Protein interactions are crucial to the life of a cell. The analysis of such interactions is allowing biologists to determine the function of uncharacterized proteins and the genes that encode them. The yeast two-hybrid system has become one of the most popular and powerful tools to study protein–protein interactions. With the advent of proteomics, the two-hybrid system has found a niche in interactome mapping. However, it is clear that only by combining two-hybrid data with that from complementary approaches such as mass spectrometry (MS) can the interactome be analyzed in full. This review introduces the yeast two-hybrid system to those unfamiliar with the technique, and discusses how it can be used in combination with MS to unravel the network of protein interactions that occur in a cell.


Keywords: interactome; mass spectrometry; protein–protein interaction; proteomics; yeast two-hybrid

I. INTRODUCTION

One of the key challenges for biologists today is the interpretation of the large amounts of genome sequence data that are being produced for numerous organisms. In particular, efforts are focusing on assigning functions to the many uncharacterized genes, and the proteins they encode. Because protein–protein interactions are fundamental to all cellular processes, it is often possible to infer the
function of an unknown protein by identifying the proteins with which it interacts. Two powerful approaches have become popular in the study of protein-protein interactions. The first is the purification of protein complexes coupled with mass spectrometry (MS), and the second is the yeast two-hybrid system: a sensitive in vivo assay for interaction analysis. In the 14 years since its introduction, the two-hybrid system has developed from a simple assay for the analysis of interactions between known proteins, into a routine system for the identification of novel interactions from complex libraries.

The versatility and sensitivity of two-hybrid and MS-based approaches has led to their use in the growing field of proteomics (the global analysis of all the proteins encoded by a genome). This use is perhaps best illustrated by ambitious projects to map the yeast interactome (the entire protein interaction network of a cell), using systematic, high-throughput two-hybrid (Ito et al., 2000, 2001; Uetz et al., 2000) or MS-based (Gavin et al., 2002; Ho et al., 2002) technologies. However, despite the scale of those studies, neither approach has produced a comprehensive map of the interactome. Importantly, two-hybrid and MS techniques address different aspects of protein-protein interactions (binary interactions and protein complex formation, respectively) (Gavin et al., 2002). Therefore, only by combining the data from complementary technologies such as these two, can the interactome be revealed in full.

The purpose of this review is to introduce the yeast two-hybrid system to researchers already familiar with MS methodologies. In writing this review, I have assumed that the reader has a thorough knowledge of the basic MS approaches used for protein identification, although a short review of those procedures is included. Instead, I have concentrated on reviewing the development of the two-hybrid system, how the basic technology has been adapted to study interactions between proteins and other molecules, and on the development of alternative systems designed to address some of the problems associated with the traditional two-hybrid system. The review discusses the development of the system for genome-wide interaction screens and how, in combination with MS-based approaches, biologically significant maps of the interactome can be generated.

II. THE YEAST TWO-HYBRID SYSTEM

A. A Brief History

1. Basis of Yeast Two-Hybrid System

A collection of work published in the mid-to-late 1980s that described the analyses of eukaryotic transcription factors proved to be the inspiration for the development of the yeast two-hybrid system (Fields & Sternglanz, 1994; Colas & Brent, 1998). That body of work established that the DNA-binding domain (BD) and transcription activation domain (AD) of many of those proteins were functionally and physically separable (Brent & Ptashne, 1985; Hope & Struhl, 1986; Keegan, Gill, & Ptashne, 1986). The BD acts to localize the protein to specific DNA sequences within the genome, whereas the AD contacts the transcription machinery to activate gene transcription. This modular nature was further demonstrated when an active transcription factor was created by fusing the BD of one protein to the AD of an unrelated protein (Brent & Ptashne, 1985). It was later shown that the two domains did not need to be present on the same polypeptide (Ma & Ptashne, 1988). If physically separated, then the BD and AD of a transcription factor do not physically interact and cannot activate responsive genes. Transcriptional activation can only occur if the two domains are physically linked to one another. Fields & Song (1989) realized that it was possible to bring the BD and AD of a transcription factor together by fusing each to one of a pair of physically interacting proteins. With this realization was born the yeast two-hybrid system. In their proof-of-principle paper, Fields & Song (1989) used the interacting proteins SNF1 and SNF4 to reconstitute activity of the yeast GAL4 transcription factor. Technically speaking, reconstitution of an active GAL4 protein was achieved by generating two plasmid constructs. The first encoded SNF1 fused to the C-terminus of the GAL4 BD (BD constructs are referred to as the bait), and the second encoded the GAL4 AD-SNF4 fusion protein (the prey construct). The two plasmids were transformed into a yeast strain that carried a lacZ reporter gene under the control of GAL4-responsive elements. Interaction between SNF1 and SNF4 in this strain was measured in a simple chromogenic assay that detected β-galactosidase activity—the product of the lacZ gene.

2. Development of the System

Since its inception, the yeast two-hybrid system has been developed in numerous ways to improve the technology and to suit individual users. The original system described by Fields & Song (1989) was based on the yeast GAL4 transcription factor, and is now known as the GAL4 system. A similar system was developed that utilized the BD of the bacterial repressor protein LexA in combination with the Escherichia coli B42 AD (the LexA or interaction trap system; Gyrus et al., 1993). Those systems are in routine use in many laboratories, with versions being commercially available from a number of sources.

