Computational methods of analysis of protein–protein interactions

Lukasz Salwinski* and David Eisenberg†

Computational methods play an important role at all stages of the process of determining protein–protein interactions. They are used to predict potential interactions, to validate the results of high-throughput interaction screens and to analyze the protein networks inferred from interaction databases.

Addresses
Howard Hughes Medical Institute, UCLA-DOE Institute for Genomics and Proteomics, Departments of Chemistry & Biochemistry and Biological Chemistry, Molecular Biology Institute, Box 951570, UCLA, Los Angeles, CA 90095-1570, USA
*e-mail: lukasz@mbi.ucla.edu
†e-mail: david@mbi.ucla.edu

Current Opinion in Structural Biology 2003, **13**:377–382

This review comes from a themed issue on
Sequences and topology
Edited by Mark Gerstein and Janet M Thornton

0959-440X/03/$ – see front matter

DOI 10.1016/S0959-440X(03)00070-8

Abbreviation
TAP tandem affinity purification

Introduction

New technologies have accelerated the pace of discovery of protein–protein interactions. Progress in the field was stimulated by the completion of dozens of genome sequencing projects, followed by the development of high-throughput experimental methods aimed at functional characterization of the newly discovered genes. The vast amount of data collected so far necessitates the systematic organization of the new information in a form amenable to analysis on the scale of entire genomes. It is expected that such analyses will reveal the large-scale patterns of protein interactions responsible for higher-level properties of organisms, such as adaptation, robustness and error correction, as well as shedding light on the evolutionary history of genomes. This review covers computational and experimental methods for determining protein interactions, methods for validating interactions and ways to interpret the protein networks that emerge from coupled interactions.

Experimental data sources

Until recently, information about protein–protein interactions was gathered via experiments that were individually designed to identify and validate a small number of specifically targeted interactions [1]. This traditional source of information has been augmented recently by the results of high-throughput experiments designed to exhaustively probe all the potential interactions within entire genomes (Table 1). However, the many discrepancies between the interacting partners identified in high-throughput studies and those identified in small-scale experiments highlight the need for caution when interpreting results from high-throughput studies. These discrepancies also call for the development of computational methods of data validation. Even when interactions have been validated, one needs to be aware that, as was demonstrated recently by Edwards *et al.* [2•], the majority of the experimental evidence cannot distinguish between direct interactions and those mediated by at least one intermediate protein.

Protein interaction databases

Publicly accessible databases of protein–protein interactions greatly simplify the analysis of various types of data on protein interactions. Several databases that are currently available (Table 2) provide access to both experimental data and the results of diverse computational methods of inference. Some databases also identify the most reliable subsets of the interaction data.

Further development of interaction databases is crucial for standardization of the interaction datasets and data-exchange formats, as well as for the integration of the databases with other bioinformatics resources.

Validation of high-throughput data

With the discovery of discrepancies between the results of different methods of identifying protein interactions (Table 3), attempts have been made to assess the quality of the high-throughput interaction datasets. Such assessment requires estimation of both the coverage and the accuracy of interaction data — not an easy task in the absence of a reference set of validated protein interactions. To a very limited extent, the aggregate of small-scale studies of protein interactions comprises such a validated set, but one that is vastly incomplete, having an overwhelming number of false negatives (Figure 1). Therefore, direct comparisons with this set are of limited value and provide estimates only of the lower limits on coverage and accuracy [3••].

Methods recently introduced by Mrowka *et al.* [4] and Deane *et al.* [5•] bypass this problem by analyzing the collective properties of the interaction datasets, such as the distribution of the expression distances between

Table 1

High-throughput experimental approaches to the determination of protein–protein interactions.

