Full text provided by www.sciencedirect.com

Protein interaction networks from yeast to human

Peer Bork[1,2], Lars J Jensen[1], Christian von Mering[1], Arun K Ramani[3], Insuk Lee[3] and Edward M Marcotte[3,4]

Protein interaction networks summarize large amounts of protein–protein interaction data, both from individual, small-scale experiments and from automated high-throughput screens. The past year has seen a flood of new experimental data, especially on metazoans, as well as an increasing number of analyses designed to reveal aspects of network topology, modularity and evolution. As only minimal progress has been made in mapping the human proteome using high-throughput screens, the transfer of interaction information within and across species has become increasingly important. With more and more heterogeneous raw data becoming available, proper data integration and quality control have become essential for reliable protein network reconstruction, and will be especially important for reconstructing the human protein interaction network.

Addresses
[1]European Molecular Biology Laboratory, Structural and Computational Biology Programme, Meyerhofstrasse 1, 69117 Heidelberg, Germany
[2]e-mail: bork@embl-heidelberg.de
[3]Center for Systems and Synthetic Biology, Institute for Cellular & Molecular Biology, University of Texas, Austin, Texas 78712, USA
[4]e-mail: marcotte@icmb.utexas.edu
All authors contributed equally to this work

Current Opinion in Structural Biology 2004, **14**:292–299

This review comes from a themed issue on
Sequences and topology
Edited by Peer Bork and Christine A Orengo

Available online 19th May 2004

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2004.05.003

Introduction

Although metabolic network analysis dates back to the 1940s, data-driven genome-scale analyses of gene and protein networks are recent newcomers by contrast, beginning not more than five years ago and receiving increasing attention ever since. One big boost came at the end of the 1990s from computational efforts that used the genomic context of genes (e.g. fusion, neighborhood and phylogenetic profile) to predict functional relations between gene products [1–3] and the respective networks of such associations. On the experimental side, the first direct large-scale protein interaction data were presented in 2000 [4,5]; both studies used yeast two-hybrid technology. Two years later, the first large-scale protein complex purification data sets were published [6,7]. Several other approaches that reveal functional associations between genes to various degrees were published in the same period (e.g. localization data, double knockouts, etc.) and other sources, such as spotted microarrays, have been used to extract interaction information [8]. In the light of these developments, several databases storing interaction data became very popular, derived at first mainly from small-scale experiments (e.g. [9–12]) and increasingly becoming warehouses for large-scale assay data, while novel databases continue to be developed [13].

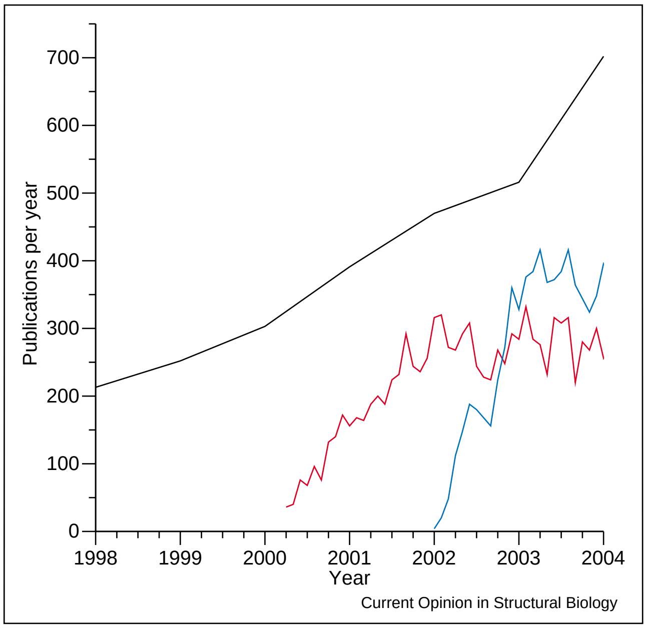
This data collection phase was accompanied by intensive analysis and comparison of networks, particularly based on the large protein interaction data sets mentioned above. The networks were compared to each other, to known protein complexes, to functional annotation and to other types of high-throughput experimental data. In particular, the topologies of the networks have received a lot of attention, as they were discovered to all be small world, scale free and modular [14•].

In 2003, we witnessed the explosive growth of research on protein interactions and networks, with new data types, data sets, analysis methods and discoveries being published at a constantly increasing pace (Figure 1). In particular, initial eukaryotic network and interaction data came almost exclusively from yeast, but we now face the challenge of deciphering the much more complex networks in metazoans. Further challenges include mapping partially complete and accurate data sets between species, with the ultimate goal of transferring the combined information to human as accurately as possible. Here, we will review recent progress by focusing on protein interaction networks in eukaryotes, but we emphasize that equally important progress is also seen for gene regulatory (transcription) and metabolic networks.

