Protein interaction databases

Ioannis Xenarios* and David Eisenberg†

Life depends on the interaction of proteins. The availability of the complete human genome sequence has highlighted the need for a tool to analyse protein interactions and several databases have been compiled for this purpose. These databases document, categorize, and analyze interacting proteins and the cellular functions of the interactions.

Addresses
UCLA-DOE Laboratory of Structural Biology & Molecular Medicine, University of California, Los Angeles, PO Box 951570, Los Angeles, California 90095-1570, USA
*e-mail: ixenario@mbi.ucla.edu
†e-mail: david@mbi.ucla.edu

Current Opinion in Biotechnology 2001, 12:334–339

0958-1669/01/$ – see front matter

Abbreviation
DIP Database of Interacting Proteins

Introduction

Virtually every cellular process is regulated by protein–protein interactions, ranging from the interactions of allosteric protomers that control metabolic fluxes to the phosphorylation of enzymes that control signal transduction cascades [1•] and the workings of complex molecular machines, such as the proteasome that mediates proteolysis. These regulatory processes have long been studied and many are well known — but many more remain uncharacterized. Recently, however, the availability of complete genome sequences and DNA microarray data has changed the way biologists study these processes, broadening the focus from a single gene or protein to the whole genome or multiple genomes.

Fully sequenced genomes lead to additional insights into the functional properties of the encoded proteins. These functional insights emerge as networks of interacting proteins. Understanding interactions between encoded proteins of a given genome is a critical step in functional genomic analysis [2•] .

To document and describe protein interactions, several databases of interacting proteins have been compiled. Interaction databases record the growing body of observations of protein–protein interactions in cells. Recently, several articles were published describing the large-scale screening of protein interactions using two-hybrid assays [3•,4•,5,6] (see the article by Pelletier and Sidhu in this issue pp 340–347), and it seems likely that a second wave of publications on protein interactions will appear as mass spectrometry becomes more widely applied in proteomics.

To date, however, the number of interactions reported from recent large-scale experiments is small compared with the number of interactions available from the thousands of small-scale experiments described over years in published articles. To gather details of these interactions into databases, several approaches have been developed to extract the information from titles and abstracts of articles indexed in the MEDLINE database. These approaches using statistical or semantic analysis have identified known interactions [7,8•,9] or identified articles describing protein interactions [10•]. Although these methods can quickly scan the millions of articles in MEDLINE, their fidelity is still far less than that achieved by a human curator who examines each article. Thus the manner in which protein interaction data should be curated into an interaction database remains a pivotal problem. Automated approaches may introduce more errors than true interactions, as has been demonstrated for automatic protein annotation [11]. We feel at present that a human curated database is probably the best choice, but the drawback of this approach is its slow growth. In the following, we describe the available databases of protein interactions.

