Annu. Rev. Biochem. 2003. 72:783–812
doi: 10.1146/annurev.biochem.72.121801.161511

PROTEOMICS

Heng Zhu,[1] Metin Bilgin,[3] and Michael Snyder[1,2]
[1]*Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520; email: heng.zhu@yale.edu*
[2]*Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520; email: michael.snyder@yale.edu*
[3]*Biological Sciences and Bioengineering, Sabanc University, Orhanli Tuzla Istanbul, 81474 Turkey; email: mbilgin@sabanciuniv.edu*

Key Words 2-D/MS, yeast two-hybrid, protein localization, proteome microarray, data integration

■ **Abstract** Fueled by ever-growing DNA sequence information, proteomics–the large scale analysis of proteins–has become one of the most important disciplines for characterizing gene function, for building functional linkages between protein molecules, and for providing insight into the mechanisms of biological processes in a high-throughput mode. It is now possible to examine the expression of more than 1000 proteins using mass spectrometry technology coupled with various separation methods. High-throughput yeast two-hybrid approaches and analysis of protein complexes using affinity tag purification have yielded valuable protein-protein interaction maps. Large-scale protein tagging and subcellular localization projects have provided considerable information about protein function. Finally, recent developments in protein microarray technology provide a versatile tool to study protein-protein, protein–nucleic acid, protein-lipid, enzyme-substrate, and protein-drug interactions. Other types of microarrays, though not fully developed, also show great potential in diagnostics, protein profiling, and drug identification and validation. This review discusses high-throughput technologies for proteome analysis and their applications. Also discussed are the approaches used for the integrated analysis of the voluminous sets of data generated by proteome analysis conducted on a global scale.

CONTENTS

INTRODUCTION

With the DNA sequences of more than 90 genomes completed, as well as a draft sequence of the human genome, a major challenge in modern biology is to understand the expression, function, and regulation of the entire set of proteins encoded by an organism—the aims of the new field of proteomics. This information will be invaluable for understanding how complex biological processes occur at a molecular level, how they differ in various cell types, and how they are altered in disease states.

A rapidly emerging set of key technologies is making it possible to identify large numbers of proteins in a mixture or complex, to map their interactions in a cellular context, and to analyze their biological activities (1). Mass spectrometry has evolved into a versatile tool for examining the simultaneous expression of more than 1000 proteins and the identification and mapping of posttranslational modifications (2, 3). High-throughput methods performed in an array format have enabled large-scale projects for the characterization of protein localization, protein-protein interactions, and the biochemical analysis of protein function (4, 5). Finally, the plethora of data generated in the last few years has led to approaches for the integration of diverse data sets that greatly enhance our understanding of both individual protein function and elaborate biological processes (6).

In this review, we discuss recent developments in various technologies for characterizing protein function at the level of the entire proteome of a given organism. Much of this work was initially established in microorganisms such as yeast but is currently being applied to multicellular organisms.

PROTEIN PROFILING

The spectrum of proteins expressed in a cell type provides that cell with its unique identity. Elucidating how the protein complement changes in a cell type during development in response to environmental stimuli and in disease states is crucial for understanding how these processes occur at a molecular level. Recent years have witnessed a revolution in the development of new approaches for

identifying large numbers of proteins expressed in cells and also for globally detecting the differences in levels of proteins in different cell states. In this section, we discuss these newly emerged technologies for profiling the proteins expressed in different cell types.

Two-Dimensional Gels and Mass Spectroscopy

Traditionally, two-dimensional (2-D) gel electrophoresis has been the primary tool for obtaining a global picture of the expression levels of a proteome under various conditions. In this method, proteins are first separated in one direction by isoelectric focusing usually in a tube gel and then in the orthogonal direction by molecular mass using electrophoresis in a slab gel containing sodium dodecyl sulfate (SDS) (7). Using this approach, several thousand protein species can be resolved in a single slab gel. However, 2-D gels are cumbersome to run, have a poor dynamic range, and are biased toward abundant and soluble proteins. Also 2-D gel analysis alone cannot provide the identity of the proteins that have been resolved.

In recent years, protein separation methods coupled with various mass spectrometry (MS) technologies have evolved as the dominant tools in the field of protein identification and protein complex deconvolution (8). The key developments were the invention of the time-of-flight (TOF) MS and relatively nondestructive methods to convert proteins into volatile ions. Two "soft ionization" methods, namely matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI), have made it possible to analyze large biomolecules, such as peptides and proteins (9–11).

In initial studies, protein mixtures were first separated using 2-D gel electrophoresis followed by the excision of protein spots from the gel. In those and more recent studies with other protein separation methods, the next step is digestion using a sequence-specific protease such as trypsin, and then the resulting peptides are analyzed by MS. When MALDI is used, the samples of interest are solidified within an acidified matrix, which absorbs energy in a specific UV range and dissipates the energy thermally. This rapidly transferred energy generates a vaporized plume of matrix and thereby simultaneously ejects the analytes into the gas phase where they acquire charge. A strong electrical field between the MALDI plate and the entrance of the MS tube forces the charged analytes to rapidly reach the entrance at different speeds based on their mass-to-charge (m/z) ratios. A significant advantage of MALDI-TOF is that it is relatively easy to perform protein or peptide identification with moderate throughput (96 samples at a time). MALDI-MS provides a rapid way to identify proteins when a fully decoded genome is available because the deduced masses of the resolved analytes can be compared to those calculated for the predicted products of all of the genes in the genomes of an organism

The ESI method is also widely used to introduce mixtures of biomolecules into the MS instrument. The unique feature of ESI is that at atmospheric pressure it allows the rapid transfer of analytes from the liquid phase to the gas phase (8).

The spray device creates droplets, which once in the MS go through a repetitive process of solvent evaporation until the solvent has disappeared and charged analytes are left in the gas phase. Normally, ESI is coupled with either a triple quadrupole, ion trap, or hybrid TOF MS. Compared with MALDI, ESI has a significant advantage in the ease of coupling to separation techniques such as liquid chromatography (LC) and high-pressure LC (HPLC), allowing high-throughput and on-line analysis of peptide or protein mixtures (12, 13). Typically, a mixture of proteins is first separated by LC followed by tandem MS (MS/MS). In this procedure, a mixture of charged peptides is separated in the first MS according to their *m/z* ratios to create a list of the most intense peptide peaks. In the second MS analysis, the instrument is adjusted so that only a specific *m/z* species is directed into a collision cell to generate "daughter" ions derived from the "parent" species (Figure 1). Using the appropriate collision energy, fragmentation occurs predominantly at the peptide bonds such that a ladder of fragments, each of which differs by the mass of a single amino acid, is generated. The daughter fragments are separated according to their *m/z*, and the sequence of the peptide can then be deduced from the resulting fragments (8, 10). By comparison with predicted sequences in the databases, the identity of the peptide is revealed.

The coupling of liquid chromatography (LC) with MS has had a great impact on small molecule and protein profiling, and has proven to be an important alternative method to 2-D gels (14). Typically, proteins in a complex mixture are separated by ionic or reverse phase column chromatography and subjected to MS analysis. Of the various ionization methods developed for coupling liquid chromatography to MS, including thermospray (15), continuous-flow fast atom bombardment (16), and particle beam (17) techniques, ESI is the most widely used interface technique (18).

LC-MS has been applied to large-scale protein characterization and identification. The Yates group (19) was able to resolve and identify 1484 proteins from yeast in a single experiment. Unlike the 2-D/MS approaches, the authors demonstrated that even low-abundance proteins could be clearly identified, such as certain protein kinases. In addition, 131 of the proteins identified have three or more predicted transmembrane domains, suggesting that this approach was able to readily detect membrane proteins. In addition to its role in protein profiling, LC-MS is perhaps the most powerful technique for the monitoring, characterization, and identification of impurities in pharmaceuticals.

An instrument that combines the benefits of high mass accuracy with highly sensitive detection is the Fourier transform ion cyclotron resonance mass spectrometer (FTICR-MS). FTICR-MS has recently been applied to identify low-abundance compounds or proteins in complex mixtures and to resolve species of closely related *m/z* ratios (20). Coupled with HPLC and ESI, FTICR-MS is able to characterize single compounds (up to 500 Da) from large combinatorial chemistry libraries and to accurately detect the masses of peptides in a complex protein sample in a high-throughput mode. For example, Nawrocki et al. studied the diversity and degeneracy of a small-peptide combinatorial library containing

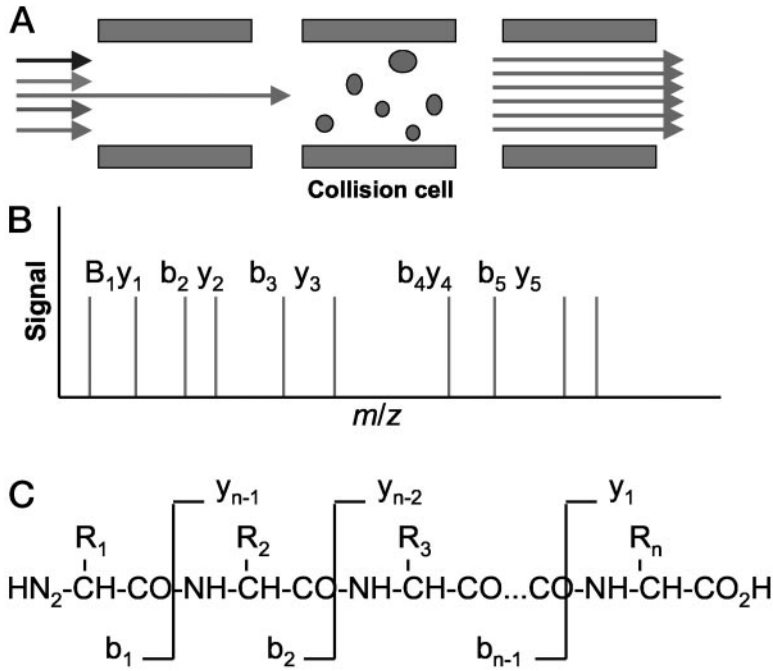


Figure 1 MS/MS analysis of peptide sequences. A protein mixture is first separated by LC followed by ESI ionization to generate fragment patterns (MS/MS spectra). In the first pass, a mixture of charged peptides, indicated as *arrows*, are separated according to their *m/z* ratios to create a list of the most intense peptide peaks. (*A*) The instrument is adjusted so that only a specific *m/z* species (indicated as the *longer arrow*) is directed into a collision cell to generate "daughter" ions derived from the "parent" species. (*B*) The newly generated fragments are separated according to their *m/z* ratio, creating the MS/MS spectrum. Using appropriate collision energy, fragmentation occurs predominantly at the peptide bonds such that a ladder of fragments, each of which differs by the mass of one amino acid, is generated. (*C*) The sequence of the peptide can then be deduced by a ladder-walk. [Adapted from (8, 100).]