Novel experimental data on protein interactions

While the majority of large-scale interaction experiments have so far been performed on yeast proteins, the past year marks the arrival of the first large-scale animal protein interaction data sets (Table 1) [15••,16••], providing insight into the molecular functions behind multicellularity and cell–cell communication. A large-scale map of approximately 4000 genetic interactions was also derived from synthetic lethal mutations [17•]. These interaction data were complemented by several types of supporting data, including large-scale yeast protein localization data (using GFP-tagged yeast proteins [18•,19]) and the quantitation of the expression levels of approximately 4500 affinity-tagged yeast proteins through western blot analysis [20•].

Figure 1



Current Opinion in Structural Biology

The growth of protein interaction literature over time. The number of publications related to protein interaction networks has been growing strongly over the past five years or so, as revealed by citation analysis of two sets of key papers (Ito *et al*. [4] and Uetz *et al*. [5], red line; Ho *et al*. [6] and Gavin *et al*. [7], blue line) and a PubMed query ('protein interaction networks OR genome interaction OR proteome network OR proteome interaction OR interactome', the total for 2004 was estimated from publications so far; black line). Although the rate at which key papers are being cited seems to have stabilized, increasing numbers of PubMed abstracts mention interaction networks, illustrating the increasing popularity of the topic.

Table 1

Interaction coverage. Current estimates, by species and type of experiment, of the volume of large-scale experimental protein–protein interaction data available in the public domain.

	Proteins	Interactions
	S. cerevisiae	
Two-hybrid assays	934 [5] 4131 [4]	854 3986
Affinity purification/ mass spectrometry	1361 [7]	3221 (spoke) 31 304 (matrix)
	1560 [6]	3589 (spoke) 25 333 (matrix)
Protein arrays	10 [24]	~30
Synthetic lethal arrays	1029 [17•]	3627
DIP [10] (small scale)	~400	~3000
	C. elegans	
Two-hybrid assays	2898 [15••]	~4000
	D. melanogaster	
Two-hybrid assays	7048 [16••] (4679 core)	20 405 (4780 core)
	H. sapiens/M. musculus	
Affinity purification/ mass spectrometry	32 [71]	221
Protein arrays	49 [24]	~450
DIP [10] (small scale)	1177	1312
HPRD [70•] (small scale)	2750	10 534

Along with the large-scale data, strategies were refined for more accurate protein interaction mapping, such as the use of isotope labeling techniques to estimate protein enrichment during affinity purification of complexes [21,22]. Also, many smaller regions of protein interaction networks were explored in detail, such as the EGF (epidermal growth factor) signaling pathway [23] and interactions among bZIP proteins [24].

Advances in methods to predict interactions

Computational methods for predicting interactions also advanced in the past year, with completely new approaches and sophisticated 'mining' of existing interaction data to infer additional interactions. One new trend to be exploited was the tendency of proteins that can functionally substitute for one another to have anti-correlated distribution patterns across organisms [25•], allowing both discovery of non-obvious components of pathways and precise function prediction of uncharacterized proteins. Another new trend was the tendency of interacting proteins to exhibit similar phylogenetic trees [26•]; quantitative algorithms for assigning interaction partners involved analyzing trees of families of interacting proteins, such as a ligand and receptor tree, and finding proteins that occupy similar positions in two trees [27•,28].

Computational approaches for predicting novel interactions from known interactions have been developed too: interactions were inferred between pairs of proteins whose sequences are compatible with known X-ray crystal structures of heterodimers [29,30•] and between pairs of proteins with domains that are often observed in interacting proteins [31]. The specificity of the interactions predicted by the latter approach can, in some cases, be improved by looking for correlated mutations in the domains using an approach dubbed '*in silico* two hybrid' [32]. Structure-based interaction prediction (including protein complex prediction and the prediction of cross-talk between complexes) has recently culminated in the delineation of the first network of modeled protein complexes in yeast [30•].

The ability to better predict protein interactions has matured to the point at which online services are now conveniently publicly accessible, such as STRING [33], PLEX (http://bioinformatics.icmb.utexas.edu/plex), Bioverse [34] and Predictome [35]. An advantage of such tools is that they allow the prediction of interactions in organisms with no experimental interaction data, permitting systematic searches for new protein systems [36•] and the genome-wide characterization of functional modules [37•].

Quality assurance, benchmarking and data integration

Although experimentalists appreciate that all data are error prone, strategies for rigorously evaluating the reliability of large-scale protein interaction data sets have emerged only recently. The essential problem is that only a relatively small fraction of interactions in networks are known with any certainty, which leads to difficulties in estimating the rate of both false positives and false negatives. Furthermore, the number of true interactions is considerably larger than the results of typical experiments suggest, implying that the failure of individual experiments to agree on interactions may stem from either poor specificity or poor coverage.