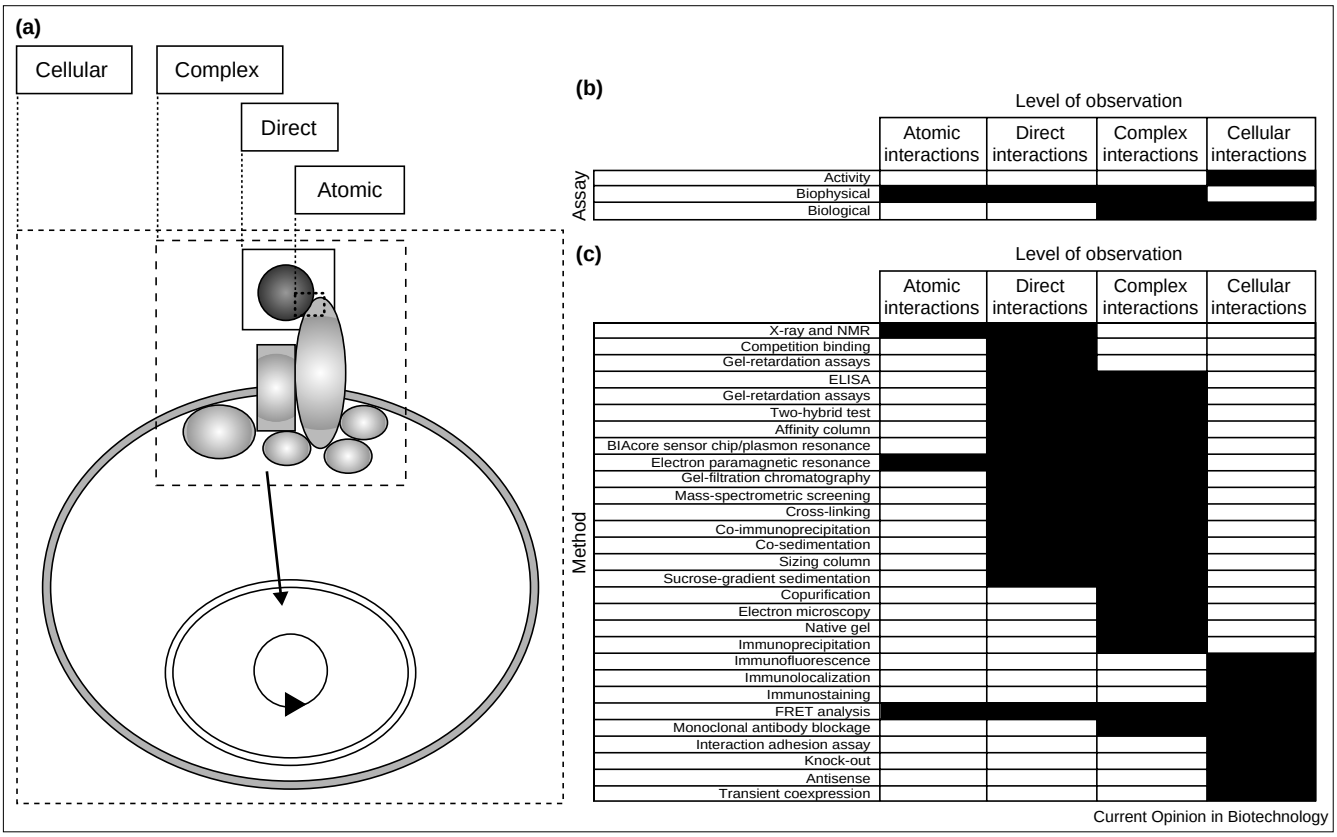
How to encode the various types of protein interactions

Protein interaction databases ideally should accommodate the full range of protein functions and interactions observed in biology. Protein interactions fall into three functional categories: metabolic and signaling (genetic) pathways; morphogenic pathways in which groups of proteins participate in the same cellular function during a developmental process; and structural complexes and molecular machines in which numerous macromolecules are brought together. The inherent complexity of interaction data leads to the design of various data structures to store interaction information [12]. Examples include the protein-based interaction partners described in the Database of Interacting Proteins (DIP) [13•] and the inclusion of non-protein partners (e.g. RNA, DNA and small molecules) as described in the Biomolecular Interaction Network Database (BIND) [14•].

Experimental methods and protein interactions

Experimental methods available to detect protein interactions vary in their level of resolution (Figure 1a). These observations can be classified into four categories. The first comprises an ‘atomic observation’ in which the protein interaction is detected using, for example, X-ray crystallography. These experiments can yield specific information on the atoms or residues involved in the interaction. Second, is a ‘direct interaction observation’ where protein interaction between two partners can be detected as in a BIAcore measurement or a two-hybrid experiment. At a third level of observation, multiprotein complexes can be detected using methods such as immunoprecipitation or mass-spectral analysis. This type of experiment does not

Figure 1



Methods to detect protein interactions. (a) Different levels of observation of an interaction. A hypothetical complex is shown where a protein ligand (dark gray circle) interacts with a receptor (light gray ellipse). The atomic level is represented by a thick dotted box, the direct level of observation by a solid box, the complex level by the dashed box and the cellular level by the thin dotted box. The nucleus and its associated cell cycle are depicted together with a hypothetical signaling cascade. (b) Categorization of protein interaction detection methods depending on their levels of observation. The three main assay categories – activity, biophysical and biological – are represented and their particular range of observation is indicated (filled boxes). (c) Examples of different methods and their associated level of observation (filled boxes). ELISA, enzyme-linked immunosorbent assay; FRET, fluorescense resonance energy transfer.

unveil the chemical detail of the interactions or even reveal which proteins are in direct contact but gives information as to which proteins are found in a complex at a given time. The fourth category comprises measurements at the cellular level, where an 'activity bioassay' is used to observe an interaction; for example, proliferation assays of cells stimulated by a receptor–ligand interaction. Here again, the exact nature of the interaction is not known but the biological readout allows inference of the potential function of a given interaction. Some experimental methods span more than one level of observation (Figure 1b). In terms of experimental data available in the literature, the complex interaction category is the most commonly represented, followed by the cellular interaction, the direct interaction, and finally the atomic observation category.

