Protein–protein interaction maps: a lead towards cellular functions

Pierre Legrain, Jérôme Wojcik and Jean-Michel Gauthier

The availability of complete genome sequences now permits the development of tools for functional biology on a proteomic scale. Several experimental approaches or *in silico* algorithms aim at clustering proteins into networks with biological significance. Among those, the yeast two-hybrid system is the technology of choice to detect protein–protein interactions. Recently, optimized versions were applied at a genomic scale, leading to databases on the web. However, as with any other 'genetic' assay, yeast two-hybrid assays are prone to false positives and false negatives. Here we discuss these various technologies, their general limitations and the potential advances they make possible, especially when in combination with other functional genomics or bioinformatics analyses.

P. Legrain*
J. Wojcik
J-M. Gauthier
Hybrigenics, 180 Avenue Daumesnil, Paris 75012, France.
*e-mail plegrain@hybrigenics.fr

In the past five years, more than 30 bacterial and four eukaryotic genomes (*Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Arabidopsis thaliana*) have been fully sequenced and the two eukaryotic genomic sequences of *Homo sapiens* and *Mus musculus* are becoming available this year. Although a lively discussion is ongoing concerning the real number of genes in these genomes, it is clear that we have now access to the coding sequences for tens of thousands of proteins for which very little functional information is available. There is therefore an urgent need for high-throughput technologies that elucidate protein function.

The availability of fully sequenced genomes led to large-scale studies of protein–protein interactions to establish complete protein interaction maps ('interactome'). Protein–protein interaction mapping identifies a putative function to uncharacterized proteins and can provide information, such as interacting domains, to direct further experiments. Ultimately, combining data such as protein–protein interactions, transcription analyses and bioinformatic analysis of protein sequences should permit the assignment of functional annotations or even a biochemical function to as yet uncharacterized proteins.

The yeast two-hybrid system[1] (Y2H) can detect interactions between two known proteins or polypeptides and can also search for unknown partners (prey) of a given protein (bait) (Fig. 1; for review, see Ref. 2). Nevertheless, due to its intrinsic properties (i.e. measuring interactions between two chimeric and heterologous proteins in a yeast cell nucleus) a Y2H assay cannot apply to all protein–protein interactions, giving rise to a certain proportion of false-positive and false-negative results. During the past ten years, a few partial protein–protein interaction maps for viruses, bacteria and eukaryotes have been produced using two different strategies: the matrix and the whole-library approach (Fig. 2). The experiments differ considerably both in the type of result and timescale (see Table 1 for a summary of most published studies).

Protein–protein interaction maps built through protein arrays: the matrix approach

This approach (referred to here as the 'matrix approach') uses a collection of predefined open reading frames (ORFs), usually full-length proteins, as both bait and prey for interaction assays. The experimental approach is to amplify ORFs by PCR, to clone them into two-hybrid vectors (specific for bait or prey) and express the fusion proteins individually in yeast cells of opposite mating type. Yeast cells transformed with bait plasmids or prey plasmids are then collected, stored and assayed after mating. Combinations of bait and prey can be assessed individually (a one by one approach for bait and prey, or 'the protein array') or after pooling cells expressing different bait or prey proteins.

The intrinsic limitation of this strategy is that it tests only predefined proteins. It was first used to explore interactions among *Drosophila* proteins involved in the control of cell cycle[3]. Last year, large-scale approaches were published for the vaccinia virus[4] (266 predicted ORFs) and for the yeast proteome[5,6] (around 6000 ORFs). In the vaccinia virus study, all possible combinations (roughly 70 000) between encoded proteins were examined (Table 1). One of the yeast studies was a pilot study[5] that has recently been completed at the proteome scale[7]. The other study aimed to assess all potential combinations (36 million) between yeast ORFs (Ref. 6).