up to 10^4 compounds using FTICR-MS (21). Lipton et al. (22) developed a high-throughput and LC-coupled FTICR-MS approach to characterize the proteome of a radiation-resistant bacterium, *Deinococcus radiodurans.* The authors combined global enzymatic digestion of the whole cell lysates, high-resolution LC separation, and analysis by FTICR-MS to resolve 6997 peptides [termed accurate mass tags (AMT)] with high confidence. The 6997 AMTs corresponded to 1910 predicted open reading frames (ORFs), which covered 61% of the *D. radiodurans* proteome. Others have used a similar strategy to characterize proteins in human body fluids (23).

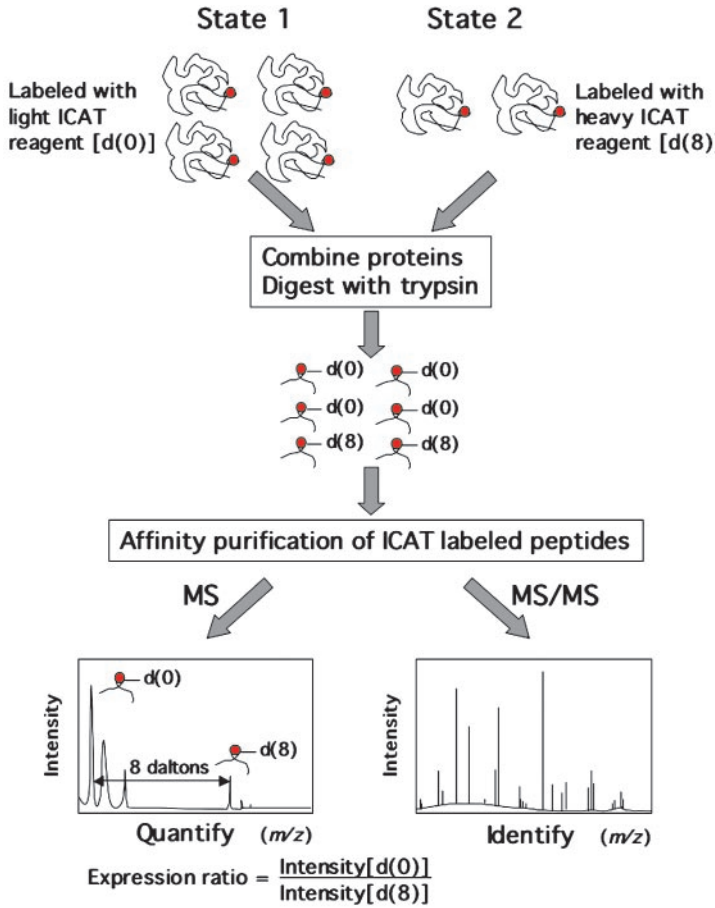


Figure 2 Schematic illustration of ICAT technology. Equal amounts of proteins extracted from two different biological states are separately labeled with heavy [d(8)] and light [d(0)] ICAT reagents. The samples are combined, digested with protease, and separated with multidimensional chromatography, and then analyzed by MS and MS/MS for quantification and identification, respectively. The relative abundances of labeled peptides are determined by comparison of peak intensities between the light and heavy forms of the peptides, which are separated by 8 Da.

Because of the complexity of any given proteome and the separation limits of both 2-D gel electrophoresis and liquid chromatography, only a fraction of that proteome can be analyzed. An alternative approach is to reduce the complexity prior to protein separation and characterization. The Aebersold group (24) designed a pair of isotope-coded affinity tag (ICAT) reagents to differentially label protein samples on their cysteine residues (Figure 2). The ICAT reagent

contains a biotin moiety and a linker chain with either eight deuterium or eight hydrogen atoms. Two samples, each labeled with the ICAT reagent carrying one of the two different isotopes, were mixed and subjected to site-specific protease digestion. The labeled peptides containing Cys can be highly enriched by binding the biotin tags to streptavidin, resulting in a greatly simplified peptide mixture. Characterization of the peptide mixture was carried out by the LC-MS approach as described above. Quantitation of differential protein expression level can be achieved by comparing the areas under the doublet peaks that are separated by eight mass units. The authors demonstrated that they could follow the differential expression of more than 1400 different proteins in yeast. When dealing with the human proteome, Han et al. (25) further simplified the protein mixture by focusing on microsomal proteins that were isolated from cells untreated or treated with a differentiation-inducing stimulus and then labeled the proteins with ICAT reagents. They were able to detect the abundance ratios of 149 proteins in the microsomal fraction of human myeloid leukemia HL-60 cells. Thus, the ICAT method works well for the differential analysis of many proteins in a complex mixture. The obvious limitation of the ICAT labeling approach is that a protein has to contain at least one cysteine residue to be detected.

Antibody Microarrays

Although mass spectroscopy has demonstrated considerable promise for examining simultaneously the expression of large numbers of proteins in a complex mixture, such as a cell lysate, antibody microarrays hold potential promise for the high-throughput profiling of a smaller number of proteins (Figure 3). Briefly, antibodies (or other affinity reagents directed against defined proteins) are spotted onto a surface such as a glass slide; a complex mixture, such as a cell lysate or serum, is passed over the surface to allow the antigens present to bind to their cognate antibodies (or targeted reagents). The bound antigen is detected either by using lysates containing fluorescently tagged or radioactively labeled proteins, or by using a secondary antibodies against each antigen of interest. Low-density antibody arrays have been constructed that measure the levels of several proteins in blood (26) and sera (27, 28). In high-density arrays constructed recently, Sreekumar et al. (29) spotted 146 distinct antibodies on glass to monitor the changes in quantity of a number of antigens expressed in LoVo colon carcinoma cells. They found that radiation treatment of the cells up-regulated the levels of many interesting proteins, including p53, DNA fragmentation factors 40 and 45, and tumor necrosis factor–related ligand, and down-regulated the levels of other proteins.

The biggest problem with antibody arrays is antibody specificity. Haab and colleagues (30) analyzed the reactivity of 115 antibodies with their respective antigens. Protein microarrays containing either immobilized antigen or immobilized antibody were probed with antibodies or antigens, respectively. Only 30% of the antibody/antigen pairs showed the linear relationships expected for specific

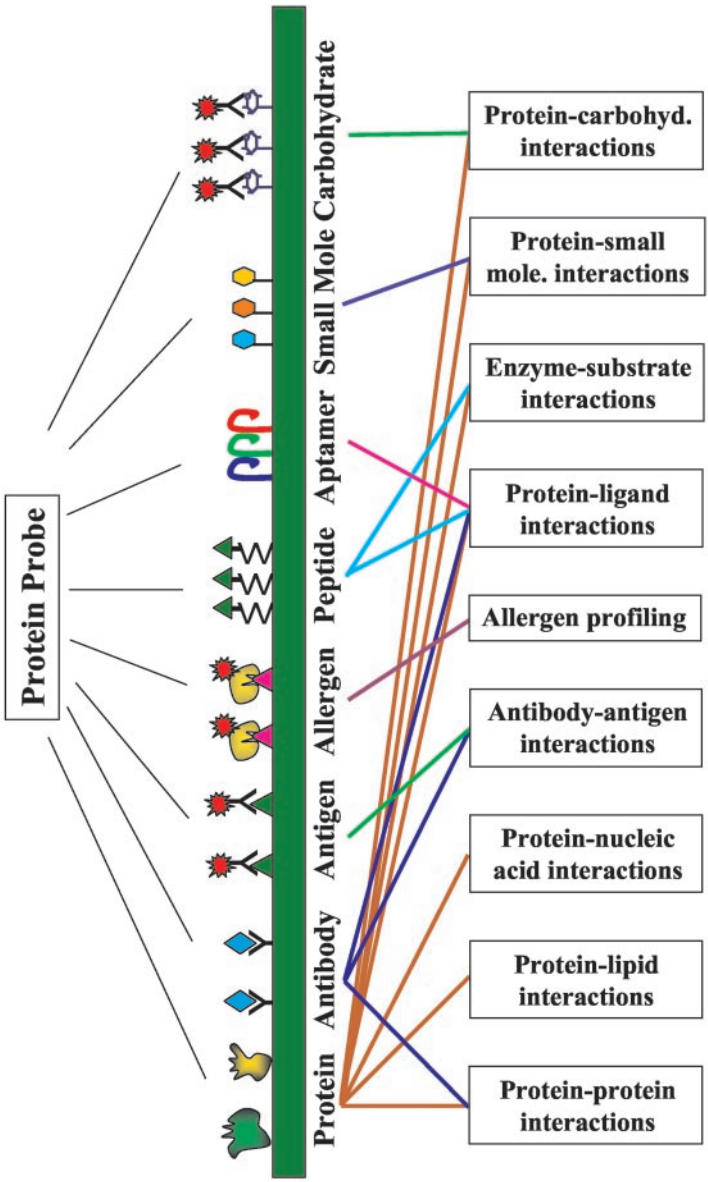


Figure 3 Protein microarrays and their applications. Ligands, such as proteins, peptides, antibodies, antigens, allergens, and small molecules, are immobilized in high density on modified surfaces to form functional and analytical protein microarrays. These protein microarrays can also be used for various kinds of biochemical analysis.

binding, indicating that most antibodies are not suitable for quantitative detection. Nonetheless, for those antibodies that are specific, quantitative detection of antigen abundance in a complex mixture could be determined. In a follow-up report, they showed that antibody microarrays could be applied to obtain serum profiles (31).

