Protein analysis on a proteomic scale

Eric Phizicky*, Philippe I. H. Bastiaens†, Heng Zhu‡, Michael Snyder‡ & Stanley Fields§

*University of Rochester School of Medicine, Department of Biochemistry and Biophysics, Box 712, 601 Elmwood Avenue, Rochester, New York 14642, USA (e-mail: eric_phizicky@urmc.rochester.edu)
†European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany (e-mail: philippe.bastiaens@embl-heidelberg.de)
‡Department of Molecular, Cellular, and Developmental Biology, PO Box 208103, Yale University, New Haven, Connecticut 06520, USA (e-mail: heng.zhu@yale.edu; michael.snyder@yale.edu)
§Howard Hughes Medical Institute, Departments of Genome Sciences and Medicine, University of Washington, Box 357730, Seattle, Washington 98195, USA (e-mail: fields@u.washington.edu)

The long-term challenge of proteomics is enormous: to define the identities, quantities, structures and functions of complete complements of proteins, and to characterize how these properties vary in different cellular contexts. One critical step in tackling this goal is the generation of sets of clones that express a representative of each protein of a proteome in a useful format, followed by the analysis of these sets on a genome-wide basis. Such studies enable genetic, biochemical and cell biological technologies to be applied on a systematic level, leading to the assignment of biochemical activities, the construction of protein arrays, the identification of interactions, and the localization of proteins within cellular compartments.

Proteomics — the analysis of genomic complements of proteins — has burst onto the scientific scene with stunning rapidity over the past few years, perhaps befitting a discipline that can enjoy the virtually instantaneous conversion of a genome sequence to a set of predicted proteins. But whereas every fragment of DNA behaves biochemically much like any other, proteins possess unique properties, and such individuality creates an enormous hurdle for methodologies that seek to assign an activity to sets of proteins that may number in the thousands[1]. Yet the confluence of breakthroughs in cloning and expression technologies, biochemical and genetic strategies, and the instrumentation of mass spectrometry and microscopy has made such global assays increasingly common.

We describe some of these technologies and strategies here, along with a discussion of their advantages and disadvantages, and a brief consideration of new technologies still at the design stage.

Protein expression and purification

The development of methods for parallel analysis of the proteome has relied on the rapid identification of open reading frames (ORFs) and their facile cloning and manipulation. An ORF is defined as the amino acid codons between the initiation codon at the start and the termination codon at the end. ORF identification can be complicated by uncertainties in defining translation start sites, small size and, in particular, the signals for splicing, polyadenylation and editing that can lead to multiple messenger RNA species from a single DNA sequence. Even for a simple and well-studied eukaryote such as the yeast *Saccharomyces cerevisiae*, in which RNA processing is relatively uncomplicated, the number of ORFs has been revised several times as a result of transcriptional analysis and the comparative analysis of genomes of close relatives[2,3].

Although cloning of a genomic set of ORFs enables the technologies discussed here to be performed, it is important to note that such a step entails the loss of much of the natural diversity of proteins. For example, a single spliced mRNA is generally chosen as a template for each gene, and the many other mRNA species that result in protein isoforms are not considered. Similarly, post-translational modifications, including phosphorylation, glycosylation, methylation, acetylation and a host of others, may be neglected. Some of this variation can be captured by mass spectrometric approaches (see review in this issue by Aebersold and Mann, page 198) and some by increasing the number of constructs that are generated for each gene. Another limitation to large-scale protein production is that the substantial class of membrane proteins is generally not amenable to the standardized procedures of genome-wide approaches.

Cloning of ORFs for subsequent expression requires a genomic set of gene-specific primers that is suitable for amplification by the polymerase chain reaction (PCR) and for subsequent insertion of the PCR products into appropriate plasmids. This latter requirement is met by use of forward and reverse primers that contain common 5′ ends. The first example of this methodology was the genomic-scale PCR amplification of the ~6,000 *S. cerevisiae* ORFs for cloning into yeast plasmids[4]. A similar strategy was applied to more than 1,200 *Caenorhabditis elegans* ORFs predicted solely by a gene analysis programme, and ~70% of these were verified by sequence analysis of the PCR products[5]. Efforts are also under way for sets of mouse and human ORFs. With modern methods of high-throughput synthesis, primers can be made with high fidelity, at reasonable cost, and in 96-well format amenable for robotic manipulation. Insertion of the amplified ORFs into vectors generally uses any of several recombination-based methods that are now in widespread use (Box 1).

For biochemical analysis of proteins, their expression in a homologous system is ideal because the proteins are in their natural environment, are subject to native modifications, and can interact with their natural partners. This has been possible for proteins from yeast and from bacteria such as *Escherichia coli*, but heterologous expression is usually used for proteins from other organisms. In most cases, expression is attempted in *E. coli*, in which upwards of 60% of likely soluble proteins may be expressed in soluble form[6]. The most common alternative for expression is the use of insect cells, which results in modifications that are usually similar to mammalian cells. However, heterologous expression inevitably can lead to problems of expression and solubility for many proteins.

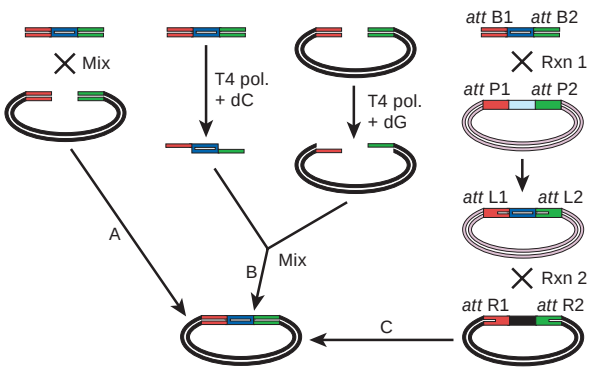
A primary goal of the genome-wide plasmid constructions is to incorporate a fusion tag, a short peptide or protein

Box 1
Recombinational strategies for rapid and precise cloning of amplified DNA

For each strategy shown in the figure opposite, amplified open reading frame (ORF) DNAs (blue) all have a common 5′ end sequence (red) and a different common 3′ end sequence (green). The earliest method (strategy A) used gap repair-mediated recombination in yeast. Amplified ORFs are mixed with a linearized vector with ends that are identical to those of the amplified DNA, and the mixture is transformed directly into yeast, where gap repair-mediated recombination occurs *in vivo* with high efficiency. The result is precisely integrated DNA, which allows simple construction of in-frame fusions of genes.

Strategy B is ligation-independent cloning. Both 3′ ends of amplified ORFs are constructed to lack a specific nucleotide (for example, dC), such that treatment with T4 DNA polymerase and dCTP removes the 3′ ends until the first dC residue and leaves 5′ overhangs of 12 or more bases. The corresponding plasmid with matching ends is treated with T4 DNA polymerase and dGTP, creating a corresponding overhang. After enzyme removal, DNAs are mixed and transformed directly into the bacterium *Escherichia coli*[72,73]. This method has recently been used for the generation of 400 clones in a 3-day automated procedure[74]. Thus, cloning the ORFs of a genome about the size of yeast might require only a few months of work.