In the exhaustive yeast study performed by Uetz and colleagues[6], two experimental designs were used: a low-throughput protein-array approach and then a high-throughput approach using pooled prey clones. In the low-throughput array, 192 bait proteins were tested against the complete set of 6000 prey proteins, identifying a total of 281 interacting protein pairs. Eighty-seven of these 192 bait proteins identified interacting proteins reproducibly (i.e. in two independent identical

http://tig.trends.com 0168–9525/01/$ – see front matter PII: S0168-9525(01)02323-X

experiments). The second, high-throughput approach used the pool of 6000 prey clones mated with cells transformed with one given bait plasmid and selected for interactions. The baits were the complete set of 6000 yeast proteins. This identified 817 proteins involved in putative protein–protein interactions (as bait or prey), leading to 692 interacting protein pairs, of which 41% were found reproducibly in two identical independent experiments. Of the 87 bait proteins that identified interactions in the first low-throughput assay, 12 formed interactions with prey proteins in the second, high-throughput experiment. This indicates that the high-throughput strategy considerably increases the number of false negatives.

Another large-scale study using a matrix approach was conducted by Ito and colleagues[5,7]. In the pilot study[5], collections of yeast cells transformed with bait or prey plasmids were prepared, clones were pooled in groups of 96, and pools of bait and prey tested against each other. More than 4 million combinations were tested, obtaining 866 positive colonies. Sequence analysis of both bait and prey plasmids (from interaction sequence tags, ISTs) identified the interacting proteins. This matrix resulted in the identification of 175 pairs of interacting proteins, 12 of which were already known.

More recently, this group completed the study for the whole yeast proteome[7]. A total of 4549 interactions were detected involving 3278 proteins. Among those, several were identified by more than three hits (841). These were compared with the set of interactions identified previously by the other large-scale approach[6] (692 interactions). Unexpectedly, only 141 interactions were common to both, suggesting that the detection of an interaction depended on the specific selection scheme. Another explanation is that the definition of the threshold for significance of the identified interaction might be a key parameter in the definition of potential false positives and/or false negatives: the 841 interactions found in the former study[7] correspond to a subset identified by more than three hits, whereas the 692 interactions from the latter[6] correspond to the those identified only once in one experiment (220) plus those identified more than once in one out of two experiments (186) plus, ultimately, those identified in two independent experiments (286).

Reproducibility is a general problem in Y2H assays. To circumvent this problem in the vaccinia interaction map[4], interactions were systematically assayed in quadruplicate, and only those corresponding to at least three positive colonies were counted as positive. This meant 20 out of 56 interactions were discarded. In one of the yeast studies, using a similar protein-array assay[6], 20% of positives were found reproducibly in a duplicated assay. Nevertheless, discarding interactions that were not confirmed in a second identical assay could increase the rate of false negatives, especially when the design of the screen does not allow the testing of a complete set of possible combinations.

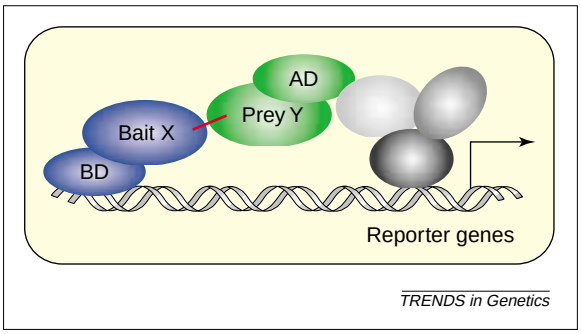

Fig. 1. The yeast two-hybrid (Y2H) assay. The Y2H system[1] detects the interaction between two proteins through an assay involving transcriptional activation of one or several reporter genes (for review, see Ref. 2). Polypeptide X is fused to a protein domain that binds specifically a DNA sequence in the promoter of the reporter gene (the DNA-binding domain; BD). Polypeptide Y is fused to a domain that recruits the transcription machinery (the activation domain; AD). Transcription of the reporter gene will occur only if X and Y interact together.

Protein–protein interaction maps built through screening of fragment libraries

Although it was originally designed to detect a physical association between two known proteins, the Y2H assay rapidly became the most widely used system to screen libraries for proteins interacting with a known protein (bait). Repeating such library screening experiments with a series of proteins involved in the same biochemical process led to the concept of specific functional protein-interaction maps that could identify other previously uncharacterized proteins involved in the same pathway. This experimental strategy was extended to a proteome-wide approach. It was first applied to determine protein networks for the T7 phage proteome, which contains 55 proteins[8].