To profile the biological activities of a living cell or tissue, however, the capture molecules immobilized on the surface are not restricted to antibodies or antibody mimics. They can be short peptides, aptamers (DNA, RNA, or protein molecules selected for their ability to bind nucleic acid, proteins, small organic compounds, or cells), polysaccharides, allergens, or synthetic small molecules (2–4, 41). To profile antibodies in human sera, Robinson et al. (32) fabricated microarrays of autoantigen by arraying hundreds of such autoantigens, including proteins, peptides, and other biomolecules. These arrays were incubated with sera from patients to study the specificity and pathogenesis of autoantibody responses. In a similar approach, Hiller et al. (33) robotically arrayed 94 purified allergens on glass to monitor the IgE activity profiles of allergy patients. Using serum samples of minute amounts, they could profile an allergic patient's IgE reactivity in a single measurement. By comparing the reactivity to controls, specific IgE profiles could then be related to a large number of disease-causing allergens. Some of the findings from the allergen microarrays were further validated by classical skin tests. As another example, Joos and colleagues (34) used 18 diagnostic markers for autoimmune diseases to form a microarray of autoantigens and used it to monitor antigen-antibody interactions. Thus, protein microarrays can be used to profile the presence of a limited number of proteins and to analyze the antibody reactivity profile of individuals.

PROTEIN POSTTRANSLATIONAL MODIFICATION

Covalent modifications to protein structures, which occur either co- or posttranslationally, play a pivotal role in regulating protein activity. Identification of the type of modification and its location often provide crucial information for understanding the function or regulation of a given protein in biological pathways. So far, more than 200 different modifications have been reported, many of which are known to control signaling pathways and cellular processes (35). Many strategies have been developed to analyze protein modifications, however, most of them focus on only a specific type of modification, such as protein phosphorylation. For example, one strategy to identify phosphoproteins is to enrich the phosphorylated peptides using either immunoprecipitation with phosphopeptide-specific antibodies or by metal-chelate affinity chromatography (36). The latter uses resins with chelated trivalent metal ions such as Fe(III) and Ge(III) to bind the phosphopeptides or phosphoproteins (37–40). The enriched proteins are then subjected to trypsin digestion, and the resulting fragments are identified using MS techniques. This approach can provide important information on the sites of

phosphorylation in proteins. This method was recently applied on a large scale to proteins of a yeast lysate (42). Phosphopeptides were purified using metal-chelated columns and subjected to MS/MS; 383 sites of phosphorylation were identified on 216 peptides.

To identify multiple types of modifications in a single experiment, MacCoss et al. (43) also employed a so-called shotgun MS approach. This approach used multidimensional liquid chromatography (LC/LC), tandem mass spectrometry (MS/MS), and database-searching algorithms. In brief, the protein mixture was first digested with one site-specific and two nonspecific proteases; the resulting peptides were separated by multidimensional liquid chromatography; and finally their identities were revealed by MS/MS. The digestion with multiple proteases produced overlapping peptides of a given protein thereby providing thorough coverage of the protein and increasing the chance of pinpointing a modification on a specific amino acid residue. Using this strategy to analyze protein samples from the human lens tissue, the researchers identified 270 proteins. Further analysis of a family of 11 lens crystallins proteins revealed a total of 73 modifications including phosphorylation, methylation, oxidation, and acetylation in these proteins. Thus, although lens tissue is not extremely complex, this method has demonstrated its great potential for revealing a more comprehensive picture of protein modifications in a complex sample.

In summary, MS has played an important role in identifying posttranslational protein modifications. Protein microarrays, described below, have played an important role in identifying the enzymes responsible for many modifications and the substrate specificity of the modifying enzymes.

PROTEIN LOCALIZATION

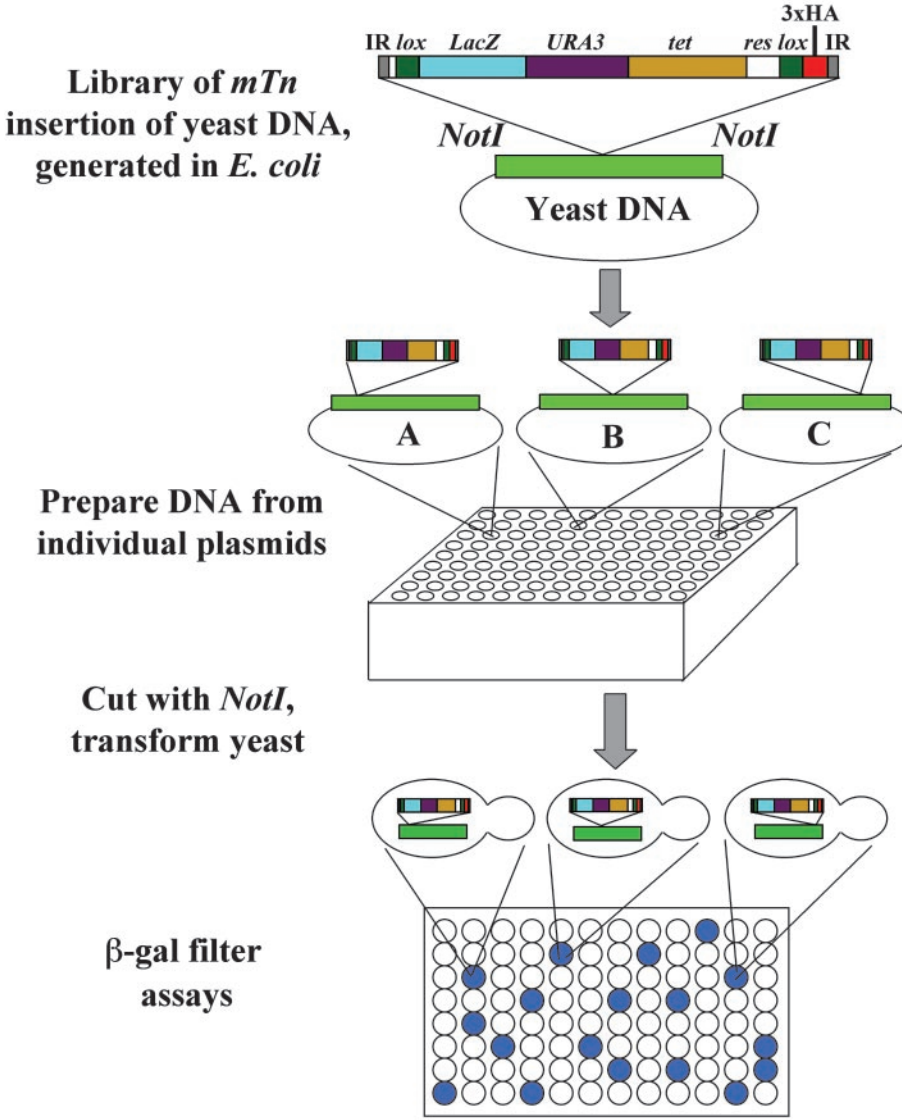
Protein localization data provide valuable information in elucidating eukaryotic protein function. To monitor the relative levels of protein expression and obtain a snapshot of protein localization in yeast, our laboratory has developed a random transposon tagging strategy to generate a library of expressed ORFs fused to coding sequences for an epitope tag (Figure 4) (44). Briefly, a transposon containing an *Escherichia coli lacZ* gene lacking its ATG translation initiation codon and promoter lies adjacent to a *lox* site at one end of the transposon; a coding sequence for three copies of a hemagluttinin epitope tag lies adjacent to another *lox* site at the other end of the transposon (Figure 4) (44). The transposition is mediated in *E. coli* and the mutagenized DNA is then shuttled into yeast. When the *lacZ* cassette is inserted in-frame in an ORF in yeast, its transcription and translation can be detected. A large portion of the inserted cassette can be looped out in yeast via recombination between the *lox* sites (mediated by the phage *Cre* recombinase), leaving behind the ORF with a short in-frame epitope tag. Using high-throughput immunostaining, the subcellular locations of 2744 yeast proteins have been determined over the years (44–47).

Using this approach as well as an approach in which 2000 ORFs were fused directly to an epitope tag, Kumar et al. (48) localized approximately 55% of the proteome and described the first "localizome"—the subcellular localization of most proteins of an organism. They showed that 47% of yeast proteins were cytoplasmic, 13% mitochondrial, 13% exocytic, and 27% nuclear/nucleolar. A subset of nuclear proteins was further analyzed by using surface-spread preparations of meiotic chromosomes, and 38% were found associated with chromosomal DNA. The major shortcoming of the transposon approach is that the library is not complete yet—it contains roughly 60% of the 6300 annotated genes (48). Furthermore, because the tagged proteins are expressed from their native promoters, the localization information is biased toward abundant proteins.