The original system relied on only a single reporter gene for the detection of an interaction. However, one of the most powerful aspects of the yeast two-hybrid system is the ability to screen prey-expression libraries for proteins that
interact with a bait protein. Library screens have the potential to generate many interesting interactions. Unfortunately, the flip side is that a large number of false positives can be isolated. Consequently, the systems have been refined to detect protein–protein interactions at a high stringency. Increased stringency has been achieved by developing new yeast strains that carry several reporter genes (common reporters include lacZ, HIS3, LEU2, URA3, LYS2, ADE2, gusA, GFP, MEL1), each with a unique promoter structure. In the case of the GAL4 system, the stringency of the interaction assay can be increased by the addition of a competitive inhibitor of the HIS3 gene product. 3-Amino-1,2,4-triazole (3-AT) can be added to growth media at millimolar concentrations to reduce the background caused by basal HIS3 expression. In the LexA system, the stringency of the assay is governed by the use of yeast strains with reporter genes that carry different numbers of LexA operator elements in the promoters of the reporter genes (usually lacZ and LEU2). The more-sensitive yeast strains have up to six LexA-binding sites compared to the less-sensitive strains that have only two. Other measures to reduce the incidence of false positives include the modification of the plasmids that encode the hybrid proteins. New-generation plasmids have been designed to maintain expression at low levels to prevent toxicity problems and the generation of false positives.

The development of haploid yeast strains of opposite mating type (MATa and MATa) for use in the yeast two-hybrid system has proved to be a particularly important step, especially for high-throughput two-hybrid approaches (see Yeast Two-Hybrid). Yeast mating simplifies the analysis of interactions among large numbers of proteins. Bait constructs are transformed into one mating type, and prey into the opposite type. Bait and prey strains are mated to generate diploids that contain both constructs, and assays for protein interactions are conducted. It has been suggested that one is more likely to detect interactions in yeast-mating screens than in standard co-transformation screens (Uetz et al., 2000) because in the mating procedure, the bait and prey proteins are already expressed at a steady state level before mating, whereas in the co-transformation procedure, expression begins after transformation, adding one constraint to the selection process. Commercial systems are available that allow yeast two-hybrid library screens to be performed, using a yeast-mating approach.

B. Current Systems

1. The GAL4 and LexA Systems

The GAL4 and LexA systems both rely on a transcriptional readout for the detection of protein–protein interactions, through the reconstitution of a functional transcription factor (Fig. 1), and both offer a number of advantages over many in vitro approaches. Firstly, the GAL4/LexA system is relatively simple to set up, requires little optimization, and is inexpensive to use. Another advantage is the scale on which interactions can be investigated. The system has the potential to map individual amino acid residues that are involved in a specific protein–protein interaction as well as to identify novel interactions from complex libraries of expressed proteins. This adaptability means that the system can be used to identify binary interactions between known proteins, to screen expression libraries for new interaction partners, and to identify protein domains involved in a particular interaction. Consequently, the system has become a popular tool in many laboratories for the identification and characterization of protein–protein interactions. However, the system is not without its pitfalls. As discussed above, the system is prone to contamination by false positives, and a number of countermeasures have been developed to reduce this problem. Another problem is that of false negatives (i.e., known interactions that are not detected), and there are a number of reasons why interactions can be missed. For instance, the system depends on the analyzed proteins localizing to the yeast cell nucleus. Although many yeast two-hybrid constructs encode nuclear localization signals (NLS), problems may still occur with proteins that carry strong signals for localization to other compartments of the cell or that have strongly hydrophobic domains. In the LexA system, the repression assay has been developed to determine whether the bait localizes to the nucleus and binds the LexA operator sequences (Brent & Ptashne, 1984). Another reason is that the yeast two-hybrid system relies on the transcriptional activation of reporter genes for the detection of an interaction. Proteins that repress gene expression may prove to be unsuitable. Equally, bait proteins that alone activate expression of the reporter genes may be problematic. For example, problems of auto-activation are often encountered when looking at the interactions among transcription factor proteins, although recent data suggest that the problem may be more far-reaching because up to 10% of randomly generated cDNAs inserted into a bait vector can activate reporter gene expression (Fashena, Serebriskii, & Golemis, 2000). As a result, it is important that all baits are tested for auto-activation prior to use in two-hybrid interaction studies. Detailed protocols to test for auto-activation, plus other aspects of yeast two-hybrid approaches, including yeast transformations and library and high-throughput screening, have been published extensively in the literature (e.g., Walhout & Vidal, 2001; Causier & Davies, 2002).

To address a number of the problems encountered with the GAL4/LexA systems, some alternative systems have been developed. The key technologies are outlined in the following sections.
2. The Sos Recruitment System

A cytoplasm-based system for the detection of protein–protein interactions has been developed (Aronheim et al., 1997). It utilizes a temperature-sensitive yeast strain (cdc25H) that carries a mutation in a gene encoding a Ras guanyl nucleotide exchange factor (GEF). The cdc25 phenotype can be rescued by the human Ras GEF hSos, if it is recruited to the yeast plasma membrane, activates yRas, and thus reconstitutes the damaged signaling pathway. hSos is expressed as a fusion with one test protein, and the second test protein is expressed with a v-Src myristylation signal that targets it to the yeast plasma membrane. Interaction between the two test proteins recruits hSos to the plasma membrane. Protein–protein interactions are detected by growth of the cdc25H yeast strain at the restrictive temperature (see Fig. 2). A similar yeast two-hybrid system, which uses the recruitment of a modified human mRas to rescue the cdc25 strain, has also been developed (Broder, Katz, & Aronheim, 1998).

Those systems do not depend on a transcriptional readout for the detection of protein–protein interactions, which suggests that they can be used in place of GAL4/LexA systems where problems with transcriptional repressors and activators may be envisaged. Furthermore, those systems are based on interactions that occur in the yeast cytoplasm. The detection of interactions in the cytoplasm offers two advantages: those systems may be more suited to proteins that may not target well to the nucleus, and in the cytoplasm, unlike the nucleus, the test proteins may undergo post-translational modifications that may be important for some interactions.