Method	References	Features
Yeast two-hybrid	Uetz *et al*. [33]	The first comprehensive study in yeast
	Ito *et al*. [34]	Broad coverage in yeast
	Boulton *et al*. [35], Walhout *et al*. [36]	Combined analysis of yeast two-hybrid interactions together with phenotype and expression data
Affinity purification/mass spectrometric identification	Ho *et al*. [37]	Purification of overexpressed, epitope-tagged proteins in yeast
	Gavin *et al*. [38••]	TAP purification of complexes expressed at physiological levels in yeast
Protein chips	Zhu *et al*. [39]	High-throughput detection of interactions with proteins over-expressed and immobilized on microscope slides to form a proteome microarray
Phage display	Tong *et al*. [24]	Phage display identification of binding motifs followed by computational identification of potential interacting partners and a yeast two-hybrid validation step
Synthetic lethals	Tong *et al*. [40]	High-throughput identification of synthetic lethal double mutants. Synthetic lethal mutants often correspond to physically interacting protein pairs.

Table 2

Databases of protein interactions.

Database	URL	Experimental links	Predicted links	Data validation	Species specific	Comments	Reference
DIP	http://dip.doe-mbi.ucla.edu	+	–	+	–	Collections of experimentally determined protein–protein interactions	[41•]
BIND	http://www.bind.ca	+	–	–	–		[42•]
MINT	http://cbm.bio.uniroma2.it/mint	+	–	–	–		[43•]
MIPS	http://mips.gsf.de	+	–	–	+	*S. cerevisiae* specific; also provides information on genetic interactions	[44]
The GRID	http://biodata.mshri.on.ca/grid/servlet/Index	+	–	–	+	Compilation of BIND, MIPS and several genome-scale datasets; *S. cerevisiae* specific	
LiveDIP	http://dip.doc-mbi.ucla.edu/ldip.html	+	–	–	–	Extension of DIP providing access to information on functional states of protein complexes	[45]
PREDICTOME	http://predictome.bu.edu	+	+	–	–	Compilation of functional link predictions with experimental, genome-scale yeast two-hybrid data	[46]
STRING	http://www.bork.embl-heidelberg.de/STRING	–	+	–	–	Compilations of functional link predictions based on gene proximity [48,49], common evolutionary history (phylogenetic profiles [50]) and domain fusion events (Rosetta stone method [51])	[47]
InterDOM	http://InterDom.lit.org.sg	–	+	–	–		[52]

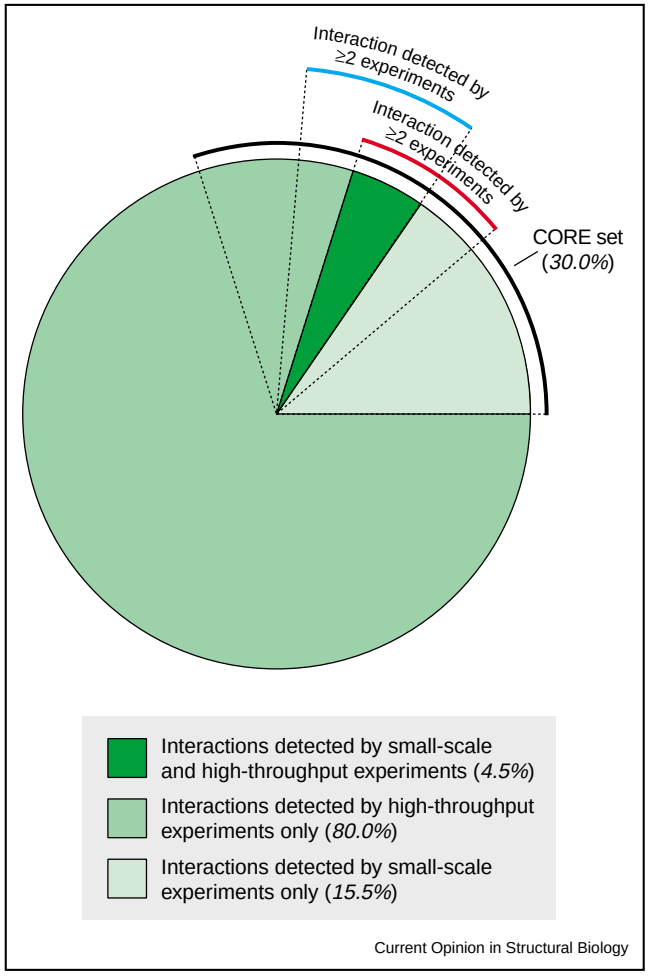
Table 3

Overlap of interactions identified in various high-throughput protein–protein interaction datasets*.

