the FEBS Journal

MINIREVIEW

High-throughput two-hybrid analysis

The promise and the peril

Stanley Fields

Howard Hughes Medical Institute, Departments of Genome Sciences and Medicine, University of Washington, Seattle, WA, USA

Keywords
computational methods; *Plasmodium falciparum*; protein–protein interaction; proteomics; yeast

Correspondence
S. Fields, Howard Hughes Medical Institute, Departments of Genome Sciences and Medicine, University of Washington, Box 357730, Seattle, WA 98195, USA
Fax: +1 206 5430754
Tel: +1 206 616 4522
E-mail: fields@u.washington.edu

(Received 27 May 2005, revised 9 August 2005, accepted 12 August 2005)

doi:10.1111/j.1742-4658.2005.04973.x

The two-hybrid method detects the interaction of two proteins by their ability to reconstitute the activity of a split transcription factor, thus allowing the use of a simple growth selection in yeast to identify new interactions. Since its introduction about 15 years ago, the assay largely has been applied to single proteins, successfully uncovering thousands of novel protein partners. In the last few years, however, two-hybrid experiments have been scaled up to focus on the entire complement of proteins found in an organism. Although a single such effort can itself result in thousands of interactions, the validity of these high-throughput approaches has been questioned as a result of the prevalence of numerous false positives in these large data sets. Such artifacts may not be an obstacle to continued scale-up of the method, because the classification of true and false positives has proven to be a computational challenge that can be met by a growing number of creative strategies. Two examples are provided of this combination of high-throughput experimentation and computational analysis, focused on the interaction of *Plasmodium falciparum* proteins and of *Saccharomyces cerevisiae* membrane proteins.

The yeast two-hybrid (YTH) assay is an example of a technology developed for the biological sciences in the last few decades that has followed a progression of four stages leading to genomic scale use. The first stage is the initial description of a method: the prototype version. Typically, the methodology is demonstrated by a single example that is performed under defined and optimal conditions. For a large fraction of new technologies, few other examples beyond this prototype are ever described. However, some methods prove of value in solving problems that confront biologists and enter a second stage: widespread application. During this period, quality improvements come into play, as experimentalists throughout the community add their own adaptations. Often at this stage, the use of a methodology is further spread by commercialization of reagents or equipment. Some fraction of these broadly applied technologies then prove sufficiently robust to be scaled up in scope, leading to a third stage: high-throughput (HT) usage. This stage is made possible by advances in some combination of automation, miniaturization, reduction in reagent costs and further refinements of the approach. As throughput escalates, so does the amount of data generated, sometimes to a staggering degree. Thus, the maturing of this scale-up phase elicits a fourth stage: a computational phase. Here, novel algorithms, which could not have been imagined given previous technologies, are developed to deal with these huge data sets. The HT and computational stages continue hand-in-hand, constituting many of the approaches of what has been termed functional genomics or systems biology.

The quintessential example of this progression is DNA sequence analysis. The major prototype

Abbreviations
HT, high-throughput; SVM, support vector machine; YTH, yeast two-hybrid.

sequencing method was the introduction of dideoxynucleotide chain terminators in a synthesis reaction with DNA polymerase I [1]. Although this version of the dideoxy procedure led to widespread use and the accumulation of many more DNA sequences that had been accomplished heretofore, it was the conversion of the method to a fluorescence-based and machine-readable format, combined with the assembly line style of the modern genome center, that made possible the deciphering of the tens of billions of sequenced bases now available [2]. As the sequence data accumulated, ever more sophisticated computational approaches were devised to examine coding capacity, repeats, duplications, mutations, recombinations, sequences of related genomes, and many other properties. Although sequence data continue to flow in at a prodigious rate, much of the key literature that relates to genome sequences now consists of novel computational analyses. Progressions similar to that for DNA sequence analysis can be outlined for transcriptional profiling via DNA arrays, the identification of regulatory regions in DNA, and the detection of human DNA sequence polymorphisms.