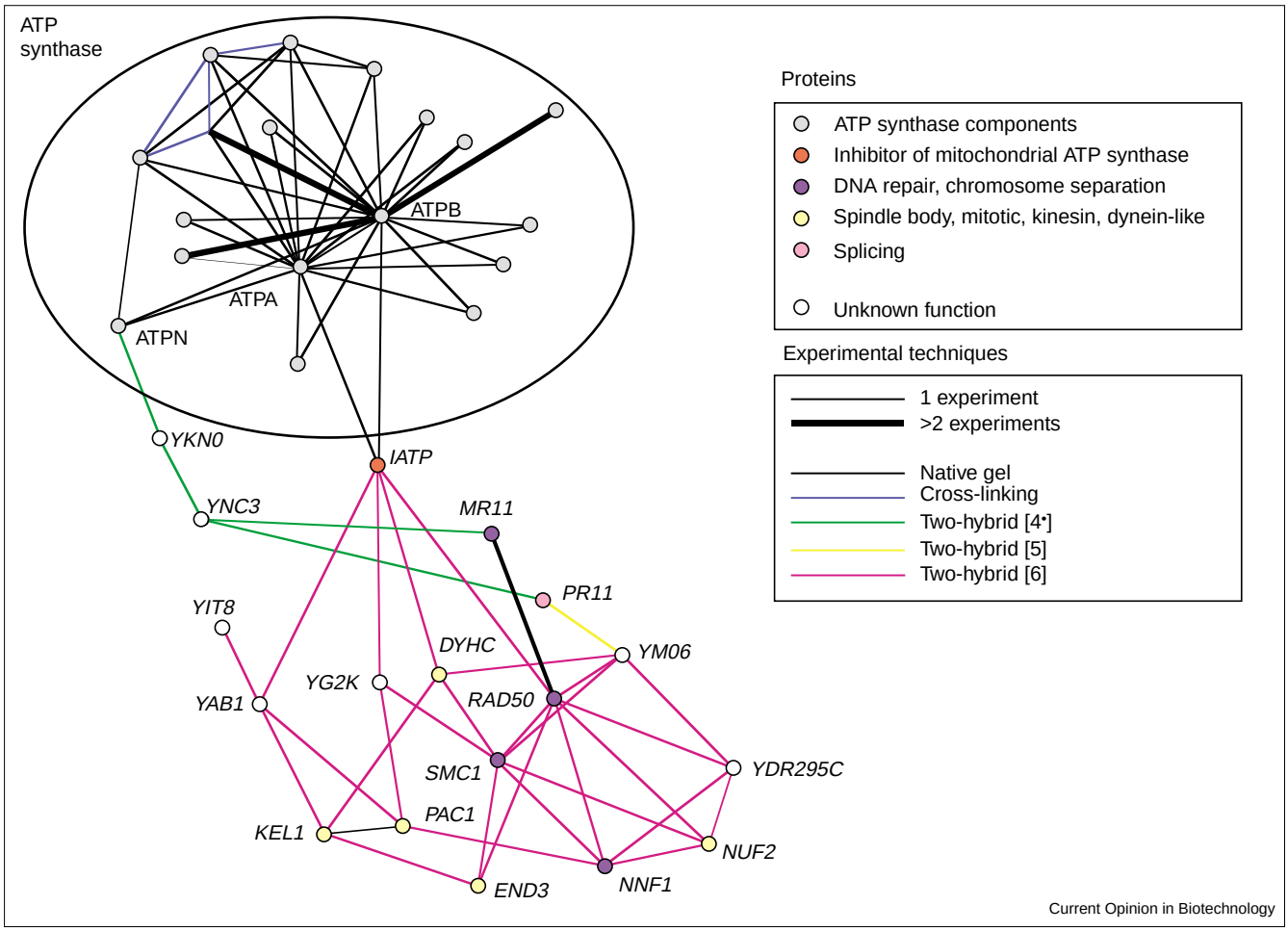
Need for confidence levels

Proteins interact with one another with a wide-range of affinities and timescales. As of the year 2000, relatively few interactions have been characterized in terms of energetics and even fewer in terms of kinetics. Many physiological interactions are transient and weak. Consequently, detection of interaction is often at the margin of observation, and non-physiological interactions result in noise. Hence, it is desirable to assign a confidence level to each observation.