Strategy C uses recombination by the Gateway cloning system of Invitrogen. This system is based on integration/excision of phage λ in *E. coli*. During integration, recombination between the λ *att* P site and the host *att* B site results in an integrated phage with ends called *att* L and *att* R. During excision, recombination between *att* L and *att* R sites regenerates *att* P and *att* B. In the Gateway application of this reaction, recombination is effected at each end of the DNA by pairs of *att* sites that are similar but not identical. Thus, amplified ORF DNAs with ends bearing *att* B1 and *att* B2 sequences recombine *in vitro* with a donor vector containing *att* P1 and *att* P2 sequences in the presence of components of λ integration to form the resulting entry plasmid, now bearing *att* L1 and *att* L2 sites. A second recombination



reaction *in vitro* with a destination vector containing *att* R1 and *att* R2 sites (black) and components of λ excision allows transfer of the ORF into the expression vector, regenerating *att* B sites. A significant advantage of this approach is that the cloned ORF DNA can be transferred easily from one plasmid to a number of other plasmids by a simple set of further recombination reactions *in vitro*. This approach has recently been applied for the analysis of ORFs from humans and *Caenorhabditis elegans*[5,6].

In yeast, a powerful alternative to plasmid expression of ORFs can be provided by chromosomal expression of tagged ORFs. This approach uses transformation of linear DNA into yeast with selection for recombination at a homologous chromosomal site. It has been used to construct a library of 1,548 strains[45] in which the coding sequence for a purification tag was inserted at the end of the ORFs. The resulting strains each express an ORF with a C-terminal fusion tag under control of its natural promoter, thereby ensuring as normal an expression pattern as possible, and stable propagation of the inserted gene.

domain that becomes linked to each member of a set of proteins. The use of these tags has continued to revolutionize biochemical analysis. For purification of biochemically active proteins, protein affinity tags (see Box 2) feature high affinity and selectivity for binding to specific resins to facilitate purification and elution under conditions that retain activity[7]. The recent application of genomic high-throughput purification illustrates the utility of such tags. Through the use of manual methods in 96-well format, 5,800 individual yeast glutathione *S*-transferase (GST) fusion proteins were purified 1,152 at a time and used successfully for biochemical analysis[8]. Current approaches now apply automation to parallel purification. But expression of each fusion protein and the purification of the corresponding tagged proteins require the use of a generic scheme. Inevitably, there will be members of a protein set that cannot be expressed, solubilized or purified under these generic conditions, because of the loss of a cofactor, inappropriate buffers, or other incompatible conditions. Additionally, proteins may be functionally inactive as fusion proteins.

Probing protein activity on a proteomic scale

The ultimate value of genomic sets of strains expressing tagged proteins, or of the corresponding purified proteins, is their potential for parallel analysis of the proteome. In this way one can, in principle, identify all of the proteins with a particular function or property in a single systematic experiment.

Biochemical genomics and functional protein microarrays

Two very different methods have been used to probe genomic sets of proteins for biochemical activity. One method has been termed a biochemical genomics approach, which uses parallel biochemical analysis of a proteome comprised of pools of purified proteins in order to identify proteins and the corresponding ORFs responsible for a biochemical activity[9]. As applied to *S. cerevisiae*, this approach involved the generation of a set of 6,144 yeast strains, each expressing a distinct *S. cerevisiae* ORF as a GST–ORF fusion protein, followed by purification of the fusion proteins in pools. A biochemical activity is mapped to a specific ORF by assaying the pools for an activity, and then deconvoluting positive pools by preparation and analysis of subpools of the proteins. This method has been used to rapidly identify a number of yeast genes whose products co-purify with activities, including two proteins implicated in the metabolism of an NAD derivative produced during transfer RNA splicing[9], a cytochrome *c* methyltransferase[9], a tRNA dihydrouridine synthase[10], both members of a tRNA m^7G methyltransferase complex[11], and a new DNA-binding protein implicated in the transcriptional regulation of the yeast *SUC2* gene[12].

Important features of this approach include its speed at assigning catalytic function to ORFs, its generality for virtually any type of catalytic activity, and its sensitivity. High sensitivity is obtained both because the fusion proteins are overexpressed, and because background proteins are removed during purification. The lack of background proteins allows activities to be assayed for hours without destruction of product, substrate or proteins, yielding a huge increase in sensitivity for catalytic activities[13]. Additionally, it allows the detection of complexes of more than one protein, which otherwise cannot be detected by overproduction of a single component[11]. Because the average protein in these preparations is present at

concentrations of ~20 nM, this approach is also suitable for the detection of protein–ligand complexes, which, unlike enzymatic activities, do not benefit from prolonged incubation[13]. The requirements of this method for a functional amino-terminal ORF fusion, and for effective solubilization and purification of the GST–ORF fusion proteins in active form, are often satisfied. However, the library has some bias against larger proteins, and those that retard growth during propagation[13,14].

The second approach for analysing genomic sets of proteins is the use of functional protein microarrays, in which individually purified proteins are separately spotted on a surface such as a glass slide and then analysed for activity. This approach has huge potential for rapid high-throughput analysis of proteomes and other large collections of proteins, and promises to transform the field of biochemical analysis (Fig. 1).

A critical first step in generating these arrays has been the development of general methods for arraying a genomic set of proteins on a solid surface without denaturing the proteins, and at high enough density for detection of activity. Recently, arrays have used both glass slides and chips with modified surfaces engineered to carry pads, films, nanowells or microfluidic channels[8,15–20]. Although such modified surface structures require sophisticated engineering, they reduce evaporation and denaturation during drying, increase protein-binding capacity, and prevent cross-contamination because of the physical boundaries separating each sample.

A comprehensive microarray screening of a class of proteins was described by Zhu *et al.*[16], who analysed the substrate specificities of 119 yeast protein kinases using 17 different test substrates that were adhered to the surface of nanowell microarrays. The experiments of MacBeath and Schreiber[15] further demonstrated the potential of functional protein microarrays. In this study, proteins were tethered covalently to chemically activated glass slides, and then shown to be active for different classes of activities. Thus, three well-studied protein–protein interactions could be detected with fluorescently labelled protein probes, three different substrate proteins were shown to be phosphorylated specifically by protein kinases known to act on them, and three types of protein–small-molecule interactions could be detected in the micromolar range using small molecules bound to fluorescently labelled beads, which allows greater sensitivity owing to avidity effects. Finally, it was shown that a single protein could be detected at high resolution on a single glass slide in the midst of 10,799 identical spots of another protein. Taken together, Zhu *et al.*[16] and MacBeath and Schreiber[15] showed the huge potential of protein microarrays for parallel biochemical analysis.