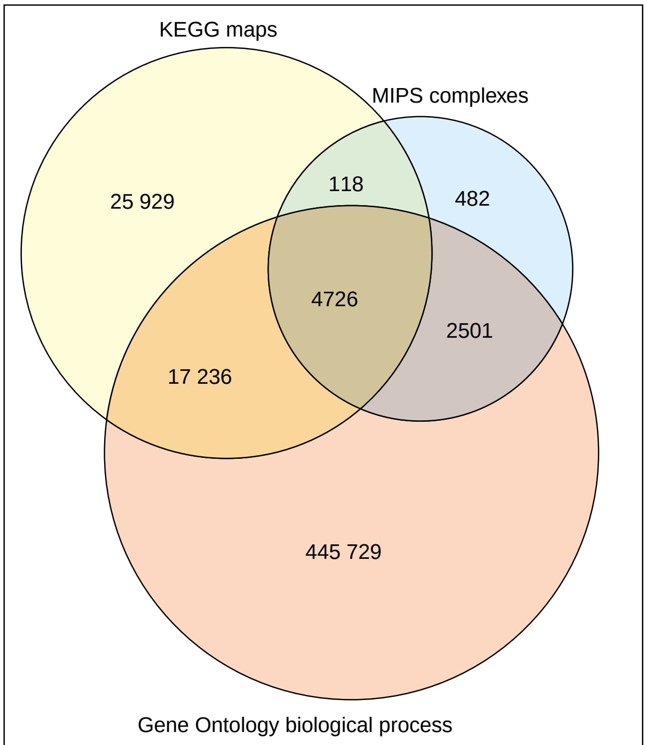
In practice, the various interaction data were tested for accuracy on confident sets of interactions [38]. The rate of false positives for the various large-scale experimental approaches varied widely, but was always larger than that for confident small-scale experiments. However, high-quality subsets could often be chosen on the basis of additional criteria [15••,16••,38,39], such as the degree to which mRNAs of interacting proteins are co-expressed in microarray experiments [40,41], topological properties of the resulting networks [42,43], shared pathways or subcellular localization [36•,39], or combinations of these various approaches [44].

Benchmarks allowing the accuracy of interaction data sets — or, better, individual interactions — to be judged are a prerequisite for the successful integration of data from multiple sources. However, if data of several types are to be integrated, the choice of benchmark set becomes less obvious as not all data will be directly related to physical interactions. This is especially problematic because the agreement among different benchmark sets is surprisingly poor (see Figure 2).

Several approaches to integration have been tried, ranging from simple intersections [8,45] or unions [46] of sets of interactions to more sophisticated probabilistic approaches [33,47•]. The past year also saw the appearance of 'meta-analyses', in which the combination of existing interaction networks suggests additional interactions from the context of the protein network [48,49].

Two key lessons emerge from the benchmarking and integration studies. First, the measurement of accuracy is critical for the integration of large-scale experiments, which rarely reach the accuracies of experiments done on a small scale. Second, the reconstruction of protein interaction networks seems to be a continuous process in which all data, even noisy data, increase the quality of the network — provided they are integrated appropriately. Thus, the current interaction maps represent the first steps on the way to accurate networks, and should continue to improve in both accuracy and sensitivity.