Screening randomly generated protein fragments also permits the determination of interacting domains. Large libraries are required to take into account the fact that only a fraction of the genomic or cDNA fragments will encode genuine protein-interaction domains, due to the location of the fragment, its orientation or its reading frame. The common sequence shared by the selected overlapping prey fragments defines the smallest selected docking site of the bait (Fig. 2b), thus allowing the precise mapping of a functionally interacting domain[9].

In 1997, a pilot experiment[10] described a mating strategy to achieve full coverage of a prey fragment library with a complexity of over 5 million independent clones, tested with a dozen bait proteins known to be involved in RNA splicing. In this exhaustive library screening procedure, all selected positive prey fragments were identified by sequencing, allowing the prey proteins to be classified according to a 'heuristic' value (in this case, their

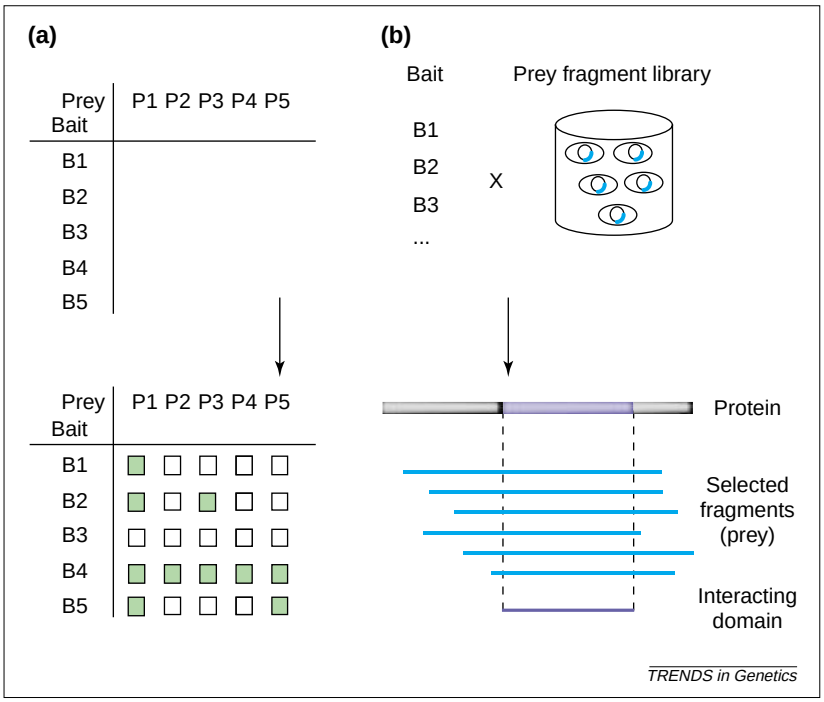

Fig. 2. The matrix and library screening approaches to build large-scale protein interaction maps. The matrix approach (a) uses the same collection of proteins (1–5) used as bait (B1–B5) and prey (P1–P5). The results can be drawn in a matrix. The bait auto-activators (for example, B4) and 'sticky' prey proteins (for example, P1 interacts with many proteins) are identified and discarded. The final result is summarized as a list of interactions that can be heterodimers (B2–P3) or homodimers (B5–P5). The library screening approach (b) identifies the domain of interaction for each prey protein interacting with a given bait. Sticky prey proteins are identified as fragments of proteins that are often selected regardless of the bait protein. An auto-activator bait can be used in the screening process with more stringent selective conditions.

potential biological significance). Briefly, this value was determined on the basis of the experimental results; that is, the number of independent overlapping fragments, the size of the fragments and the number of their occurrence in the set of prey fragments. The most convincing prey proteins (with the highest heuristic values, taking into account the reproducibility issue) were then used as bait proteins in iterative screens. In total, 170 interactions were found connecting 145 different yeast proteins, leading to the identification of new RNA splicing factors.

