatory role in transport as opposed to being absolutely essential. One needs to remember that most mutations are not deleterious but rather show no, or only subtle, defects. This is due to the fact that gene functions can be substituted on the single gene level by duplicate or redundant genes. Such special circumstances usually cannot be identified by HTS and thus have to be analysed by a painstaking hypothesis-driven approach, where the hypothesis is refined by each additional experiment.

As an interesting new development, King and coworkers have devised algorithms to automate such hypothesis-driven research [42]. Computer algorithms can replace human reasoning to a certain extent and it may be possible to push HTS to a degree that its experimental conditions can be automatically refined based on previous experiments and therefore, do simulate hypothesis-driven experimentation.

Figure 5. Comparison and evolution of protein interactions. (A) The comparison of protein interaction networks of different species reveals conserved pathways. PathBlast, an algorithm for the alignment of protein interaction networks, was used to identify conserved pathways between Helicobacter pylori and yeast [38]. As an example, a protein degradation/DNA replication pathway is shown. Proteins with a certain sequence similarity are placed in one row. Oriented protein interactions appear as solid lines and gaps or mismatches are dotted. This pathway alignment demonstrates an association of two pathways which were not previously known to be linked. The network contains proteins associated with DNA polymerase (Rfc2, 3, 4, 6) and subunits of the 19S proteasome regulatory cap (Rpt1, 2, 3, 4, 6) and thereby provides evidence that the protein degradation and the DNA replication pathways associate in vivo. This method can be helpful for predicting protein functions and identifying functional orthologs from among multiple homologous sequences. Furthermore, the comparison of pathways and functional modules helps to understand and visualize protein network evolution [38].

(B) Interacting proteins show coevolution. The phylogenetic tree of the GyrA and ParC look strikingly similar to the trees of their interaction partners, GyrB and ParE (i.e., GyrA and GyrB form a complex as do ParC and ParE). Ramani and Marcotte used that similarity to predict interaction partners because the evolution of interacting proteins often shows a similar pattern [51].

C. crescentus: Caulobacter crescentus; E. coli: Escherichia coli; H. influenzae: Haemophilus influenzae; N. meningitidis: Neisseria meningitidis; P. multocida: Pasteurella multocida; R. solanacearum: Ralstonia solanacearum; S. typhimurium: Salmonella typhimurium; Y. pestis: Yersinia pestis.
Protein interaction networks for medical research

Most diseases are caused by malfunctioning proteins in one way or another. However, there are only a few known examples of disease-causing defects in protein interactions. The best-studied cases are probably receptors that bind or do not bind to peptide hormones or oncoproteins, such as Ras, which may cause cancer when their signaling interactions are defective.

When analyzing mutant proteins it is usually not easy to tell an impaired protein interaction apart from some unrelated effect, such as a folding problem. Hence it is difficult to say if a certain phenotype arises from a defective protein interaction or some indirect cause, such as an instability that prevents a protein from interacting.

For a detailed understanding of disease-causing mutations it would be desirable to have the crystal structures of proteins and their mutants. This would tell us if the structure is really unaffected by a mutant or if the mutant affects an exposed interaction surface.

Interestingly, Cize and colleagues present a human disease protein view in their Drosophila PIM, in which proteins with sequence similarity to human disease genes are highlighted. The PIM database has strong matches (BLAST e-value <10^-10) to one or more sequences in the Drosophila database [13]. This clearly shows the utility of PIMs in model organisms for medical research.

Using protein interaction networks for drug discovery

The goal of drug discovery is to design or identify small molecular compounds which help to cure or at least ameliorate disease. Protein interaction mapping can be useful at several levels of the drug discovery process. The first step should be the drug target identification. PIMs can help to identify proteins of relevant molecular pathways or complexes which are involved in a specific disease. For example, a highly connected protein (hub) may be a suitable target for an antibiotic whereas a more peripheral protein with few interactions may be more appropriate for a highly specific drug that needs to avoid side effects.

Proteins and their protein-protein interaction surfaces are promising targets for specific drugs, although only a few published examples of interaction inhibitors are available. One example are agents which inhibit the interaction between the BCL-2 domain and Bcl-xl [14].

Another recently published example is the hepatitis C virus protease, which cleaves the virus encoded polyprotein. Lamivudine and nucleoside analogues which act as a weak enzyme inhibitor, three amino acid inhibitors were then selected. These short peptides could then be used to design a...
specific chemical inhibitor of similar structure. The inhibitor had to be designed to enter the cell and appears to have antiviral activity in preliminary clinical trials. Theoretically this approach can also be applied to other interactions. The limiting problem is to find compounds that mimic peptides and are able to enter epithelia or cells.

The diversity of interactions of a targeted protein could also help to estimate or explain side effects of a drug. PIMs indicate immediately which other proteins or processes may be affected by inhibiting a certain interaction. Therefore PIMs can help to design selective agents which target specific interactions of a protein but do not affect others.