Because the transposon tagging approach visualizes proteins via indirect immunostaining on fixed cells, the dynamics of protein localization and transportation cannot be analyzed. To develop a real-time detection method, Ding et al. (49) attempted to tag *Schizosaccharomyces pombe* proteins using green fluorescent protein (GFP) in a genomic library. The tagged plasmid library was transformed into the *S. pombe* cells, and 6954 transformants exhibiting GFP fluorescence were obtained, 728 of which showed distinct localization patterns. By recovering plasmids from these strains, 250 unique genes were confirmed to have GFP tags in-frame. For mammalian cells, systematic GFP tagging of complementary DNA (cDNA) clones has been accomplished. Simpson et al. (50) cloned 107 novel human cDNAs to produce both N- and C-terminal fusions to GFP; ~100 proteins showed a clear pattern of intracellular localization. On the basis of sequence homology, they were able to predict the locations of 47% of these novel cDNAs; these predictions were in good agreement with the experimental results. Although considerable effort is needed to cover the entire proteome of any organism, these studies indicate that it is feasible to analyze protein localization globally in microbes and multicellular organisms.

INTERACTION PROTEOMICS AND PATHWAY BUILDING

It is widely acknowledged that proteins rarely act as single isolated species when performing their functions in vivo (1). The analysis of proteins with known functions indicates that proteins involved in the same cellular processes often interact with each other (6). Following this observation, one valuable approach for elucidating the function of an unknown protein is to identify other proteins with which it interacts, some of which may have known activities. On a large scale, mapping protein-protein interactions has not only provided insight into protein function but facilitated the modeling of functional pathways to elucidate the molecular mechanisms of cellular processes.


Library of mTn insertion of yeast DNA, generated in E. coli
3xHA
IR lox
LacZ
URA3
tet
res lox
IR
NotI
NotI
Yeast DNA
A
B
C
Prepare DNA from individual plasmids
Cut with NotI, transform yeast
β-gal filter assays

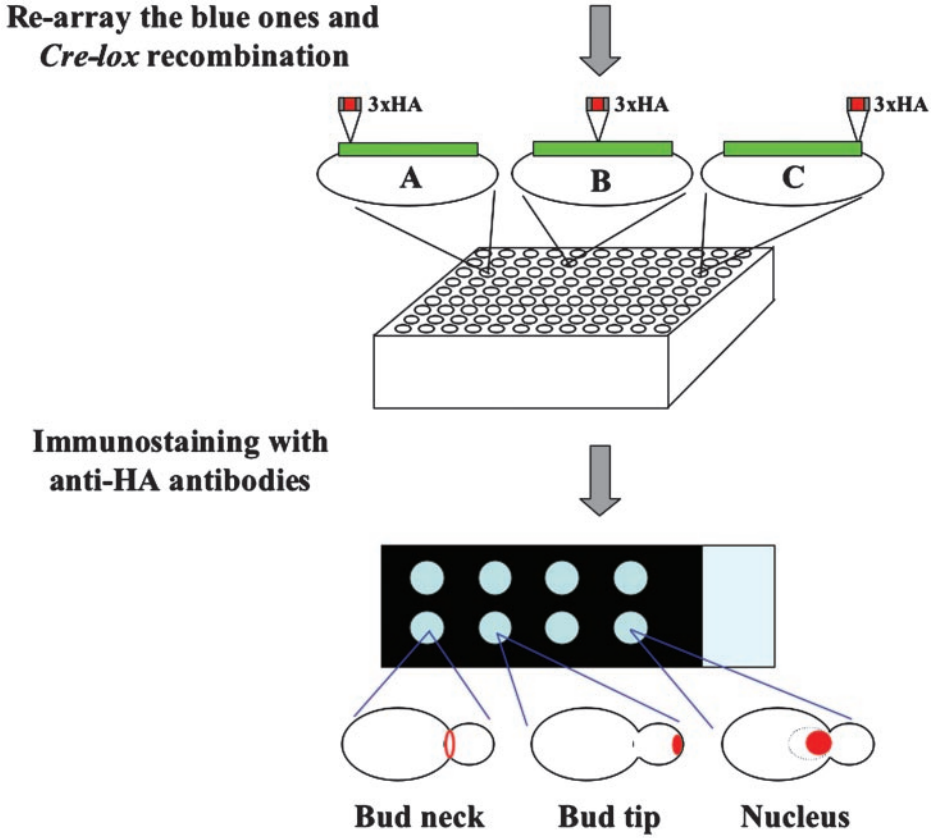


Figure 4 *Continued.*

←

Figure 4 Transposon tagging strategy for protein localization. To monitor the relative levels of protein expression, a transposon tagging strategy using a mini-transposon (mTn) was used to generate a library of random insertions in yeast DNA in an *E. coli* plasmid. The mTn contains a promoterless and 5′-truncated *lacZ* gene near one end of the transposon and coding sequence for three copies of an epitope tag at the other end. The mutagenized yeast DNA is prepared in a 96 well format, cleaved with the restriction endonuclease NotI to free the yeast DNA for the plasmid, and individually transformed into a diploid yeast strain. The insertion allele replaces one of the chromosomal copies. When the *lacZ* gene is inserted in-frame within a yeast ORF, its transcription and translation can be visualized in yeast cells using assays for β-galactosidase; such clonies will turn blue. A large portion of the inserted cassette is then excised at the *lox* sites via *Cre*-mediated recombination, leaving behind the ORF with a short in-frame epitope tag coding sequence. The subcellular location of the protein is then determined by indirect immunofluorescence. [Adapted from (44).]

Two-Hybrid Studies

One of the best-established in vivo approaches to map protein-protein interactions is the yeast two-hybrid method (51). In this technique, a component of interest (bait) is typically fused to a DNA-binding domain. Other proteins (preys), which are fused to a transcription-activating domain, are screened for physical interactions with the bait protein using the activation of a transcription reporter construct as the detection method (read out). This approach is scalable and can be fully automated.

Recently, systematic two-hybrid projects have been undertaken to analyze protein-protein interactions at a global level in budding yeast, the nematode, *Caenorhabditis elegans*, and human gastric bacterial pathogen, *Helicobacter pylori* (51–55). More than 4500 interactions have been identified in yeast, unraveling a host of unexpected and interesting interactions (51, 52). Likewise, Walhout et al. (53) and Boulton et al. (54) were able to map protein-protein interaction networks involved in *C. elegans* vulval development and DNA-damaging response components, respectively. Finally, Rain et al. (55) have recently built a large-scale protein-protein interaction map of the human gastric bacterial pathogen *Helicobacter pylori*, whose genome encodes 1590 predicted coding sequences (56). A total of 261 bait proteins were used to identify 1280 interactions, resulting in a protein interaction map covering much of the proteome. In general, the two-hybrid approach is especially powerful when analyzing smaller genomes because most of genes are easily known from the genome sequence and because the method has the potential to uncover all the possible interaction combinations quickly and relatively easily.

However, there are several drawbacks to the two-hybrid studies. First, they are not comprehensive. Based on the observations that most well-characterized proteins interact with 5–7 other proteins, estimates of the number of interactions expected in yeast are ~30,000–40,000 (59), significantly higher than the 4500 identified thus far. This is likely due to the fact that most of the studies performed thus far have been carried out in pools of yeast strains; this probably fails to detect all possible individual interactions. Direct tests of all possible individual interactions have not yet been performed, and thus the screens are not saturated.

A second disadvantage of two-hybrid methods is that they identify a large number of false positives, presumably through spurious interactions between proteins that do not normally occur in vivo (6). Approximately 50% of the interactions are estimated to be false positives.

A third drawback of the conventional two-hybrid is that interaction occurs in the nucleus and uses a transcriptional readout. Consequently, the interaction of many membrane proteins and transcription factors cannot be measured. To circumvent this problem, other two-hybrid methods have been developed. One promising method is the "split ubiquitin system," which appears to be especially useful for detecting interactions among membrane proteins (57). The split ubiquitin system involves bringing together two halfs of ubiquitin; the N

terminus ubiquitin fragment is fused to one protein (e.g., the bait) and the C-terminal fragment, which is fused to a transcription factor, is also fused to a potential interaction protein. When the proteins interact, the ubiquitin fragments interact, and the transcription factor is released by cleavage from the C-terminal fragment and activates a reporter construct. This system has been used successfully to detect interaction among a variety of test proteins, including yeast oligosaccharyl transferases (57), sucrose transporters (57a), proteins involved in viral replication in *Arabidopsis* (57b), and transmembrane proteins normally present in the endoplasmic reticulum (57c, 57d).

In spite of the drawbacks of two-hybrid studies, the data have proven to be exteremely valuable, and even more so when integrated with protein-protein interaction data from other sources. Schwikowski et al. (58) compiled a list of 2709 published yeast protein interactions (including two hybrid) and found that 1548 of them could be mapped in a single network containing 2358 interactions. Based on this network, a putative function category can often be assigned to many novel proteins. Thus, although the data are incomplete, they have enormous utility.

Affinity Tagging and Mass Spectroscopy

Affinity purification has long been used to identify protein-protein interactions or validate their existance in cell extracts on a one-protein-at-a-time basis. However, the throughput has been dramatically improved by two recent reports on the systematic characterization of protein complexes in yeast (60, 61). In one study, endogenous protein coding genes were fused to the coding sequences of a tandem affinity purification (TAP) tag. The tagged proteins were then purified under gentle conditions along with their associated partners, and subsequently separated by gel electrophoresis. The copurifying proteins were identified by MS. Starting with a set of more than 500 chromosomal tagged genes, Gavin et al. (60) were able to purify and subsequently resolve 232 protein complexes encompassing 1440 distinct proteins in yeast. In another study, Ho et al. (61) constructed 725 inducible FLAG epitope–tagged fusions, which they overexpressed and purified along with their associated proteins. This study identified more than 3000 protein-protein interactions involving 1578 individual yeast proteins. The differences in the results likely reflect the fact that one study used endogenous protein levels whereas the latter used overexpressed proteins. The overexpressed proteins are likely to interact with more proteins, but they may also associate with proteins that they do not normally bind and thereby yield false positives. By combining both data sets, much greater accuracy is achieved (see below).