This strategy has now been applied to many proteins in yeast. On the basis of this approach and genetics studies, a model of the RNA polymerase III pre-initiation complex has been proposed[11]. Interaction domains were defined for many components of the complex, filling gaps between 3D structures of monomers and the functional definition of the active complex. More than 100 yeast proteins known to be involved in RNA metabolism have been screened for protein interactions, leading to a network of interactions involving several hundreds of proteins[12]. This network has incorporated links between RNA splicing factors and mRNA processing complexes, which have recently been corroborated by biological evidence.

These large-scale proteomics studies have now been applied to several other genomes. Another recent study dealt with hepatitis C virus (HCV) polypeptide interactions[13]. The HCV genome encodes a single polyprotein that is post-transcriptionally cleaved into ten polypeptides. A matrix approach using the ten mature polypeptides failed to detect any interaction between HCV polypeptides, not even for the well-known capsid oligomer or the heterodimer between the NS3A protease and its cofactor NS4A. This suggests again that predefined fusion proteins might not always be suitable for Y2H assays, probably because of

Table 1. Large-scale datasets for protein–protein interaction maps

Organism	Technology	Number of assays (bait × prey)	Detected interactions	Already known interactions	Refs
Vaccinia virus (~266 ORFs)	Protein array	Proteome × proteome	37	9	4
S. cerevisiae (~6000 ORFs)	Protein arrays	192 × proteome	281		6
	Pools of prey	Proteome × proteome	692	109[a]	
S. cerevisiae (~6000 ORFs)	Pools of baits and prey	430 assays of pools (96 × 96)	175	12	5
S. cerevisiae (~6000 ORFs)	Pools of baits and prey	3.844 assays of pools (96 × 96)	841[b]	105	7
C. elegans (~20 000 ORFs)	Protein array	29 × 29	8	6	15
	Library screening	27 × proteome	124	3	
HCV (10 ORFs)	Protein array	10 × proteome	0	2	13
	Library screening	22 fragments × proteome	5	2	
S. cerevisiae (~6000 ORFs)	Library screening	15 × proteome	170	3	10
S. cerevisiae (~6000 ORFs)	Library screening	11 × proteome	113	34	12
H. pylori (~1600 ORFs)	Library screening	261 × proteome	1524	0[c]	14

[a]Total number for both studies.
[b]This number corresponds to highly significant interactions (more than three hits).
[c]No complexes or interactions were formally reported in *H. pylori*, although many interactions were reported in other bacterial organisms, especially *E. coli*, for homologous proteins.

incorrect folding of the chimeric proteins. However, screening for the interactions of randomly generated HCV genomic fragments revealed the expected capsid homodimer and viral protease heterodimer, as well as novel interactions.

Finally, a similar exhaustive proteome-wide approach for building the protein interaction map in *Helicobacter pylori* was completed recently[14]. The map links half of the proteins of the proteome in a comprehensive network of protein–protein interactions. This study identifies complexes that have been demonstrated or postulated to exist in other organisms. For example, *H. pylori* proteins were identified that are homologous to *E. coli* proteins that form functional heterodimers and homodimers (verified by other experimental methods). When these *H. pylori* proteins were used as bait in a Y2H assay, they formed ~65% of the heterodimers and ~50% of the homodimers expected from the homologous *E. coli* proteins. In some cases, the interacting domains thus identified were mapped on 3D structures of proteins and assigned to a functional domain. These interacting domains also constitute a first step towards the construction of dominant–negative mutants or the development of an assay for interaction modulation.

Another protein–protein interaction map, involved in vulval development in *C. elegans*, has been published[15]. This study used a set of 29 proteins implicated in this developmental pathway combined with a protein-array assay and a library screening to identify other proteins potentially involved in this pathway (Table 1). Indeed, the library screening identified many novel potential protein–protein interactions.

The major limitation of the library approach is the preparation of highly complex prey fragment libraries and the cost of interaction screens. In most cases, a specialized technological platform (including robots for lab work, specialized computer software and algorithms) is required to cover the library exhaustively, to identify the prey fragments by high-throughput sequencing and to represent the interaction data appropriately.

