Comparative assessment of large-scale data sets of protein–protein interactions

Christian von Mering*, Roland Krause†, Berend Snel*, Michael Cornell‡, Stephen G. Oliver‡, Stanley Fields§ & Peer Bork*

* *European Molecular Biology Laboratory, Meyerhofstrasse 1, 69012 Heidelberg, Germany*
† *Cellzome AG, Meyerhofstrasse 1, 69117 Heidelberg, Germany*
‡ *School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Oxford Road, Manchester M13 9PT, UK*
§ *Departments of Genome Sciences and Medicine, Howard Hughes Medical Institute, University of Washington, Box 357730, Seattle, Washington 98195, USA*

Comprehensive protein–protein interaction maps promise to reveal many aspects of the complex regulatory network underlying cellular function. Recently, large-scale approaches have predicted many new protein interactions in yeast. To measure their accuracy and potential as well as to identify biases, strengths and weaknesses, we compare the methods with each other and with a reference set of previously reported protein interactions.

For an increasing number of organisms, we can now list the genes and encoded proteins[1,2]. It is the proteins that execute the genetic programme, and those that are actually produced by a cell at any given time constitute its 'proteome'[3]. The proteome is much more dynamic than the genome: it changes during development and in response to external stimuli, and the proteins form large interaction networks, in which they regulate and support each other[4–6]. To fully understand the cellular machinery, simply listing the proteins is not enough—all the interactions between them need to be delineated as well.

Traditionally, protein interactions have been studied individually by genetic, biochemical and biophysical techniques. However, the speed with which new proteins are being discovered or predicted has created a need for high-throughput interaction-detection methods. Consequently, in the last two years, methods have been introduced that can globally tackle the problem, resulting in a vast amount of interaction data[7–14]. To use these data efficiently, a critical evaluation of their accuracy, biases, overlaps and complementarities is essential. Here, we analyse data sets from yeast two-hybrid systems[7,8], protein complex purification techniques using mass spectrometry[10,11], correlated messenger RNA expression profiles[15,16] and genetic interaction data[9,17], as well as '*in silico*' (computed) interaction predictions derived from gene context analysis (gene fusion[18,19], gene neighbourhood[14,20] and gene co-occurrences or phylogenetic profiles[21,22]). We compare the methods with each other, focusing on the yeast proteome (for details of the methods, see Box 1). Although all of these techniques can be used for interaction prediction, their goals are different. Yeast two-hybrid and mass spectrometry techniques aim to detect physical binding between proteins, whereas genetic interactions, mRNA coexpression and *in silico* methods seek to predict functional associations, for example, between a transcriptional regulator and the pathway it controls. In many cases, however, such functional associations do take the form of physical binding[20,23].

Comparing interaction data is difficult, because they are often derived under different conditions, come in different formats (see Box 2), and need to be benchmarked against a trusted reference set. For this study, we chose binary interactions as the common unit of analysis, and we relied on manually curated catalogues of known protein complexes (Munich Information Center for Protein Sequences, MIPS[17], and the Yeast Proteome Database, YPD[24]) as the trusted reference.

Overlaps and complementarities

About 80,000 interactions between yeast proteins are currently available from the different high-throughput methods (Fig. 1) (the exact number depends on filtering criteria). Of these, only a surprisingly small number (~2,400) is supported by more than one method. There are three possible explanations for this: the methods may not have reached saturation; many of the methods may produce a significant fraction of false positives; and some methods may have difficulties for certain types of interactions, resulting in complementarities between the methods.

We note, for example, that each technique produces a unique distribution of interactions with respect to functional categories of interacting proteins (Fig. 1). These differences in coverage suggest that the methods have specific strengths and weaknesses. The data sets based on purified complexes, for example, predict relatively few interactions for proteins involved in transport and sensing (possibly because these are enriched in transmembrane proteins, which are more difficult to purify). Similarly, interactions detected by the yeast two-hybrid technology largely fail to cover certain categories; for example, proteins involved in translation are found comparatively less often than by other methods.