ANALYSIS OF PROTEIN BIOCHEMICAL ACTIVITIES

One of the most direct methods for elucidating protein function and regulation is to determine its biochemical activity. The past few years have brought several powerful approaches for the large-scale analysis of biochemical activities in

yeast (summarized in Table 1). These approaches involve overexpressing the protein-coding genes from the organism of interest and then screening the expressed proteins for biochemical activities of interest. Much of this work is possible because of the development of efficient methods to clone sets of open reading frames (ORFs) into plasmids for expression in an appropriate host. As a result, large-scale biochemical analysis of proteins has been initiated.

Analysis of Biochemical Activities Using Pooling Strategies

In a pioneering work, the Phizicky group (62) applied a recombination-based cloning strategy to fuse >85% of the yeast ORFs to glutathione-*S*-transferase (GST) on a plasmid under the control of an inducible promoter. Yeast strains containing these fusion plasmids were grouped into 64 pools, each containing 96 clones, and induced to produce fusion proteins. GST fusion proteins from each pool were purified and screened for biochemical activities. Individual strains from positive pools were screened again to identify the specific clone expressing the activity of interest. A number of ORFs carrying new biochemical activities were identified, including tRNA ligase, 2′-phosphotransferase, phosphodiesterase, and cytochrome *c* methyltransferase. Compared to conventional approaches, this method allows a rapid and sensitive assignment of catalytic function to ORFs and is generally applicable for detection of virtually any type of activity. The disadvantages of this approach are that it requires several steps and that prevalent enzymes or activities, such as kinases and phosphatases, must usually be screened in smaller pools.

Functional Protein Microarrays for the Analysis of Biochemical Activities

A more direct approach for the global identification of biochemical activities of interest is using functional protein microarrays. In this technique, sets of proteins of interest or an entire proteome is overexpressed, purified, distributed in an addressable array format, and then assayed (Figure 3). These arrays can be used not only to screen for biochemical activities of interest but also to examine posttranslational modifications and to detect binding to small molecules, proteins, antibodies, and drugs. The latter feature has potentially powerful applications in the discovery and development of pharmaceuticals.

There are several types of functional protein array formats [reviewed in (63)]: nanowells, which are miniature wells; solid surface supports such as glass slides; and thick absorbent surfaces, such as hydrogels. The latter have not been used extensively and thus are not discussed here.

Nanowell arrays typically contain wells 1 mm or less in diameter; they can be made of a plastic such as polydimethylsiloxane (PDMS) by using a mold; alternatively, wells can be etched in glass. The wells compartmentalize samples and reduce evaporation. Using this format, Zhu et al. (64) analyzed the kinase-substrate specificity of almost all (119 of 122) yeast kinases using 17 different

TABLE 1 Comparison of different technologies for interaction proteomics

Approach	Application	Advantage	Disadvantage
Yeast two-hybrid	Protein-protein interactions, protein-DNA interactions	High-throughput and systematic to reveal protein interactions	No control over interaction condition; interactions are usually in the nucleus
Affinity tagging/MS	Dissecting protein complexes	In vivo interactions that involve multiple partners	May miss transient or weak interactions, hard to identify false positives
Antibody array	Protein profiling, protein detection, clinical diagnostics	Very sensitive and low sample consumption, great potential in biomarker and drug development	Highly restricted by the quantity and quality of available antibodies; semiquantitative protein detection
Functional protein array	Diverse, e.g., protein-protein, protein-lipid, protein–small molecule, enzyme-substrate interactions as well as drug discovery and posttranslational modifications	Great potentials for analyzing biochemical activities of proteins and high-throughput drug and drug target screening	In vitro assays
Peptide array	Enzyme-substrate interaction and drug discovery	Sensitive and straightforward way to identify epitopes	Expensive to fabricate; in vitro assays
Carbohydrate array	Carbohydrate-mediated molecular recognition and anti-infection response	A new and sensitive way to study carbohydrate-mediated molecular events	In vitro arrays; tough to acquire carbohydrate molecules in pure forms
Small molecule array	Protein–small molecule interaction, drug discovery, enzyme specificity profiling	Minimum small molecule consumption and high sensitivity	In vitro assays; necessary to improve throughput to cover 10^6 molecules in a normal combinatorial chemistry library

substrates. The substrates were first covalently immobilized on the surface of individual nanowells, and individual protein kinases in kinase buffer with [^{33}P]ATP were incubated with the substrates. After washing away the kinases and unincorporated ATP, the nanowell chips were analyzed for phosphorylated substrates using a Phosphoimager (Molecular Dynamics, Inc.). Not only were known kinase-substrate interactions identified but many novel activities were revealed. These studies showed that approximately one fourth of yeast protein kinases are capable of phosphorylating tyrosine residues on an artificial substrate (poly Glutamine-Tyrosine), even though by sequence they are all members of the Ser-Thr family of protein kinases. Thus, many kinases are capable of phosphorylating tyrosine in vivo. Thus far, at least two of the in vitro phosphorylating enzymes have been shown to phosphorylate their substrates on tyrosine in vivo, suggesting that many of them are also tyrosine kinases in vivo (64a; M. Snyder, unpublished).

The more common approach for functional protein microarrays is to use glass microscope slides, as these are compatible with many commercial scanners. Proteins are attached to the surface using either direct covalent methods, linkers, or affinity tags (63, 65, 66). The bound proteins are then assayed for binding or enzymatic activities. MacBeath & Schreiber (65) used this format to demonstrate that they could detect antibody-antigen interactions, protein kinase activities, and protein interaction with small molecules using several test systems.

The major limitation in functional protein microarrays has been the preparation of proteins to analyze. This requires high-quality and comprehenisve expression libraries and methods that yield a large number of functional active proteins. This problem was recently surmounted. For example, as mentioned above, it was possible to produce in functional forms nearly all yeast protein kinases. More recently, the first eukaryotic proteome chip was prepared. This microarray is composed of $>$5800 individually cloned, overexpressed, and purified proteins (66). A high-throughput protein purification protocol was developed to purify 80% of the yeast proteome as full-length proteins (Figure 5). In initial studies, the proteome chips were used to identify protein-protein interactions by screening for binding targets of calmodulin (Figure 6). Calmodulin is a highly conserved protein that regulates signaling pathways and other cellular processes. The proteome chip was probed with biotinylated calmodulin and washed stringently; the bound calmodulin was detected by binding of streptavidin, which was labeled with a cyanine dye, Cy3. The identities of the calmodulin-interacting proteins on the proteome chips were deconvoluted using a laser scanner and the known addresses on the array. In addition to six known targets, 33 potential new binding partners of calmodulin were identified. Sequence comparison revealed that 14 of the 39 calmodulin-binding proteins shared a common motif, which is similar to a previously known calmodulin-binding motif, called the IQ sequence.

To explore the possibility of using proteome chips to identify the binding targets of secondary messengers, the chips were probed with phosphatidylinosi-

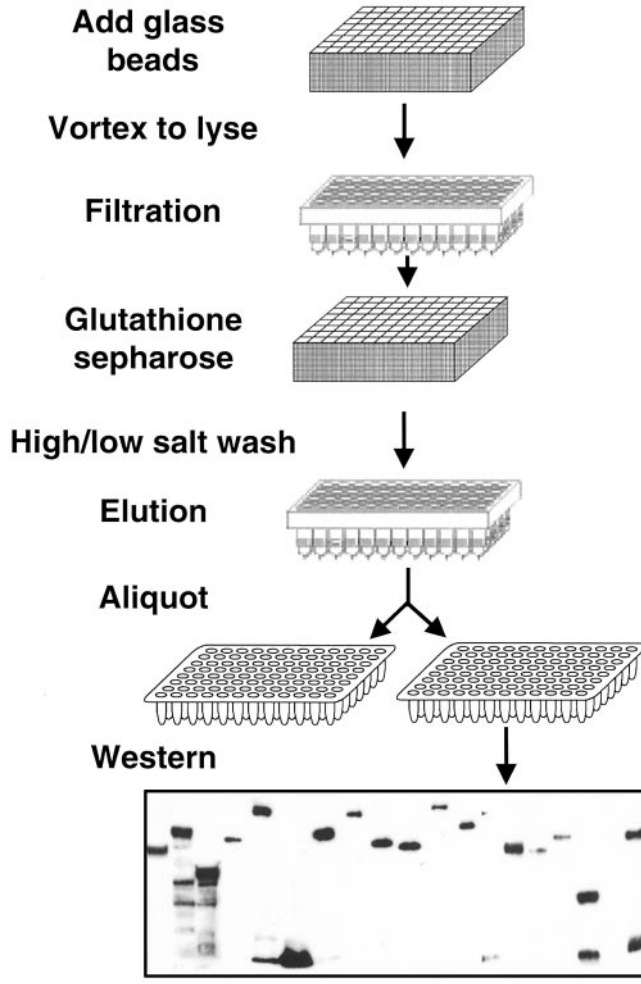


Figure 5 High-throughput procedure for purification of protein from yeast cells. Yeast strains each carrying a different fusion protein expression vector were grown and purified in a 96-well format. Western analysis revealed that 80% of the purified proteins were full length.

tides incorporated in liposomes. These liposomes should represent the most relevant physiological binding environment and also contain 1% biotinylated lipid as a detection tag. A total of 150 proteins were identified as binding either phosphotidylcholine or phosphotidylcholine vesicles containing one of five different phosphatidylinositides. Of the 98 annotated proteins, 45 are membrane associated and 8 more are involved in lipid metabolism. One interesting result from this study is that many protein kinases bind liposomes containing specific phosphatidylinositides.