The first full-scale genomic protein microarray was demonstrated by Zhu *et al.*[8]. In this experiment, 5,800 (94%) of the predicted yeast ORFs were cloned, and greater than 80% of these produced detectable amounts of protein, after purification in a high-throughput protocol. The proteins were spotted onto nickel-coated glass slides and used for the analysis of two different binding activities. First, a biotinylated calmodulin probe detected 6 of the known calmodulin-binding proteins that were present in the purified collection, as well as 33 new, potentially interacting proteins. Second, biotinylated liposomes detected 150 proteins that bind different phosphoinositides[8]. These experiments opened a new field in which entire proteomes can be screened for binding and other biochemical assays.

This approach can be extended in several different ways. Binding can be studied in real-time by use of a surface plasmon resonance (SPR) biosensor surface with 64 individual immobilized sites in a single flow cell, which can be scaled to 400 assays per day[21]. Peptides can also be analysed using microarrays. Recently, a monolayer-coated gold chip was shown to be useful for immobilization of peptides for biochemical analysis using detection by a phosphorimager, SPR and fluorescence microscopy[22]. Synthesis of peptide microarrays may become more practical with the development of methods for *in situ* synthesis of high-density peptide microarrays, using photolithography or light-directed synthesis[23]. Carbohydrate and small-molecule microarrays have also shown great potential for characterizing protein–small-molecule binding activities[24,25].

Both the biochemical genomics approach and protein microarrays have advantages and disadvantages. Use of biochemical genomics for yeast requires only 64 assays to cover the genome, is flexible for many types of assays, and is particularly useful for enzymatic activities. But use of pools does not allow easy assessment of the quality of each individual protein in a pool, can cause interference by the 95 other proteins present in the mixture, is not well-suited to binding assays with fluorescent probes, and cannot easily handle multiple positives at once. Use of microarrays to probe activity allows individual assessment of the quality of each protein, the immediate identification of the source ORF responsible for a particular activity, the identification of multiple positives in a single round, and high-throughput analysis of activities via automated arraying, assaying and scanning. However, it requires individual growth of 6,000 strains (for yeast) and 6,000 individual purifications of proteins, and is best suited at present for binding assays using fluorescent probes or activity assays with tethered substrates.

A second type of protein microarray, which is early in development, is the analytical microarray. Here, a genomic set of protein-specific ligands such as antibodies, nucleic acid aptamers or chemical probes is spotted on a microarray, and then the levels of different proteins in an extract are quantified in parallel by binding extract proteins to the microarray. Analytical protein microarrays are starting to realize their potential for monitoring protein expression on a proteome-wide scale and in medical diagnostics. Microarrays

Box 2
Commonly used affinity tags

Each affinity tag used for purification of fused proteins has its unique features. Glutathione *S*-transferase binds tightly to glutathione-agarose and is eluted with glutathione, allowing very high levels of purification in one step; however, it is known to dimerize and this may interfere with protein function. His6 binds immobilized metal-ion columns and is eluted with imidazole; this tag is very simple, but it results in more modest purification than some other tags. Calmodulin-binding peptide binds calmodulin-agarose columns and is eluted with EGTA[75]; purification is efficient, but proteins with required divalent metal ions may be affected.

An extremely useful variation of these commonly used tags uses very-high-affinity purification tags coupled with protease cleavage sites to remove the purification tag and simultaneously elute the proteins. An example is the generation of fusion proteins in which the target protein is fused to a tobacco etch virus (TEV) protease site linker and a protein A domain. The protein A domain binds tightly to immunoglobulin-γ columns, following which the target protein is released using treatment with the highly specific TEV protease[75]. This results in extremely efficient purification of proteins.

Additional fusion tags used for other purposes include maltose-binding protein, which increases the solubility of fused proteins[76]; epitope tags such as haemagglutinin, myc and FLAG tags, for immunoprecipitation and detection; and green fluorescent protein and its derivatives, which have been used for localizing proteins in living cells and for detection of protein interactions through fluorescence resonance energy transfer. Often tags or domains are used in concert. Examples include the use of two tags to effect very high purification of complexes of proteins for mass spectrometry analysis[45], the use of nuclear localization tags in concert with DNA-binding domains or transcription-activation domains for two-hybrid experiments, the use of purification tags coupled with a His6 tag for arraying on a microscope slide[8], and the use of purification tags coupled with tags for immunodetection.

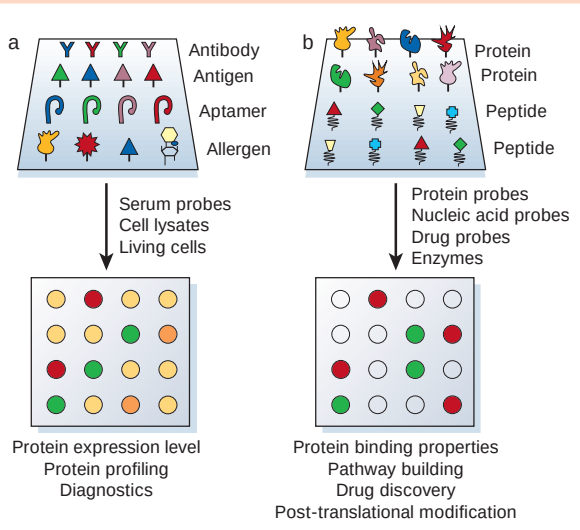


Figure 1 Analytical versus functional protein microarrays. **a**, Analytical protein microarray. Different types of ligands, including antibodies, antigens, DNA or RNA aptamers, carbohydrates or small molecules, with high affinity and specificity, are spotted down onto a derivatized surface. These chips can be used for monitoring protein expression level, protein profiling and clinical diagnostics. Similar to the procedure in DNA microarray experiments, protein samples from two biological states to be compared are separately labelled with red or green fluorescent dyes, mixed, and incubated with the chips. Spots in red or green colour identify an excess of proteins from one state over the other. **b**, Functional protein microarray. Native proteins or peptides are individually purified or synthesized using high-throughput approaches and arrayed onto a suitable surface to form the functional protein microarrays. These chips are used to analyse protein activities, binding properties and post-translational modifications. With the proper detection method, functional protein microarrays can be used to identify the substrates of enzymes of interest. Consequently, this class of chips is particularly useful in drug and drug-target identification and in building biological networks.

containing antibodies, antigens or in some cases peptides and other biomolecules have been used to monitor differential expression of proteins in colon carcinoma cells[26], cell-surface antigens specific for particular cell types[27], and autoantibodies in patient sera[28,29]. The main problems with antibody-mediated analytical protein microarrays are specificity and quantitation. Most antibodies cross-react with proteins other than the antigen of interest, which leads to poor quantification. Haab and colleagues[17] showed that only 23% of 115 well-characterized antibody–antigen pairs could be accurately quantified at the level of 1 $\mu g\ ml^{-1}$ soluble antigen, although 60% of the binding interactions could be estimated qualitatively. Nonetheless, it seems likely that better and more efficient methods will be developed in the coming years to quantitatively assay the amounts of proteins in a high-throughput, parallel manner.