As a specific example for the complementarity between the data sets, we considered glycine decarboxylase, which is a well-characterized multi-enzyme complex needed when glycine is used as a one-carbon source[25–27]. It consists of the four proteins Gcv1, Gcv2, Gcv3 and Lpd1. This complex is not detected in the systematic purification of complexes[10] (using tandem affinity purification (TAP)-tagged Gcv3 as a bait), presumably because it is not expressed in yeast cells grown on rich medium. Three of the four proteins, however, can be confidently linked to each other through the remaining data sets, by a total of nine links involving synexpression (correlated mRNA expression) and *in silico* predictions (see Supplementary Information).

A converse example is the protein PPH3, for which neither *in silico* methods nor synexpression studies predict any interaction, but which is nevertheless consistently detected in a protein complex by mass spectrometry. PPH3 is a protein phosphatase distantly related to PP2A; very little is known about its function or interaction partners[28,29]. In the high-throughput data, two uncharacterized proteins (YBL046W and YNL201C) are invariably found with PPH3 in four independent mass-spectrometry purifications, thereby clearly defining a protein complex (the two are also joined by a two-hybrid interaction).

Even data sets based on the same technique can complement each other to some extent. The two mass-spectrometry approaches, for

example, differ in how they express the tagged 'baits': the TAP approach[10] relies on genomic integration (using the endogenous promoter), whereas the approach of high-throughput mass-spectrometry protein complex identification (HMS-PCI)[11] employs inducible overexpression. Both systems have their advantages: the endogenous promoter will often maintain the stoichiometry of interacting proteins and minimize artificial overexpression, whereas inducible expression allows the purification of proteins that are normally not expressed under laboratory conditions, and can be used in species where genomic integration is not feasible, such as the human. Of the proteins that could not be expressed in the TAP approach, 90 were successfully purified by HMS-PCI, producing more than 600 interactions.

Benchmarking high-throughput interactions

When assessing the quality of interaction data, coverage and accuracy need to be considered together. A data set of high coverage is not very useful if its accuracy is low (that is, it contains many false positives), and vice versa. Comparing the data with a reference set of trusted interactions allows the estimation of lower limits for accuracy and coverage. Figure 2 summarizes how these values relate to each other for the different data sets. Such a comparison can provide only a very rough picture, because it is a snapshot of ongoing efforts, because it is based on a particular framework for

Box 1 High-throughput methods for detecting protein interactions

Yeast two-hybrid assay. Pairs of proteins to be tested for interaction are expressed as fusion proteins ('hybrids') in yeast: one protein is fused to a DNA-binding domain, the other to a transcriptional activator domain. Any interaction between them is detected by the formation of a functional transcription factor[7,8,41,42]. Benefits: it is an *in vivo* technique; transient and unstable interactions can be detected; it is independent of endogenous protein expression; and it has fine resolution, enabling interaction mapping within proteins. Drawbacks: only two proteins are tested at a time (no cooperative binding); it takes place in the nucleus, so many proteins are not in their native compartment; and it predicts possible interactions, but is unrelated to the physiological setting.

Mass spectrometry of purified complexes. Individual proteins are tagged and used as 'hooks' to biochemically purify whole protein complexes. These are then separated and their components identified by mass spectrometry. Two protocols exist: tandem affinity purification (TAP)[10,43], and high-throughput mass-spectrometric protein complex identification (HMS-PCI)[11,44]. Benefits: several members of a complex can be tagged, giving an internal check for consistency; and it detects real complexes in physiological settings. Drawbacks: it might miss some complexes that are not present under the given conditions; tagging may disturb complex formation; and loosely associated components may be washed off during purification.