In the proteomics arena, the appreciation that proteins exert virtually all of their activities via interactions with other molecules – be they other proteins, nucleic acids, lipids, carbohydrates or small molecules – has driven the development of technologies to examine these macromolecular associations. The most realized of these methods detect protein–protein interactions, with two approaches proving to be most widespread: the YTH assay and the biochemical purification of tagged proteins followed by identification of associated proteins via mass spectrometry. Here, I focus on the yeast assay, but note that many of the conclusions apply equally well to the biochemical approach [3,4].

YTH progression

The original description of the YTH assay [5] introduced the idea of splitting into two domains a site-specific transcription factor, whose activity could then be reconstituted via the interaction of heterologous proteins fused to these two domains. Although the test case for this assay was only a single example of yeast proteins previously known to interact, the results led to the suggestion that the approach might be applicable to the identification of new interactions via a search of a library of activation domain-tagged proteins. Such a search procedure was shown to be feasible [6], and the assay was soon adopted by numerous laboratories and converted to the 'kit'-based format that is popular with molecular biologists. The yeast system had the advantages of speed, sensitivity and simplicity in addressing an important biological question at a time when other methods were far more laborious, and when the identification of an interacting protein following its purification was difficult. The two-hybrid assay also proved to have utility with proteins from essentially any organism and involved in any biological process, although certain types of proteins, such as membrane or extracellular proteins, were less amenable to this approach. The two-hybrid concept also proved remarkably malleable, with adaptations appearing that detected protein–DNA, protein–RNA, or protein–small molecule interactions, as well as protein–protein interactions that are dependent on post-translational modifications, that occur in compartments of the cell other than the nucleus, or that yield signals other than transcription of a reporter gene [7].

The typical two-hybrid experiment, during most of the 15 years that the method has been around, focused on a single protein or, at most, a few proteins implicated in the same process. An experimenter carrying out the method might find that for the protein fused to the DNA-binding domain (often termed the 'bait') used as the target in the search, the assay yielded a handful of candidate interactors (often termed the 'prey') as activation domain-fused proteins. These candidates were generally analyzed individually to evaluate their authenticity, by using experimental methods that might include co-immunoprecipitation, *in vitro* biochemical binding, protein localization, or transfection of the genes encoding the candidate proteins in cell lines. From these follow-up experiments, a common outcome was that all but one of the candidates were eliminated from consideration as relevant partners, and an eventual publication highlighted only this survivor. The other candidates, which proved not to be *bona fide* interactors, were considered false positives and were never reported.

It is important to distinguish two types of proteins, not of biological relevance, that can be recovered from a two-hybrid screen. The first type consists of those that do indeed bind to the bait protein in the context of the YTH assay, but not in the normal *in vivo* context. Such proteins might be, for example, members of a family with the requisite recognition property but not the specific member that recognizes the bait protein in its normal cellular milieu. The second type represents artifacts of the YTH assay itself, wherein transcriptional activity occurs independently of any protein–protein interaction. Such artifacts include proteins that when fused to a DNA-binding or activation

domain can activate transcription on their own, plasmid rearrangements or copy number changes that generate such auto-activators, or alterations at a reporter gene that result in constitutive expression. These false positives, while not reflecting binding in the context of the yeast assay, may still be highly reproducible. Our own experience is that the great bulk of false positives which arise in two-hybrid searches and that are generally eliminated when they occur in small-scale experiments, fall into this second class.

Beginning in the mid-1990s, efforts were initiated to apply the two-hybrid assay on a HT basis, first with bacteriophage T7 [8], then with yeast [9–11], and more recently with *Drosophila melanogaster* [12,13] and *Caenorhabditis elegans* [14]. Several developments, such as the availability of genome sequences, reduced primer and sequencing costs, array and pooling strategies, and the increasing use of robotics, made such scale-ups possible. As anticipated, the number of protein interactions present in biological databases [15] increased enormously, with a curve not unlike that of DNA sequence accumulation (Fig. 1). Despite this considerable ramp-up in the number of interactions detected, it is likely that only a small fraction of the total number that occur in a cell has been uncovered. Parallel efforts in either yeast [10,11] or *D. melanogaster* [12,13] show little overlap in their data sets, and approaches based on using small fragments in the assay yield different interactions than those based on full-length proteins.