Other large-scale activity-based assays

Other activity assays have been used that address functional classes of activities within the proteome. The goal of one approach was to assess all of the DNA targets of the known DNA-binding protein regulators of yeast under one defined growth condition[30]. To this end, a series of strains was constructed in which each of the 141 known yeast regulators was epitope-tagged at its carboxy terminus and expressed under control of its normal promoter at its appropriate chromosomal locus (see Box 1). After growth of each strain, chromatin immunoprecipitation analysis was carried out, in which each tagged protein was purified along with its population of bound DNA, and the identity and amount of the DNA was determined with conventional DNA microarrays. The technique was used with 106 of the 141 known transcription factors, and the study allowed not only a genomic view of the regulatory modules of each gene, but also a description of a number of different networks of transcription regulation in the cell, and a functional assessment of the role of each transcription factor in yeast.

Another general method for assessing catalytic activity of the proteome is activity-based protein profiling. In this method, an extract is treated with a chemical probe that reacts covalently with any protein having a specific class of activity, and modified proteins are detected with a second tag such as biotin that is present on the reactive chemical[31–33]. The key to the approach is the use of a probe that is specific for the activity, but general for all proteins with that class of activity. The method has been applied to probe cysteine proteases, resulting in the identification of two previously known caspase species in cells induced for apoptosis and evidence for several candidates in another cell line[31,34]. It has also resulted in the identification of three previously known cathepsins and several other reactive proteins in rat kidney extracts, and demonstrated distinct labelling patterns during the progression of skin cancer in mice[35]. Activity-based protein profiling has also been applied to probe serine hydrolases, resulting in the identification of two such hydrolases from rat brain and the detection of a number of others in different tissues[36]. It is evident from these studies that this method is remarkably useful for profiling extracts to define the number of different activities of a particular type, the amounts of each protein in the active state, and the onset of the activity in different cell states.

Recently, this technology has been extended in three ways. First, a general isolation procedure was developed to purify and identify multiple reacted proteins in parallel. Denatured proteins were captured with avidin beads and then subjected to SDS-polyacrylamide gel electrophoresis, trypsin treatment and mass spectrometry[37]. Second, a panel of different fluorescent derivatives of activity-based probes of the papain family of cysteine proteases was used to monitor active proteases in living cells, and to enable facile *in vivo* screening of small-molecule inhibitors for their activity and specificity[38]. Third, small-molecule probes have been developed that are active against multiple types of enzymes, which allows profiling of several species simultaneously[39]. The unique ability of activity-based protein profiling to monitor active species of a panel of enzymes in cells gives this method huge potential in profiling signal transduction pathways in development and differentiation, as demonstrated by the recent analysis of the activity, subcellular distribution and glycosylation state of the serine hydrolase superfamily in cancer cells[40].

A related activity-based probe involves the specific targeting of a single protein kinase *in vitro* or *in vivo* to elucidate its function. Identification of the natural targets of a protein kinase is of enormous importance because there are so many protein kinases in the proteome, a large fraction of the proteins in the cell are phosphorylated, and phosphorylation often has significant effects on protein function. To accomplish this, Shokat and colleagues[41] re-engineered a highly conserved region of the ATP-binding site of protein kinases to allow the use of ATP analogues and kinase inhibitors that would not normally be active. Thus, a specific kinase can be retailored such that it alone is inhibited *in vivo*, allowing an assessment of its function[41], or such that it is the only active kinase in extracts, allowing facile identification of potential substrates[42]. For example, the specifically activated kinase JNK was used to identify a new substrate in crude extracts by isolation of the corresponding phosphorylated protein from two-dimensional gels, followed by mass spectrometry[42]. This approach is generally applicable to many serine/threonine protein kinases and tyrosine protein kinases[43], and promises to have a prominent role in deducing the range and scope of function of this broad class of cellular activity.

Protein interaction analysis

One powerful method for deducing protein function is to identify the interacting partners of proteins, as proteins that interact with one

another or are part of the same complex are generally involved in the same cellular processes. As such, there have been intensive efforts in the past few years to identify protein–protein interaction on a large scale. Two types of approaches have been used: the two-hybrid system described below, which is used to detect binary interactions *in vivo*, and biochemical co-purification of complexes using affinity tags, coupled with protein identification using mass spectrometry, which defines the total spectrum of complexes for a particular tagged protein[44,45]. The latter is reviewed by Aebersold and Mann on page 198 of this issue and will not be discussed. Fluorescent-based interaction assays have also been developed, but have not been used on a high-throughput basis.

Genome-wide two-hybrid approaches

The yeast two-hybrid assay[46] provides a genetic approach to the identification and analysis of protein–protein interactions. It relies on the modular nature of many eukaryotic transcription factors, which contain both a site-specific DNA-binding domain and a transcriptional-activation domain that recruits the transcriptional machinery. In this assay, hybrid proteins are generated that fuse a protein X to the DNA-binding domain and protein Y to the activation domain of a transcription factor (Fig. 2a). Interaction between X and Y reconstitutes the activity of the transcription factor and leads to expression of reporter genes with recognition sites for the DNA-binding domain. In the typical practice of this method, a protein of interest fused to the DNA-binding domain (the so-called 'bait') is screened against a library of activation-domain hybrids ('preys') to select interacting partners.

Key advantages of the two-hybrid assay are its sensitivity and flexibility. The sensitivity derives in part from overproduction of proteins *in vivo*, their designed direction to the nuclear compartment where the interactions are monitored, the large number of variable inserts of the interacting proteins that can be examined at once, and the potency of the genetic selections. This sensitivity leads to the detection of interactions with dissociation constants around 10^{-7} M, in the range of most weak protein interactions found in the cell, and is more sensitive than co-purification, which requires stability of a complex through dilution from cell lysis, and through subsequent purification steps. This sensitivity also allows detection of certain transient interactions or those that might affect only a subpopulation of the hybrid proteins.

Flexibility of the assay is provided by calibration to detect interactions of varying affinity by altering the expression levels of the hybrid proteins, the number and nature of the DNA-binding sites, and the composition of the selection media. Disadvantages of the yeast assay include the unavoidable occurrence of false negatives and false positives. False negatives include proteins such as membrane proteins and secretory proteins that are not usually amenable to a nuclear-based detection system, proteins that activate transcription when fused to a DNA-binding domain, proteins that fail to fold correctly, and interactions dependent on domains occluded in the fusions or on post-translational modifications. False positives include colonies not resulting from a bona fide protein interaction, as well as colonies resulting from a protein interaction not indicative of an association that occurs *in vivo*. Predominantly, false positives seem to be due to spurious transcription that does not derive from any interaction occurring between the hybrid proteins.