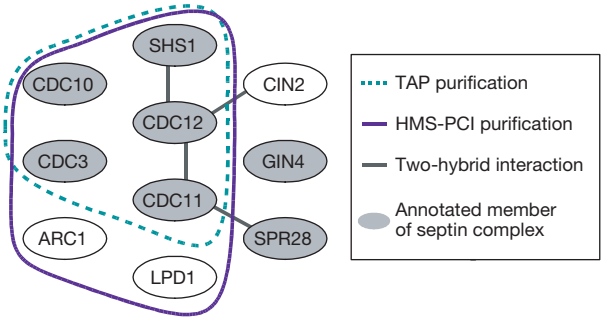
Correlated mRNA expression (synexpression). mRNA levels are systematically measured under a variety of different cellular conditions, and genes are grouped if they show a similar transcriptional response to these conditions. These groups are enriched in genes encoding physically interacting proteins[23]. Benefits: it is an *in vivo* technique, albeit an indirect one; and it has much broader coverage of cellular conditions than other methods. Drawbacks: it is a powerful method for discriminating cell states or disease outcomes, but is a relatively inaccurate predictor of direct physical interaction; and it is very sensitive to parameter choices and clustering methods during analysis.

Genetic interactions (synthetic lethality). Two nonessential genes that cause lethality when mutated at the same time form a synthetic lethal interaction. Such genes are often functionally associated and their encoded proteins may also interact physically. This type of genetic interaction is currently being studied in an all-versus-all approach in yeast[9]. Benefits: it is an *in vivo* technique, albeit an indirect one; and it is amenable to unbiased genome-wide screens.

***In silico* predictions through genome analysis.** Whole genomes can be screened for three types of interaction evidence: (1) in prokaryotic genomes, interacting proteins are often encoded by conserved operons[13,14,20]; (2) interacting proteins have a tendency to be either present or absent together from fully sequenced genomes[21,22], that is, to have a similar 'phylogenetic profile'; and (3) seemingly unrelated proteins are sometimes found fused into one polypeptide chain. This is an indication for a physical interaction[18,19]. Benefits: fast and inexpensive *in silico* techniques; and coverage expands as more genomes are sequenced. Drawbacks: it requires a framework for assigning orthology between proteins, failing where orthology relationships are not clear; and so far it has focused mainly on prokaryotes.

Box 2 Counting interactions

Interaction data come in two formats: binary interactions or groups of interacting partners, shown here for a known protein complex, and the high-throughput data that support it (the septin complex, not all interactions are indicated).



For any quantitative comparison of data sets, a common unit of analysis needs to be defined, which in this case means focusing either on proteins or on binary interactions. It makes a difference: for the example shown, the two-hybrid data have a false-positive rate of 1 in 5 when counting proteins, but it is 1 in 4 when counting interactions. Counting proteins seems more intuitive at first glance, but it becomes complicated when comparing more than two partially overlapping experiments. Also, counting proteins has a lower resolution because it does not discriminate whether a protein is involved in one or more interactions.

When focusing on interactions, data that come in groups need to be expanded to all possible binary interactions within a group, and can then be compared with other data sets. For example, consider the overlap of the two mass-spectrometry approaches[10,11]. When counting interactions, the overlap stands at 1,728 shared binary interactions (which is 27.5% of the TAP interactions[10] and 19.2 % of the HMS-PCI interactions[11], considering only proteins present in both data sets). This overlap is more difficult to define when counting proteins, and can be as high as 42% of the TAP data and 33% of the HMS-PCI data. (This depends on how one counts (see Supplementary Information). The largest overlap between two purifications is 18 proteins.) Importantly, the overall trend stays the same, irrespective of the unit of analysis.