In addition, the problem of false positives did not disappear with the accelerating scale of two-hybrid experiments. Instead, what disappeared was the ability

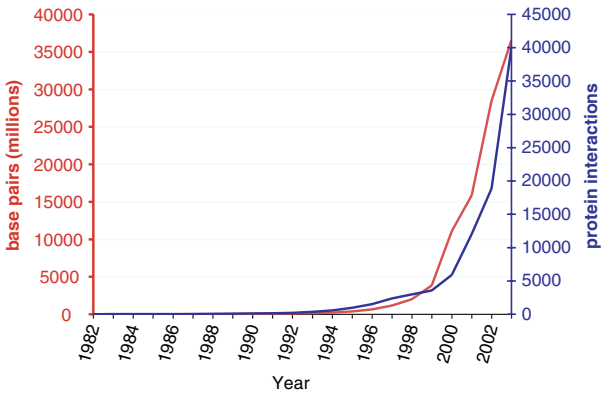

Fig. 1. A comparison of DNA sequence and protein interaction data. Cumulative base pairs of DNA sequence available in GenBank (http://www.ncbi.nlm.nih.gov/) are shown in red, and cumulative protein interactions available in the DIP database (http://dip.doe-mbi.ucla.edu/) are shown in blue. Interaction data are from L. Salwinski and D. Eisenberg, University of California, CA, USA. (personal communication).

to validate each candidate interaction by another experimental approach, given that the number of interactions which would need to be tested overwhelmed the capacity of other methodologies to do so. Thus, data sets were published that included many pairs of proteins that seemed biologically implausible. Furthermore, when computational approaches (see below) were applied both to the HT data sets and to the combined data set generated by many small-scale experiments (the 'community' data set), clear differences were found in the properties of these two data sets [16]. The conclusion from some of these studies was that 50% or more of the HT data were likely to be false positives.

If such a high prevalence of spurious data was indeed the case, it is fair to ask: are these large-scale efforts worth undertaking? I would argue that they are unquestionably of value, for at least four reasons. First, the goal of a complete description of an organism's protein network warrants whatever attempts, however early in their maturation, are needed to accomplish it. As projects to decipher such a network for human cells begin ramping up, they may eventually result in data as significant as the human genome sequence. Second, when researchers tackle ambitious goals, they reveal the limitations of the technology, thus enabling subsequent improvements. For example, YTH analysis can be accompanied by increasingly large scale co-immunoprecipitation approaches, such as that attempted for 143 pairs of *C. elegans* interactions [14], and other robust experimental means of validating interactions can be envisioned. For example, in the approach of Tong *et al.* [17], two-hybrid positives were compared with protein interactions derived from phage display experiments, with the intersection of these noisy (but independent) data sets yielding pairs of higher confidence. Third, protein interaction data can be highly useful to biologists, simply as lists of candidates; in this way, these data are similar to the noisy results of other approaches, such as large-scale chromatin immunoprecipitation [18,19] or synthetic lethality studies [20]. Those with sophisticated knowledge of a particular protein may have observed one or another two-hybrid candidate also arise in a complementary approach; they may home in on a candidate based on other large-scale data, such as expression profiles; or they may be able to test a defined set of candidates in another experimental assay. Fourth, computational biologists have applied an ever-burgeoning set of approaches to examine protein interaction data, often with the aim of discriminating biologically likely examples from false positives.

Computational assessments and insights

As protein interaction data accumulated via HT approaches, an increasing number of papers appeared from computational biologists in which the quality of interaction maps was analyzed. Just as the experimental approaches to generate these data focused on yeast, the computational ones also centered on this organism. The basis for most of these computational strategies is a test of the correlation between the interaction data and other properties known about the proteins, protein networks, or the corresponding genes.