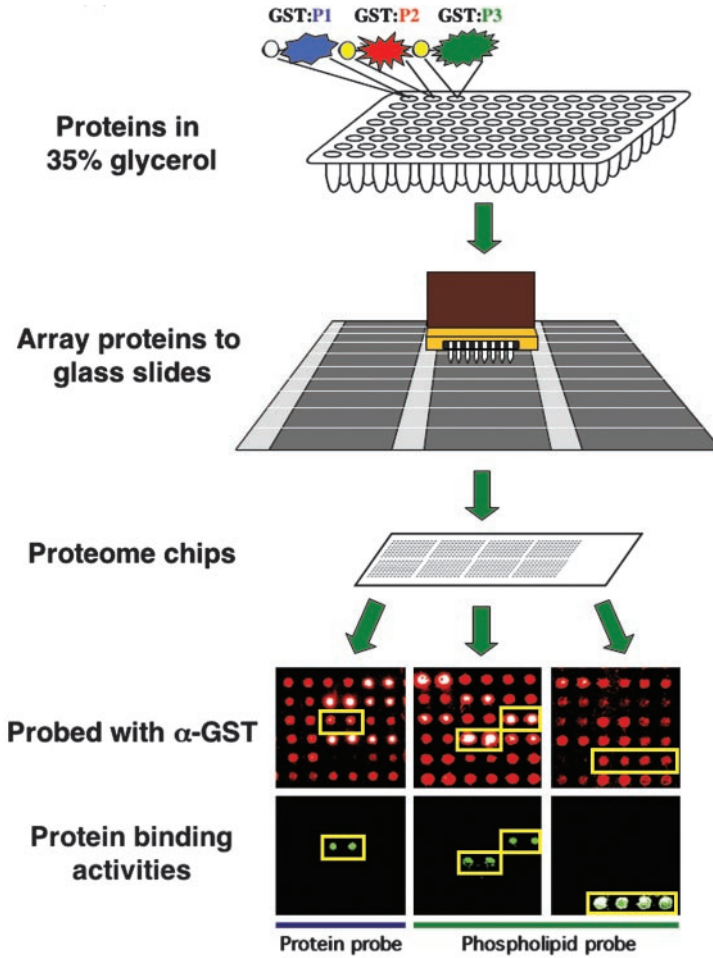


Figure 6 Protein activity analysis using the yeast proteome microarrays. GST-purified proteins were arrayed in duplicate onto nickel-coated glass slides at high density using a commercially available microarrayer. These chips were subsequently probed with biotinylated calmodulin and several phospholipid-containing liposomes. The binding activities were then detected using fluorescently labeled streptavidins. Positive signals are indicated in green and encapsulated in a yellow box. For comparison, the corresponding section of an array from the same printing was probed with anti-GST antibodies. All of the proteins present on the array react with this antibody (positives shown in red).

The proteome chips can also be used to study other binding activities, such as protein–nucleic acid, protein–small molecule, and protein-drug interactions. Although these are all in vitro binding assays, the advantage is that the

experimental conditions can be well controlled. For example, different cofactors or inhibitors can be included in the binding assays to adjust the strigency of the binding activities. Another advantage is that these highly parallel assays are not biased toward abundant proteins. In addition, with proper detection methods, proteome chips can be used to identify the downstream targets of various enzymes such as protein kinases, phosphatases, methyl transferases, and proteases (4, 63). Finally, protein microarrays can be used to identify in vivo posttranslational modifications by probing for specific modifications, such as glycosylation or phosphorylation, using lectins or antibodies, respectively (M. Bilgin, H. Zhu, & M. Snyder, unpublished observations). This potentially allows the identification of all of the proteins that carry those modifications in a single experiment.

Protein microarrays also have the potential advantage of permitting analysis of the kinetics of protein-protein interactions via real-time detection methods. Surface plasmon resonance (SPR) has matured as a versatile detection tool to analyze the kinetics of protein-ligand interactions over a wide range of molecular weights, affinities, and binding rates (67–69). Although the commercially available SPR chips are not yet high throughput, Myszka & Rich (70) recently described a sensor surface with 64 individual immobilization sites in a single flow cell. Alternatively, Sapsford et al. (71) used a planar waveguide as the detection method to develop an antibody array biosensor and studied the kinetics of antigen-antibody interactions. More importantly, they demonstrated that significant signal intensity could be achieved from spots as small as 200 μm in diameter. However, to date, most of these alternative approaches have been applied successfully on a small scale using only a handful of samples. Some of these techniques may prove robust enough to be applied in a fully automated fashion.

A modified version of protein microarrays is peptide microarrays, which can be used as substrates for enzymatic activites and as potential ligands when probing with proteins or other molecules. In one recent example, Houseman et al. (72) characterized the substrate specificity of a nonreceptor tyrosine kinase, the *c-Src* product, using immobilized 9-mer peptide substrates arrayed at high density on a gold-coated glass surface. They characterized the kinase–peptide substrate interactions using SPR and the phosphorylation events using ATP derivatives, fluorescent labeling, and phosphoimaging methods. They could also quantitatively evaluate the effect of three known inhibitors of the kinase. Although the study was still primitive, the authors demonstrated the potential of coupling peptide chips with various detection methods to quantitatively study the dynamics of enzyme-substrate interactions, with obvious applications in drug discovery. In another example, Lizcano et al. (73) studied the molecular basis for the substrate specificity of a human protein kinase Nek6 using peptide microarrays harboring $>$1000 peptide species. They observed that a Leu located three residues N-terminal to the phosphorylation site in a substrate was important for the phosphorylation of this enzyme. However, genetic analysis indicated that selectivity for a Leu-X-X-Ser/Thr site might not occur in cells, which raises the

challenge of using consensus sequences determined in vitro as a means to identify physiologically relevant substrates for Nek6 or any other kinase.

Finally, in addition to protein and peptide arrays, carbohydrate arrays have been introduced recently. Wang et al. (74) used carbohydrate-based microarrays to analyze the different types of anticarbohydrate antibodies in human and mammalian sera. An array of 48 different carbohydrate macromolecules was prepared and probed with sera from 20 normal individuals. A variety of different reactivities were observed. Interestingly, many of the carbohydrates that react with the sera are normally present in pathogen microbes, suggesting that the individuals may have acquired these antibodies during a microbial infection. Carbohydrate arrays can also be used to profile other types of binding activities, as demonstrated by Houseman & Mrksich (75). The authors conjugated and self-assembled a monolayer of 10 monosaccharides on a glass surface. The arrays were then used to profile the binding specificities to several lectins. Both SPR and competition experiments demonstrated that the carbohydrate-protein interactions were highly specific. Although the density of the chips was not high, it is scalable for large-scale and high-throughput analysis in the future.

In summary, protein microarrays can be used to globally analyze the activities of proteins including their binding to proteins, nucleic acids, lipids, carbohydrates, and small molecules. Because of its miniaturized and versatile nature, microarray technology is expected to flourish in the field of proteomics.

PROTEIN ENGINEERING

Another way to probe protein function is through protein engineering. The Shokat group (76) developed an elegant approach for analyzing kinase activities in a variety of organisms. Protein kinases were engineered to have a “hole” in the ATP-binding pocket. A modified ATP analogue containing an extra side chain (i.e., a “bump”) is used to selectively inhibit that kinase in vivo. This approach was used to analyze the role of three yeast protein kinases, Cdc28p, Cla4p, and Pho85p. The inhibitor was added to cells containing the engineered protein kinase, and the resulting phenotype examined (76–78). In each case, in addition to their known function, a new role of each affected protein kinase could be deduced.

A modified version of this approach has been used to identify substrates of protein kinases in vitro (79). In this case, a different ATP analogue that can be used only by the engineered kinases is employed. The ATP analogue is added to a cell extract containing the engineered protein kinase, and the resulting phosphorylated proteins are likely to be the substrates of the engineered kinase. This technology can theoretically be applied to analyze other classes of enzymes for which detailed structural information about active site geometry and substrate recognition determinants are available.

PROTEOMICS AND DRUG DISCOVERY

In recent years, a number of approaches have been developed to identify molecules that bind or inhibit protein function. Molecules that bind a protein of interest can be used to develop diagnostic tests for that protein, or they can be used to probe that protein's function. Molecules that inhibit a protein of interest have the potential to be used directly for therapeutic applications. There are two basic types of protein binding agents: macromolecules such as proteins or nucleic acids and small molecules.

Macromolecular Inhibitors

Antibodies, both monoclonal and polyclonal have been used for many years as affinity reagents to probe and inhibit protein function. Recently, alternative methods, such as phage antibody-display, ribosome display, and mRNA display, have been developed to expedite the process of drug discovery (2, 80). All these approaches involve the construction of large repertoires of folded domains with potential binding activity, which are then selected by multiple rounds of affinity purification. The binding affinity of the resulting candidate clones can be further improved using subsequent mutagenesis and selection strategies.

Oligonucleotides can also be selected to bind proteins. Using a protocol for in vitro evolution, called systematic evolution of ligands by exponential enrichment (SELEX), nucleic acids that bind a wide variety of molecules have been selected. Using this method, many powerful antagonists of proteins have been found with K_d (equilibrium dissociation constant) values in the range of 1 pM to 1 nM. For example, Biesecker et al. (81) used the SELEX procedure to identify specific aptamer inhibitors of the human complement C5 component. In an initial round of selection, seven aptamers were isolated; these formed a closely related family based on sequence homology with a K_d values of 20–40 nM. The binding affinity was improved by a second round of SELEX, which produced an aptamer with a K_d of 2–5 nM.

Small Molecules

For several decades, natural and synthetic small molecules have provided powerful and invaluable means to dissect protein functions and regulatory mechanisms (82). Overwhelmed by the ever-growing sequence information, a more efficient and systematic approach to identify these small molecules is needed. Recently, several groups have devised methods for identifying small molecules using a microarray format.

MacBeath et al. (83) immobilized small organic compounds from a combinatorial chemistry library to form a high-density small molecule microarray. These microarrays were probed with fluorescently labeled target proteins to identify new ligands for these proteins. Similarly, Winssinger et al. (84) constructed a library of small molecules by tethering them to a peptide nucleic acid

(PNA) tag. These PNA tags provide the address for the structure of the corresponding small molecules as well as immobilize them at specific sites on the chip. The study further proved that the immobilized small molecule could withstand stringent washing conditions. As a test case, these arrays were used to identify a small molecule inhibitor of a capase.