The two-hybrid system evolved to a proteomics strategy by the construction of ordered arrays of strains expressing either DNA-binding domain or activation-domain fusion proteins, the implementation of improved selection methods and plasmids, the use of mating to introduce pairs of plasmids for testing, and the use of automation.

Different genome-wide two-hybrid strategies have been used to analyse protein interactions in *S. cerevisiae*. One approach involved screening a large number of individual proteins against a comprehensive library of randomly generated fragments (Fig. 2b), as was used to identify numerous interactions for proteins implicated in RNA splicing[47]. A second approach used systematic one-by-one testing of every possible combination of proteins using a mating assay with a comprehensive array of strains. In this way, 192 baits were screened against an array of essentially all activation-domain fusions of full-length yeast ORFs to identify 281 putative interactions[48], and ~1,000 proteins have been screened to date (S.F., unpublished data). A third approach used a one-by-many mating strategy in which each member of a nearly complete set of strains expressing yeast ORFs as DNA-binding domain hybrids was mated to a library of strains containing activation-domain fusions of full-length yeast ORFs (Fig. 2d), resulting in 692 positives[48]. A fourth variation involved mating of defined pools of strain arrays[49]. This approach required cloning all of the yeast ORFs into both two-hybrid vectors, followed by pooling sets of 96 transformants each. Matings were conducted for the 62×62 combinations of pools, and positives were sequenced (Fig. 2e), resulting in a total of 4,549 positives, of which the 841 that were identified more than three times form a core data set.

In addition to the analyses of yeast proteins, large-scale two-hybrid studies have been carried out for proteins of *Helicobacter pylori*[50], *C. elegans*[51] and *Drosophila melanogaster* (R. Finley, personal communication).

Notably, these approaches are not exclusive; for example, full-length ORFs are often used in screens of random libraries, and protein fragments can be tested in a one-by-one format against an

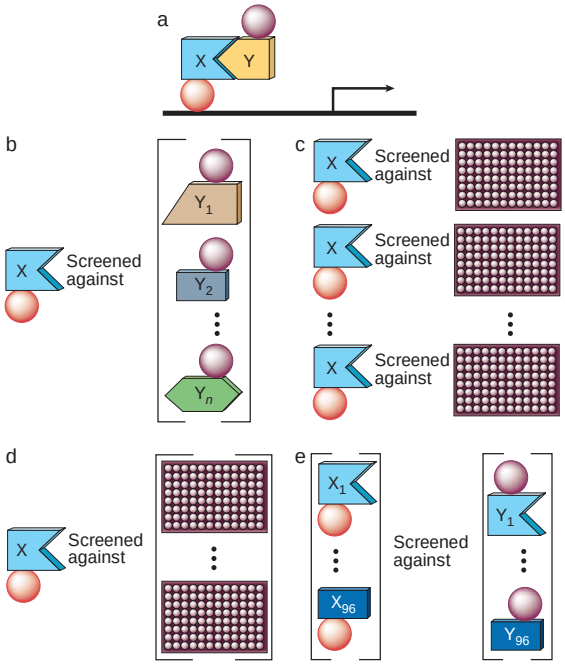


Figure 2 Yeast two-hybrid approaches. **a**, The yeast two-hybrid system. DNA-binding and activation domains (circles) are fused to proteins X and Y; the interaction of X and Y leads to reporter gene expression (arrow). **b**, A standard two-hybrid search. Protein X, present as a DNA-binding domain hybrid, is screened against a complex library of random inserts in the activation-domain vector (square brackets). **c**, A two-hybrid array approach. Protein X is screened against a complete set of full-length open reading frames (ORFs) present as activation-domain hybrids (shown as yeast transformants spotted onto microtitre plates). **d**, A two-hybrid search using a library of full-length ORFs. The set of ORFs as activation-domain hybrids (microtitre plates in square brackets) is combined to form a low-complexity library. **e**, A two-hybrid pooling strategy. Pools of ORFs as both DNA-binding domain and activation-domain hybrids (square brackets) are screened against each other.

activation-domain array. Compared to systematic mating, random insert or defined ORF libraries require more statistical sampling to ensure adequate coverage of the interactions. They also require sequencing of plasmids to identify interacting partners and tend, on average, to yield fewer interactions than systematic mating, although throughput is faster. Random fragment libraries may also reveal domains that might be masked, and smaller fusion proteins work better in the assay and provide direct information about interaction domains.

Unlike the case for a single two-hybrid experiment conducted by an individual laboratory dedicated to the investigation of a specific biological question, the proteomic two-hybrid projects produce potential interactions at a rate too rapid to allow individual testing for confirmation. Small-scale experiments generally allow the elimination of false positives, yielding a literature focused on a few interactions that have often been validated by additional experimentation; by contrast, genome-wide projects necessarily report all of their putative interactions. This raises the question of the accuracy of genomic data in general, and of two-hybrid data in particular.

Several analyses of genomic two-hybrid results suggest that about 50% are correct[52–56]. These studies have set the tone for how other large proteomic data sets can be mined to retrieve biologically significant findings. For example, one approach is based on the fact that genes encoding proteins involved in the same function tend to be co-expressed[53]. A second strategy[53] assesses reliability by determining whether two proteins that interact putatively have paralogues that also interact. A third uses information about protein localization (that is, which proteins lie in the same subcellular compartment) to increase the accuracy of the two-hybrid interaction data[54]. These analyses indicate that the data from small-scale studies are of considerably greater reliability than that from high-throughput studies. Additionally, they show how computational assessment of large-scale data that relies on a different property of proteins can find the most reliable interactions (for example, Deane *et al.*[53] identified ~1,400 interactions of yeast proteins that are likely to be correct). Computational analysis also indicates that experimental corroboration of protein interactions by a combination of methods is likely to yield data that are substantially more reliable[54,56,57]. Finally, the large number of false negatives in proteomic studies suggests that most of the studies completed so far are far from saturated and that the universe of protein–protein interactions is likely to be several times higher than those currently known.

The principle of using hybrid proteins to analyse interactions has been extended to examine DNA–protein interactions, RNA–protein interactions, small-molecule–protein interactions, and interactions dependent on bridging proteins or post-translational modifications[58]. Additionally, the reconstitution of proteins other than transcription factors, such as ubiquitin, has been used to establish reporter systems to detect interactions[58], and these may enable the analysis of proteins not generally suitable for the traditional two-hybrid assay, such as membrane proteins. Although some of these alternative methods may be robust enough for high-throughput proteomic analysis, so far most of these approaches have been demonstrated only for their initial proof of principle, or in screens of a small number of proteins.

