The yeast two-hybrid assay: still finding connections after 25 years

Marc Vidal & Stanley Fields

The idea of using hybrid proteins containing transcription factor domains to analyze protein-protein interactions was described in 1989. Over the past 25 years, this method has begun to reveal the complex protein networks that underlie cellular behavior.

This year marks 25 years since the publication of the first paper describing the yeast two-hybrid assay[1]. Twenty-five years is a long survival time for a biological technique; many methods published in this and similar journals perish in much shorter order. Upon receiving the original submission in 1989, *Nature* swiftly returned it unreviewed because it was not of sufficient general interest, but reversed course upon appeal. Only with the method's widespread use, continuing over a quarter of a century, has *Nature*'s decision to publish been vindicated.

The two-hybrid method arose from the efforts of one of us (S.F.) to devise a technology that would satisfy his university's request for grant applications that led to products with potential commercialization. The concept originated from knowledge about the domain structure of transcription factors, which suggested that hybrid proteins based on these factors could be exploited for a novel purpose. This purpose—turning on a reporter gene in yeast via the interaction of two proteins: one fused to a DNA-binding domain and one fused to a transcriptional activation domain (**Fig. 1a**)—was proposed to have commercial possibilities such as sales of libraries encoding hybrid proteins and licenses to practice the method. But like *Nature* initially, the review panel for technology grants was unpersuaded of the generality of the approach and did not fund the two-hybrid proposal.

Following the publications describing first the method and then the demonstration that library searches could turn up interacting partners[2], the two-hybrid assay began moving into general use. The identification of partners for widely studied proteins—mostly in the cancer field—caught the attention of molecular biologists and geneticists and further catalyzed the method's use. Evidence had emerged in the 1980s that protein interactions are key to our understanding of biology, underlying biological processes from the formation of molecular machines and enzymatic complexes to the regulation of signal transduction pathways and cell-cell interactions. Furthermore, such interactions are perturbed in cancer, heart disease, neurodegeneration and every other malady. In short, the two-hybrid assay arrived to address a fundamental property of biological circuitry just as it was becoming apparent that protein interactions constitute a vast jigsaw puzzle of interlocking cellular components. Thus, although its origin—like that of several methods— was not in the tackling of a specific biological problem, the two-hybrid approach became popular precisely because it filled such an obvious biological need.

In addition, the two-hybrid assay possessed numerous virtues, many of which were not immediately obvious upon its launch. First, it was simple to perform even though it revealed detailed information about protein binding sites. The assay required not much more than a couple of plasmids and a yeast strain along with modest DNA sequencing capacity, and it benefited from cloning and sequencing advances, new reporter strains with Gal4- or LexA-driven reporter genes, and a proliferation of cDNA and genomic activation-domain libraries (reviewed in ref. 3). Second, the approach exploited optimal DNA-binding sites, potent activation domains and metabolic enzymes as reporters to amplify signals from weak interactions. Third, the method succeeded with an impressive array of diverse proteins from any organism, especially human proteins. Fourth, the strategy of analyzing interactions through the use of transcription driven by hybrid proteins had wide applicability, extending to the detection of interactions between proteins and DNA[4,5], proteins and RNA[6], proteins and small molecules[7], and other combinations (reviewed in ref. 3). Protein interactions not amenable to a transcription-based assay could be analyzed by the functional reconstitution of other proteins such as ubiquitin[8], dihydrofolate reductase[9] and others (reviewed in ref. 3), and the concept proved to be powerful in mammalian cells[10] and even *in vitro*[11]. Fifth, the concept could be run in its opposite mode to identify reagents for studying and manipulating interactions. Reverse two-hybrid and one-hybrid assays were developed to identify *cis*-acting mutations and *trans*-acting reagents that affect binding[12], with the goal of connecting biophysical interactions directly to phenotypic readouts[13].