Confidence in the validity of a reported interaction is enhanced by observations of the same interaction by other methods. For example, if an interaction is measured with two distinct experiments, one using the two-hybrid assay and another using immunoprecipitation, the joint observation increases our confidence in this particular interaction. The higher confidence level is indicated in the DIP by a thicker line between two proteins (see Figure 2). Some methods of observation are more prone to false results, however, and so in the DIP the method of detection is given as an essential field for evaluation by the user.

Another approach to validating a given interaction is the use of subcellular colocalization. If an interaction occurs between two proteins that are known to be both localized to the same subcellular compartment, the likelihood of

Figure 2



Reconstruction of the yeast ATP synthase protein network (enclosed within the circle). Shown below are two different interaction pathways: one through the inhibitor of mitochondrial ATP synthase (IATP) and an alternative interaction through two proteins of unknown function (YKN0 and YCN3). Line thickness represents the number of experiments describing a given interaction. Lines are color-coded according to the experiment: native-gel electrophoresis (black), cross-linking (blue). Three classes of two-hybrid experiment are represented depending on the group that generated the data (green, Uetz *et al.* [4]; yellow, Newman *et al.* [5]; magenta, Fromont-Racine *et al.* [6]). Each protein is labeled using the first letters of its SwissProt accession code (e.g. YNC3_YEAST). The different functional groups are color-coded as indicated in the key. Data came from the DIP database (http://dip.doe-mbi.ucla.edu).

physiological interaction is increased. This type of validation might be useful in increasing confidence levels in protein interactions detected by error-prone methods [15•].

Another approach used in the case of the two-hybrid assay evaluates the interacting protein fragments, as described by Rain *et al.* [16]. In this approach, fragments of proteins are used in the yeast two-hybrid screening process; the relative signal given by each protein fragment can be correlated with its propensity to interact with another protein fragment. There are two advantages to this method: a minimally interacting region can be determined and fragments that bind to a higher fraction of proteins than expected at random can be removed or down-weighted as being promiscuous [17]. This approach relies on the ability of the protein fragment to refold in the same conformation as in the context of the full-length protein [18]. The database curator or the user can employ such considerations to ultimately filter interaction data.

Where can we find protein interaction information?

Several biological databases are dedicated to protein interactions; we summarize some of these interaction databases below and in Table 1.

DIP

The DIP is a database that documents experimentally determined protein–protein interactions. It provides a comprehensive and integrated tool for browsing and extracting information about protein interactions. It contains pairwise interactions between proteins [13•,19]. The basic structure on which the data are stored has been extended to add additional information on the position of an interaction in a

Table 1

Protein interaction databases available on the World Wide Web.

Database name	Acronyms	URL	Reference	Content	Number of interactions	Academic version	Commercial version
Database of Interacting Proteins	DIP	dip.doe-mbi.ucla.edu	[13•]	Catalog of protein–protein interactions	9700	Yes	Yes
Biomolecular Interaction Network Database	BIND	www.bind.ca	[14•]	Molecular interaction complexes and pathways	5800	Yes	No
Munich Information Center for Protein Sequences	MIPS	www.mips.biochem.mpg.de	[21]	List of protein interactions with description of the methods involved	2400	Yes	No
Proteome	PROTEOME	www.proteome.com	[23••]	Integration of protein information, function, localization and interactions	NP	Yes	Yes
Protein interaction on the Web	PRONET	pronet.doubletwist.com		Protein interaction and signaling database	NP	Yes	Yes
Hybrigenics	PIM	www.hybrigenics.fr	[16]	Protein interactions for *Helicobacter pylori*	1400*	Yes	Yes
Curagen	CURAGEN	www.curagen.com	[4•]	Protein interactions from two-hybrid screen using full-length yeast proteins	1500†	Yes	Yes

The last two columns indicate whether an academic or commercial user can access the data. NP; not published; *published at the time of [14]; †published at the time of [4•].

pathway and on specific post-translational modifications. The DIP allows the visual representation and navigation of protein-interaction networks. The quality of a given interaction can be assessed visually by the thickness of the lines between two proteins and the selection of a specific method can be applied to show the results from only a given method (Figure 2). The DIP allows the integration of a diverse body of information onto a protein-interaction network, such as the predominance of certain domains or the different subcellular compartments in which a protein can be found. The data files are available for non-commercial users who are interested in studying protein interactions. So far, roughly 9700 interactions among 5700 proteins are available.