Figure 2



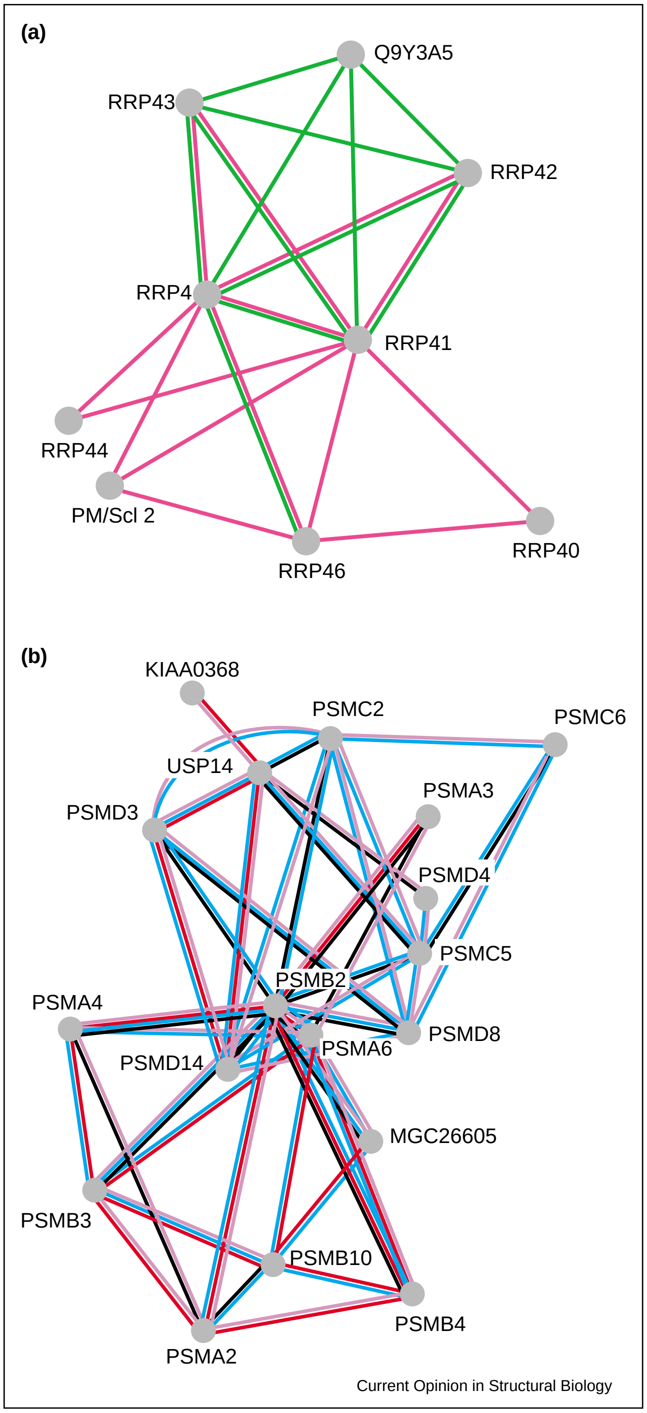
Current Opinion in Structural Biology

Comparison of benchmark data sets. The number of interactions/associations between yeast proteins is shown for each of three different benchmark data sets. The agreement of these three benchmarks sets is surprisingly poor, as illustrated by the fact that less than half of all pairs in the KEGG benchmark set are present in the Gene Ontology biological process benchmark set.

New analyses and interpretations of networks

The topological properties of protein interaction networks have been intensely studied since the first large-scale data sets were published. Interaction networks have been shown by numerous groups to be so-called 'small-world' networks, an architecture previously observed for several other types of networks (e.g. co-appearance of actors, the US power grid and *Caenorhabditis elegans* neuron connectivity [50]). Another topological term frequently attributed to interaction networks is 'scale free'. Although small-world networks need not be scale free, or vice versa, protein interaction networks have both properties [14•]. However, the biological consequences of these topological properties are not clear — in fact, there

Figure 3



Examples of human functional modules. The modules were predicted by automated transfer of interaction evidence from other species to human. **(a)** All proteins but one (Q9Y3A5) are known members of the exosome complex. Q9Y3A5 is the product of a human disease gene (Shwachman-Bodian-Diamond syndrome protein) that has been suggested to function in RNA processing [73,74]. This network was derived using STRING [33]; RRP41 was the query protein; settings were modified to exclude interaction evidence derived from PubMed abstracts or expression analysis. A total of 15 organisms contributed to this network — none of the data stem from *Homo sapiens*. Pink line: high-throughput interaction data, interactions transferred from yeast and/or fruit fly. Green line: conserved genomic neighborhood, information transferred from 13 archaeal genomes. **(b)** Proteins in this module form part of the proteasome core and regulatory particles, two of the better characterized protein complexes, as reconstructed from a combination of small-scale and high-throughput protein interaction data in yeast, as well as mRNA co-expression data and phylogenetic profiling using PLEX. One gene, KIAA0368, is uncharacterized in the human genome, belonging to 'uncharacterized conserved protein family' KOG0915, although the yeast ortholog Ecm29 has been suggested to tether the core particle to the regulatory particle [75]. Purple line: high-throughput interaction data, interactions transferred from yeast. Cyan line: mRNA co-expression of yeast orthologs. Red line: phylogenetic profile, based on 89 genomes. Black line: small-scale interaction assays, transferred from yeast.

might not be any, as both small-world and scale-free behavior can be explained by well-known evolutionary events without the need for any selective pressure acting on the network topology itself [51]. Still, the over-representation of genetic interactions between hubs in protein interaction networks supports the hypothesis that hubs play an important role [52].

In addition to the study of the global topology of interaction networks, the existence of recurring local topological features, known as network motifs, has been shown first in transcriptional networks [53] and later in protein networks [54]. Although some of these motifs (particularly in regulatory networks) make biological sense, the biological significance of network motifs remains to be studied.