However, although the systems have many advantages, they also have a number of associated problems. The cdc25H yeast strain has been reported to revert, and consequently the temperature sensitivity of the strain has to be confirmed. Checking the phenotype is particularly important because the detection of a protein–protein interaction relies solely on the growth of the yeast strain at the restrictive temperature; there is no other marker to confirm an
interaction. In addition, expression libraries used with these systems encode proteins that are fused to the myristylation signal. Consequently, Ras proteins are a common false positive isolated in library screens. Procedures to screen through library positives to eliminate those that encode Ras have been published (e.g., Huang et al., 2001).

3. The Split-Ubiquitin System

Ubiquitin is a small polypeptide that associates with proteins and targets them for degradation (Varshavsky, 1997). Once tagged, the protein is transported to the 26S proteasome, where ubiquitin is released by ubiquitin-specific proteases (UBPs) and is recycled back to the cytoplasm (Auerbach et al., 2002 and references therein). Johnsson & Varshavsky (1994) took advantage of the specific nature of the UBP s to develop the split-ubiquitin system for the analysis of protein–protein interactions. One test protein (X) was expressed as a fusion with a C-terminal polypeptide of ubiquitin (Cub), to which a reporter protein was attached. The second test protein (Y) was fused to a modified N-terminal region of ubiquitin (Nub). This modification prevents correct folding of Nub and thus its association with Cub. However, if X and Y interact, then an active ubiquitin molecule ("split-ubiquitin") is reconstituted. Split-ubiquitin is recognized by UBPs, and the reporter protein is released. In the original system, the reporter protein was detected by using immunological techniques. A number of variations of the system have since been published. In the first, designed for the study of interactions between membrane proteins, the transcription factor A-LexA-VP16 was used as the reporter protein (Stagljar et al., 1998). Using a LexA two-hybrid yeast strain, protein–protein interactions can be detected through the activation of the lacZ and HIS3 reporter genes (see Fig. 3). In the second derivation of the system, Ura3 was used as the reporter protein (Wittke et al., 1999). Ura3 converts 5-fluoroorotic acid (5-FOA) into the toxic compound 5-fluorouracil. Because the Cub polypeptide alone is not recognized by UBPs, the Cub-Ura3 protein is toxic to yeast cells that grow on media supplemented with 5-FOA. However, if Cub-Ura3 is expressed as a fusion with protein X, which interacts with protein Y fused to Nub, then split-ubiquitin is formed, Ura3 is released by UBPs and is rapidly degraded within the cell. Consequently, yeast cells that express the interacting proteins X and Y do not metabolize 5-FOA and can grow in its presence.

Reports of protein–protein interactions using split-ubiquitin systems are becoming commonplace in the literature. The system is versatile and has been used to demonstrate interactions between nuclear proteins, cytoplasmic proteins, and integral membrane proteins. In terms of use, it is second only to the GAL4/LexA systems (Auerbach et al., 2002).

4. The Dual-Bait System

The dual-bait system uses two sets of reporter genes, bound by different BDs, to detect protein–protein interactions. The first set of reporter genes (lacZ and LEU2) contains elements to which LexA binds, whereas the promoters of the second set (LYS2 and gusA) are bound by λcl (Serberiiskii, Khazak, & Gomelis, 1999). Two bait constructs are prepared. The first encodes protein X1 fused to the LexA BD, the second expressing protein X2 as a fusion with the λcl BD. The constructs are co-transformed into an appropriate yeast strain, and interaction studies are performed with prey proteins fused to the transcription AD of B42 (see Fig. 4).

The dual-bait system can be used to screen libraries with two bait proteins in a single experiment. In addition, the system also facilitates the analysis of domains required for an interaction, and allows screening of proteins or other molecules that disrupt a particular interaction (reviewed by Fashena, Serberiiskii, & Golemis, 2000).

5. Other Two-Hybrid Systems

Alternative yeast two-hybrid technologies are being developed all the time. Systems such as the reverse Ras recruitment system (Hubsman, Yudkovsky, & Aronheim, 2001), and one that utilizes a G-protein-linked signal...
transduction pathway (Ehrhard et al., 2000), both developed to study integral membrane protein interactions, are among those systems that have yet to find their full potential.

The most common two-hybrid systems are based on assays performed in yeast cells. However, similar systems have been developed for use in mammalian and prokaryotic cells (Mohler & Blau, 1996; Luo et al., 1997; Joung, Ramm, & Pabo, 2000; Serebriiskii, Toby, & Golemis, 2000). In addition, a Ras-recruitment system has been developed for use in mammalian cell lines (Maroun & Aronheim, 1999).

6. Variations on a Theme
The yeast two-hybrid system revolutionized the way that researchers analyzed protein–protein interactions. Consequently, investigators soon found ways of using two-hybrid
technology to study protein interactions with other molecules (such as nucleic acids) and to study interactions that involve more than two proteins.

The first variant of two-hybrid technology was the yeast one-hybrid system, designed for the analysis of protein–DNA interactions (Wang & Reed, 1993; see Fig. 5A). In principle, the yeast one-hybrid system is similar to the two-hybrid system in that the interaction between a bait and prey is detected in vivo. However, in the case of the one-hybrid system, the bait is a known DNA sequence for which binding proteins are to be identified. In practical terms, multiple copies of the bait sequence are

![Diagram of yeast two-hybrid-based technologies.