The launching of two-hybrid technologies was propitious, too, in that systematic efforts were ramping up to generate

Marc Vidal is at the Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA; and Stanley Fields is at the Howard Hughes Medical Institute and Departments of Genome Sciences and Medicine, University of Washington, Seattle, Washington, USA.
e-mail: marc_vidal@dfci.harvard.edu or fields@uw.edu

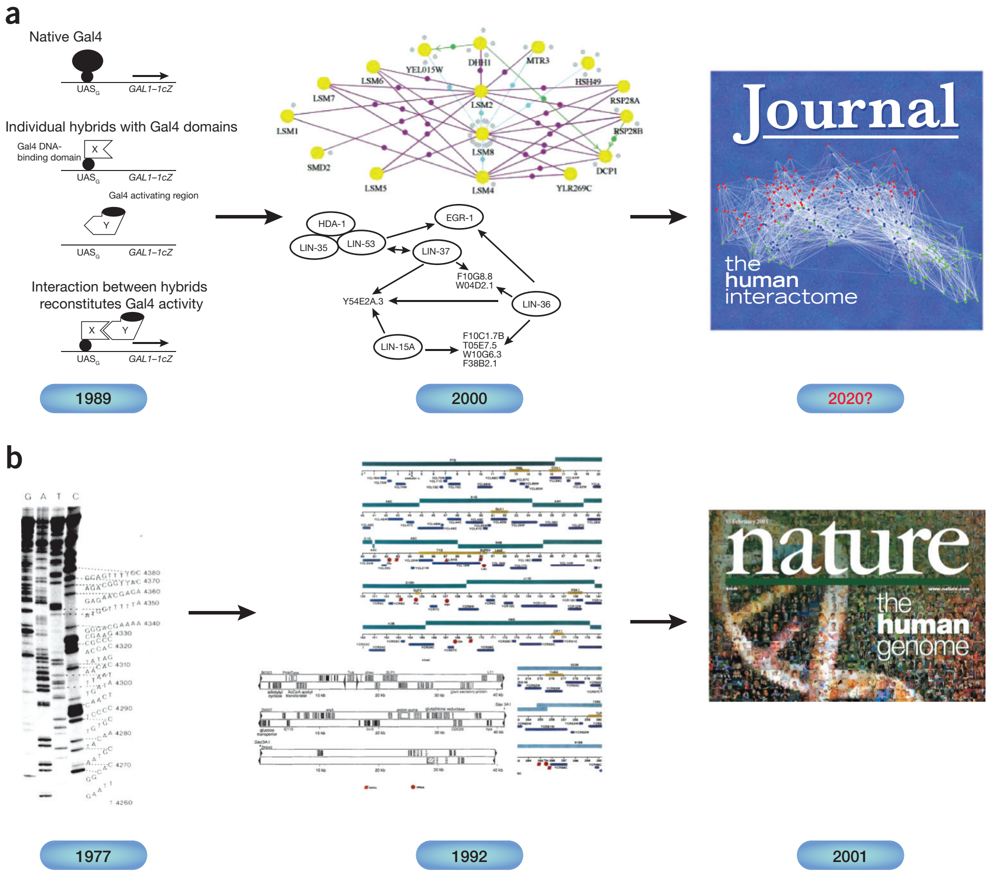

Figure 1 | A human reference interactome by 2020? (**a**) The invention of the yeast two-hybrid assay sparked the idea that protein-protein interaction networks should and could be mapped, as shown here for *Saccharomyces cerevisiae* and *Caenorhabditis elegans*. The current status of the field suggests the possibility that a reference human interactome map may become available by the end of this decade. Images adapted from ref. 1, Nature Publishing Group (left), and reproduced from ref. 19, Nature Publishing Group (center top), and ref. 21, AAAS (center bottom). (**b**) The field of protein interaction mapping shows similarity to DNA sequence analysis. A decade and a half following the invention of dideoxy sequencing, the notion of systematic genome sequencing efforts became a possibility, as shown here for portions of yeast and *C. elegans* chromosomes, which eventually led to a reference human genome sequence. Images reproduced from Sanger, F., Nicklen, S. & Coulson, A.R. *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467 (1977) (left) and refs. 14,15, Nature Publishing Group (center).