	Ito *et al*. [34]	Uetz *et al*. [33]	Gavin *et al*. [38••]	Ho *et al*. [37]	α_{EPR}
Ito *et al*.	**4363**	186	54	63	18 ± 3
Uetz *et al*.		**1403**	54	56	44 ± 6
Gavin *et al*.			**3222**	198	80 ± 6
Ho *et al*.				**3596**	6 ± 3
Small-scale experiments in DIP	442	415	528	391	

*Values in **bold** give the number of interactions found in each dataset. Elements above diagonal give the number of interacting pairs found in both datasets. The bottom row shows the overlap between the high-throughput dataset whose reference is given at the top of the column and small-scale experiments listed in the DIP database [41•]. The right-hand column gives the expression profile index, α_{EPR} [5•], which estimates from mRNA data the fraction of true positives in the large dataset whose reference is given at the left of the row. Notice that no pair of high-throughput studies of protein interactions agrees well with each other. Yet every high-throughput study finds numerous interactions that are also detected in the reliable small-scale studies.

Figure 1



Yeast protein–protein interactions in the Database of Interacting Proteins (DIP), as of January 2003. Approximately 80% of the 15 099 known binary protein–protein interactions in *S. cerevisiae* were detected by high-throughput screens only [33,34,37,38•]. Small-scale experiments detected the remaining 20% of interactions, but only 4.5% were identified in both small-scale and high-throughput experiments. Note that nearly half of the interactions detected by small-scale experiments (red arc), but only about a tenth of the interactions detected by high-throughput methods (blue arc) were identified in more than one independent experiment. The most reliable subset (CORE; black arc), composed of interactions validated by one of the methods described by Deane *et al.* [5•], constitutes 30% of all known interactions.

interacting partners [6–8]. Statistical analysis and comparison of these properties to those of a trusted reference set result in quantitative estimation of the accuracy of the high-throughput data [5•] (Table 3). In general, the results indicate that some high-throughput datasets contain a significant fraction of false positives.

In addition to these evaluations of the overall quality of the interaction datasets, attempts have been made to identify the most reliable subsets of high-throughput data. These attempts usually involve combining multiple sources of experimental information. However, because there is only a marginal overlap between datasets (Table 3), the number of interactions validated this way is very small. The number of validated interactions increases if one also takes into consideration known interactions between paralogs of the putative interacting pair. This approach, as demonstrated by Deane *et al.* [5•], allows one to identify roughly half of the true interactions within a typical high-throughput dataset. Recently, another method of quality evaluation has been proposed by Bader *et al.* [9•]. At its root, this method exploits the observation, made recently by Ravasz *et al.* [10], that interacting proteins tend to form highly connected clusters within interaction networks; it is therefore possible to assess the quality of a prospective interaction by examining the length of the shortest path that connects the potential interactors.

Protein interaction networks

One common method of analyzing the global properties of protein–protein interactions is by graph theory. Individual proteins are modeled as graph vertices connected by edges that correspond to experimentally identified binary interactions. Despite limitations that include the lack of temporal and spatial resolution, as well as the neglect of multiprotein complexes, graph-theoretical analysis has provided interesting insights into the structure of the protein interaction network. For example, Jeong and colleagues [11••] described the scale-free topology [12] of protein interaction networks; the scale-free topology of metabolic networks has also been described [13]. The most characteristic feature of scale-free networks is the presence of few highly connected nodes well separated within the network [14]. It was postulated that such topology is responsible for the robustness of the scale-free networks [15]. The finding that the essential protein-encoding genes within the protein–protein interaction network coincide with the highly connected nodes seems to confirm this interpretation of robustness [11••].