**Figure 5.** Yeast two-hybrid-based technologies. A: The one-hybrid system. Target DNA sequences are placed upstream of reporter genes. Interaction between prey proteins (fused to an AD) and the target DNA results in the activation of the reporters. B: The RNA three-hybrid system. A hybrid RNA molecule is generated, part of which is bound by the MS2 coat protein (fused to a BD) and part that contains the RNA sequence of interest. The hybrid RNA molecule is recruited to the promoters of yeast reporter genes by association with the MS2-BD fusion protein. Interaction between the target RNA sequence and proteins fused to a transcription AD, up-regulates reporter gene expression. C: The protein three-hybrid system. In the top diagram, protein Z interacts with a X–Y protein dimer. In the bottom diagram, an interaction between the X-bait and the Y-prey can only be detected in the presence of protein Z. D: The reverse two-hybrid system. Expression of the reporter gene is toxic to the yeast cells under certain conditions. In this example, the product of the URA3 gene breaks down 5-fluoroorotic acid (5-FOA) into a toxic metabolite. If proteins X and Y interact, then the yeast cells fail to grow in the presence of 5-FOA (top diagram). However, if the interaction between X and Y is somehow disrupted, then growth on 5-FOA is detected (bottom diagram).
inserted upstream of yeast reporter genes such as lacZ or HIS3. Proteins that bind the bait are isolated from a prey library by using screening procedures that are similar to those used in yeast two-hybrid screens (for review, see Vidal & Legrain, 1999). Conveniently, libraries used for yeast two-hybrid screens can often be used for one-hybrid screens.

A modification of the LexA-based yeast two-hybrid system has provided the tools for researchers to study RNA–protein interactions in vivo. Developed by SenGupta et al. (1996), and known as the RNA yeast three-hybrid system, it is used to isolate proteins that bind to a bait RNA molecule from a VP16-AD prey library (again, the same libraries used for two-hybrid can be used here). The bait RNA molecule is a hybrid between the RNA of interest and MS2 RNA that binds to the MS2 coat protein. By fusing the MS2 coat protein to the LexA BD, the bait RNA molecule is recruited to the promoters of the yeast reporter genes. Proteins that interact with the RNA of interest can easily be identified from the prey library by assaying for reporter gene activity. The basis of the RNA yeast three-hybrid system is presented in Figure 5B.

Protein yeast three-hybrid systems have been designed for the analysis of ternary complex formation (Zhang & Lautar, 1996; see Fig. 5C). Many cellular processes depend on the formation of protein complexes—gene transcription is one example. Traditional yeast two-hybrid technologies are limited because they only detect binary interactions between proteins. Protein three-hybrid systems identify proteins required for the interaction between two other proteins (X and Y) or that form a complex with X and Y. For example, if a known protein (Z) is thought to be required for the interaction between proteins X and Y, then X and Y can be expressed as bait and prey whereas Z is expressed only with a NLS. Reporter gene activity can only be detected if X, Y, and Z all interact. Alternatively, proteins that interact with the X–Y dimer can be isolated from a prey library. In this case, X is expressed as bait, and Y is expressed with a NLS. As above, reporter gene activity is only detected if X, Y, and Z form a complex. A number of variants of the protein three-hybrid system have been reported, including one that enables the detection of interactions that are dependent on non-protein molecules (for reviews see Brent & Finley, 1997; Van Crickringe & Beyaert, 1999).

A number of systems that use counter-selectable reporter genes have been developed to identify mutations that disrupt protein–protein or DNA–protein interactions. Those reverse one- and two-hybrid systems rely on the expression of reporter genes that are toxic to yeast cells if the bait and prey interact (see Fig. 5D). Such reporters include URA3 (Vidal et al., 1996), which breaks down 5-FOA into a toxic metabolite (see The Split-Ubiquitin System); LYS2, the expression of which prevents yeast growth on media containing α-aminoadipate (Brent & Finley, 1997); and CYH2 which confers sensitivity to cycloheximide (Leanna & Hannink, 1996).

C. Applications of Yeast Two-Hybrid

In contrast to in vitro studies of protein–protein interactions, the yeast two-hybrid system offers a sensitive in vivo assay to study interactions, and offers a number of advantages. The system maintains eukaryotic proteins in an environment close to their physiological norm, ensuring that some post-translational modifications will be performed and that the expressed proteins fold correctly. One other key advantage of the system, touched upon in previous sections of this review, is the scale on which interactions can be investigated. On the smallest scale, individual residues involved in an interaction can be mapped, using, for example, reverse two-hybrid technologies. On the largest scale, interactions can be identified from complex mixtures of proteins by using systematic high-throughput approaches. The system can be used routinely to test for interactions between known proteins or to identify novel interaction partners through prey library-screening approaches.

As genome sequencing projects are completed, attention is turning to the functional characterization of the predicted genes. Often, the function of a protein can be inferred from the proteins with which it interacts. A large amount of data generated by protein–protein interaction studies is available in databases (see Emerging Technologies). Those databases provide the opportunity to take proteins conserved between species, and to predict the proteins with which they may interact. As more data become available, it may be possible to construct entire protein interaction pathways for a species, completely in silico. A study by Matthews et al. (2001) provides validation for such an approach. They randomly selected 72 yeast interactions identified by the large-scale two-hybrid approaches, and tested the corresponding Caenorhabditis elegans proteins to determine whether similar interaction partnerships were formed. Thirty-one percent of the interactions detected between the yeast proteins were also detected between the corresponding C. elegans proteins.

III. STUDYING PROTEIN INTERACTIONS BY MS

The main aim of this review is to introduce yeast two-hybrid technologies to the MS community, and to demonstrate how it and MS-based approaches can be used in combination to address key proteomic questions. However, it seems appropriate at this stage to offer a brief discussion of some of the biochemical and MS techniques used to
identify components of protein complexes. Excellent reviews by Aebersold & Mann (2003), Bauer & Kuster (2003), and Lin, Tabb, & Yates (2003), provide more detailed discussion of the MS techniques used for protein identification.

The yeast two-hybrid system is one of many ways in which protein–protein interactions can be investigated. Although the two-hybrid system is an in vivo genetic assay, numerous in vitro protocols are also in common use, such as co-immunoprecipitation, pull-down assays, affinity chromatography, protein-array chips, and surface plasmon resonance. Many of these methods are now used in combination with MS to examine protein interactions.