Analysis of protein–protein interaction maps: false negatives and false positives

An intrinsic limitation of the conventional Y2H system is that it relies upon the transcriptional activation of reporter genes. Incorrect folding, inappropriate subcellular localization (bait and prey proteins must interact in a nuclear environment) or degradation of chimeric proteins and absence of certain types of post-translational modifications in yeast could lead to false negatives.

Other properties of the assay might lead to the selection of false positives. For example, bait proteins might activate the transcription of reporter genes above the threshold level by themselves (auto-activation), and some prey proteins or fragments might be selected in a Y2H assay in combination with a wide variety of bait proteins (sticky prey). These are key issues that should be addressed when selecting an experimental strategy for building large-scale protein–protein interaction maps.

False negatives

Collections of bait and prey constructs are generally prepared as batches by two-step PCR amplification, multiplying the risk of frameshift and mis-sense mutations. Each construct is not usually controlled individually. For example, in one of the yeast proteome studies[6], a careful analysis indicated that only 87% of ORFs were correctly cloned into bait and prey vectors, excluding many proteins from the study. It should be emphasized that in a matrix approach only two assays are performed for each pair of bait (B) and prey (P) proteins (i and j) encoded in the genome (i.e. $B_i \times P_j$ and $B_j \times P_i$; Fig. 2a), whereas in the library screening strategy, tens of fragments are screened for each prey protein (each nucleotide in a genome is represented in about 50 different fragments) and most selected interacting domains are defined by more than one fragment[14]. This explains the difference in numbers of potential interactions that are detected by the two approaches (Table 1). Note the most recent yeast study[7] where only a limited number of interactions were detected using the two symmetrical combinations of bait and prey proteins.

Studies that combine both matrix and library screening approaches[13,15] confirm that the library strategy yields many more potential interactions. In addition, when the matrix approach is used with prey-protein arrays[6] (low-throughput assay), ten times more interactions are detected than when using pools of prey[5,6] (high-throughput assay), suggesting that the latter approach is not suitable for building complete protein-interaction maps. This is probably due to the necessity of defining a common set of selective conditions for all bait and prey combinations (same selective medium, same reporter genes) that does not take into account the intrinsic capability of each bait protein to auto-activate reporter genes to some extent. Due to this high number of false negatives, the two exhaustive studies of the yeast proteome failed to identify as many as 90% of interactions previously described in the literature[7].

False positives

Large-scale analyses by Y2H assays might also generate false positives. For example, searching for many potential interactions, especially when screening a random fragment library, increases the chance of a selection of interacting polypeptides that are not significant biologically. Thus, it becomes necessary to score every single interacting pair for its reliability with respect to the technology. The rate of false positive interactions is difficult to

evaluate and is largely dependent on the criteria applied for the significance of the interactions, such as the reproducibility of results. In one yeast pilot study[5], baits that auto-activated reporter gene transcription were removed from pools. A very strong selective pressure was applied, and prey proteins that were selected with more than three unrelated bait proteins out of a pool of 100 bait proteins were discarded, leading to well-established, although partial results.

The same approach was followed for the full-scale study[7]. Only interacting pairs of proteins that were found more than three times were included in the core data (Table 1). In the screen of a random fragment library, the selection procedures were adapted to every single bait protein, permitting a strong selectivity for all bait proteins. In addition, screening of the complete fragment library permitted labelling of interactions through a global scoring scheme using a statistical model (for details see Ref. 14). In this case, prey fragments that were often selected with unrelated bait proteins (i.e. probably nonspecific partners) were specifically labelled and discarded for further functional analyses. Nevertheless, some fragments of a protein might be termed 'sticky' (interacting with many proteins), although other domains of the same protein are specific interactors.