One major contribution from the computational analyses was the finding that interactions that are evolutionarily conserved have a higher probability of being biologically relevant than those detected in only a single organism [21–23]. Indeed, some interactions have been experimentally observed in several different organisms. In a similar manner, computational analyses demonstrated that if two proteins implicated in an interaction have paralogues that also interact, this interaction is of increased likeliness [24].

Several studies demonstrated that *Saccharomyces cerevisiae* genes whose encoded proteins interacted are more likely than random gene pairs to be transcriptionally CO–regulated across different biological conditions [24–29]. Such a correlation allows a set of interaction data to be parsed into those of higher or lower confidence. Similar analyses have been performed for the *C. elegans* [14] and *D. melanogaster* [12] data sets. However, it must be noted that many interacting protein pairs, such as cyclins and cyclin-dependent kinases, are encoded by genes with very different transcriptional timing.

Another type of analysis has examined the connectivity of protein networks: if protein A interacts with proteins B and C, then the finding that B and C also interact forms a closed loop of three proteins and serves as a measure of interaction reliability [30–34]. Such interconnected clusters, which can be of varying size, are a feature of many biological complexes and pathways. Taken another step in analysis, a group of proteins may form a conserved module [35–37], which is reflective of these proteins performing a discrete biological activity. Such modules are often evolutionarily conserved, at least in part, among many species.

Another computational approach evaluates the functional assignments of interacting proteins. Given that a set of interacting proteins is likely to work in the same biological process, common functional annotations for such proteins support their relevance [38,39]. Other comparisons can be made between interaction data and the available set of protein structures [40,41] or protein domains [42]. Although only a very small fraction of the interaction data corresponds to protein complexes with solved structures, these examples provide a particularly good set for use for validating HT approaches. Finally, several groups have taken a combined approach, whereby interaction data are assessed according to the amount and type of supporting data [43,44]. Going beyond the experimentally derived data, computational biologists have developed novel algorithms that functionally link proteins, based on features other than sequence homology or experimental data [45], to predict protein networks. These algorithms use properties such as the conservation or loss of protein pairs during the evolution of species, the presence of a protein with two domains matching up to two separate proteins that interact, and the order of genes encoding interacting proteins.

A striking insight to emerge from analyzing the overall protein networks that result from large-scale approaches is their scale-free degree distribution, in which the number of links per protein is highly non-uniform, ranging from a few hubs with many connections to the great majority of hubs with only a few connections [37]. Another feature is their small-world property, in which any two proteins can be connected by a path with only a few links [37]. Such characteristics are also seen in other networks, such as the World Wide Web and social networks. In the case of protein networks, the evolution of this topology can be explained by the preferential attachment of new nodes to ones that already have many links, in a process related to gene duplication [37]. Highly connected proteins are more likely to interact with a protein that is duplicated than with one that has few links. Thus, these highly connected proteins gain even more links. Another intriguing aspect of the structure of protein networks is their robustness, namely the ability to respond to changes. This robustness is a consequence of the protein network topology, with random loss of proteins mostly affecting the many proteins with only a few partners rather than the small number of hubs [37]. In yeast, deletion of genes encoding highly connected proteins is three times more likely to result in a lethal phenotype than deletion of other genes [46].

Plasmodium falciparum protein interactions

In the last few years, my laboratory has taken on a few projects that required HT two-hybrid methods and that yielded data sets whose analysis required a

significant investment in computational approaches. In one, we have sought to identify protein–protein interactions on a large scale for the malaria parasite *Plasmodium falciparum* as part of a collaboration with Prolexys Pharmaceuticals, Inc. (Salt Lake City, UT, USA). One major issue with this organism is that approximately two-thirds of the genes do not bear sufficient similarity to genes of other organims to allow functional predictions to be made [47]. A second issue with the genes of *P. falciparum* is their exceptionally high A + T content (of nearly 80%), which makes protein expression in heterologous organisms, such as *Escherichia coli* or yeast, problematic. We observed in pilot two-hybrid studies that *P. falciparum* proteins generally either did not express or appeared as much smaller protein fragments than expected. To circumvent these problems, we used a strategy developed by Prolexys Pharmaceuticals in which hybrid proteins are generated that consist of the Gal4 DNA-binding or activation domain, a fragment encoded by a small segment of *P. falciparum* DNA, and a metabolic enzyme whose activity can be directly selected for in the YTH reporter strain (Fig. 2). The result of this configuration is that only recombinant plasmids bearing the *P. falciparum* insert as an in-frame fusion, and that lead to successful transcription and translation to produce the hybrid protein, are included in a two-hybrid library.