As a first step to identify proteins that associate with a particular protein (the bait), the bait protein is modified to carry an affinity tag. A number of tags are in common use such as glutathione S-transferase, His6, calmodulin-binding peptide, and common epitope tags (haemagglutinin, myc, FLAG) (Pizicky et al., 2003). Often, different tags are used in combination on the same protein to allow an extensive purification of protein complexes (tandem affinity purification (TAP); first developed by Rigaut et al., 1999). In some species, it is possible to express the bait at near physiological levels by using the native promoter. In *Saccharomyces cerevisiae*, for example, allele replacement can be used to replace the native gene in the yeast genome with the gene that encodes the bait protein (Aebersold & Mann, 2003). However, in species where allele replacement is not possible, tagged proteins are usually expressed transiently. Expression of the tagged proteins allows them to form normal physiological complexes that are isolated and purified under mild conditions, using one or more affinity steps.

Following affinity purification of protein complexes, their components are separated from one another either by gel electrophoresis or liquid chromatography (LC). In the case of gel electrophoresis, proteins can be separated in one dimension (based on molecular weight) or in two dimensions (based on charge and molecular weight). Proteins are stained, and gel slices that contain the individual proteins are excised. These slices are subjected to protease digestion, typically with trypsin or Lys-C, to produce characteristic peptide fragments that are analyzed with standard MS technologies (see below). Alternatively, complete protein complexes can be fragmented with proteases, and the peptide mixture subjected to two orthogonal LC fractionations in combination with tandem MS (Deshaies et al., 2002; Bauer & Kuster, 2003).

Two MS methods are in common use for protein identification. Peptide mass fingerprinting (PMF) uses matrix-assisted laser desorption/ionization time-of-flight MS (MALDI-TOF) to determine the masses of peptides generated by a tryptic digest. Experimentally determined protein masses are compared with a database that consists of peptide fragments derived from protein sequences digested in silico with trypsin. Experimental proteins are identified by a statistically significant overlap with the computer-generated peptide masses (Bauer & Kuster, 2003). The second method, tandem mass spectrometry (MS/MS), uses a combination of peptide mass and sequence data for protein identification. The mass of a particular peptide is measured, and that peptide is isolated from the mixture within the mass spectrometer. The accelerated peptide protonated molecule ions, (M + H)+, collide with inert gas molecules to cleave the peptide along its backbone to generate fragments that differ in length by only one amino acid. The mass of the fragments is measured, and the amino acid sequence of the peptide is thus determined. Numerous datasets (i.e., peptide sequence, peptide mass, and fragment mass) can be simultaneously queried against appropriate MS/MS databases. Analysis of a number of peptides from the protein digest can be used to confirm the identity of the protein (Bauer & Kuster, 2003).

Binary data generated by simple yeast two-hybrid assays may provide a starting point for the analysis of protein assemblies, using MS-based approaches. A single binary interaction, or a set of such interactions, may point at functional pathways in which these proteins may function. An MS-based approach would confirm those pathways by identifying the complexes of proteins of which the binary interactions are only a part. Used in this way, a combination of two-hybrid and MS-based approaches provides a powerful system for the characterization of protein-interaction networks. Certainly, those two techniques are proving to be powerful allies in the race to delineate interactomes.

### IV. GENOME-WIDE ANALYSIS OF PROTEIN–PROTEIN INTERACTIONS

#### A. Yeast Two-Hybrid

Over the years, the yeast two-hybrid system has demonstrated its versatility in the study of protein–protein interactions. Now, in the post-genomic era, the system has found a new lease of life in the field of interaction proteomics.

Two systematic methods have been employed in large-scale yeast two-hybrid screens, and each adopts a yeast-mating strategy to test for interactions in an ordered manner. In the matrix-based approach, full-length open-reading frames (ORFs) are cloned into bait and prey two-hybrid vectors that are separately transformed into MATα and MATα yeast strains. Each bait strain is mated with an array of the prey strains, and diploids that express interacting proteins are identified-based on reporter-gene activity. The position on the array is used to identify the interactor. The second approach is library-based.
Undefined prey libraries, expressing either full-length ORFs or randomly generated cDNAs, are screened with bait strains. This screen is simplified by producing pools of the library against which each bait (or pools of baits) can be screened. The identity of the interactors is determined by DNA-sequence analysis. Both methods lend themselves to automation, allowing high-throughput interaction screens.

Early genome-wide protein interaction screens were restricted to simple organisms with small genomes, such as the E. coli bacteriophage T7 (Bartel et al., 1996) and later with the proteomes of the hepatitis C virus (Flajolet et al., 2000) and the vaccinia virus (McCrath et al., 2000). However, as increasing amounts of genome sequence data became available for organisms such as S. cerevisiae, reports of large-scale yeast two-hybrid screens for more complex organisms began to appear in the literature. The first of these reports was a study of the yeast functional complex involved in pre-mRNA splicing. A total of 26 pre-mRNA splicing factors were used as baits to screen against a random yeast genomic library (representing the yeast proteome). More than 400 putative interactions were identified, including known splicing factors, new splicing factors, and proteins involved in mRNA turnover (Fromont-Racine, Rain, & Legrain, 1997; Fromont-Racine et al., 2000). Using the same yeast proteome library, Flores et al. (1999) investigated interactions that involved 12 of the protein subunits of the yeast RNA polymerase III complex. Analysis of exhaustive two-hybrid screens revealed a protein–protein interaction map of the polymerase complex, and suggested a model of the RNA polymerase III pre-initiation complex. Those large-scale interaction screens with yeast proteins demonstrated the suitability of the technique to delineate functionally important protein-interaction networks. Similar screens have also been used to answer important biological questions in organisms such as Drosophila melanogaster (Finley & Brent, 1994), C. elegans (Walhout et al., 2000; Davy et al., 2001), Helicobacter pylori (Rain et al., 2001), and Oryza sativa (Fang et al., 2002).