Analysing protein interactions by fluorescence methods

Another potentially general method to detect protein–protein interactions involves the use of fluorescence resonance energy transfer (FRET) between fluorescent tags on interacting proteins. FRET is a non-radiative process whereby energy from an excited donor fluorophore is transferred to an acceptor fluorophore that is within ~60 Å of the excited fluorophore[59]. After excitation of the first fluorophore, FRET is detected either by emission from the second fluorophore using appropriate filters, or by alteration of the fluorescence lifetime of the donor. Two fluorophores that are commonly used are variants of green fluorescent protein (GFP): cyan fluorescent protein (CFP) and yellow fluorescent protein (YFP)[60]. A number of protein interactions have been demonstrated in cells by FRET microscopy[59], including oligomerization of the Fas receptor[61], interaction between the apoptosis-regulating proteins Bcl-2 and Bax in mitochondria[62], and interaction between Pit-1 and Ets-1 transcription factors in the nucleus[63].

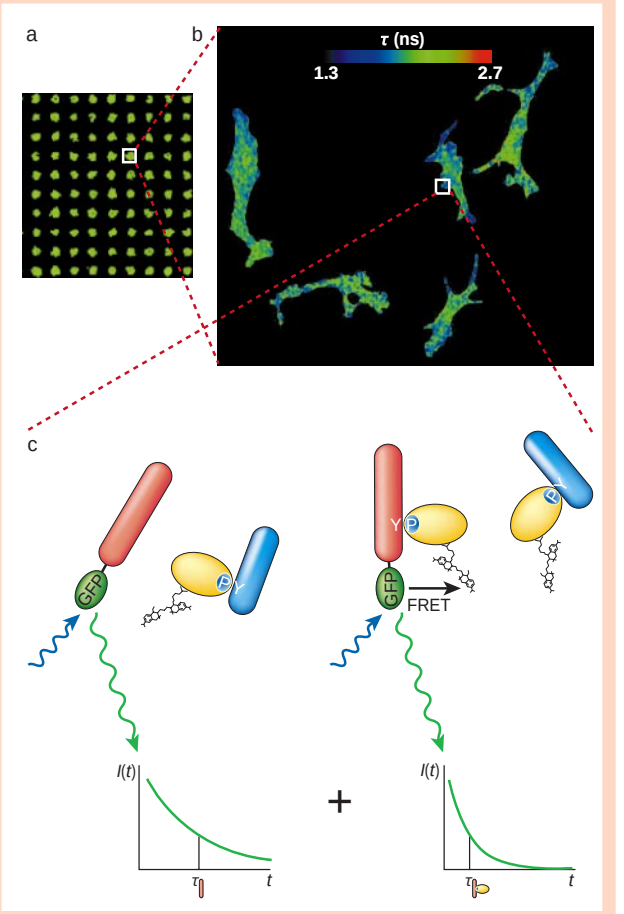


Figure 3 Principle of optical detection of protein post-translational modifications on a cell microarray. **a**, Local transfection techniques enable the expression of defined green fluorescent protein (GFP)–cDNA in clusters of cells at each location. **b**, Fluorescence lifetime imaging microscopy on the cell clusters reveals fluorescence resonance energy transfer (FRET) between GFP and an acceptor dye on a protein module or antibody directed against phosphorylated tyrosines by a drop in the fluorescence lifetime (τ) of the GFP (blue areas). **c**, Binding of the acceptor-tagged protein module or antibody (yellow) to the phosphorylated amino acid (YP) is detected only on the GFP fusion protein (red) via FRET and not on other phosphorylated proteins (blue). The extent of phosphorylation on the GFP fusion protein at each optically resolvable volume element in the cell can be determined from the linear combination of the GFP fluorescence decays in the presence and absence of FRET.

The potential of FRET is considerable, for two reasons. First, it can be used to make measurements in living cells, which allows the detection of protein interactions at the location in the cell where they normally occur, in the presence of the normal cellular milieu. For example, inducible interactions have been demonstrated, such as the binding of Grb2 to activated epidermal growth factor receptors[64] and the hormone-induced binding of co-activator proteins to nuclear receptors[65]. Second, transient interactions can be followed with high temporal resolution in single cells.

In principle, one can imagine two classes of high-throughput FRET screens that might be used. First, protein interactions within the proteome might be mapped by performing FRET screens on cell

arrays that are co-transfected with complementary DNAs bearing CFP and YFP fusion proteins. In practice, however, this may be difficult because of the high incidence of false negatives. These can arise from the lack of proper geometric orientation for FRET detection, and from the low FRET contributions in the fluorescence signals, which are difficult to detect above the background fluorescence from direct acceptor excitation or donor emission, particularly when expression levels of donor and acceptor tagged proteins are unbalanced.

Second, post-translational modifications might be detected by challenging GFP–cDNA donors with a FRET acceptor-tagged protein specific for that class of modification[59]. For example, cell microarrays expressing GFP–cDNA fusion libraries can be permeabilized and incubated with an anti-phosphotyrosine antibody conjugated to a FRET acceptor to measure tyrosine phosphorylation of any of the GFP fusion proteins by fluorescence lifetime imaging of the donor[59]. This approach allows specific detection of the signal even though the antibody binds all phosphotyrosine-containing proteins (Fig. 3). And because the acceptor fluorescence is filtered out in this approach, it permits the use of saturating amounts of labelled acceptor molecules. To boost the signal from such an experiment, the FRET acceptor-tagged protein can be tagged with several acceptor fluorophores. If this method becomes practical, similar approaches could be used to monitor other post-translational modifications.

Protein localization

A proteomics strategy of increasing importance involves the localization of proteins in cells as a necessary first step towards understanding protein function in complex cellular networks. A proteome-scale analysis of protein localization has been performed in *S. cerevisiae* by immunolocalization of epitope-tagged gene products[66]. These experiments established the subcellular localization of 2,744 proteins, 955 of which had no previously known function. The data were integrated with those previously published to identify the localization of 55% of the yeast proteome, which was extended to the full proteome by using a Bayesian estimation system[66]. This study corroborated that there is a good correlation between protein function and localization in the cell.

The discovery of GFP and the development of its spectral variants[60] has opened the door to analysis of proteins in living cells by use of the light microscope. Large-scale approaches of localizing GFP-tagged proteins in cells have been performed in the genetically amenable yeast *S. pombe*[67,68] and in *Drosophila*[69]. For the localization of proteins in mammalian cells, a strategy was developed that enables the systematic GFP tagging of ORFs from novel full-length cDNAs that are identified in genome projects[70]. This approach proved remarkably successful, showing a high correlation between prediction and the subsequent subcellular localization of targeted proteins, and could be fully automated.