Recently, it was observed that some of the characteristic parameters of the metabolic networks, such as the degree of clustering, deviate from the values expected from the scale-free model [10]. It remains to be seen if those discrepancies, which have been attributed to the modular structure of the metabolic networks, are also observed for the networks of protein–protein interactions. If this is indeed the case, the module identification approach presented by Ravasz *et al.* [10] might prove to be a useful way of identifying multiprotein complexes automatically.

Current models of network growth [10,12] can explain some, but not all, of the features of biological networks. More detailed models that attempt to take into account known mechanisms of protein evolution might ultimately explain, from the evolutionary perspective, features not addressed by the original model of Barabasi and Albert

Table 4

Methods for computational inference of protein functional linkage.

Linkage type	Method	References
Physical	Interspecies interaction transfer based on the interacting sequence motif pairs identified in yeast two-hybrid screens.	[19,53]
Physical	Interactions inferred from correlated mutations.	[54]
Physical	Co-occurrence of sequence domains.	[20,21•]
Physical	Structure assignment followed by threading-based interaction energy evaluation.	[22]
Physical	Ortholog-based transfer of interactions between species followed up by experimental validation.	[23]
Functional annotation transfer	Network-topology-based functional annotation. A function is transferred to proteins that form the shortest path connecting two proteins of the same, known function.	[25•]
Functional links	Introduction of the phylogenetic profile method. Functional links are created between proteins with a similar evolutionary history as judged by the similar pattern of their presence across multiple genomes.	[50]
	Phylogenetic profile enhancements. Measures of phylogenetic profile distance that reflect the detailed evolutionary history of the species improve performance of the method.	[55–57]

[12], such as the distribution of protein family interactions [16] and the asymmetric divergence of interactions between paralogs [17].

Computational inference of protein interactions

The function of a protein can be viewed as its position within the cellular interaction network [18]. Therefore, the inference of a protein's interaction partners is an important step towards the identification of its role within a cell.

Recently developed methods for the inference of protein–protein interactions are listed in Table 4. They cover a spectrum of approaches, including the analysis of experimentally determined sequence interaction profiles [19], the exploitation of the frequencies of specific domain–domain interactions [20,21•] and an extension of threading-based fold recognition to the prediction of protein–protein interaction contacts [22]. In the last method, an interaction is found by identifying a pair of sequences compatible with the structure of two proteins known to interact. However, as in the case of high-throughput data validation, systematic evaluation and comparison of different methods is difficult because of the lack of complete, reliable reference sets of interacting proteins. Therefore, approaches that combine a computational step with experimental validation of the results are gaining popularity. For example, Matthews and colleagues [23] inferred interactions in *Saccharomyces cerevisiae* on the basis of homology to *Caenorhabditis elegans*, and vice versa, and then used yeast two-hybrid validation to validate these inferences. In a more involved scheme, Tong *et al.* [24] began by using a phage-display method to create a library of peptides that interact with yeast SH3 domains. The library was then computationally processed to define binding motifs that were, in turn, used to scan the yeast genome for potential interacting partners. In the final step, the predicted interactions were validated by comparison with the results of a yeast two-hybrid screen.

Another set of methods attempts to address protein function directly by inferring the 'functional links' that connect proteins with similar functions [18]. Table 4 documents the recent refinements of the original methods. Additionally, an interesting method of achieving functional annotation was introduced recently by Zhou *et al.* [25•]. It exploits the topology of a protein network, using expression-distance measurements to annotate the nodes on the shortest paths connecting proteins of the same function.

Conclusions

The recent profusion of data on protein–protein interactions poses computational challenges when assessing data quality and organizing data into a consistent, easily accessible database that is useful for further studies. It is hoped that such efforts will provide a framework for analyzing the biological networks that determine the physiological properties of living cells.