As a third example, Kuruvilla et al. (85) constructed a small molecule microarray containing a collection of 3780 structurally complex 1,3-dioxane compounds to dissect the function of a yeast protein, Ure2. Ure2 is a central regulator of the nitrogen metabolic pathway. The library of small molecules was synthesized with a technology platform based on one bead–one stock solution and parsed out to form the small molecule microarray (86, 87). The microarray then probed with fluorescently labeled Ure2. One compound (uretupamine) was identified as specifically inhibiting Ure2 function in a subsequent cellular reporter assay. An analysis of gene expression profiles determined that uretupamine inhibits a particular function of Ure2 without affecting all of the functions of the protein globally. This approach uses a structurally complex small molecule library that is unbiased toward any particular protein targets (82).

The microarray format used in these approaches allows high throughput and highly parallel screening with minimum consumption of small molecules from a combinatorial chemistry library, and yet the signal-to-noise ratio is still high. This strategy is therefore expected to generate several tailored probes for every protein of interest in an entire proteome, which may greatly facilitate the development of pharmaceutical agents.

INTEGRATION OF DIVERSE DATA SETS

High-throughput methods and proteomics projects have exploded during the past few years. These projects have generated overwhelming amounts of data that help identify and characterize components of biological pathways as well as elucidate the response of these pathways (5). However, one important conclusion is that no single set of any high-throughput data is definitive. Instead, integration of multiple sets of data and verification using alternative methods is always required before drawing any firm conclusion. For example, Jansen et al. (88) investigated the relationship of protein-protein interactions with the expression level of mRNAs encoding the components in clearly defined protein complexes. The mRNA levels of the subunits in the same protein complexes showed significant coexpression patterns over a time course. By contrast, protein interactions identified by the yeast two-hybrid approach applied genome-wide had only a weak relationship with gene expression. Therefore, protein-protein interactions identified by the yeast two-hybrid approach should be independently confirmed using other methods such as protein chip and/or affinity tagging/MS

approaches (54, 60, 61, 66), or by more traditional methods (e.g., coimmunoprecipitation).

To provide a systematic approach for analyzing multiple sets of data generated by gene expression profile studies, Hughes et al. (89) constructed a reference database of expression profiles corresponding to 300 diverse mutations and chemical treatments in *Saccharomyces cerevisiae*. By pattern matching to the database, even subtle differences in profiles can be revealed. Using this approach, the authors could predict the function of eight uncharacterized genes and confirm the predictions experimentally. In addition, they showed that it can be applied to characterize pharmacological perturbations, which means that this in silico (computed) approach has great potential in drug discovery and drug target identifications.

Large sets of data generated from different types of high-throughput methods can also be integrated to obtain a more comprehensive view of a biological process. Ideker et al. (90), for example, analyzed the galactose metabolic pathway in yeast using an integrated approach that combined expression profiles, MS analysis, and known protein-protein interaction information in the database. They were able to build, test, and refine the existing model, suggesting new hypotheses about the regulation of galactose utilization and physical interactions between this and a variety of other metabolic pathways.

Global sets of data generated with high-throughout approaches can be integrated to evaluate the quality and improve the confidence of individual data sets. In a recent report, Kemmeren et al. (91) applied a collection of expression profiles to assess the quality of several high-throughput protein interaction data sets. Out of 5342 putative two-hybrid interactions, the confidence levels of 973 interactions have been dramatically increased. In addition, the integration of the expression profiles and two-hybrid interactions has functionally annotated more than 300 previously uncharacterized genes. Furthermore, such in silico predictions were validated experimentally.

The Bork group (6) has taken such in silico evaluation of data one step further by examining protein interactions revealed by high-throughput yeast two-hybrid experiments, affinity protein complex purification analyses, correlated mRNA expression profiles, genetic interactions, and in silico prediction. Many interesting and intriguing results were discovered. Of the 80,000 purported interactions between yeast proteins, only ~2400 interactions are supported by more than one method. Each technique produced a unique coverage of interactions in terms of gene categories, which suggests that these methods have their specific strengthes and weaknesses. For example, most protein interaction data sets are heavily biased toward proteins of high abundance and toward particular cellular localizations of interacting proteins. In addition, the degree of evolutionary novelty of proteins plays a role in causing biased interaction coverage. To assess the accuracy and coverage of large interaction data, the authors compared the data with a reference set of trusted interactions. Not surprisingly, the highest accuracy was achieved for interactions supported by more than one method. Therefore, to

increase coverage and accuracy for protein interactions, as many complementary methods as possible should be used; however, the ever-growing body of high-throughput data remains a challenge to integrate using data storage and analysis approaches (92–95).

In summary, the integration of multiple sets of high-throughput data has shown its power in evaluating and improving the quality of the data. The strategy also provides insight into hidden properties to better understand the molecular mechanisms of various biological pathways and processes.

CONCLUSIONS

The rapid progress in the field of large-scale biology has provided us opportunities to understand the function of biological networks as a whole. Genomics decodes sequence information of an organism and provides the "parts catalog," while proteomics attempts to elucidate the functions and relationships of the individual "parts" and predict the outcomes of the modules they form on a higher level (97). Since protein is closer to biological function than DNA, much emphasis has been devoted to the development of new tools for proteomics. The recent advances discussed in this review have demonstrated that new technologies are powering proteomics research.

Among these new methodologies, microarray technology has served as a versatile tool to analyze protein activities and holds great promise in the identification of drugs and drug targets, as well as in clinical diagnostics. As discussed above, protein microarray technologies have shown their great potential in basic proteomics research such as determining protein-protein, protein-lipid, protein-ligand, and enzyme-substrate interactions. The reduced sample consumption in the microarray format is important in both basic proteomics research and diagnostics, where only minimal amounts of samples are available (3, 63). It is expected that real-time patient monitoring during disease treatment and therapy will be developed based on this emerging technology.

With the massive amount of data produced by high-throughput assays, it has become obvious that the integration of different sets of data provided by different systematic methods will greatly enhance the understanding of biological systems. For example, efforts have been devoted to combine all the protein information to generate a uniform numerical ranking system, which provides a more comprehensive picture of the biological context of each protein (96). Davidson et al. (98) borrowed the concepts used in integrated circuit design to dissect DNA-based regulatory gene networks using the sea urchin embryo as a model system. This network was derived from the combination of large-scale perturbation analyses, *cis*-regulatory analysis, molecular embryology, and computational algorithms (99). In such networks, the relationships between the transcriptional factors and their targets are represented by "and," "or," or "non" logic. Such networks could reveal the hidden interactions between pathways and deepen our understanding

of biological processes. Thus, it is increasingly important to properly integrate multiple sets of data to form an interaction network composed of all kinds of biological components, which will help elucidate the molecular mechanisms of life.

ACKNOWLEDGMENTS

We thank Jeremy Thorner and Jesslyn Holombo for critical comments on the manuscript. H. Zhu is supported by a postdoctoral fellowship from the Damon Runyon–Walter Winchell Cancer Research Foundation. Research in the Snyder lab is supported by grants from the National Institutes of Health.

The *Annual Review of Biochemistry* is online at http://biochem.annualreviews.org

LITERATURE CITED

1. Yanagida M. 2002. *J. Chromatogr. B* 771:89–106
2. Templin MF, Stoll D, Schrenk M, Traub PC, Vohringer CF, et al. 2002. *Trends Biotechnol.* 20:160–66
3. Stoll D, Templin MF, Schrenk M, Traub PC, Vohringer CF, Joos TO. 2002. *Front. Biosci.* 7:C13–32
4. Zhu H, Snyder M. 2001. *Curr. Drug Disc.* Sept:31–34
5. Zhu H, Snyder M. 2002. *Curr. Opin. Cell Biol.* 14:173–79
6. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, et al. 2002. *Nature* 417: 399–403
7. O'Farrell PZ, Goodman HM, O'Farrell PH. 1977. *Cell* 12:1133–41
8. Figeys D, Linda D, McBroom LD, Moran MF. 2001. *Methods* 24:230–39
9. Yates JR 3rd. 1998. *J. Mass Spectrom.* 33:1–19
10. Godovac-Zimmermann J, Brown LR. 2001. *Mass Spectrom. Rev.* 20:1–57
11. Mann M, Hendrickson RC, Pandey A. 2001. *Annu. Rev. Biochem.* 70:437–73
12. Ducret A, Van Oostveen I, Eng JK, Yates JR 3rd, Aebersold R. 1998. *Protein Sci.* 7:706–19
13. Gatlin CL, Kleemann GR, Hays LG, Link AJ, Yates JR 3rd. 1998. *Anal. Biochem.* 263:93–101
14. Ermer J, Vogel M. 2000. *Biomed. Chromatogr.* 14:373–83
15. Vestal ML. 1984. *Science* 226:275–81
16. Caprioli RM. 1990. *Anal. Chem.* 62: A477–85
17. Baczynskyj L. 1991. *J. Chromatogr.* 562: 13–29
18. Covey TR, Huang EC, Henion JD. 1991. *Anal. Chem.* 63:1193–200
19. Washburn MP, Wolters D, Yates JR 3rd. 2001. *Nat. Biotechnol.* 19:242–47
20. Schmid DG, Grosche P, Bandel H, Jung G. 2000. *Biotechnol. Bioeng.* 71:149–61
21. Nawrocki JP, Wigger M, Watson CH, Hayes TW, Senko MW, et al. 1996. *Rapid Commun. Mass Spectrom.* 10: 1860–64
22. Lipton MS, Pasa-Tolic' L, Anderson GA, Anderson DJ, Auberry DL, et al. 2002. *Proc. Natl. Acad. Sci. USA* 99:11049–54
23. Bergquist J, Palmblad M, Wetterhall M, Hakansson P, Markides KE. 2002. *Mass Spectrom. Rev.* 21:2–15
24. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, et al. 1999. *Nat. Biotechnol.* 17:994–99