Predicting functional modules

As interaction networks become increasingly large and complex, there is a growing need to break them down into more manageable subnetworks or 'modules'. These modules should preferably represent groups of proteins that together contribute to the same cellular function and the modularity should be dictated largely by the topology of the network itself. Functional modules are useful for annotating uncharacterized proteins, for studying the evolution of interacting systems and for getting a general overview of the immediate, first-order functional partners of a protein. Modules are being sought for a variety of networks: metabolic networks [55], high-throughput experimental interaction data [56–60] and *in silico* predicted networks [37•,61]. Prediction accuracy for the latter type of functional module can be high — almost 90% when benchmarked against manually curated metabolic pathways in *Escherichia coli*. These modules are thus a rich source for function prediction and pathway annotation [37•]. Furthermore, functional genomics data such as mRNA expression profiles can be integrated with high-throughput interaction data to find consistent subnetworks [62].

To identify the actual modules, a variety of supervised or unsupervised clustering techniques are used — often with previous knowledge as the benchmark — but it remains to be seen to what extent the various results are consistent. In general, functional modules often encompass protein complexes, but conceptually they go beyond stable physical interactions; modules can include transient binding partners and upstream transcriptional regulators, and even proteins that never bind each other but nevertheless function in the same pathway [63]. Implicitly, modules are also the basis of approaches that functionally classify proteins according to their network neighbors [64–66]. Figure 3 shows two examples of functional modules that expand known protein complexes with either additional subunits or functionally associated proteins. Taken together, identification of network modules not only hints at new cellular systems, but might also guide the ongoing discussions concerning the definition of pathways and cellular processes.

Network comparisons: can interactions be transferred between species?

The abundance of interaction data on yeast, and now fly and worm as well, combined with the paucity of information on other organisms, has led naturally to the question of how networks compare between species and to what extent interactions in one organism are maintained in another. Not surprisingly, conserved proteins tend to have conserved interactions and comparison of the yeast network with the *Helicobacter pylori* bacterial network identified small conserved subnetworks [67•] by searching for conserved interactions between pairs of yeast/bacterial orthologs. A related approach was used to predict a protein interaction map for *E. coli* from *H. pylori* interaction data [68] and to import yeast interactions to expand the *C. elegans* interaction data set [16••], and has led to reconstructions of genetic networks based on the evolutionary conservation of gene co-expression patterns [69•]. The significant conservation of interactions confirms that a feasible strategy for reconstructing networks is to transfer interactions from organisms in which they have been measured (for examples, see Figure 3, in which all interactions have been transferred). Despite the fact that functional modules are not always present together in distant organisms (i.e. they can change within evolutionary timescales), one can imagine constructing a composite interaction network representing the union of interactions from many different cells within or between organisms, with any particular cell possessing only a subnetwork. However, a realistic characterization of metazoan interaction networks and their conservation is only now becoming feasible with the availability of large-scale data sets, such as the yeast two-hybrid screens in fly and worm [15••,16••].

Perspectives: towards the human interactome

Obtaining a reliable interaction set describing the human interactome is a milestone yet to be reached. Given the existing data sets for yeast proteins (see Table 1), we estimate a total of 10 000–30 000 pairwise interactions. This would correspond to roughly 3–10 interactions per protein in the yeast cell. In contrast to the yeast interactome, the human interactome is largely unknown: a back-of-the-envelope calculation assuming that 3–10 interactions also holds true for each of the 25 000–40 000 human proteins leads to an estimate of roughly 40 000–200 000 interactions. Beyond this obvious uncertainty, this estimate does not even take into account complicating factors such as alternative mRNA splicing or post-transcriptional modification, both of which produce many more protein species and hence more interactions. Compounding the larger scale of the human interactome is the fact that it has not yet been studied by high-throughput interaction assays and a much smaller fraction of protein interactions are known for human (perhaps ~20 000–30 000 total are recorded in the literature [70•]) than for yeast. Despite the first medium-scale studies in human centered around individual pathways [71] or machineries [72], there is a strong need for methods to predict or measure protein interactions that scale to the size of the human interactome.

Although experimental approaches are still being scaled to tackle the number of mammalian genes, as witnessed by the first animal protein interaction networks published this past year [15••,16••], computational approaches can rapidly generate initial interaction sets, mostly by transfer of information from other organisms (e.g. see Figure 3). In metazoans such as human, this harbors additional challenges due to the distinct networks in each of the various cell types, many of which have no clear correspondence in other organisms (e.g. there is no adaptive immunity in the fly and probably not too much data should be transferred from the fly exoskeleton to human). As comparative morphology and anatomy is an unfinished research field in its own right, obtaining a thorough and well-annotated benchmark for protein interactions in human is an important next step. If a common reference was to be accepted by experimentalists and computational biologists alike, we would, from the very beginning of large-scale interaction and network prediction in human, have a much better idea of how much we have to expect. We could avoid misperceptions such as the inflated human gene numbers and could more quickly concentrate on important downstream questions, such as the impact of context on networks, or their temporal and spatial dynamic changes.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

• of special interest
•• of outstanding interest

1. Galperin MY, Koonin EV: **Who's your neighbor? New computational approaches for functional genomics**. *Nat Biotechnol* 2000, **18**:609-613.