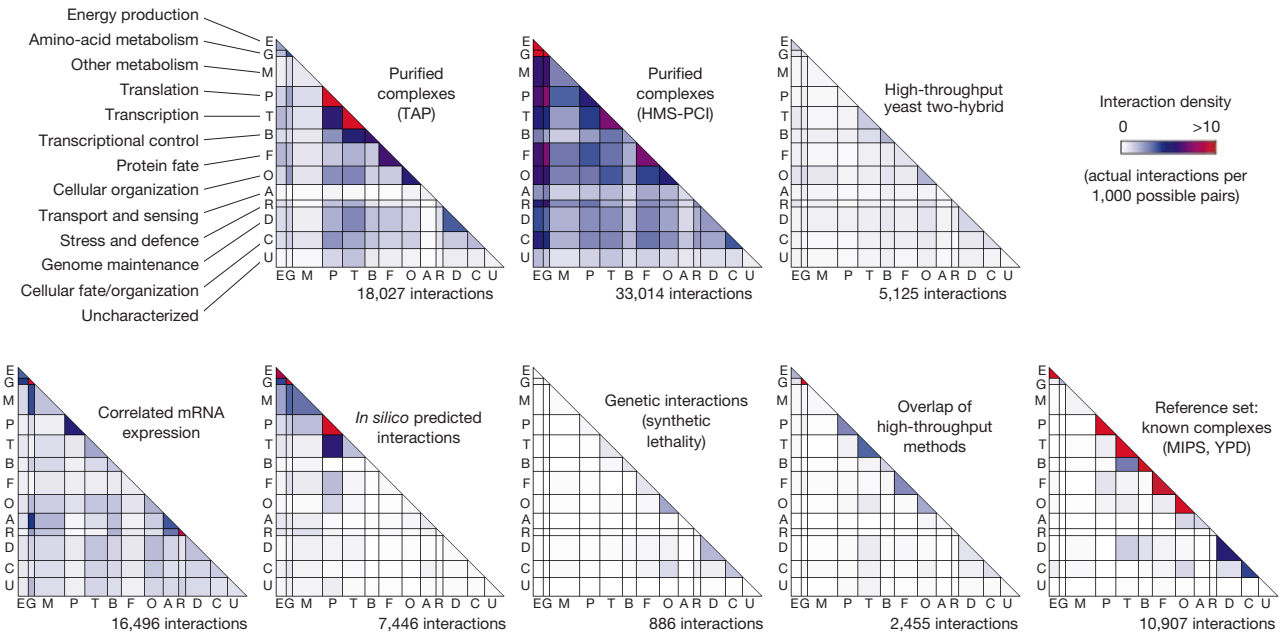


Figure 1 Large-scale interaction data and the distribution of interactions according to functional categories. Each data set is represented by a matrix showing the distribution of interactions (interaction density[23]) by colour. Each axis on a matrix represents the entire yeast genome, which has been subdivided into functional categories using a catalogue of known and predicted protein functions at MIPS[17]. The 'uncharacterized' category is not drawn to scale, because it would encompass more than a third of each axis. Some categories were fused for conciseness, and genes annotated in multiple categories were manually assigned to one. For the large-scale purification of protein complexes, two data sets are shown separately (one based on the TAP system[10], the other based on HMS-PCI[11]), because these are the largest to date and technical details vary considerably between them. The synthetic lethal interactions come from one initial high-throughput screen[9] (295 interactions), but also from individual screens and experiments compiled at the MIPS database[17] (note that these are derived from the literature and might thus not be entirely independent from the reference set). For this and subsequent figures, details on data sets, parameters and more examples are available in the Supplementary Information.

counting and defining interactions, and because the reference set is necessarily incomplete and may well have unknown biases itself. Nevertheless, it is evident that there are large differences between the methods and even within a method when parameters are changed. As noted previously[12,30], the highest accuracy is achieved for interactions supported by more than one method (Fig. 2).

There are of course many different and valid ways to count and compare interactions. In the HMS-PCI study[11], for example, only the interactions between the bait and the co-purified proteins were counted, not the interactions among all the proteins in a purification. We can confirm that this increases the accuracy (from 2 to 6.8% for HMS-PCI and from 12.5 to 27.8% for TAP), but it is concomitant with a strong decrease in coverage (see Supplementary Information).