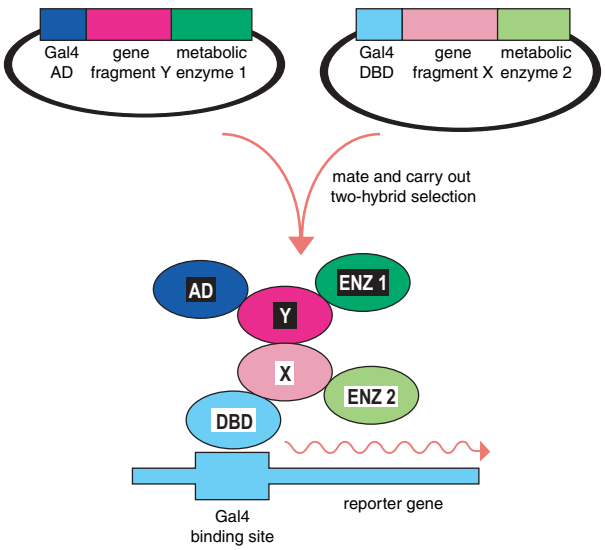

Fig. 2. A two-hybrid selection using libraries that contain only plasmids whose protein fusions are expressed in yeast. Both the Gal4 DNA-binding domain (DBD) and activation domain (AD) plasmids encode metabolic enzymes (ENZ) at the C terminus of the protein fusions whose activities complement mutations in the yeast reporter strain. Only when the gene fragments X and Y allow expression of the metabolic enzymes are the recombinant plasmids included in the DBD or AD library.

Using this strategy, we carried out > 30 000 two-hybrid searches by mating random yeast transformants expressing a DNA-binding domain fusion to an activation domain library [48]. Diploids were plated on media selecting both for expression of the two metabolic enzymes at the C termini of the fusions and for the activity of the two-hybrid reporter genes. Only at the stage when transformants grew on these selective plates were plasmids recovered from the yeast, inserts sequenced, and the identity of the two *P. falciparum* fragments determined. Nearly 14 000 pairs of plasmid inserts were sequenced, leading to an initial description of > 5000 unique protein–protein interactions.

From preliminary analyses of this data set, it was clear that some proteins had many more partners than did others, and that these partners, in the cases in which they were annotated, tended to encompass a huge diversity of biological processes. Thus, these 'promiscuous' proteins appeared to behave as typical false positives in the YTH assay. We used a computational method to eliminate the interactions involving these proteins, thereby deleting nearly half of the initial data set and resulting in a core set of ≈2800 interactions.

A large fraction of the interactions in this core set are between two proteins of uncharacterized function, or between a protein of uncharacterized function and a protein with an annotation, although some of the annotations are of limited usefulness. We carried out several computational approaches to attempt to find some of the interactions most likely to be biologically relevant. By searching for regions of the overall network that are statistically more connected than others, we identified a subnetwork around the *P. falciparum* Gcn5 protein, the homologue of a yeast histone acetylase. The subnetwork also contained a number of other proteins implicated in chromatin metabolism, DNA replication, transcription and ubiquitin metabolism. By comparing transcriptional profiles of the genes encoding interacting pairs, we found another subnetwork of interactions encompassing proteins implicated in host cell invasion. These included several known merozoite surface proteins and others probably expressed on the surface of the parasite. By analyzing interactions for those enriched in specific protein domains, we identified probable interactions involved in processes such as RNA processing, transcription, translation and ubiquitin metabolism. None of these approaches is certain to validate two-hybrid interactions as ones that occur in the parasite; however, they provide test cases for the parasitology community for additional research efforts.