The first eukaryotic genome to be completely sequenced, and annotated, was that of S. cerevisiae (Goffeau et al., 1996); that genome led to the first comprehensive analysis of the interactome of a cell (Ito et al., 2000, 2001; Uetz et al., 2000). Uetz et al. (2000) used matrix- and library-based approaches to examine the interactions between the proteins encoded by the 6,000 predicted yeast ORFs. Using the matrix-based approach (Fig. 6A), 192 bait proteins were screened against an array of approximately 6,000 preys, and 281 interacting protein pairs were identified. In the library-based approach (Fig. 6B), designed to be an exhaustive analysis of protein interactions, the 6,000 preys were pooled to generate a complex library. The library was screened, using a partially automated system, with individual baits that represented almost all of the 6,000 yeast ORFs; 692 protein–protein interactions were detected through this approach. A comparison of the two approaches suggested that the matrix-based method, which yielded 3.4 interactions per bait, was more sensitive than the library-based methods, which yielded only 1.8 interactions per bait (Uetz & Hughes, 2000). A library-based approach was also used by Ito et al. (2000, 2001) in their comprehensive study of the yeast interactome (Fig. 6C). The 6,000 yeast ORFs were individually cloned into bait and prey vectors. Baits were individually transformed into MATα yeast cells and preys into MATα. Each transformant was propagated in the well of a 96-well plate, and the yeast strains of each plate (minus auto-activators) were collected into a single tube to generate 62 bait and 62 prey libraries. All possible interactions between the pools were analyzed by mating each bait pool against each prey pool, resulting in a total of 3,844 mating reactions. That procedure yielded 841 highly significant interactions (i.e., interactions that were independently isolated more than three times).

B. Mass Spectrometry

Advances in the field of MS have, for the first time, allowed the analysis of complete interactomes, essentially in vitro. In contrast to yeast two-hybrid approaches that detect binary interactions between proteins, MS-based approaches are effective at characterizing large protein complexes. Two such studies on the yeast interactome were recently published (Gavin et al., 2002; Ho et al., 2002; see also Fig. 7). Each study attempted to identify all of the components that were present in ‘naturally’ generated protein complexes. Practically speaking, this goal requires essentially pure preparations of each complex (Kumar & Snyder, 2002). To obtain pure complexes each group generated yeast ORFs fused to DNA sequences that encoded specific tags. In the case of Gavin et al. (2002), 1,739 genes were expressed with the TAP tag, allowing a tandem purification regime. The tagged ORFs were expressed in yeast cells, and were allowed to form ‘natural’ complexes under appropriate physiological conditions. Protein complexes were purified from the cell lysates of 1,167 yeast strains that expressed the modified proteins to detectable levels; first using a high-affinity purification step with mild elution, followed by a second affinity step. Using a similar approach, Ho et al. (2002) expressed 725 proteins (approximately 10% of the predicted yeast proteins) modified to carry the FLAG epitope. Again, the proteins were expressed in yeast cells and complexes were purified, using a single immunoaffinity purification step. Both groups resolved the components of each purified complex with a one-dimensional denaturing polyacrylamide gel electrophoresis (PAGE) step. Protein bands were excised from stained gels, subjected to proteolytic cleavage, and analyzed by either MALDI-TOF MS (Gavin et al., 2002) or...
Proteins were identified through the interrogation of the appropriate databases. From the 1,167 yeast strains generated by Gavin et al. (2002), 589 protein complexes were purified, 232 of which were unique. Ho et al. (2002) used 725 protein baits, and detected 3,617 interactions that involved 1,578 different proteins. Among their set of baits were 86 proteins involved in the DNA-damage response (DDR). Much of the DDR network was delineated with their approach, and contained many known interactions as well as some previously unknown interactions that they predicted to be of biological significance.
Protein-interaction networks, such as those involved in essential cellular activities, should be largely conserved between species. By delineating network organization in one species, an understanding of similar networks in other species can be gained (Kumar & Snyder, 2002). Using their TAP/MS approach, Gavin et al. (2002) performed a parallel analysis of human and yeast complexes, including a cytoskeleton-associated assembly, a nuclear complex, and a Golgi-associated complex. As expected, the complex composition between human and yeast was highly similar, and suggested that, as discussed above, predictions about protein interactions and the formation of complexes can be applied across species based on findings from a single organism such as yeast.

**C. Comparison of the Two Approaches**

The preparation of biologically relevant protein interaction maps requires dependable interaction data. However, a number of published analyses suggest that the current data
are not completely reliable. The first problem is the large number of false positives generated by the yeast two-hybrid and MS-based interactome studies. Ito et al. (2002) analyzed a subset of their two-hybrid data, composed of 415 interactions, and found that approximately 50% of the interactions were biologically significant. Similarly, Gavin et al. (2002) estimated that 30% of their interactions might be spurious. Both approaches also missed a large number of known interactions. In the case of the two-hybrid screens, up to 90% of known interactions were not detected (Ito et al., 2002). Matthews et al. (2001) independently attempted to recapitulate a number of the interactions detected in the large-scale two-hybrid screens, and a reanalysis of their data suggested that ~90% of the interactions that they failed to recapitulate were originally only identified in one of the large-scale screens (Ito et al., 2002). By taking into account the cellular localization and known functional roles of interacting proteins, Sprinzak, Sattath, & Margalit (2003) estimated that the reliability of the high-throughput two-hybrid assays was ~50%. Finally, a comparison of the interactions detected in both yeast two-hybrid screens revealed that they share a surprisingly small number of interactions (141 or ~10% of the total number of independent interactions) (Ito et al., 2001, 2002), possibly because of the differences in techniques used or because neither screen was exhaustive. Taken together, those results imply that a complete picture of the interactome would require the combination of several independent two-hybrid screening projects.