large DNA sequencing data sets, which would eventually lead to complete genome sequences. This information turned out to be crucial, not only for small-scale experiments aimed at identifying interactions but also for an entirely novel way of comprehending cellular organization. The publication of a full yeast chromosome sequence[14] and three adjacent *Caenorhabditis elegans* cosmids[15] in 1992 was eye opening for those who were starting to imagine studying biology beyond the limitations of one-gene-at-a-time approaches. Although most of the newly identified genes had no assigned function, these papers suggested that similar systematic strategies could be used to functionally characterize the full complement of proteins, or the 'proteome' as we now call it. By the early 1990s, it became clear to both of us, independently, that with proteomes being predicted from genome sequences, the yeast two-hybrid assay might be able to test all pairwise combinations of proteins for interaction; the resulting data might provide interacting networks at the scale of whole cells.

The first large-scale efforts with random libraries of bacteriophage T7 (ref. 16), yeast genomic DNA[17] and mouse cDNA[18] suggested a better approach: the systematic mapping of cellular networks should employ cloned ORFeomes, i.e., organized resources of cloned open reading frames representing the full protein-coding potential of an organism. Not long thereafter, first-generation systematic interactome maps of binary protein interactions became available for yeast[19,20], roundworms[21,22], fruit flies[23] and, eventually, humans[24,25]. Thus, the assay continued to change from its initial one-by-one process to an increasingly high-throughput strategy (**Fig. 1a**).

In addition to binary interactions, indirect associations of proteins due to membership in the same cellular complex also need to be understood at the scale of the whole proteome. As two-hybrid approaches were scaling up, so was the complementary idea of using hybrid proteins in biochemical purifications and then identifying the copurified proteins by mass spectrometry. Provided that a protein was first fused to a conveniently purifiable tag, complexes associating with this protein could be readily identified. The affinity purification–mass spectrometry concept became automated and applied systematically at the proteome scale, with its first major successes in yeast[26–29]. Not unexpectedly, maps of binary interactions and maps of co-complex associations give rise to different views of the protein-protein interaction space, or 'interactome', but their combined use led to improved network models.

Throughout its 25-year journey from a single interaction (of the yeast Snf1 and Snf4 proteins) to current interactome networks, the yeast two-hybrid assay has had its reliability questioned. Claims of high rates of false positives arising from the assay abound in the literature. And yet, today, among all highly reliable interactions published by the scientific community[30], upwards of three quarters are supported by at least one yeast two-hybrid experiment. How has a method that nominally generates so many false positives produced such credible data? Most early problems with the assay were primarily due to the way it was performed, not to any fundamental flaw of the original concept. Because most readouts are based on the ability of yeast cells to grow, artifacts can arise from growth that is independent of the two-hybrid reconstitution of a transcription factor. A few examples illustrate this point. Mutations or rearrangements can result in a DNA-binding domain hybrid that activates transcription on its own, i.e., in the absence of any activation-domain fusion. Expression of activation-domain out-of-frame fusions, as occurs in five out of six clones in randomly generated cDNA libraries, can correspond to irrelevant peptides leading to adventitious binding. Finally, cases occur in which two activation-domain plasmids, one corresponding to a genuine interactor and one irrelevant, cotransform a yeast cell. However, once these artifacts are understood, they can be dealt with by appropriate experimental controls. In short, if pairs of hybrid proteins reconstitute a bona fide transcription factor in yeast, considerable evidence from orthogonal assays in mammalian cells and *in vitro* has shown that these pairs indeed correspond to proteins that can biophysically interact with each other[31].

However, the phrase 'false positives' is often misused and, instead of pointing to experimental artifacts that should be avoided at all costs, actually refers to a more fundamental aspect of biology: one that lies beyond experimental controls to remove artifacts and deals with the fascinating question of whether all proteins that can interact with each other in artificial assays do so in their *in vivo* setting[31]. Referred to as pseudointeractions, such hypothetical genuine biophysical interactions lacking physiological relevance have been hypothesized to represent evolutionary remnants of past functional interactions or reservoirs to evolve new ones, a concept reminiscent of how and why pseudogenes are still present in contemporary genomes. Although it remains unclear what fraction of genuine biophysical interactions corresponds to physiologically relevant interactions versus pseudointeractions, computational analyses that combine biophysical networks with networks of functional relationships have been developed to extract the most biologically relevant information from interactome maps[32].