2. Valencia A, Pazos F: **Computational methods for the prediction of protein interactions**. *Curr Opin Struct Biol* 2002, **12**:368-373.

3. Huynen MA, Snel B, von Mering C, Bork P: **Function prediction and protein networks**. *Curr Opin Cell Biol* 2003, **15**:191-198.

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

5. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al.*: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae***. *Nature* 2000, **403**:623-627.

6. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K *et al.*: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry**. *Nature* 2002, **415**:180-183.

7. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM *et al.*: **Functional organization of the yeast proteome by systematic analysis of protein complexes**. *Nature* 2002, **415**:141-147.

8. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function**. *Nature* 1999, **402**:83-86.

9. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW: **BIND–The Biomolecular Interaction Network Database**. *Nucleic Acids Res* 2001, **29**:242-245.

10. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update**. *Nucleic Acids Res* 2004, **32**:D449-D451.

11. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B: **MIPS: a database for genomes and protein sequences**. *Nucleic Acids Res* 2002, **30**:31-34.

12. Csank C, Costanzo MC, Hirschman J, Hodges P, Kranz JE, Mangan M, O'Neill K, Robertson LS, Skrzypek MS, Brooks J *et al.*: **Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD)**. *Methods Enzymol* 2002, **350**:347-373.

13. Breitkreutz BJ, Stark C, Tyers M: **The GRID: the General Repository for Interaction Datasets**. *Genome Biol* 2003, **4**:R23.

14. • Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization**. *Nat Rev Genet* 2004, **5**:101-113.
This authoritative review on the network topology of biological systems discusses hierarchical modularity, network motifs and the biological implications of both.

15. •• Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E *et al.*: **A protein interaction map of *Drosophila melanogaster***. *Science* 2003, **302**:1727-1736.
The first large-scale yeast two-hybrid study of protein interactions in a metazoan. The data were carefully benchmarked, including cross-check with interactions of homologous yeast proteins.

16. •• Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T *et al.*: **A map of the interactome network of the metazoan *C. elegans***. *Science* 2004, **303**:540-543.
The first large-scale yeast two-hybrid analysis in *C. elegans*. Together with [15••], this represents a milestone towards network analysis in metazoans, including human.

17. • Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M *et al.*: **Global mapping of the yeast genetic interaction network**. *Science* 2004, **303**:808-813.
The first genome-scale analysis of double mutants (synthetic lethals), capturing about 1000 genes and 4000 interactions. On a smaller scale, this methodology has been among the most accurate for interaction prediction.

18. • Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast**. *Nature* 2003, **425**:686-691.
Systematic GFP tagging was used to obtain information on the subcellular localization(s) of 75% of the yeast proteome.

19. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y *et al.*: **Subcellular localization of the yeast proteome**. *Genes Dev* 2002, **16**:707-719.

20. • Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS: **Global analysis of protein expression in yeast**. *Nature* 2003, **425**:737-741.
Systematic GFP tagging was used to quantify the expression level of 80% of the yeast proteome during normal growth conditions.

21. Griffin TJ, Lock CM, Li XJ, Patel A, Chervetsova I, Lee H, Wright ME, Ranish JA, Chen SS, Aebersold R: **Abundance ratio-dependent proteomic analysis by mass spectrometry**. *Anal Chem* 2003, **75**:867-874.

22. Ranish JA, Yi EC, Leslie DM, Purvine SO, Goodlett DR, Eng J, Aebersold R: **The study of macromolecular complexes by quantitative proteomics**. *Nat Genet* 2003, **33**:349-355.

23. Blagoev B, Kratchmarova I, Ong SE, Nielsen M, Foster LJ, Mann M: **A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling**. *Nat Biotechnol* 2003, **21**:315-318.

24. Newman JR, Keating AE: **Comprehensive identification of human bZIP interactions with coiled-coil arrays**. *Science* 2003, **300**:2097-2101.

25. • Morett E, Korbel JO, Rajan E, Saab-Rincon G, Olvera L, Olvera M, Schmidt S, Snel B, Bork P: **Systematic discovery of analogous enzymes in thiamin biosynthesis**. *Nat Biotechnol* 2003, **21**:790-795.
A novel way to apply genome context knowledge to predict protein interactions and, implicitly, very precisely predict the function of genes. By identifying anti-correlation of gene occurrence in various genomes, analogous functions are predicted.