The parallel functional analysis of many proteins in cells has become possible by a microarray-driven gene expression system[71]. In this system, mammalian cells are cultured on glass slides printed in defined locations with different DNAs specifying, for example, different defined cDNA–GFP fusions. The local transfections of cells growing over the DNA spots allow the simultaneous observation of many different fusion constructs, which can be correlated with the coordinates to link the images with the identity of any particular DNA. In principle, this approach can be applied to steady-state imaging for localization, to dynamic imaging to monitor changes during signal transduction, and to FRET to monitor changes in interactions.

Outlook

The promise of proteomics is the precise definition of the function of every protein in the cell, and how that function changes in different environmental conditions, with different modification states of the protein, in different cellular locales, and with different interacting partners. Just in the past few years, tremendous progress has been made in dissecting the functions of proteins using a battery of newly developed, sophisticated genome-wide approaches. Yet there is still a need both for additional high-throughput technologies and for computational methods to analyse large data sets and to integrate complex and disparate kinds of protein information. Another challenge will be for the proteomics community to work hand in hand with those focused on biological problems in order to best convert the broad but shallow proteomic data into deeper understanding. Within the next decade, we might have a reasonably complete picture of the proteome of a simple model organism such as yeast. This picture, in turn, will provide a blueprint for understanding the proteomes of other more complex model organisms and of humans. □

doi:10.1038/nature01512

1. Kenyon, G. L. *et al.* Defining the mandate of proteomics in the post-genomics era. Workshop Report: National Academy of Sciences, Washington DC, USA. *Mol. Cell. Proteomics* **1,** 763–780 (2002).
2. Cliften, P. F. *et al.* Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* **11,** 1175–1186 (2001).
3. Kumar, A. *et al.* An integrated approach for finding overlooked genes in yeast. *Nature Biotechnol.* **20,** 58–63 (2002).
4. Hudson, J. R. Jr *et al.* The complete set of predicted genes from *Saccharomyces cerevisiae* in a readily usable form. *Genome Res.* **7,** 1169–1173 (1997).
5. Reboul, J. *et al.* Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. *Nature Genet.* **27,** 332–336 (2001).
6. Braun, P. *et al.* Proteome-scale purification of human proteins from bacteria. *Proc. Natl Acad. Sci. USA* **99,** 2654–2659 (2002).
7. Nilsson, J., Stahl, S., Lundeberg, J., Uhlen, M. & Nygren, P. A. Affinity fusion strategies for detection, purification, and immobilization of recombinant proteins. *Protein Exp. Purif.* **11,** 1–16 (1997).
8. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293,** 2101–2105 (2001).
9. Martzen, M. R. *et al.* A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286,** 1153–1155 (1999).
10. Xing, F., Martzen, M. R. & Phizicky, E. M. A conserved family of *Saccharomyces cerevisiae* synthases effects dihydrouridine modification of tRNA. *RNA* **8,** 370–381 (2002).
11. Alexandrov, A. V., Martzen, M. R. & Phizicky, E. M. Two proteins that form a complex are required for 7-methylguanosine modification of yeast tRNA. *RNA* **8,** 1253–1266 (2002).
12. Hazbun, T. R. & Fields, S. A genome-wide screen for site-specific DNA-binding proteins. *Mol. Cell. Proteomics* **1,** 538–543 (2002).
13. Phizicky, E. M. *et al.* Biochemical genomics approach to map activities to genes. *Methods Enzymol.* **350,** 546–559 (2002).
14. Grayhack, E. J. & Phizicky, E. M. Genomic analysis of biochemical function. *Curr. Opin. Chem. Biol.* **5,** 34–39 (2001).
15. MacBeath, G. & Schreiber, S. L. Printing proteins as microarrays for high-throughput function determination. *Science* **289,** 1760–1763 (2000).
16. Zhu, H. *et al.* Analysis of yeast protein kinases using protein chips. *Nature Genet.* **26,** 283–289 (2000).
17. Haab, B. B., Dunham, M. J. & Brown, P. O. Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol.* **2,** RESEARCH0004.1–0004.13 (2001).
18. Zhu, H. & Snyder, M. Protein arrays and microarrays. *Curr. Opin. Chem. Biol.* **5,** 40–45 (2001).
19. Weng, S. *et al.* Generating addressable protein microarrays with PROfusion covalent mRNA- protein fusion technology. *Proteomics* **2,** 48–57 (2002).
20. Templin, M. F. *et al.* Protein microarray technology. *Trends Biotechnol.* **20,** 160–166 (2002).
21. Myszka, D. G. & Rich, R. L. Implementing surface plasmon resonance biosensors in drug discovery. *Pharmacol. Sci. Technol. Today* **3,** 310–317 (2000).
22. Houseman, B. T., Huh, J. H., Kron, S. J. & Mrksich, M. Peptide chips for the quantitative evaluation of protein kinase activity. *Nature Biotechnol.* **20,** 270–274 (2002).
23. LeProust, E. *et al.* Digital light-directed synthesis. A microarray platform that permits rapid reaction optimization on a combinatorial basis. *J. Comb. Chem.* **2,** 349–354 (2000).
24. Wang, D., Liu, S., Trummer, B. J., Deng, C. & Wang, A. Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. *Nature Biotechnol.* **20,** 275–281 (2002).
25. Kuruvilla, F. G., Shamji, A. F., Sternson, S. M., Hergenrother, P. J. & Schreiber, S. L. Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. *Nature* **416,** 653–657 (2002).
26. Sreekumar, A. *et al.* Profiling of cancer cells using protein microarrays: discovery of novel radiation-regulated proteins. *Cancer Res.* **61,** 7585–7593 (2001).
27. Belov, L., de la Vega, O., dos Remedios, C. G., Mulligan, S. P. & Christopherson, R. I. Immunophenotyping of leukemias using a cluster of differentiation antibody microarray. *Cancer Res.* **61,** 4483–4489 (2001).
28. Joos, T. O. *et al.* A microarray enzyme-linked immunosorbent assay for autoimmune diagnostics. *Electrophoresis* **21,** 2641–2650 (2000).
29. Robinson, W. H. *et al.* Autoantigen microarrays for multiplex characterization of autoantibody responses. *Nature Med.* **8,** 295–301 (2002).
30. Lee, T. I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298,** 799–804 (2002).
31. Faleiro, L., Kobayashi, R., Fearnhead, H. & Lazebnik, Y. Multiple species of CPP32 and Mch2 are the major active caspases present in apoptotic cells. *EMBO J.* **16,** 2271–2281 (1997).
32. Cravatt, B. F. & Sorensen, E. J. Chemical strategies for the global analysis of protein function. *Curr. Opin. Chem. Biol.* **4,** 663–668 (2000).
33. Adam, G. C., Sorensen, E. J. & Cravatt, B. F. Chemical strategies for functional proteomics. *Mol. Cell. Proteomics* **1,** 781–790 (2002).