As shown recently, even simplified models of interaction networks, based on vastly incomplete data, have provided insights into the patterns of organization and evolution of living matter [11••,16]. To fully utilize the wealth of available information, models of biological networks will have to be extended to incorporate information on the dynamics of the cellular components, including spatial and temporal changes in gene expression levels, post-translational protein modifications and the activity of protein degradation pathways. The initial attempts to analyze such diverse data have demonstrated that a combination of different types of data can result in a more complete picture of living cells, leading to a better understanding of biological processes [26,27].

There exists a growing body of experimental evidence that confirms ubiquitous interconnections and interdependencies between the different components of a cell. It raises questions about the methodologies and computational resources required to study such complex systems.

However, recent discoveries [10,28–30] support the idea that living cells are, as suggested by Hartwell *et al.* [31], formed by several tightly organized modules connected more loosely to one another. If so, the identification of such modules [10,28] is all the more important because it will ultimately provide a means of identifying smaller subsystems amenable to detailed computational analysis and simulation [32].

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. Golemis E (Ed): *Protein–Protein Interactions. A Molecular Cloning Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2002.

2. • Edwards A, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M: **Bridging structural biology and genomics: assessing protein interaction data with known complexes**. *Trends Genet* 2002, **18**:529-536.
This paper evaluates the differences between the results of biochemical and structural methods of interaction determination.

3. •• 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.
Comprehensive evaluation of the results of genome-wide interaction screens.

4. Mrowka R, Patzak A, Herzel H: **Is there a bias in proteome research?** *Genome Res* 2001, **11**:1971-1973.

5. • Deane CM, Salwinski L, Xenarios I, Eisenberg D: **Protein interactions: two methods for assessment of the reliability of high throughput observations**. *Mol Cell Proteomics* 2002, **1**:349-356.
This paper describes two publicly available (URL: http://dip.doe-mbi.ucla.edu/Services.cgi) methods of evaluation of high-throughput protein interaction screens.

6. Grigoriev A: **A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae***. *Nucleic Acids Res* 2001, **29**:3513-3519.

7. 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.

8. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein–protein interactions**. *Genome Res* 2002, **12**:37-46.

9. • Bader JS, Chaudhuri A, Chant J: **Computational analysis of genome-scale protein interaction data to define biologically coherent pathways.** *Nat Biotechnol* 2003, in press.
This paper describes a network-topology-based method of assessment of protein interaction data.

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

11. •• Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks**. *Nature* 2001, **411**:41-42.
A demonstration that topological localization of proteins within interaction networks correlates with the phenotype they code for.

12. Barabasi AL, Albert R: **Emergence of scaling in random networks**. *Science* 1999, **286**:509-512.

13. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks**. *Nature* 2000, **407**:651-654.

14. Maslov S, Sneppen K: **Specificity and stability in topology of protein networks**. *Science* 2002, **296**:910-913.

15. Albert R, Jeong H, Barabasi AL: **Error and attack tolerance of complex networks**. *Nature* 2000, **406**:378-382.

16. Rzhetsky A, Gomez SM: **Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome**. *Bioinformatics* 2001, **17**:988-996.

17. Wagner A: **Asymmetric functional divergence of duplicate genes in yeast**. *Mol Biol Evol* 2002, **19**:1760-1768.

18. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: **Protein function in the post-genomic era**. *Nature* 2000, **405**:823-826.

19. Wojcik J, Schachter V: **Protein-protein interaction map inference using interacting domain profile pairs**. *Bioinformatics* 2001, **17(suppl 1)**:S296-S305.

20. Sprinzak E, Margalit H: **Correlated sequence-signatures as markers of protein–protein interaction**. *J Mol Biol* 2001, **311**:681-692.

21. • Deng M, Mehta S, Sun F, Chen T: **Inferring domain–domain interactions from protein–protein interactions**. *Genome Res* 2002, **12**:1540-1548.
A maximum likelihood estimation method for the prediction of protein–protein interaction that utilizes conserved patterns of domain–domain interactions.