26. • Goh CS, Cohen FE: **Co-evolutionary analysis reveals insights into protein-protein interactions**. *J Mol Biol* 2002, **324**:177-192.
This article explores the notion, introduced earlier by the same authors, that similarity of phylogenetic trees of interacting proteins could form the basis of a method for identifying protein interaction partners.

27. • Ramani AK, Marcotte EM: **Exploiting the co-evolution of interacting proteins to discover interaction specificity**. *J Mol Biol* 2003, **327**:273-284.
A novel algorithm for predicting specific protein interaction partners based on comparing the position of proteins in their respective phylogenetic trees, exploiting the trend described in [26•]. A similar algorithm is described in [28].

28. Gertz J, Elfond G, Shustrova A, Weisinger M, Pellegrini M, Cokus S, Rothschild B: **Inferring protein interactions from phylogenetic distance matrices**. *Bioinformatics* 2003, **19**:2039-2045.

29. Russell RB, Alber F, Aloy P, Davis FP, Korkin D, Pichaud M, Topf M, Sali A: **A structural perspective on protein–protein interactions**. *Curr Opin Struct Biol* 2004, **14**:in press.

30. • Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin AC, Bork P, Superti-Furga G, Serrano L *et al.*: **Structure-based assembly of protein complexes in yeast**. *Science* 2004, **303**:2026-2029.
A first structure-based network of (partially modeled) protein complexes in yeast. The authors combine protein complex modeling with experimental data to reveal many molecular details of interactions and also show the interconnectivity of complexes.

31. Ng SK, Zhang Z, Tan SH: **Integrative approach for computationally inferring protein domain interactions**. *Bioinformatics* 2003, **19**:923-929.

32. Pazos F, Valencia A: ***In silico* two-hybrid system for the selection of physically interacting protein pairs**. *Proteins* 2002, **47**:219-227.

33. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B: **STRING: a database of predicted functional associations between proteins**. *Nucleic Acids Res* 2003, **31**:258-261.

34. McDermott J, Samudrala R: **Enhanced functional information from predicted protein networks**. *Trends Biotechnol* 2004, **22**:60-62.

35. Mellor JC, Yanai I, Clodfelter KH, Mintseris J, DeLisi C: **Predictome: a database of putative functional links between proteins**. *Nucleic Acids Res* 2002, **30**:306-309.

36. • Date SV, Marcotte EM: **Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages**. *Nat Biotechnol* 2003, **21**:1055-1062.

A test of the idea that computational inference of protein systems is now sufficiently accurate (e.g. see [38]) that it can be used to systematically survey genomes for experimentally uncharacterized systems as a guide for future experiments.

37. • von Mering C, Zdobnov EM, Tsoka S, Ciccarelli FD, Pereira-Leal JB, Ouzounis CA, Bork P: **Genome evolution reveals biochemical networks and functional modules**. *Proc Natl Acad Sci USA* 2003, **100**:15428-15433.

Genomic-context-based *in silico* methods are used to predict an interaction network in *E. coli*, which is then clustered to reveal functional modules. The modules are benchmarked against previously known metabolic pathways and are shown to be highly accurate.

38. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions**. *Nature* 2002, **417**:399-403.

39. Sprinzak E, Sattath S, Margalit H: **How reliable are experimental protein-protein interaction data?** *J Mol Biol* 2003, **327**:919-923.

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

41. Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A, Holstege FC: **Protein interaction verification and functional annotation by integrated analysis of genome-scale data**. *Mol Cell* 2002, **9**:1133-1143.

42. Goldberg DS, Roth FP: **Assessing experimentally derived interactions in a small world**. *Proc Natl Acad Sci USA* 2003, **100**:4372-4376.

43. Saito R, Suzuki H, Hayashizaki Y: **Construction of reliable protein-protein interaction networks with a new interaction generality measure**. *Bioinformatics* 2003, **19**:756-763.

44. Bader JS, Chaudhuri A, Rothberg JM, Chant J: **Gaining confidence in high-throughput protein interaction networks**. *Nat Biotechnol* 2004, **22**:78-85.

45. Strong M, Mallick P, Pellegrini M, Thompson MJ, Eisenberg D: **Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach**. *Genome Biol* 2003, **4**:R59.

46. Yanai I, DeLisi C: **The society of genes: networks of functional links between genes from comparative genomics**. *Genome Biol* 2002, **3**:R64.

47. • Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data**. *Science* 2003, **302**:449-453.

This paper presents a probabilistic approach to integrating protein interaction data in yeast. A network of yeast proteins is derived from both high-throughput interaction data and interactions inferred from mRNA co-expression, essentiality data and functional classification schemes. The combined network is shown to be more accurate than each separate network and a few of the predictions are confirmed experimentally.