Although many interactions were also missed in the MS-based screens, by comparing their data with interactions reported in the literature, Ho et al. (2002) suggested that, on average, the detection of known interactions was threefold higher with their MS-based approach than by the large-scale two-hybrid approaches. However, the limitations of the two MS-based approaches were demonstrated by a comparison of the common purification data sets from each group (Ito et al., 2002). Of the protein complexes purified in each MS-based screen, 93 were purified by both groups using the same bait. Furthermore, of the 93 shared complexes, 48 contained at least one protein detected by both groups, whereas the remaining 45 complexes had no components in common. Coincidently, the overlap between the two large-scale MS projects was also approximately 10% of the entire proteins detected collectively by the two groups (Ito et al., 2002).

Deng, Sun, & Chen (2003) compared the data from all of the large-scale interaction screens. They developed a maximum likelihood estimation method to assess the reliability of interaction data, and found that the Uetz data were more reliable than that of Ito, and that the Gavin data were more reliable than the Ho data. In addition, they suggested that the MS-based analysis of protein complexes performed better in function predictions than the two-hybrid data—thus validating the theory that each component of a complex can be assigned a function based on that of the whole complex.

Protein-complex formation is seen as more than the sum of its binary interactions (Gavin et al., 2002). Because the two-hybrid system is suited to the characterization of binary interactions, it may not be adequate for the comprehensive analysis of protein complexes. In contrast, MS-based approaches allow for the isolation of large protein complexes and for the detection of networks of protein interactions. However, MS-based approaches are biased towards highly abundant, stable complexes, whereas the two-hybrid system is particularly useful for the detection of weak or transient interactions.

It is clear that yeast two-hybrid and MS-based techniques have both independently made significant impacts on our understanding of the interactome. However, both techniques have many associated problems that, if used alone, limit the information that they can provide. For example, analysis of the data from the large-scale interaction studies demonstrated that each method used produced “a unique distribution of interactions with respect to functional categories of interacting proteins” (von Mering et al., 2002). Our view of the interactome can only be enhanced when data from yeast two-hybrid and MS approaches, in combination with other data, are integrated. This combined approach is discussed further in the following section.

D. Placing Interactions in Context

The generation of false positives and false negatives by both large-scale yeast two-hybrid and MS-based approaches seriously hampers the construction of meaningful interaction maps. The yeast interactome is predicted to range from 10,000–30,000 interactions (Bader & Hogue, 2002; Kumar & Snyder, 2002; Sprinzak, Sattath, & Margalit, 2003). The combined data from all the large-scale protein–protein interaction studies indicates that currently ~15,000 interactions have been identified in yeast (Bader & Hogue, 2002; Kumar & Snyder, 2002), and of these, a large proportion (>50%) may be false positives (von Mering et al., 2002). Continual development of the technologies used to create interactome maps will be required to reduce the incidence of false positives and negatives. However, to gain a complete, functionally significant picture of the interactome, interaction data generated from as many different approaches as possible should be integrated (Uetz & Hughes, 2000; Bader & Hogue, 2002; von Mering et al., 2002). An elegant approach taken by Tong et al. (2002) goes some way to illustrate this point. They used two orthogonal approaches to identify a network of biologically relevant interactions. In the first, phage-display technology was used to isolate peptide sequences
bound by SH3 domains (commonly found in proteins involved in signal transduction or actin cytoskeleton re-organization). Consensus sequences were used to identify proteins within the yeast proteome that contained SH3-binding domains and a protein interaction network was predicted. In the second approach, the yeast two-hybrid system was used to pull out SH3-binding proteins from prey libraries. Fifty-nine significant interactions common to the phage display (394 interactions among 206 proteins) and two-hybrid (233 interactions among 145 proteins) networks were identified. However, although interaction data may provide some insight into function, used alone it is not sufficient to assign function on the scale of the genome. Such data should be combined with large-scale expression pattern analysis of the genes that encode interacting proteins and with the phenotypic analysis of cells that contain mutations in those genes (Legrain, Wojcik, & Gauthier, 2001; Ito et al., 2002; von Mering et al., 2002). For example, recent studies have shown that proteins that interact and function in the same pathway have similar expression patterns and produce similar phenotypes when mutated. Statistical analyses have shown that proteins encoded by genes with similar expression profiles are more likely to interact with one another (Grigoriev, 2001; Ge et al., 2001; Walhout et al., 2002). Furthermore, comparison of the interactions among C. elegans germline proteins with the phenotypes of nematodes that carry null alleles of the genes that encode these proteins has shown that essential proteins have a tendency to interact with one another (Walhout et al., 2002). Those results clearly illustrate that the integration of data from a number of different approaches can be used to link, with high confidence, protein–protein interactions to developmental processes.