25. Han DK, Eng J, Zhou H, Aebersold R. 2001. *Nat. Biotechnol.* 19:946–51
26. Wiese R, Belosludtsev Y, Powdrill T, Thompson P, Hogan M. 2001. *Clin. Chem.* 47:1451–57
27. Moody MD, Van Arsdell SW, Murphy KP, Orencole SF, Burns C. 2001. *BioTechniques* 31:186–90
28. Huang RP, Huang R, Fan Y, Lin Y. 2001. *Anal. Biochem.* 294:55–62
29. Sreekumar A, Nyati MK, Varambally S, Barrette TR, Ghosh D, et al. 2001. *Cancer Res.* 61:7585–93
30. Haab BB, Dunham MJ, Brown PO. 2001. *Genome Biol.* 2:RESEARCH0004
31. Miller JC, Butler EB, Teh BS, Haab BB. 2001. *Dis. Markers* 17:225–34
32. Robinson WH, DiGennaro C, Hueber W, Haab BB, Kamachi M, et al. 2002. *Nat. Med.* 8:295–301
33. Hiller R, Laffer S, Harwanegg C, Huber M, Schmidt WM, et al. 2002. *FASEB J.* 16:414–16
34. Joos TO, Schrenk M, Hopfl P, Kroger K, Chowdhury U, et al. 2000. *Electrophoresis* 21:2641–50
35. Krishna RG, Wold F. 1993. *Adv. Enzymol. Relat. Areas Mol. Biol.* 67:265–98
36. Griffin TJ, Aebersold R. 2001. *J. Biol. Chem.* 276:45497–500
37. Andersson L, Porath J. 1986. *Anal. Biochem.* 154:250–54
38. Neville DC, Rozanas CR, Price EM, Gruis DB, Verkman AS, et al. 1997. *Protein Sci.* 6:2436–45
39. Posewitz MC, Tempst P. 1999. *Anal. Chem.* 71:2883–92
40. Stensballe A, Andersen S, Jensen ON. 2001. *Proteomics* 1:207–22
41. Seetharaman S, Zivarts M, Sudarsan N, Breaker RR. 2001. *Nat. Biotechnol.* 19:336–41
42. Ficarro SB, McCleland ML, Stukenberg PT, Burke DJ, Ross MM, et al. 2002. *Nat. Biotechnol.* 20:301–5
43. MacCoss MJ, McDonald WH, Saraf A, Sadygov R, Clark JM, et al. 2002. *Proc. Natl. Acad. Sci. USA* 99:7900–5
44. Ross-Macdonald P, Sheehan A, Roeder GS, Snyder M. 1997. *Proc. Natl. Acad. Sci. USA* 94:190–95
45. Ross-Macdonald P, Coelho PS, Roemer T, Agarwal S, Kumar A, et al. 1999. *Nature* 402:413–18
46. Ross-Macdonald P, Sheehan A, Friddle C, Roeder GS, Snyder M. 1999. *Methods Enzymol.* 303:512–32
47. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, et al. 2002. *Genes Dev.* 16:707–19
48. Kumar A, Cheung KH, Tosches N, Masiar P, Liu Y, et al. 2002. *Nucleic Acids Res.* 30:73–75
49. Ding DQ, Tomita Y, Yamamoto A, Chikashige Y, Haraguchi T, et al. 2000. *Genes Cells* 5:169–90
50. Simpson JC, Wellenreuther R, Poustka A, Pepperkok R, Wiemann S. 2000. *EMBO Rep.* 1:287–92
51. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. 2000. *Nature* 403: 623–27
52. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. 2001. *Proc. Natl. Acad. Sci. USA* 98:4569–74
53. Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, et al. 2000. *Science* 287:116–22
54. Boulton SJ, Gartner A, Reboul J, Vaglio P, Dyson N, et al. 2002. *Science* 295: 127–31
55. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, et al. 2001. *Nature* 409: 211–15
56. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, et al. 1997. *Nature* 388:539–47
57. Stagljar I, Korostensky C, Johnsson N, te Heesen S. 1998. *Proc. Natl. Acad. Sci. USA* 95:5187–92
57a. Reinders A, Schulze W, Kuhn C, Barker L, Schulz A, et al. 2002. *Plant Cell* 14:1567–77
57b. Tsujimoto Y, Numaga T, Ohshima K, Yano MA, Ohsawa R, et al. 2003. *EMBO J.* 22:335–43

57c. Wang B, Nguyen M, Breckenridge DG, Stojanovic M, Clemons PA, et al. 2003. *J. Biol. Chem.* In press.

57d. Wittke S, Dunnwald M, Albertsen M, Johnsson N. 2002. *Mol. Biol. Cell*. 13:2223–32

58. Schwikowski B, Uetz P, Fields S. 2000. *Nat. Biotechnol.* 18:1257–61

59. Snyder M, Kumar A. 2002. *Funct. Integr. Genomics* 2:135–37

60. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. 2002. *Nature* 415: 141–47

61. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. 2002. *Nature* 415:180–83

62. Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, et al. 1999. *Science* 286:1153–55

63. Zhu H, Snyder M. 2001. *Curr. Opin. Chem. Biol.* 5:40–45

64. Zhu H, Klemic JF, Chang S, Bertone P, Casamayor A, et al. 2000. *Nat. Genet.* 26:283–89

64a. Malathi K, Xiao Y, Mitchell AP. 1999. *Genetics* 153:1145–52

65. MacBeath G, Schreiber SL. 2000. *Science* 289:1760–63

66. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, et al. 2001. *Science* 293: 2101–5

67. McDonnell JM. 2001. *Curr. Opin. Chem. Biol.* 5:572–77

68. Nieba L, Nieba-Axmann SE, Persson A, Hamalainen M, Edebratt F, et al. 1997. *Anal. Biochem.* 252:217–28

69. Salamon Z, Brown MF, Tollin G. 1999. *Trends Biochem. Sci.* 24:213–19

70. Myszka DG, Rich RL. 2000. *Pharm. Sci. Technol. Today* 3:310–17

71. Sapsford KE, Liron Z, Shubin YS, Ligler FS. 2001. *Anal. Chem.* 73:5518–24

72. Houseman BT, Huh JH, Kron SJ, Mrksich M. 2002. *Nat. Biotechnol.* 20:270–74

73. Lizcano JM, Deak M, Morrice N, Kieloch A, Hastie CJ, et al. 2002. *J. Biol. Chem.* 277:27839–49

74. Wang D, Liu S, Trummer BJ, Deng C, Wang A. 2002. *Nat. Biotechnol.* 20:275–81

75. Houseman BT, Mrksich M. 2002. *Chem. Biol.* 9:443–54

76. Bishop AC, Ubersax JA, Petsch DT, Matheos DP, Gray NS, et al. 2000. *Nature* 407:395–401

77. Weiss EL, Bishop AC, Shokat KM, Drubin DG. 2000. *Nat. Cell Biol.* 2:677–85

78. Carroll AS, Bishop AC, DeRisi JL, Shokat KM, O'Shea EK. 2001. *Proc. Natl. Acad. Sci. USA* 98:12578–83

79. Habelhah H, Shah K, Huang L, Burlingame AL, Shokat KM, et al. 2001. *J. Biol. Chem.* 276:18090–95

80. Haab BB. 2001. *Curr. Opin. Drug Discov. Dev.* 4:116–23

81. Biesecker G, Dihel L, Enney K, Bendele RA. 1999. *Immunopharmacology* 42: 219–30

82. Chen J. 2002. *Chem. Biol.* 9:543–44

83. MacBeath G, Koehler AN, Schreiber SL. 1999. *J. Am. Chem. Soc.* 121:7967–68

84. Winssinger N, Ficarro S, Schultz PG, Harris JL. 2002. *Proc. Natl. Acad. Sci. USA* 99:11139–44

85. Kuruvilla FG, Shamji AF, Sternson SM, Hergenrother PJ, Schreiber SL. 2002. *Nature* 416:653–57

86. Blackwell HE, Perez L, Stavenger RA, Tallarico JA, Eatough EC, et al. 2001. *Chem. Biol.* 8:1167–82

87. Clemons PA, Koehler AN, Wagner BK, Sprigings TG, Spring DR, et al. 2001. *Chem. Biol.* 8:1183–95

88. Jansen R, Greenbaum D, Gerstein M. 2002. *Genome Res.* 12:37–46

89. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. 2000. *Cell* 102:109–26

90. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, et al. 2001. *Science* 292:929–34

91. Kemmeren P, van Berkum NL, Vilo J, Bijma T, Donders R, et al. 2002. *Mol. Cell* 9:1133–43

92. Xenarios I, Salwinski L, Duan XJ,

Higney P, Kim SM, et al. 2002. *Nucleic Acids Res.* 30:303–5
93. Mellor JC, Yanai I, Clodfelter KH, Mintseris J, DeLisi C. 2002. *Nucleic Acids Res.* 30:306–9
94. Bader GD, Hogue CW. 2000. *Bioinformatics* 16:465–77
95. Paton NW, Khan SA, Hayes A, Moussouni F, Brass A, et al. 2000. *Bioinformatics* 16:548–57
96. Qian J, Stenger B, Wilson CA, Lin J, Jansen R, et al. 2001. *Nucleic Acids Res.* 29:1750–64
97. Noble D. 2002. *Science* 295:1678–82
98. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, et al. 2002. *Dev. Biol.* 246:162–90
99. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, et al. 2002. *Science* 295:1669–78
100. Goodlett DR, Yi EC. 2002. *Funct. Integr. Genomics* 2:138–53