BIND

The BIND database possesses a data structure that can store a variety of interactions between molecular compounds and includes protein–protein, protein–RNA, protein–DNA and protein–small-molecule interactions. It presents protein interactions from the molecular level to the pathway level. The structure of the database has been published [20] and the first data released [14•]. At present 5800 protein interactions are present; its inherent structure allows a wide variety of information to be included.

MIPS

The Munich Information Center for Protein Sequences (MIPS) has long provided protein sequence information for yeast and other model organisms [21]. This database also contains compiled protein interaction data for yeast, which is available for download. So far, roughly 2400 unique protein interactions are deposited in the database. Multiple experimental techniques for a given interaction are also available [22].

PROTEOME

The PROTEOME database contains information regarding many biological aspects of proteins, including cellular function, localization and protein interactions. Although it is not strictly an interaction database, a wide variety of information is combined in this impressive resource with a strong emphasis on protein function [23••]. Interactions between proteins and methods used to define an interaction are obtainable for each protein, but no visualization tools are available. Proteins from organisms such as yeast, worms and humans are currently available in this database.

PRONET

PRONET is a non-academic project funded by Doubletwist™ and Myriad™. This database contains compiled information about protein interactions and will add data produced by in-house screening.

CURAGEN

Yeast has been the preferred genetic system for many biologists. A comprehensive study of protein interactions among yeast proteins has been produced by the laboratory of S Field and the Curagen company [4•]. Data are available at the Curagen website for academic users. Visualization tools are available for the protein network.

PIM

Recently the pathogenic bacterium *Helicobacter pylori* has been studied using the genome-wide two-hybrid

assay [16]. Roughly 1400 interactions have been described in the first article, and are available on the web for academic users. One feature of this database is that it includes all the different protein fragments that were observed during the screening, and assigns a likelihood score.

Conclusions

In contrast to the protein sequence databases for which a simple structure can be defined, the diverse nature of protein interactions has hindered representation. It is therefore not surprising that interaction databases appeared only in recent years. Although biologists have revealed many protein interactions, few have been integrated in a database structure. Interaction databases fulfill various needs: first, the interaction database provides a centralized data repository allowing users to validate protein interactions by comparing results with previous experiments. For example, in the case of the two-hybrid assay, many experiments have been performed but not systematically collected and analyzed. In retrospect some interactions might have been considered as false positives or noise, even though they had been reproduced several times by other laboratories. Second, the collection and organization of known protein interactions allows navigation of the protein-interaction network and the discovery of new pathways and modes of regulation. Third, general properties of networks can be studied, as already described for biochemical networks [24].

One computationally difficult problem is the integration of data produced in various laboratories into interaction databases. One way would be to provide each data-producing laboratory with a software tool that integrates their new data with previously deposited information. In order to be successful this tool should be easy to use and should be able to encompass different types of data. For example, many laboratories have generated data on mass-spectra of peptides derived from different immunoprecipitation experiments. These data could be compared to known protein complexes with the software tool [25]. In a sense this type of resource sharing is analogous to the controversial Napster sharing, in which each user (experimentor) is connected to the Internet and distributes his/her music (interactions).

The publication of the draft human genome with only 30,000–40,000 genes emphasizes the importance of interaction databases [26,27]. The complexity and richness of human biology can be achieved by the combinatorial possibilities offered by protein interactions. Hence, interaction databases are a necessary tool for the biology of the 21st century.

Update

Since submission of this review another large dataset of protein interactions has been deposited [28]. The overlap between the different studies is limited [28,29]. It is becoming clear that scientists need to control all such studies with a common set of identical interactions. Without this control, the reported differences and limited overlap may arise from detailed differences in screening procedures and screen efficiencies.

Acknowledgements

We thank Thomas Graeber, Ralf Landgraf, Mike Sawaya and Joyce Duan for helpful discussions and critical reading of the manuscript and the Department of the Environment and National Institutes of Health for support.

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. • Pawson T, Nash P: Protein–protein interactions define specificity in signal transduction. *Genes Dev* 2000, 14:1027-1047.
Review of the different molecular pathways that control cellular behavior and explanation of the specific mechanisms that allows certain pathways to be separated or permissive to other signaling components.

2. • Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: Protein function in the post-genomic era. *Nature* 2000, 405:823-826.
Review of the different *in silico* function prediction methods and their use to infer function in combination with known protein interaction networks.

3. • Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y: Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci USA* 2000, 97:1143-1147.
Analysis of yeast protein interactions using two-hybrid experiments (see also [4•]).

4. • 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.
Analysis of yeast protein interactions using two-hybrid experiments (see also [3•]).

5. Newman JR, Wolf E, Kim PS: A computationally directed screen identifying interacting coiled coils from ***Saccharomyces cerevisiae***. *Proc Natl Acad Sci USA* 2000, 97:13203-13208.

6. Fromont-Racine M, Rain JC, Legrain P: Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat Genet* 1997, 16:277-282.

7. Thomas J, Milward D, Ouzounis C, Pulman S, Carroll M: Automatic extraction of protein interactions from scientific abstracts. *Pac Symp Biocomput* 2000, 541-552.

8. • Blaschke C, Andrade MA, Ouzounis C, Valencia A: Automatic extraction of biological information from scientific text: protein–protein interactions. *Proc Int Conf Intell Syst Mol Biol* 1999, 60-67.
An efficient approach to information extraction from MEDLINE abstracts.

9. Stapley BJ, Benoit G: Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac Symp Biocomput* 2000, 529-540.

10. • Marcotte E, Xenarios I, Eisenberg D: Mining literature for protein–protein interactions. *Bioinformatics* 2001, 17:1-7.
This paper describes the use of a dataset of known protein interactions to train a statistical classifier to identify abstracts discussing protein interactions.

11. Devos D, Valencia A: Practical limits of function prediction. *Proteins* 2000, 41:98-107.

12. Eilbeck K, Brass A, Paton N, Hodgman C: INTERACT: an object oriented protein–protein interaction database. *Proc Int Conf Intell Syst Mol Biol* 1999, 87-94.

13. • Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: DIP: the database of interacting proteins. *Nucleic Acids Res* 2000, 28:289-291.
A description of a database of protein interactions.

14. • Bader GD, Hogue CW: BIND — a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* 2000, 16:465-477.
This paper describes the generation of a database structure capable of encoding a wide variety of protein/RNA/DNA interactions.

15. Schwikowski B, Uetz P, Fields S: A network of protein–protein
• interactions in yeast. *Nat Biotechnol* 2000, 18:1257-1261.
Analysis of 2709 published protein interactions in yeast and the use of subcellular localization and functional annotation to assess the reliability of interactions.

16. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S, Lenzen G, Petel F, Wojcik J, Schachter V *et al.*: The protein–protein interaction map of ***Helicobacter pylori***. *Nature* 2001, 409:211-215.

17. Legrain P, Selig L: Genome-wide protein interaction maps using two-hybrid systems. *FEBS Lett* 2000, 480: 32-36.

18. Legrain P, Jestin JL, Schachter V: From the analysis of protein complexes to proteome-wide linkage maps. *Curr Opin Biotechnol* 2000, 11:402-407.

19. Xenarios I, Fernandez E, Salwinski L, Duan XJ, Thompson MJ, Marcotte EM, Eisenberg D: DIP: the database of interacting proteins: 2001 update. *Nucleic Acids Res* 2001, 29:239-241.

20. 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.

21. Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, Lemcke K, Mannhaupt G, Pfeiffer F, Schuller C *et al.*: MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 2000, 28:37-40.

22. Fellenberg M, Albermann K, Zollner A, Mewes HW, Hani J: Integrative analysis of protein interaction data. *Proc Int Conf Intell Syst Mol Biol* 2000, 8:152-161.

23. Costanzo MC, Crawford ME, Hirschman JE, Kranz JE, Olsen P,
•• Robertson LS, Skrzypek MS, Braun BR, Hopkins KL, Kondu P *et al.*: YPD™, PombePD™ and WormPD™: model organism volumes of the BioKnowledge™ library, an integrated resource for protein information. *Nucleic Acids Res* 2001, 29:75-79.
Describes the different ways biological knowledge can be searched and stored, in this most comprehensive database of protein function.

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

25. Mann M, Pandey A: Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. *Trends Biochem Sci* 2001, 26:54-61.

26. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al.*: The sequence of the human genome. *Science* 2001, 291:1304-1351.

27. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.*: Initial sequencing and analysis of the human genome. *Nature* 2001, 409:860-921.

28. 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.

29. Hazbun TR, Fields S: Networking proteins in yeast. *Proc Natl Acad Sci USA* 2001, 98:4277-4278.