Such integrated network models obtained by adding other large-scale data sets— from studies of transcription, systematic gene knockouts, gene knockdowns and synthetic loss-of-function phenotypes—significantly improve our understanding of interactomes. Pairs of coexpressed genes and of genes sharing similar phenotypic profiles can be extremely helpful for interpreting biophysical networks. Consider, for example, that interacting proteins are 100 times more likely to correspond to genes with high synthetic similarity than expected by chance. Properties of the combined interactome network models have provided fundamental answers to questions of global cellular organization[33].

Although computational analyses can be useful, formal evidence can be provided by only an *in vivo* experiment that tests the functional consequences of perturbing an interaction. An exciting prospect will be to use the power of the reverse yeast two-hybrid assay to generate the necessary reagents to study large numbers of interactions in their natural setting, be that human cells or cells of model organisms. For example, the process to identify specific interaction-defective alleles in the yeast assay could be scaled up by automation, and these alleles could be subsequently reintroduced into the genomes of mammalian cells or model organisms by gene-editing techniques such as the clustered, regularly interspaced, short palindromic repeats (CRISPR)-Cas9 system. The systematic phenotypic analyses of these alleles should reveal the functions of many physiologically important interactions.

Entering the second 25-year phase of the yeast two-hybrid assay, we can speculate on where the journey is likely to take us. Just as reference genome sequences revolutionized genetics, reference maps of interactome networks will be critical to fully understand cellular systems. With a unified, high-quality and centralized source of information for most of the human interactome at hand, our understanding of biology is likely to be greatly accelerated. The human protein-protein binary interactome is expected to contain on the order of a few hundred-thousand interactions. The combined efforts of the scientific community to map these interactions one or several at a time has led so far to about 10,000 high-quality interactions. Systematic efforts to map the interactome network have now just surpassed this number by about 70% (ref. 30), with equally high-quality information. Thus, the current tally of interactions suggests that on the order of 90% of the task is still ahead of us.

Such a massive amount of data still to be obtained appears an insurmountable challenge. However, the history of the Human Genome Project provides some guidance. It took a long time for sequencing centers to ramp up their activities to produce high-quality sequence at reasonable cost. But by the late 1990s, these centers started producing more sequence data than the rest of the world combined. Just as the exponential growth of production from the sequencing centers relied on comprehensive sets of genomic clones, highly automated processes and stringent quality scores (**Fig. 1b**), the necessary infrastructure of ORFeome collections, assay automation and a stringent empirical framework to control quality have been assembled to ramp up human interactome reference mapping. Dedicated centers now produce more protein interaction data than all other labs combined. The first reference interactome map could very well be released by the end of this decade (**Fig. 1a**).

Similarly to how only a limited portion of the human genome is functionally relevant in any given cell at any given moment, a reference map of most of the protein-protein interactions that can take place in an organism will have to be dissected to unravel dynamic and condition-specific interactions. Beyond a first-version interactome reference, it is harder to speculate on what else will need to be developed to achieve this next goal, but there are signs of 'next-generation interactome mapping' techniques around the corner. For example, attaching DNA barcodes covalently to proteins might make it possible to test millions of proteins pairs on microscopic devices and identify interactions by modern DNA sequencing techniques[34]. This in turn would increase our capabilities of characterizing the interactome at a deeper level, including variations between alternatively spliced polypeptides encoded by the same genes, differences between disease-associated alleles and common variants, and analyses of specific protein domains responsible for macromolecular interactions. In addition, increasingly efficient high-throughput approaches might become available to study the *in vivo* dynamic and functional aspects of the interactome, one cell type at a time, including as yet unimaginable improvements of fluorescence resonance energy transfer (FRET)–based technology and nanoscopy instruments. The availability of an interactome reference map may well serve as a catalyst for such inventions, not unlike the way in which the Human Genome Project was instrumental in providing the right environment for the development of revolutionary techniques such as microarrays, RNA interference and high-throughput sequencing.