To evaluate false positives and reproducibility, access to primary data is necessary. A simple list of interactions such as those present in most public databases do not take into account the reproducibility of results. For example, data accessible through the web corresponding to one study of the complete yeast proteome[6] do not indicate which interactions were detected in the protein array versus the pool strategies or, more importantly, which ones were reproducible. The second full-scale study of the yeast proteome[7] is linked to a website that presents these primary data as tables, permitting in depth exploration and comparison. To give access to primary data generated in a library screening experiment (such as the number of selected prey fragments and their position in the ORFs) a new type of database was created[14], which also provides a graphical display of protein–protein interaction maps. Queries made on proteins and interactions are filtered according to their reliability score. Thus, bioinformatics tools might also contribute to identifying false positives.

Protein interaction maps can now be built at the scale of a proteome. Extrapolations from experiments made on the yeast proteome suggest the total number of yeast protein–protein interactions is between 7000, the number of all known interactions including the novel interactions identified in the two full-size genomic Y2H analyses[6,7] and 70 000, assuming that these studies have missed over 90% of already known interactions[7]. A fair estimate is probably in the range of 15 000 to 20 000 significant interactions. Any reliable proteome-wide strategy should aim at the detection of the majority of protein–protein interactions, while keeping false positives as low as possible (hopefully below the number of biologically significant interactions). However, the usefulness of such a dataset is limited, without specific experiments for biological validation or without cross-referencing the proteomics data with independent data obtained from unrelated technologies. Biological validation and/or integration of data from other sources will help predict the biological relevance of these interactions; for example, taking into account that Y2H assays could detect interactions between proteins that are never co-localized in the cell.

Additional genomic approaches: large-scale functional studies

Recently, other technologies were developed to tackle the functions of genes and proteins at the level of the genome and the proteome.

Global gene expression

Because analytical methods using RNA are well adapted to large-scale investigation, studies examining global gene expression ('transcriptomics') are the most popular of the new functional genomics[16,17]. Based on cDNA or oligonucleotide arrays, and other systems such as serial analysis of gene expression (SAGE), these technologies simultaneously monitor the rates of mRNA expression of large sets of genes. However, although they give valuable information in terms of biological meaning, they measure changes in abundance of mRNA and not necessarily the final and functional products of the genes – the proteins.

The successes of global gene-expression experiments in the past three years include the discovery of gene-expression markers associated with transcriptional alterations in cancers, the effect of overexpression or knockout of regulatory genes (transcription factors, kinases, etc.) and global transcriptional changes during biological processes (mitosis, induction by a growth factor, etc.). The bioinformatics analysis of large-scale expression data frequently clusters genes sharing a similar transcriptional profile. These genes are supposedly co-regulated or involved in the same biological process. Thus, it is possible to infer a biological function for an unannotated gene by matching its expression patterns to annotated genes with the same profile.

Genome-wide mutagenesis

Another approach in functional genomics is to alter systematically many or all genes in a genome, one by one, and to observe the resulting phenotype. Several groups carried out genome-wide mutagenesis programmes on cellular and animal models. These include systematic gene disruption in *S. cerevisiae*

Box 1. Useful databases

Protein–protein interaction and functional clustering databases can be found at the following Internet addresses.

Yeast
http://depts.washington.edu/sfields/yplm/data/index.htm
http://portal.curagen.com
http://www.mips.biochem.mpg.de/proj/yeast/tables/interaction/physical_interact.html
http://www.pnas.org/cgi/content/full/97/3/1143/DC1
http://www.proteome.com/databases/YPD
http://dip.doe-mbi.ucla.edu/
http://genome.c.kanazawa-u.ac.jp/Y2H

C. elegans
http://cancerbiology.dfci.harvard.edu/cancerbiology/ResLabs/Vidal/

H. pylori
http://pim.hybrigenics.com

Drosophila
http://gifts.univ-mrs.fr/FlyNets/FlyNets_home_page.html

by deletion and transposon-tagging[18,19], chromosome-wide RNA-mediated interference (RNAi) in *C. elegans*[20,21] (see Ref. 22 for review) and large-scale generation of mutant mice by ENU mutagenesis[23–25]. Coupled with a systematic phenotyping, these projects promise to clarify gene and protein functions by involving the disrupted genes in biological processes. The phenotypes screened in these studies depend on the cellular model. For instance, the growth abilities of the *S. cerevisiae* mutants were tested on different growth conditions. On *C. elegans*, developmental phenotypes and appearance of various cells were carefully observed. For the mouse projects, an impressive effort is made to check for several phenotypes including developmental, behaviour, immunological or allergic abnormalities and some metabolites or proteins contained in the blood.