An independent measure of quality is the degree to which interacting proteins are annotated with the same functional category (Fig. 1): for highly accurate data sets, t a intercetiTns e ndeto gusoer on the diagonal, which shows that proteins of broadly related functions preferentially interact with each other. This correlation suggests that the interactions outside of the diagonal consist largely of false positives. We note that the reference set is particularly well clustered on the diagonal, as is the overlap of high-throughput data (Fig. 1).

Biases in interaction coverage

None of the methods covers more than 60% of the proteins in the yeast genome. Are there common biases as to which proteins are covered? We identify three areas where the high-throughput interaction data are indeed biased. First, there is a bias towards proteins of high abundance. There are no genome-wide measurements of protein abundance in yeast, but mRNA levels can be used as a crude substitute[31,32]. A plot of interaction coverage versus mRNA abundance (Fig. 3) shows that most protein interaction data sets (including the curated complexes) are heavily biased towards proteins of high abundance. However, the two genetic approaches

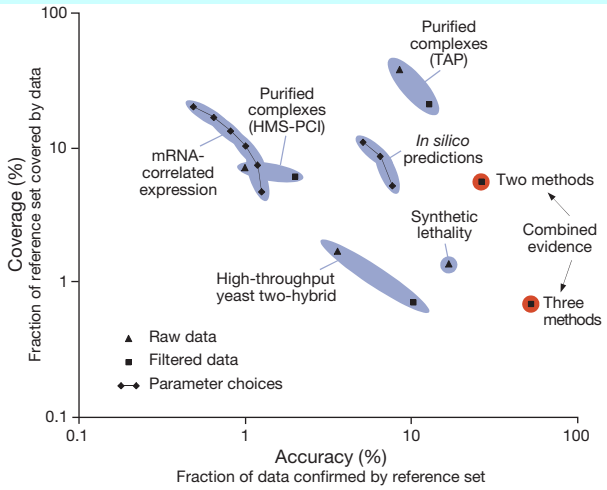


Figure 2 Quantitative comparison of interaction data sets. The various data sets are benchmarked against a reference set of 10,907 trusted interactions, which are derived from protein complexes annotated manually at MIPS[17] and YPD[24]. Coverage and accuracy are lower limits owing to incompleteness of the reference set. Each dot in the graph represents an entire interaction data set, and its position specifies coverage and accuracy (on a log–log scale). For the combined evidence, we considered only interactions supported by an agreement of two (or three) of any of the methods shown. For most data sets, raw and filtered data are shown, demonstrating the trade-off between coverage and accuracy achieved by filtering (see Supplementary Information for details on the filtering).

(two-hybrid and synthetic lethality) appear relatively unbiased. This is intriguing because these are the two methods that are, by design, largely independent of endogenous protein levels. Moreover, these two methods are especially capable of detecting transient or indirect interactions. The observed bias in the other data sets could thus reflect mainly experimental limitations, which would indicate that a large body of interactions remains undiscovered for proteins of low abundance.

Second, the data sets are biased towards particular cellular localizations of interacting proteins (Fig. 4a), for example, towards mitochondrial proteins in the case of the *in silico* predictions. As well as identifying biases, protein localization data provide an independent measure of quality for the different data sets, because proteins known to interact are usually localized similarly (Fig. 4b). Third, there is a bias in interaction coverage that relates to the degree of evolutionary novelty of proteins. We note that proteins restricted to yeast are less well covered than ancient, evolutionarily conserved proteins (see Supplementary Information).