48. Schlitt T, Palin K, Rung J, Dietmann S, Lappe M, Ukkonen E, Brazma A: **From gene networks to gene function**. *Genome Res* 2003, **13**:2568-2576.

49. Tornow S, Mewes HW: **Functional modules by relating protein interaction networks and gene expression**. *Nucleic Acids Res* 2003, **31**:6283-6289.

50. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks**. *Nature* 1998, **393**:440-442.

51. Aloy P, Russell RB: **Taking the mystery out of biological networks**. *EMBO Rep* 2004, **5**:349-350.

52. Ozier O, Amin N, Ideker T: **Global architecture of genetic interactions on the protein network**. *Nat Biotechnol* 2003, **21**:490-491.

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

54. Wuchty S, Oltvai ZN, Barabasi AL: **Evolutionary conservation of motif constituents in the yeast protein interaction network**. *Nat Genet* 2003, **35**:176-179.

55. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks**. *Science* 2002, **297**:1551-1555.

56. Spirin V, Mirny LA: **Protein complexes and functional modules in molecular networks**. *Proc Natl Acad Sci USA* 2003, **100**:12123-12128.

57. Rives AW, Galitski T: **Modular organization of cellular networks**. *Proc Natl Acad Sci USA* 2003, **100**:1128-1133.

58. Krause R, von Mering C, Bork P: **A comprehensive set of protein complexes in yeast: mining large scale protein-protein interaction screens**. *Bioinformatics* 2003, **19**:1901-1908.

59. Pereira-Leal JB, Enright AJ, Ouzounis CA: **Detection of functional modules from protein interaction networks**. *Proteins* 2004, **54**:49-57.

60. Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks**. *BMC Bioinformatics* 2003, **4**:2.

61. Snel B, Bork P, Huynen MA: **The identification of functional modules from the genomic association of genes**. *Proc Natl Acad Sci USA* 2002, **99**:5890-5895.

62. Ideker T, Ozier O, Schwikowski B, Siegel AF: **Discovering regulatory and signalling circuits in molecular interaction networks**. *Bioinformatics* 2002, **18(suppl 1)**:S233-S240.

63. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology**. *Nature* 1999, **402**:C47-C52.

64. Vazquez A, Flammini A, Maritan A, Vespignani A: **Global protein function prediction from protein-protein interaction networks**. *Nat Biotechnol* 2003, **21**:697-700.

65. Brun C, Chevenet F, Martin D, Wojcik J, Guenoche A, Jacq B: **Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network**. *Genome Biol* 2003, **5**:R6.

66. Letovsky S, Kasif S: **Predicting protein function from protein/protein interaction data: a probabilistic approach**. *Bioinformatics* 2003, **19(suppl 1)**:I197-I204.

67. • Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T: **Conserved pathways within bacteria and yeast as revealed by global protein network alignment**. *Proc Natl Acad Sci USA* 2003, **100**:11394-11399.

The protein interaction networks of two remote organisms are systematically compared and aligned to reveal conserved pathways. Both well-described and novel pathways are discovered, as are previously unknown relations between pathways.

68. Wojcik J, Boneca IG, Legrain P: **Prediction, assessment and validation of protein interaction maps in bacteria**. *J Mol Biol* 2002, **323**:763-770.

69. • Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules**. *Science* 2003, **302**:249-255.

This paper demonstrates that gene co-expression networks are often conserved during evolution. This is used to improve the accuracy of interaction prediction using microarrays, by demanding co-expression in more than one organism.

70. • Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M *et al.*: **Development of human protein reference database as an initial platform for approaching systems biology in humans**. *Genome Res* 2003, **13**:2363-2371.

This manually curated collection of current knowledge concerning human proteins contains a significant fraction of the currently described protein–protein interactions in human. The interaction data set is now publicly available and, although it is yet to be benchmarked wholesale, it is a promising resource against which to compare upcoming large-scale interaction networks.

71. Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, Croughton K, Cruciat C, Eberhard D, Gagneur J, Ghidelli S *et al.*: **A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway**. *Nat Cell Biol* 2004, **6**:97-105.

72. Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, Mann M: **Proteomic characterization of the human centrosome by protein correlation profiling**. *Nature* 2003, **426**:570-574.

73. Boocock GR, Morrison JA, Popovic M, Richards N, Ellis L, Durie PR, Rommens JM: **Mutations in SBDS are associated with Shwachman-Diamond syndrome**. *Nat Genet* 2003, **33**:97-101.

74. Koonin EV, Wolf YI, Aravind L: **Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach**. *Genome Res* 2001, **11**:240-252.

75. Leggett DS, Hanna J, Borodovsky A, Crosas B, Schmidt M, Baker RT, Walz T, Ploegh H, Finley D: **Multiple associated proteins regulate proteasome structure and function**. *Mol Cell* 2002, **10**:495-507.