Yeast membrane protein interactions

Another large-scale project that we have tackled focused on the interactions of membrane proteins, a difficult class of proteins to work with in interaction studies because they are ill-suited for the nuclear localization that is required for a transcription-based approach (such as the two-hybrid assay) and they are difficult to purify in biochemical approaches that rely on tagged proteins and mass spectrometry identification. We attempted to identify interactions for a set of ≈700 *S. cerevisiae* proteins annotated as integral membrane proteins by applying a modified membrane two-hybrid assay [49,50] on a large scale. This assay relies on the fusion of two membrane proteins to the two halves of ubiquitin, an N-terminal domain (N-Ub) and a C-terminal domain (C-Ub) (Fig. 3A). The C-Ub domain, in turn, is fused to a LexA-VP16 transcription factor. Interaction of the membrane proteins reconstitutes a quasi-native ubiquitin, leading to cleavage by cellular ubiquitin-specific proteases after the ultimate ubiquitin residue. The cleavage releases the transcription factor, which enters the nucleus and activates expression of reporter genes, detected in our case by expression of the *HIS3* gene, and thus growth on media lacking histidine.

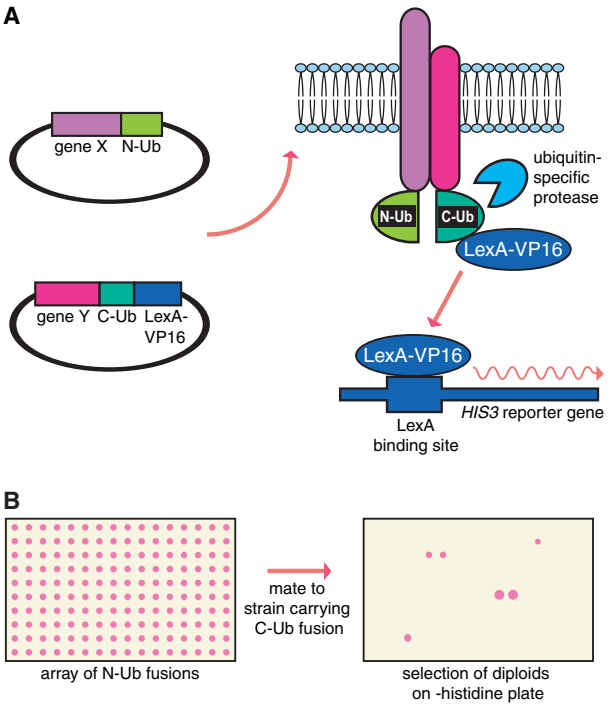

Fig. 3. Array-based split-ubiquitin approach. (A) Plasmids encode protein fusions to the N-terminal (N-Ub) and C-terminal (C-Ub) halves of ubiquitin. The C-Ub plasmid additionally encodes the LexA-VP16 transcription factor. Interaction of the X and Y proteins leads to reconstitution of ubiquitin, cleavage by cellular proteases, and transcriptional activation of the *HIS3* reporter gene by LexA-VP16. (B) An array of transformants was generated that includes ≈700 yeast membrane proteins fused to N-Ub. Mating of these transformants to a strain carrying a protein fused to the C-Ub domain allows selection of diploids on a plate deficient in histidine.

We generated an array of ≈1400 colonies, representing two transformants for each annotated membrane protein, as a fusion to the N-Ub domain (Fig. 3B). About half of the set of ≈700 C-Ub fusions were suitable to screen against this array by the use of a mating assay. From a duplicate set of screens, we identified a total of nearly 2000 putative protein interactions [51]. But, as with the traditional two-hybrid approach, we realized that many of the interactions detected by the split ubiquitin assay were likely to be false positives.