Outlook

How many protein–protein interactions can be expected in yeast? A minimum estimate can be made by comparing the currently annotated interactions to those high-throughput interactions that are supported by more than one method. This overlap of high-throughput data is around 20 times larger than would be expected by chance (compared with randomized data sets; see Supplementary Information), which indicates a good signal-to-noise ratio. Furthermore, it consists mainly of interactions in which both partners have the same functional category and cellular localization (Figs 1 and 4). Both observations suggest that the overlap consists largely of true positives. We note, however, that less than a third of these are annotated as previously known, indicating that there should be at least three times as many interactions in yeast as there are described today. Roughly 10,000 interactions are currently known[17,24], which leads to a lower estimate of 30,000 interactions in yeast. The actual number, however, will probably be much higher because protein expression and interaction patterns will change during development or morphogenesis, or in response to the many different external conditions to which yeast may be exposed in its natural environment.

Among the interactions proposed by high-throughput methods will be many false positives. In fact, we estimate that more than half of all current high-throughput data are spurious. This estimate is based on the frequent linkage of functionally unrelated proteins, often from distinct cellular compartments, which is in contrast to the reference set and the overlap data. (On average, only 21% of high-throughput interactions link proteins of the same functional category. In reality, this fraction should be higher: the overlap data has 48% and the reference set more than 80%. A similar argument can be made for the data concerning proteins from distinct cellular compartments (see Fig. 4).) For a filtered yeast two-hybrid data set, which shows medium accuracy in our benchmark (Fig. 2), the fraction of false positives has also been predicted[8,33] to be of the order of 50%.

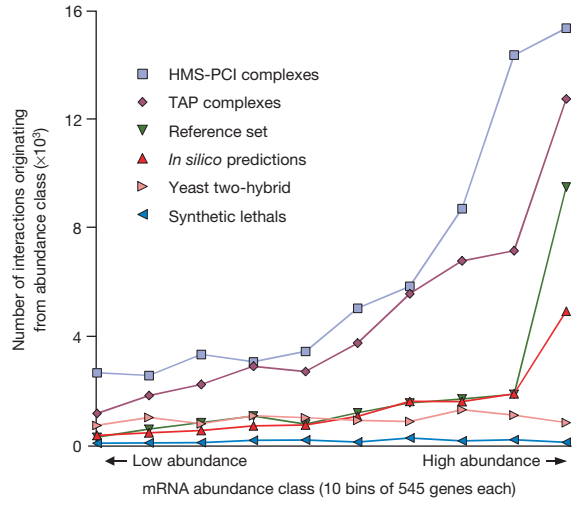


Figure 3 A bias in interaction coverage from mRNA abundance data. Using data from a survey of mRNA abundances in yeast[32], we divided the yeast genome into ten mRNA abundance classes (bins) of equal size. For each data set and abundance class, we recorded the number of interactions having at least one protein in that class. Each interaction (A–B) is counted twice: once under the abundance class of partner A, and once under the abundance class of partner B.