However, although they are informative, the phenotypes screened in these studies are still rather descriptive and do not provide precise biochemical functions or mechanisms of action. Nevertheless, phenotypic clustering allows functional predictions. Through protein–protein interaction maps, global gene-expression experiments and genome-wide phenotype-driven mutagenesis approaches, experimental genomics is now building the tools to decipher function on an unprecedented scale. Linkage of these heterogeneous data and others (structure or motif predictions, human genetics, etc.) provides a powerful means of inferring new cellular and molecular functions to unannotated or already-studied proteins. This is a challenge for bioinformatics and will require appropriate biological databases (see Box 1).

Exploring databases and predicting functions: bioinformatics tools

Protein interaction maps are a new and potentially rich source to assign function to uncharacterized gene products by bioinformatic techniques. The first attempts use 'guilt-by-association' methods to annotate proteins on the basis of the annotations of their interacting partners or, more generally, of the proteins sharing a common property in a given cluster. These emergent bioinformatics algorithms are promising, but should be used with caution because of low quality and incomplete data. Moreover, such techniques suffer from a lack of independent validation methods.

For example, all yeast protein interactions described in the literature or revealed by large-scale Y2H screens were recently analyzed through a clustering method[26] based on cellular role and subcellular localization annotations from the Yeast Proteome Database (YPD)[27] (Box 1). The function of an uncharacterized protein is assigned on the basis of the known functions of its interacting partners. A function was assigned to 29 proteins that have two or more interacting proteins with at least one common function. This prediction is highly dependent on the YPD functional annotations, which are often reductive and sometimes false, and on the protein interaction map, which is far from complete (with possibly 90% of interactions being missed). Thus, poorly defined annotations can gather different concepts and induce clustering that is not significant biologically. Furthermore, 'functional clustering' methods are also very sensitive to false positives. Indeed, Y2H false positives represent highly connected nodes in the network of proteins. Such nodes can greatly disturb the general shape and characteristics of a network[28]. Nevertheless, the most recent study on yeast interactions[7] also compare their experimental large-scale Y2H data with interaction data extracted from YPD on the basis of literature (with the exclusion of Y2H experiments). Both datasets exhibit one large cluster of proteins (half of the proteins and two-thirds of the interactions), suggesting a possible intrinsic biological property of this huge network.

To improve the prediction quality, a combination of independent data were also used[29]. In this study, three bioinformatic prediction methods (analysis of related metabolic function, analysis of related phylogenetic profiles, and the Rosetta stone method) were used with two sources of experimental data [interactions from the Munich Information Center for Protein Sequences (MIPS; Ref. 30) and Database of Interacting Proteins (DIP; Ref. 31) yeast protein-interaction databases and data from mRNA expression] to build a compound protein network. Function assignments then followed a similar guilt-by-association rule using 'high-confidence' links obtained from one experimental source or two different prediction methods. The 29 function

assignments made in the former study[26] were used comparison in the corresponding high-confidence links of the latter[29], although they were themselves partly predicted from interactions listed in the MIPS database used in the former study. This emphasizes the fact that predictions must be used with caution: the oversight of the initial hypothesis and the deficiency in independent data sources could lead to biased conclusions. One major hurdle in bioinformatics prediction algorithms is clearly the lack of independently validated methods.

Bioinformatics clustering of protein interactions still represents a powerful annotation tool that will become more and more useful as the interaction data accumulate. However, to be used successfully for appropriate functional annotation, the data need to be stored in elaborate structures that allow each individual scientist to test his/her own hypothesis against complex heterogeneous primary data and then to design further experimental setting to validate the functional assignment[14,31,32].