22. Lu L, Lu H, Skolnick J: **MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading**. *Proteins* 2002, **49**:350-364.

23. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, Garrels J, Vincent S, Vidal M: **Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or ''interologs''**. *Genome Res* 2001, **11**:2120-2126.

24. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S *et al.*: **A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules**. *Science* 2002, **295**:321-324.

25. • Zhou X, Kao MC, Wong WH: **Transitive functional annotation by shortest-path analysis of gene expression data**. *Proc Natl Acad Sci USA* 2002, **99**:12783-12788.
A network-topology-based method of functional annotation.

26. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network**. *Science* 2001, **292**:929-934.

27. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al.*: **Transcriptional regulatory networks in *Saccharomyces cerevisiae***. *Science* 2002, **298**:799-804.

28. 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.

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

30. 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.

31. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology**. *Nature* 1999, **402**:47-52.

32. Tyson JJ, Chen K, Novak B: **Network dynamics and cell physiology**. *Nat Rev Mol Cell Biol* 2001, **2**:908-916.

33. 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.

34. 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.

35. Boulton SJ, Gartner A, Reboul J, Vaglio P, Dyson N, Hill DE, Vidal M: **Combined functional genomic maps of the *C. elegans* DNA damage response**. *Science* 2002, **295**:127-131.

36. Walhout AJ, Reboul J, Shtanko O, Bertin N, Vaglio P, Ge H, Lee H, Doucette-Stamm L, Gunsalus KC, Schetter AJ *et al.*: **Integrating interactome, phenome and transcriptome mapping data for the *C. elegans* germline**. *Curr Biol* 2002, **12**:1952-1958.

37. 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.

38. •• 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.
A comprehensive study of protein–protein interactions in yeast that uses tandem affinity purification (TAP) methodology to purify tagged protein complexes followed by mass spectrometric identification of individual components of the complexes. The native expression levels and stringent complex purification protocol ensure high-quality results.

39. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T *et al.*: **Global analysis of protein activities using proteome chips**. *Science* 2001, **293**:2101-2105.

40. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H *et al.*: **Systematic genetic analysis with ordered arrays of yeast deletion mutants**. *Science* 2001, **294**:2364-2368.

41. • Xenarios I, Salwinski L, Duan XQJ, Higney P, Kim SM, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions**. *Nucleic Acids Res* 2002, **30**:303-305.
See annotation [43•].

42. • Bader GD, Donaldson I, Wolting C, Ouellette BFF, Pawson T, Hogue CWV: **BIND - The biomolecular interaction network database**. *Nucleic Acids Res* 2001, **29**:242-245.
See annotation [43•].

43. • Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database**. *FEBS Lett* 2002, **513**:135-140.
These three references [41•–43•] describe the most comprehensive protein–protein interaction databases available publicly online.

44. 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.

45. Duan XJ, Xenarios I, Eisenberg D: **Describing biological protein interactions in terms of protein states and state transitions: the LiveDIP database**. *Mol Cell Proteomics* 2002, **1**:104-116.

46. 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.

47. 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.

48. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact**. *Trends Biochem Sci* 1998, **23**:324-328.

49. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling**. *Proc Natl Acad Sci USA* 1999, **96**:2896-2901.

50. 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* 1999, **96**:4285-4288.

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

52. Ng SK, Zhang Z, Tan SH, Lin K: **InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes**. *Nucleic Acids Res* 2003, **31**:251-254.

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

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

55. Pazos F, Valencia A: **Similarity of phylogenetic trees as indicator of protein-protein interaction**. *Protein Eng* 2001, **14**:609-614.

56. Liberles DA, Thoren A, Heijne GV, Elofsson A: **The use of phylogenetic profiles for gene predictions**. *Current Genomics* 2002, **3**:131-137.

57. Vert JP: **A tree kernel to analyse phylogenetic profiles**. *Bioinformatics* 2002, **18(suppl 1)**:S276-S284.