34. Martins, L. M. *et al.* Activation of multiple interleukin-1β converting enzyme homologues in cytosol and nuclei of HL-60 cells during etoposide-induced apoptosis. *J. Biol. Chem.* **272,** 7421–7430 (1997).
35. Greenbaum, D., Medzihradszky, K. F., Burlingame, A. & Bogyo, M. Epoxide electrophiles as activity-dependent cysteine protease profiling and discovery tools. *Chem. Biol.* **7,** 569–581 (2000).
36. Liu, Y., Patricelli, M. P. & Cravatt, B. F. Activity-based protein profiling: the serine hydrolases. *Proc. Natl Acad. Sci. USA* **96,** 14694–14699 (1999).
37. Kidd, D., Liu, Y. & Cravatt, B. F. Profiling serine hydrolase activities in complex proteomes. *Biochemistry* **40,** 4005–4015 (2001).
38. Greenbaum, D. *et al.* Chemical approaches for functionally probing the proteome. *Mol. Cell. Proteomics* **1,** 60–68 (2002).
39. Adam, G. C., Sorensen, E. J. & Cravatt, B. F. Proteomic profiling of mechanistically distinct enzyme classes using a common chemotype. *Nature Biotech.* **20,** 805–809 (2002).
40. Jessani, N., Liu, Y., Humphrey, M. & Cravatt, B. F. Enzyme activity profiles of the secreted and membrane proteome that depict cancer cell invasiveness. *Proc. Natl Acad. Sci. USA* **99,** 10335–10340 (2002).
41. Bishop, A. C. *et al.* A chemical switch for inhibitor-sensitive alleles of any protein kinase. *Nature* **407,** 395–401 (2000).
42. Habelhah, H. *et al.* Identification of new JNK substrate using ATP pocket mutant JNK and a corresponding ATP analogue. *J. Biol. Chem.* **276,** 18090–18095 (2001).
43. Bishop, A. C., Buzko, O. & Shokat, K. M. Magic bullets for protein kinases. *Trends Cell Biol.* **11,** 167–172 (2001).
44. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415,** 180–183 (2002).
45. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147 (2002).
46. Fields, S. & Song, O. A novel genetic system to detect protein–protein interactions. *Nature* **340,** 245–246 (1989).
47. Fromont-Racine, M. *et al.* Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast* **17,** 95–110 (2000).
48. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403,** 623–627 (2000).
49. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98,** 4569–4574 (2001).
50. Rain, J. C. *et al.* The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409,** 211–215 (2001).
51. Walhout, A. J. M. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287,** 116–122 (2000).
52. Mrowka, R., Patzak, A. & Herzel, H. Is there a bias in proteome research? *Genome Res.* **11,** 1971–1973 (2001).
53. Deane, C. M., Salwinski, L., Xenarios, I. & Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* **1,** 349–356 (2002).
54. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **12,** 37–46 (2002).
55. Kemmeren, P. *et al.* Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell* **9,** 1133–1143 (2002).
56. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417,** 399–403 (2002).
57. Edwards, A. *et al.* Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.* **18,** 529–536 (2002).
58. Fashena, S. J., Serebriiskii, I. & Golemis, E. A. The continued evolution of two-hybrid screening approaches in yeast: how to outwit different preys with different baits. *Gene* **250,** 1–14 (2000).
59. Wouters, F. S., Verveer, P. J. & Bastiaens, P. I. H. Imaging biochemistry inside cells. *Trends Cell Biol.* **11,** 203–211 (2001).
60. Tsien, R. Y. The green fluorescent protein. *Annu. Rev. Biochem.* **67,** 509–544 (1998).
61. Siegel, R. M. *et al.* Fas preassociation required for apoptosis signaling and dominant inhibition by pathogenic mutations. *Science* **288,** 2354–2357 (2000).
62. Mahajan, N. *et al.* Bcl-2 and Bax interactions in mitochondria probed with green fluorescent protein and fluorescence resonance energy transfer. *Nature Biotechnol.* **16,** 547–552 (1998).
63. Day, R. N. Visualization of Pit-1 transcription factor interactions in the living cell nucleus by fluorescence resonance energy transfer microscopy. *Mol. Endocrinol.* **12,** 1410–1419 (1998).
64. Sorkin, A., McClure, M., Huang, F. & Carter, R. Interaction of EGF receptor and Grb2 in living cells visualized by fluorescence resonance energy transfer (FRET) microscopy. *Curr. Biol.* **10,** 1395–1398 (2000).
65. Llopis, J. *et al.* Ligand-dependent interactions of coactivators steroid receptor coactivator-1 and peroxisome proliferator-activated receptor binding protein with nuclear hormone receptors can be imaged in live cells and are required for transcription. *Proc. Natl Acad. Sci. USA* **97,** 4363–4368 (2000).
66. Kumar, A. *et al.* Subcellular localization of the yeast proteome. *Genes Dev.* **16,** 707–719 (2002).
67. Ding, D. Q. *et al.* Large-scale screening of intracellular protein localization in living fission yeast cells by the use of a GFP-fusion genomic DNA library. *Genes Cells* **5,** 169–190 (2000).
68. Sawin, K. E. & Nurse, P. Identification of fission yeast nuclear markers using random polypeptide fusion with green fluorescent protein. *Proc. Natl Acad. Sci. USA* **94,** 15146–15151 (1996).
69. Morin, X., Daneman, R., Zavortink, M. & Chia, W. A protein trap strategy to detect GFP-tagged proteins expressed from their endogenous loci in *Drosophila*. *Proc. Natl Acad. Sci. USA* **98,** 15050–15055 (2001).
70. Simpson, J. C., Wellenreuther, R., Poustka, A., Pepperkok, R. & Wiemann, S. Systematic subcellular localisation of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep.* **1,** 287–292 (2000).
71. Ziauddin, J. & Sabatini, D. M. Microarrays of cells expressing defined cDNAs. *Nature* **411,** 107–110 (2001).
72. Aslanidis, C. & de Jong, P. J. Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res.* **18,** 6069–6074 (1990).
73. Aslanidis, C., de Jong, P. J. & Schmitz, G. Minimal length requirement of the single-stranded tails for ligation-independent cloning (LIC) of PCR products. *PCR Methods Appl.* **4,** 172–177 (1994).
74. Dieckman, L., Gu, M., Stols, L., Donnelly, M. I. & Collart, F. R. High throughput methods for gene cloning and expression. *Protein Exp. Purif.* **25,** 1–7 (2002).
75. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17,** 1030–1032 (1999).
76. Kapust, R. B. & Waugh, D. S. *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* **8,** 1668–1674 (1999).

Acknowledgements We thank T. Davis and E. Grayhack for comments on the manuscript. This work was supported by grants from the National Center for Research Resources and National Human Genome Research Institute of the National Institutes of Health. S.F. is an investigator of the Howard Hughes Medical Institute.