Unlike the case for *P. falciparum* proteins, for yeast proteins there is a wealth of available data from both small-scale and HT studies. Thus, a large fraction of yeast proteins have been classified in the Gene Ontology system [52] for their biological process, molecular activity, or subcellular localization. Furthermore, virtually all of the genes of yeast have been individually deleted and the resulting phenotypes of the deletion strains examined for numerous properties [53]. Transcriptional profiles of yeast genes have been carried out under many different environmental conditions [54]. In addition, we considered a number of features of the interactions derived from the assay itself, such as whether one or both of the transformants in the array yielded a signal, whether positives were observed in one or both of the duplicate screens, whether an interaction was found in both of the reciprocal orientations of the vectors, and how much growth on a histidine-deficient plate a transformant displayed.

To use all of this information for classifying the interactions by likelihood, we collaborated with William Noble's group at the University of Washington to apply a support vector machine (SVM) approach [55]. The SVM is an algorithm that is trained on a set of positive examples and a set of negative examples to 'learn' the features that discriminate these two sets. It then uses this knowledge to separate the uncharacterized examples into two groups that resemble either the positive or the negative training set.

Our SVM analysis of 100 trials revealed that ≈7% of the interactions were of the highest confidence and always grouped with the positive set, 11% were of next highest confidence, and 24% of the next highest confidence. Slightly more than half of the interactions were never grouped with the positive set and were of low confidence. We identified examples of small clusters of proteins implicated in such processes as

insertion of proteins into the endoplasmic reticulum, vesicle transport, sterol biosynthesis, and phosphate transport. In each cluster, we were able to define interactions of high likelihood (including positive training set examples), as well as those of decreased likelihood. The interactions defined in these clusters, like the subnetworks found for *P. falciparum* proteins, ultimately must be experimentally validated by biologists and biochemists whose research focuses on these processes.

Conclusions and perspectives

The YTH assay turned out to be highly general in its applicability to a wide range of different proteins, and easily mastered by the biological community. However, early efforts with this method demonstrated that in addition to *bona fide* interactors, false positives were readily observed. These undesirable positives could be eliminated, but to do so required labor-intensive experimental means. As the assay became more widespread, it proved amenable to HT approaches, first for simple organisms and then for fruit flies and nematodes. These large-scale approaches resulted in huge data sets of interactions, but they also led to the inclusion of a significant degree of false positives. Because experimental methods could no longer be used to weed out false positives, new computational methods to classify interactions by their confidence were developed.

The next frontier is likely to be protein interaction maps of human cells, with efforts already underway to accomplish this goal. The yeast assay, along with the complementary biochemical approach, will be used to yield enormous riches of interaction pairs, but continuing computational efforts will be needed both to assess the reliability of the data and to reveal the implications of these interactions for insights into biological processes.

Acknowledgements

I thank Eric Phizicky for comments on the manuscript and Marissa Vignali for help with the figures. This work was supported by NIH grants from the National Center for Research Resources (RR11823) and the National Institute of General Medical Sciences (GM64655). I am an investigator of the Howard Hughes Medical Institute.

References

1 Sanger F, Nicklen S & Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**, 5463–5467.

2 Fields S (2001) The interplay of biology and technology. *Proc Natl Acad Sci USA* **98**, 10051–10054.

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

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

5 Fields S & Song O (1989) A novel genetic system to detect protein–protein interactions. *Nature* **340**, 245–246.

6 Chien CT, Bartel PL, Sternglanz R & Fields S (1991) The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci USA* **88**, 9578–9582.

7 Vidal M & Legrain P (1999) Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Res* **27**, 919–929.

8 Bartel PL, Roecklein JA, SenGupta D & Fields S (1996) A protein linkage map of *Escherichia coli* bacteriophage T7. *Nat Genet* **12**, 72–77.

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

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

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

12 Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736.

13 Formstecher E, Aresta S, Collura V, Hamburger A, Meil A, Trehin A, Reverdy C, Betin V, Maire S, Brun C *et al.* (2005) Protein interaction mapping: a *Drosophila* case study. *Genome Res* **15**, 376–384.

14 Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T *et al.* (2004) A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543.

15 Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM & Eisenberg D (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* **30**, 303–305.