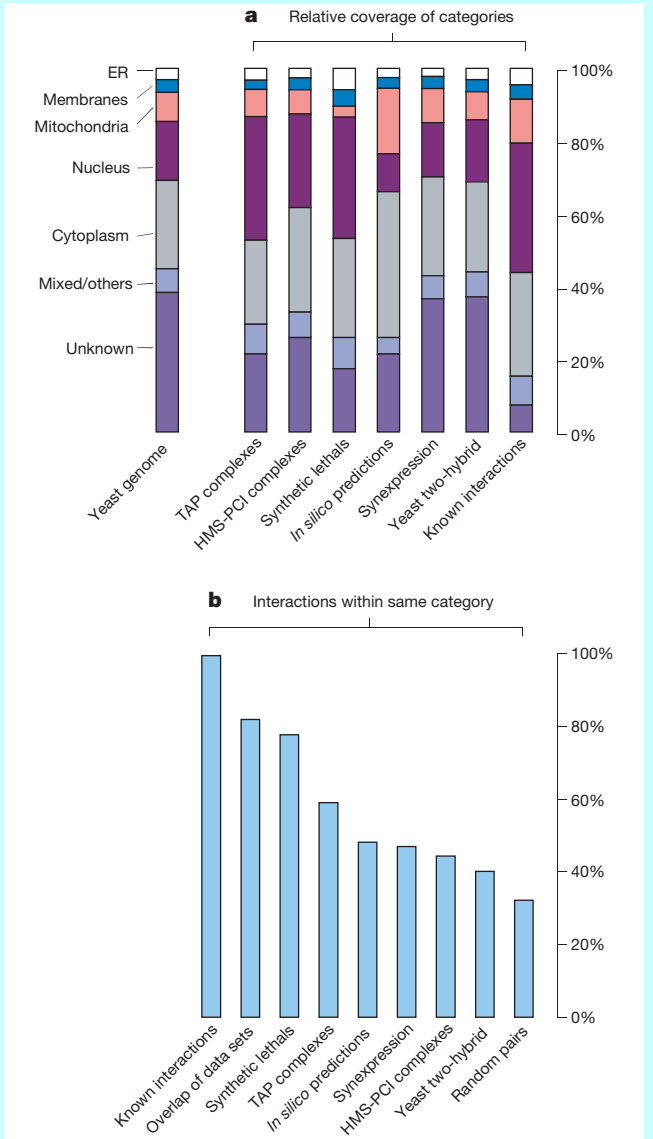


Figure 4 Protein localization and interaction coverage. Protein localizations are derived from the MIPS[17] and TRIPLES (Transposon-insertion Phenotypes, Localization, and Expression in *Saccharomyces*)[40] databases. **a**, The distribution of protein localization among the proteins covered by a data set (relative coverage). ER, endoplasmic reticulum. **b**, The fraction of interactions in which both partners have the same protein localization. Here, only proteins clearly assigned to a single category are considered.

To increase coverage and to improve confidence in detected or predicted protein interactions, as many complementary methods as possible should be used, including those studied here as well as new approaches such as protein microarrays[34,35]. Together with improvements in current technologies, transparent quality control and benchmarking, this will lead to a vast and expanding body of reliable high-throughput interaction data. Exploitation of the non-overlapping interactions will remain a challenge and will require integrated database approaches[17,36–39] as well as careful curation and validation, as has long been good practice for individual experiments in the laboratory. □

Received 27 February; accepted 17 April 2002.
Published online 8 May 2002, DOI 10.1038/nature750.

1. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30,** 38–41 (2002).
2. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **30,** 17–20 (2002).
3. Kahn, P. From genome to proteome: looking at a cell's proteins. *Science* **270,** 369–370 (1995).
4. Tucker, C. L., Gera, J. F. & Uetz, P. Towards an understanding of complex protein networks. *Trends Cell Biol.* **11,** 102–106 (2001).
5. Zamir, E. & Geiger, B. Molecular complexity and dynamics of cell-matrix adhesions. *J. Cell Sci.* **114,** 3583–3590 (2001).
6. Vidal, M. A biological atlas of functional maps. *Cell* **104,** 333–339 (2001).
7. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403,** 623–627 (2000).
8. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98,** 4569–4574 (2001).
9. Tong, A. H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294,** 2364–2368 (2001).
10. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147 (2002).
11. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415,** 180–183 (2002).
12. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402,** 83–86 (1999).
13. Huynen, M., Snel, B., Lathe, W. 3rd & Bork, P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10,** 1204–1210 (2000).
14. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA* **96,** 2896–2901 (1999).
15. Cho, R. J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2,** 65–73 (1998).
16. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102,** 109–126 (2000).
17. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30,** 31–34 (2002).
18. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402,** 86–90 (1999).
19. Marcotte, E. M. *et al.* Detecting protein function and protein–protein interactions from genome sequences. *Science* **285,** 751–753 (1999).
20. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23,** 324–328 (1998).
21. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96,** 4285–4288 (1999).
22. Huynen, M. A. & Bork, P. Measuring genome evolution. *Proc. Natl Acad. Sci. USA* **95,** 5849–5856 (1998).
23. Ge, H., Liu, Z., Church, G. M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genet.* **29,** 482–486 (2001).
24. Costanzo, M. C. *et al.* YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.* **29,** 75–79 (2001).
25. Piper, M. D., Hong, S. P., Ball, G. E. & Dawes, I. W. Regulation of the balance of one-carbon metabolism in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **275,** 30987–30995 (2000).
26. Sinclair, D. A. & Dawes, I. W. Genetics of the synthesis of serine from glycine and the utilization of glycine as sole nitrogen source by *Saccharomyces cerevisiae*. *Genetics* **140,** 1213–1222 (1995).
27. Douce, R., Bourguignon, J., Neuburger, M. & Rebeille, F. The glycine decarboxylase system: a fascinating complex. *Trends Plant Sci.* **6,** 167–176 (2001).
28. Hoffmann, R., Jung, S., Ehrmann, M. & Hofer, H. W. The *Saccharomyces cerevisiae* gene *PPH3* encodes a protein phosphatase with properties different from PPX, PP1 and PP2A. *Yeast* **10,** 567–578 (1994).
29. Kalhor, H. R., Luk, K., Ramos, A., Zobel-Thropp, P. & Clarke, S. Protein phosphatase methyltransferase 1 (Ppm1p) is the sole activity responsible for modification of the major forms of protein phosphatase 2A in yeast. *Arch. Biochem. Biophys.* **395,** 239–245 (2001).
30. Gerstein, M., Lan, N. & Jansen, R. Proteomics. Integrating interactomes. *Science* **295,** 284–287 (2002).
31. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19,** 1720–1730 (1999).
32. Holstege, F. C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95,** 717–728 (1998).
33. Mrowka, R., Patzak, A. & Herzel, H. Is there a bias in proteome research? *Genome Res.* **11,** 1971–1973 (2001).
34. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293,** 2101–2105 (2001).
35. Zhou, H., Roy, S., Schulman, H. & Natan, M. J. Solution and chip arrays in protein profiling. *Trends Biotechnol.* **19,** S34–S39 (2001).
36. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30,** 303–305 (2002).
37. Mellor, J. C., Yanai, I., Clodfelter, K. H., Mintseris, J. & DeLisi, C. Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.* **30,** 306–309 (2002).
38. Bader, G. D. *et al.* BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Res.* **29,** 242–245 (2001).
39. Paton, N. W. *et al.* Conceptual modelling of genomic information. *Bioinformatics* **16,** 548–557 (2000).
40. Kumar, A. *et al.* Subcellular localization of the yeast proteome. *Genes Dev.* **16,** 707–719 (2002).
41. Fields, S. & Song, O. A novel genetic system to detect protein–protein interactions. *Nature* **340,** 245–246 (1989).
42. Oliver, S. Guilt-by-association goes global. *Nature* **403,** 601–603 (2000).
43. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17,** 1030–1032 (1999).
44. Mann, M., Hendrickson, R. C. & Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70,** 437–473 (2001).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com).

Acknowledgements

We thank members of the Bork group, N. Paton, A. Brass, S. Cohen, A.-C. Gavin and B. Kuster for critical discussions, and Cellzome AG for access to their interaction data before publication. C.v.M. and P.B. are supported by Bundesministerium fur Bildung und Forschung, Germany, and work in Manchester is supported by the Biotechnology and Biological Sciences Research Council through its IGF (Investigating Gene Function) initiative. S.F. is supported by the National Center for Research Resources of the National Institutes of Health, and is an investigator of the Howard Hughes Medical Institute. S.F. and P.B. are supported by the Human Frontier Science Program.

Competing interests statement

The authors declare competing financial interests: details accompany the paper on *Nature*'s website (http://www.nature.com).

Correspondence and requests for materials should be addressed to P.B. (e-mail: bork@embl-heidelberg.de).