In summary, the development of the yeast two-hybrid assay has helped molecular and cell biologists in their quest to understand the cell, at the level of single proteins or a few proteins at a time—a use of the method that is still the predominant one. But the assay also proved fundamental in characterizing interactions among the full complement of proteins in a cell. We anticipate in the next 25 years that the approach and its derivatives will reveal many more unexpected protein and cellular systems properties. The idea of using pairs of hybrid proteins to solve biological questions is likely to outlive the current generation of biologists.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Fields, S. & Song, O. *Nature* **340**, 245–246 (1989).
2. Chien, C.T., Bartel, P.L., Sternglanz, R. & Fields, S. *Proc. Natl. Acad. Sci. USA* **88**, 9578–9582 (1991).
3. Vidal, M. & Legrain, P. *Nucleic Acids Res.* **27**, 919–929 (1999).
4. Li, J.J. & Herskowitz, I. *Science* **262**, 1870–1874 (1993).
5. Wang, M.M. & Reed, R.R. *Nature* **364**, 121–126 (1993).
6. SenGupta, D.J. *et al. Proc. Natl. Acad. Sci. USA* **93**, 8496–8501 (1996).
7. Licitra, E.J. & Liu, J.O. *Proc. Natl. Acad. Sci. USA* **93**, 12817–12821 (1996).
8. Johnsson, N. & Varshavsky, A. *Proc. Natl. Acad. Sci. USA* **91**, 10340–10344 (1994).
9. Pelletier, J.N., Campbell-Valois, F.X. & Michnick, S.W. *Proc. Natl. Acad. Sci. USA* **95**, 12141–12146 (1998).
10. Eyckerman, S. *et al. Nat. Cell Biol.* **3**, 1114–1119 (2001).
11. Ramachandran, N. *et al. Science* **305**, 86–90 (2004).
12. Vidal, M., Brachmann, R.K., Fattaey, A., Harlow, E. & Boeke, J.D. *Proc. Natl. Acad. Sci. USA* **93**, 10315–10320 (1996).
13. Dreze, M. *et al. Nat. Methods* **6**, 843–849 (2009).
14. Oliver, S.G. *et al. Nature* **357**, 38–46 (1992).
15. Sulston, J. *et al. Nature* **356**, 37–41 (1992).
16. Bartel, P.L., Roecklein, J.A., SenGupta, D. & Fields, S. *Nat. Genet.* **12**, 72–77 (1996).
17. Fromont-Racine, M., Rain, J.C. & Legrain, P. *Nat. Genet.* **16**, 277–282 (1997).
18. Walhout, A.J. & Vidal, M. *Genome Res.* **9**, 1128–1134 (1999).
19. Uetz, P. *et al. Nature* **403**, 623–627 (2000).
20. Ito, T. *et al. Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
21. Walhout, A.J. *et al. Science* **287**, 116–122 (2000).
22. Li, S. *et al. Science* **303**, 540–543 (2004).
23. Giot, L. *et al. Science* **302**, 1727–1736 (2003).
24. Rual, J.-F. *et al. Nature* **437**, 1173–1178 (2005).
25. Stelzl, U. *et al. Cell* **122**, 957–968 (2005).
26. Gavin, A.C. *et al. Nature* **415**, 141–147 (2002).
27. Gavin, A.C. *et al. Nature* **440**, 631–636 (2006).
28. Ho, Y. *et al. Nature* **415**, 180–183 (2002).
29. Krogan, N.J. *et al. Nature* **440**, 637–643 (2006).
30. Rolland, T. *et al. Cell* (in the press).
31. Venkatesan, K. *et al. Nat. Methods* **6**, 83–90 (2009).
32. Ge, H., Walhout, A.J. & Vidal, M. *Trends Genet.* **19**, 551–560 (2003).
33. Han, J.D. *et al. Nature* **430**, 88–93 (2004).
34. Gu, L. *et al. Nature* doi:10.1038/nature13761 (21 September 2014).