V. EMERGING TECHNOLOGIES

The comprehensive analysis of any interactome will require the continued development of the technologies used to examine it. Increased sensitivity, coupled with a reduction in the incidence of false positives, will allow complete protein-interaction networks to be delineated. In terms of MS-based approaches, a number of recent developments may facilitate such studies. For example, weak or transient interactions may be more readily detected by the use of methods based on stable-isotope labeling (Aebersold & Mann, 2003). Such methods allow for fewer and milder purification steps. In addition, they can be used to distinguish between true complexes and those that contain contaminating proteins, thus reducing the number of false positives detected. This technique has already proved fruitful in the study of the protein-interactions involved in the epidermal growth factor receptor pathway (Blagoev et al., 2003). Other advances include multidimensional protein identification technology (MudPIT) (Deshaies et al., 2002 and references therein), which uses a multidimensional LC step in combination with tandem MS to analyze peptides that are generated from protease-treated protein mixtures (instead of from single proteins—as in traditional MS analyses). MudPIT provides snapshots of protein complexes, and can be used to compare complexes under different physiological conditions. Furthermore, it is reported to be suitable for the analysis of ‘megacomplexes’ (very large protein assemblies such as the 2S proteasome) and, because little bias is introduced during the sampling of proteins, it can be used to characterize dynamic complexes that may contain low-abundance proteins, small proteins (<10 kDa), or other proteins missed by traditional MS approaches (Deshaies et al., 2002; Lin, Tabb, & Yates, 2003). However, it remains to be seen whether the sensitivity of systems such as MudPIT is compromised in favor of the ability to cope with complexity (Bauer & Kuster, 2003).

Protein-array chips are becoming a powerful tool for the analysis of protein–protein interactions, particularly because comprehensive sets of individual proteins can be screened by using a wide range of in vitro conditions. For example, in a trial run of the technology, Zhu et al. (2001) generated a yeast proteome array that contained ~80% of the yeast proteins. The chips were probed for interactions with calmodulin in the presence of calcium. Six known interactors were identified as well as 33 novel interactions, demonstrating the suitability of the technology for interactome studies. Indeed, the system could be made high-throughput by coupling the protein-chip approach with MS-based protein identification technologies.

In vivo technologies for the investigation of protein interactions are also advancing rapidly. Systems that simplify yeast two-hybrid analyses and facilitate automation of the technique have been reported (e.g., see Diaz-Camino et al., 2003; Reboul et al., 2003), and robotics systems designed specifically for high-throughput two-hybrid screening are now available. One technology, that detects binary interactions in a manner analogous to the two-hybrid system, is that of fluorescence resonance-energy transfer (FRET) (for review, see Hink, Bisseling, & Visser, 2002). FRET measures the interaction between two proteins, one fused to the cyan fluorescent protein (CFP) fluorophore and the other to the yellow fluorescent protein (YFP), by detecting the transfer of energy from the excited donor (CFP) to the acceptor (YFP). FRET can be used to monitor protein interactions in living cells, is suitable for the analysis of transient interactions, and has the potential to be used in a high-throughput manner (Phizicky et al., 2003).
Elaborate databases are required for the storage and retrieval of interactome data (Legrain, Wojcik, & Gauthier, 2001). Currently, databases such as the yeast proteome database (YPD; Costanzo et al., 2000), those at the Munich Information Centre for Protein Sequences (MIPS; Mewes et al., 2000), and the Biomolecular Interaction Network Database (BIND; Bader, Betel, & Hogue, 2003) contain annotations on thousands of protein–protein interactions. However, obtaining biologically significant information from such databases is dependent on the stored information being free from false positives and other misleading data. As databases become refined and integrate data obtained from many different screening procedures, more reliable interaction maps and predictions can be made. For example, Bader & Hogue (2002) identified 15,143 non-redundant yeast protein interactions among 4,825 proteins (representing ~75% of the yeast genome) from data compiled from different sources in the BIND database.

Bioinformatically generated interaction maps provide a means of assigning a function to uncharacterized proteins. In addition, such technologies are also being used to make broad comparisons of protein interactions among kingdoms, and to produce predictive conserved interactions between species.

VI. CONCLUSIONS

Many challenges lie ahead in our quest for a complete understanding of the interactome. Yeast two-hybrid and MS-based technologies have already demonstrated that they are each a powerful tool for the exploration of protein interaction networks. One can envisage, based on the large-scale interaction studies already performed using these approaches, that these complimentary techniques will spearhead the further unraveling of the interactome.

Much of the work has centered on the model organism: yeast. Although yeast is predicted to express around 6,000 genes, only half of all the predicted protein interactions have thus far been identified. As technology improves, and other functional data become available, comprehensive protein-interaction maps for yeast, and other organisms, will be generated. Biologists will then be charged with the task of associating biological function with these interactions.

ACKNOWLEDGMENTS

The author thanks Dr. Brendan Davies (University of Leeds) for helpful comments during the preparation of the manuscript, and the BBSRC for financial support.

ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD</td>
<td>activation domain</td>
</tr>
<tr>
<td>3-AT</td>
<td>3-amino-1,2,4-triazole</td>
</tr>
<tr>
<td>BD</td>
<td>DNA-binding domain</td>
</tr>
<tr>
<td>CFP</td>
<td>cyan fluorescent protein</td>
</tr>
<tr>
<td>DDR</td>
<td>DNA-damage response</td>
</tr>
<tr>
<td>5-FOA</td>
<td>5-fluoroorotic acid</td>
</tr>
<tr>
<td>FRET</td>
<td>fluorescence resonance-energy transfer</td>
</tr>
<tr>
<td>GEF</td>
<td>guanyl nucleotide-exchange factor</td>
</tr>
<tr>
<td>LC</td>
<td>liquid chromatography</td>
</tr>
<tr>
<td>MALDI-TOF</td>
<td>matrix-assisted laser desorption/ionization time-of-flight mass spectrometry</td>
</tr>
<tr>
<td>MS</td>
<td>mass spectrometry</td>
</tr>
<tr>
<td>MS/MS</td>
<td>tandem mass spectrometry</td>
</tr>
<tr>
<td>MudPIT</td>
<td>multidimensional protein identification technology</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>PAGE</td>
<td>polyacrylamide gel electrophoresis</td>
</tr>
<tr>
<td>TAP</td>
<td>tandem affinity purification</td>
</tr>
<tr>
<td>YFP</td>
<td>yellow fluorescent protein</td>
</tr>
</tbody>
</table>

REFERENCES

Reboul C 2003. for R.