Conclusions
High-throughput protein–protein interaction data provide a starting point for the analysis of complexity, signaling and the structural and dynamic organization of cells. In addition, it illuminates an important aspect of the evolution of molecular systems. If combined with results from other high-throughput methods, such as microarray analysis, a systematic, global view of the molecular functioning of organisms can be gained which for the first time gives us a glimpse of an organism as a whole. By contrast, conventional biological methods are hardly comprehensive, no matter how detailed they are since they always have to focus on certain selected aspects.

Knowledge about biological networks will help us to understand the complexity of biological systems not just as an intellectual achievement. Systems biology will eventually facilitate the simulation and even manipulation of living systems, for example to cure diseases or for the generation of safe and healthy food.

Expert opinion & five-year view
Today, only a limited number of protein interaction studies have been completed. Moreover, the available studies are far from being complete. The most comprehensive data set is available for the yeast proteome, for which several Y2H and pull-down studies have been published. Protein interaction data sets for other organisms would not only provide insights into the biology of these organisms, but would also tell us about the evolution and general structure of protein interaction networks. In 5 years, many new PIMs for viruses and bacteria as well as other eukaryotes will become available. This will permit the assessment of the diversity of organisms from a systems perspective.

Of course, the protein–protein interaction map of the human proteome is an important goal since this knowledge would promote the understanding of human biology and the therapy of diseases. In 5 years, the human protein interaction map will be far from being complete but there will be several partial interaction maps which elucidate specific pathways and modules, such as those related to human diseases.

While a plethora of data are already available, today’s protein interaction networks only give a static view of the molecular organization of the cell. In contrast, the dynamic regulation of protein interactions, for example, in signal transduction cascades, is central to the understanding of biological processes. Small-scale studies have succeeded in analyzing the dynamics of single protein-protein interactions, for example, of the bacterial chemotaxis system. In 2009, the investigation of the dynamics of several biological subsystems (e.g., specific signal transduction cascades) will provide a deeper insight into the complex temporal regulation of the interactome. In addition, new high-throughput techniques which capture these dynamic properties of protein-protein interactions will be available and thus make it possible to initiate projects to understand their dynamics on a proteome-wide scale.

Last but not least, improved databases and visualization tools are urgently needed to make available data more accessible, ideally even to nonspecialists. Only when we have a clear idea of what is known can we imagine what we do not know.

Key issues
- Protein interaction maps (PIMs) are generated either by the yeast two-hybrid system or by mass spectrometric analysis of protein complexes.
- Both methods produce a certain number of false negatives and false positives. However, the reliability can be improved by combining several data sets.
- Visualization of protein interaction networks as a graph with nodes (proteins) and edges (interactions) reveals the scale-free topology of these networks.
- Biological, meaningful and functional modules can be identified in these networks and interconnections between these modules can be explored.
- A function can be assigned to an unknown protein by examining its binding partners (guilt-by-association approach).
- Protein interaction networks help to identify evolutionary conserved pathways.
- PIMs can be applied in drug discovery to identify target proteins and to minimize side effects.
References

Papers of special note have been highlighted as:

** of considerable interest

References

1. Ghe W, Wulff JB, Vidal M. Integrating "omics" information: a bridge between genomics and systems biology. Trends Genet. 19(10), 552–560 (2003).

2. Alberts B, Johnson A, Lewis J et al. Molecular Biology of the Cell. Fourth Edition. Garland Sciences, NY, USA (2002).

3. Fields S, Hartley P. The two-hybrid system. A personal view. Methods Mol. Biol. 177, 3–8 (2001).

4. Mann M, Hendrickson RC, Pandey A. Analysis of protein and protein networks by mass spectrometry. Annu. Rev. Biochem. 70, 437–473 (2001).

5. Fields S, Song P. The yeast two-hybrid system. Nature 340(6248), 245–246 (1989).

6. Uetz P, Bader JS, Giot L et al. A comprehensive analysis of protein--protein interactions in Saccharomyces cerevisiae. Nature 403(6770), 623–627 (2000).

7. Ito T, Chiba T, Ozawa R et al. A comprehensive two-hybrid analysis to explore the yeast proteome. Proc. Natl Acad. Sci. USA 98(8), 4934–4939 (2001).

8. Giot L, Bader JS, Ver内在G et al. Protein complexes in yeast. Science 295(5557), 1770–1774 (2002).

9. Uetz P, Hartling T, Rapaport M et al. The interactome network of the metazoan Drosophila melanogaster: Science 302(5651), 790–792 (2003).

10. Jeong H, Mason SP, Barabási AL et al. The large-world problem: how to map the internet and other very large networks. Science 298(5591), 66–73 (2002).

11. Alberts B, Johnson A, Lewis J et al. Molecular Biology of the Cell. Fourth Edition. Garland Sciences, NY, USA (2002).

12. Tong AH, Drees B, Nardelli G et al. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295(5557), 32–34 (2002).

13. Hartwell LH, Hopfield JJ, Leibler S et al. From molecular to modular cell biology. Proc. Natl Acad. Sci. USA 95(9), 676–683 (1998).

14. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc. Natl Acad. Sci. USA 100(19), 10783–10788 (2003).

15. Hartuv H, Shamir R. A map of the human proteome. Science 297(5580), 262–267 (2002).

16. Jeon H, Hwang YI, Lai S-X et al. A random walk approach to predict protein complexes in metazoans. Science 295(5557), 463–466 (2002).

17. Tong AH, Drees B, Nandi R et al. A comprehensive experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295(5557), 32–34 (2002).

18. Hartwell LH, Hopfield JJ, Leibler S et al. From molecular to modular cell biology. Proc. Natl Acad. Sci. USA 95(9), 676–683 (1998).

19. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc. Natl Acad. Sci. USA 100(19), 10783–10788 (2003).

20. Jeon H, Hwang YI, Lai S-X et al. A random walk approach to predict protein complexes in metazoans. Science 295(5557), 463–466 (2002).

21. Hartwell LH, Hopfield JJ, Leibler S et al. From molecular to modular cell biology. Proc. Natl Acad. Sci. USA 95(9), 676–683 (1998).

22. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc. Natl Acad. Sci. USA 100(19), 10783–10788 (2003).

23. Jeon H, Hwang YI, Lai S-X et al. A random walk approach to predict protein complexes in metazoans. Science 295(5557), 463–466 (2002).

24. Hartwell LH, Hopfield JJ, Leibler S et al. From molecular to modular cell biology. Proc. Natl Acad. Sci. USA 95(9), 676–683 (1998).

25. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. Proc. Natl Acad. Sci. USA 100(19), 10783–10788 (2003).

26. Jeon H, Hwang YI, Lai S-X et al. A random walk approach to predict protein complexes in metazoans. Science 295(5557), 463–466 (2002).

27. Tong AH, Drees B, Nandi R et al. A comprehensive experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295(5557), 32–34 (2002).

28. Hartwell LH, Hopfield JJ, Leibler S et al. From molecular to modular cell biology. Proc. Natl Acad. Sci. USA 95(9), 676–683 (1998).
Jordán BE, Wu YI, Kosint Evans J. No simple dependence between protein evolution rate and the number of protein–protein interactions: only the most prolific interactors tend to evolve slowly. BMC Biol. 5(1), 1 (2003).

Kelley BF, Naranjo R, Karp RM et al. Conserved pathways within bacterium and yeast as revealed by global protein network alignment. Proc. Natl Acad. Sci. USA. 104(4), 11394–11399 (2007).

The protein interaction networks of yeast and Helicobacter pylori are aligned using an algorithm combining sequence identity and network topology. This approach can be used to identify and extend conserved pathways in different species.

Hazbun TR, Malmstrom L, Anderson S et al. Assigning function to yeast proteins by integration of technologies. Mol. Cell. 12(6), 1353–13565 (2003).

Jansen R, Lan N, Qian J et al. Integration of genomic data sets to predict protein complexes in yeast. J. Struct. Funct. Genomics 2(2), 71–81 (2002).

Bonangelino CJ, Chavez EM, Bonifacino JS. Genomic screen for vacuolar protein sorting genes in Saccharomyces cerevisiae. Mol. Biol. Cell. 13(7), 2486–2501 (2002).

King BD, Wehren LR, Jones FM et al. Protein interaction-targeted drug discovery: evaluating critical issues. Biotechniques 62(2), 51–61 (2002).

Bourguignon CJ, Chair CE, Benfatto S. Genetic screen for vacuolar protein sorting genes in Saccharomyces cerevisiae. Mol. Biol. Cell. 13(7), 2486–2501 (2002).

Chung S, Bevit E, Bar T et al. Homophilic human disease gene regulates its Drosophila Niagara. Cell. 101(1), 149–151 (2002).

Degnan A, Ibrahim A, Cardozo M et al. Identification of small-molecule inhibitors of interaction between the BRI-I kinase and BRS-1. Nature Cell. Biol. 5(2), 173–182 (2003).

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

Websites

101 NCBI www.ncbi.nlm.nih.gov:80//entrez/query.fcgi?db=Genome (Viewed May 2004)

102 Whitehead Institute www.pathblast.org (Viewed May 2004).

Affiliations

Björn Titz, MS Institut für Genetik, Forschungszentrum Karlsruhe, Box 3640, D-76021 Karlsruhe, Germany Tel.: +49 724 782 2148 Fax: +49 724 782 3354

Matthias Schlesner, MS Max-Planck-Institut für Biochemie, Am Klopferspitz 18a, 82152 Martinsried, Germany

Peter Uetz, PhD Assistant professor in genetics Institut für Genetik, Forschungszentrum Karlsruhe, Box 3640, D-76021 Karlsruhe, Germany Tel.: +49 724 782 6103 Fax: +49 724 782 3354 peter.uetz@itg.fzk.de