16 von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S & Bork P (2002) Comparative assessment

of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403.
17 Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S *et al*. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **295**, 321–324.
18 Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804.
19 Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M & Snyder M (2002) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev* **16**, 3017–3033.
20 Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M *et al*. (2004) Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813.
21 Stuart JM, Segal E, Koller D & Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**, 249–255.
22 Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M & Gerstein M (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res* **14**, 1107–1118.
23 Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR & Ideker T (2004) PathBLAST: a tool for alignment of protein interaction networks. *Nucleic Acids Res* **32**, W83–W88.
24 Deane CM, Salwinski L, Xenarios I & Eisenberg D (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* **1**, 349–356.
25 Ge H, Liu Z, Church GM & Vidal M (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat Genet* **29**, 482–486.
26 Grigoriev A (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res* **29**, 3513–3519.
27 Mrowka R, Patzak A & Herzel H (2001) Is there a bias in proteome research? *Genome Res* **11**, 1971–1973.
28 Jansen R, Greenbaum D & Gerstein M (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res* **12**, 37–46.
29 Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, Brazma A & Holstege FC (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell* **9**, 1133–1143.
30 Saito R, Suzuki H & Hayashizaki Y (2002) Interaction generality, a measurement to assess the reliability of a protein–protein interaction. *Nucleic Acids Res* **30**, 1163–1168.
31 Wuchty S, Oltvai ZN & Barabasi AL (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet* **35**, 176–179.
32 Goldberg DS & Roth FP (2003) Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci USA* **100**, 4372–4376.
33 Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U & Margalit H (2004) Network motifs in integrated cellular networks of transcription-regulation and protein–protein interaction. *Proc Natl Acad Sci USA* **101**, 5934–5939.
34 King AD, Przulj N & Jurisica I (2004) Protein complex prediction via cost-based clustering. *Bioinformatics* **20**, 3013–3020.
35 Strogatz SH (2001) Exploring complex networks. *Nature* **410**, 268–276.
36 Bader GD & Hogue CW (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat Biotechnol* **20**, 991–997.
37 Barabasi AL & Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101–113.
38 Hishigaki H, Nakai K, Ono T, Tanigami A & Takagi T (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast* **18**, 523–531.
39 Letovsky S & Kasif S (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* **19** (Suppl. 1), 197–204.
40 Aloy P & Russell RB (2002) The third dimension for protein interactions and complexes. *Trends Biochem Sci* **27**, 633–638.
41 Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J & Gerstein M (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet* **18**, 529–536.
42 Sprinzak E & Margalit H (2001) Correlated sequence-signatures as markers of protein–protein interaction. *J Mol Biol* **311**, 681–692.
43 Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF & Gerstein M (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* **302**, 449–453.
44 Asthana S, King OD, Gibbons FD & Roth FP (2004) Predicting protein complex membership using probabilistic network reliability. *Genome Res* **14**, 1170–1175.
45 Eisenberg D, Marcotte EM, Xenarios I & Yeates TO (2000) Protein function in the post-genomic era. *Nature* **405**, 823–826.

46 Jeong H, Mason SP, Barabasi AL & Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* **411**, 41–42.
47 Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S *et al*. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511.
48 LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth J, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C *et al*. (2005) A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* in press.
49 Johnsson N & Varshavsky A (1994) Split ubiquitin as a sensor of protein interactions in vivo. *Proc Natl Acad Sci USA* **91**, 10340–10344.
50 Stagljar I, Korostensky C, Johnsson N & te Heesen S (1998) A genetic system based on split-ubiquitin for the analysis of interactions between membrane proteins in vivo. *Proc Natl Acad Sci USA* **95**, 5187–5192.
51 Miller JP, Lo RS, Ben-Hur A, Desmarais C, Stagljar I, Noble WS & Fields S (2005) Large-scale identification of yeast integral membrane protein interactions. *Proc Natl Acad Sci USA* **102**, 12123–12128.
52 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*. (2000) Gene ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29.
53 Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B *et al*. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391.
54 Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126.
55 Boser BE, Guyon IM & Vapnik VN (1992) A training algorithm for optimal margin classifiers. 5th Annual ACM Workshop on COLT. (Haussler, D. ed), pp 144–152, ACM Press, Pittsburgh, PA, USA.