Concluding remarks

Large-scale protein interaction maps are, with gene-expression profiles, among the first examples of datasets generated without specific knowledge about the functions of genes. These are technology-driven experiments rather than hypothesis-driven experiments. They are valuable tools for protein function prediction, despite the occurrence of typical artefacts. These approaches are still in their early stages. Related bioinformatics tools are also primitive and require much more independent experimental validation before becoming useful predictive tools. Finally, functional annotations based on predictions cannot replace primary experimental information, which will be also accessible through functional databases on the web. Ultimately, functional annotation will certainly move towards a more precise description of the characteristics of every single biological entity, helping users of databases to build new hypotheses that still will have to be experimentally proven.

References

1 Fields, S. and Song, O. (1989) A novel genetic system to detect protein–protein interactions. *Nature* 340, 245–246
2 Vidal, M. and Legrain, P. (1999) Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Res.* 27, 919–929
3 Finley, R.L., Jr and Brent, R. (1994) Interaction mating reveals binary and ternary connections between *Drosophila* cell cycle regulators. *Proc. Natl. Acad. Sci. U. S. A.* 91, 12980–12984
4 McCraith, S. *et al.* (2000) Genome-wide analysis of vaccinia virus protein–protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* 97, 4879–4884
5 Ito, T. *et al.* (2000) 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. U. S. A.* 97, 1143–1147
6 Uetz, P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627
7 Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4569–4574
8 Bartel, P.L. *et al.* (1996) A protein linkage map of *Escherichia coli* bacteriophage T7. *Nat. Genet.* 12, 72–77
9 Siomi, M.C. *et al.* (1998) Functional conservation of the transportin nuclear import pathway in divergent organisms. *Mol. Cell. Biol.* 18, 4141–4148
10 Fromont-Racine, M. *et al.* (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.* 16, 277–282
11 Flores, A. *et al.* (1999) A protein–protein interaction map of yeast RNA polymerase III. *Proc. Natl. Acad. Sci. U. S. A.* 96, 7815–7820
12 Fromont-Racine, M. *et al.* (2000) Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast* 17, 95–110
13 Flajolet, M. *et al.* (2000) A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene* 242, 369–379
14 Rain, J.C. *et al.* (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature* 409, 211–216
15 Walhout, A.J. *et al.* (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 287, 116–122
16 Cho, R.J. and Campbell, M.J. (2000) Transcription, genomes, function. *Trends Genet.* 16, 409–415
17 Lockhart, D.J. and Winzeler, E.A. (2000) Genomics, gene expression and DNA arrays. *Nature* 405, 827–836
18 Ross-Macdonald, P. *et al.* (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* 402, 413–418
19 Winzeler, E.A. *et al.* (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285, 901–906
20 Fraser, A.G. *et al.* (2000) Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* 408, 325–330
21 Gonczy, P. *et al.* (2000) Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature* 408, 331–336
22 Hammond, S.M. *et al.* (2001) Post-transcriptional gene silencing by double-stranded RNA. *Nat. Rev. Genet.* 2, 110–119
23 Justice, M.J. *et al.* (1999) Mouse ENU mutagenesis. *Hum. Mol. Genet.* 8, 1955–1963
24 Hrabe de Angelis, M.H. *et al.* (2000) Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nat. Genet.* 25, 444–447
25 Nolan, P.M. *et al.* (2000) A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nat. Genet.* 25, 440–443
26 Schwikowski, B. *et al.* (2000) A network of protein–protein interactions in yeast. *Nat. Biotechnol.* 18, 1257–1261
27 Costanzo, M.C. *et al.* (2000) The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Res.* 28, 73–76
28 Albert, R. *et al.* (2000) Error and attack tolerance of complex networks. *Nature* 406, 378–382
29 Marcotte, E.M. *et al.* (1999) A combined algorithm for genome-wide prediction of protein function. *Nature* 402, 83–86
30 Mewes, H.W. *et al.* (2000) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 28, 37–40
31 Xenarios, I. *et al.* (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.* 28, 289–291
32 Sanchez, C. *et al.* (1999) Grasping at molecular interactions and genetic networks in *Drosophila melanogaster* using FlyNets, an Internet database. *Nucleic Acids Res.* 27, 89–94