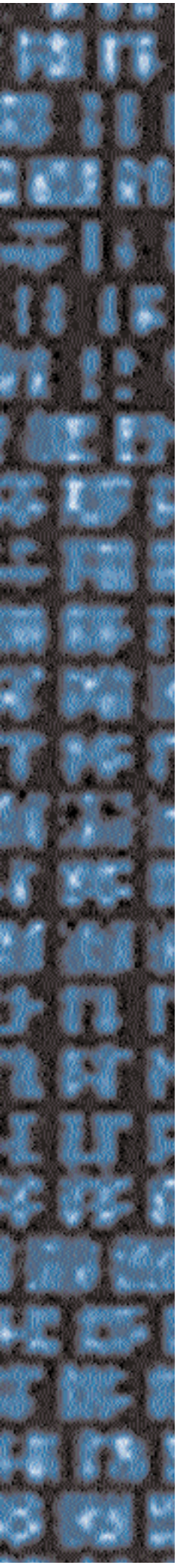

Proteomics to study genes and genomes

Akhilesh Pandey* & Matthias Mann†

**Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, and Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA*
†Protein Interaction Laboratory (PIL), University of Southern Denmark, Campusvej 55, DK-5230 Odense M, and Denmark and Protana A/S, Staermosegaardsvej 16, DK-5230 Odense M, Denmark (e-mail: mann@cebi.sdu.dk)

Proteomics, the large-scale analysis of proteins, will contribute greatly to our understanding of gene function in the post-genomic era. Proteomics can be divided into three main areas: (1) protein micro-characterization for large-scale identification of proteins and their post-translational modifications; (2) 'differential display' proteomics for comparison of protein levels with potential application in a wide range of diseases; and (3) studies of protein–protein interactions using techniques such as mass spectrometry or the yeast two-hybrid system. Because it is often difficult to predict the function of a protein based on homology to other proteins or even their three-dimensional structure, determination of components of a protein complex or of a cellular structure is central in functional analysis. This aspect of proteomic studies is perhaps the area of greatest promise. After the revolution in molecular biology exemplified by the ease of cloning by DNA methods, proteomics will add to our understanding of the biochemistry of proteins, processes and pathways for years to come.

Large-scale DNA sequencing has transformed biomedical research in a short span of time. With the discovery of most human genes, it is now apparent that a 'factory approach' to address biological problems is desirable if we are to gain a comprehensive understanding of complex biological processes. In this article we will review how proteomics is similarly making a crucial contribution to our understanding of biology and medicine through the global analysis of gene products.

Defining proteomics

Proteomics is the large-scale study of proteins, usually by biochemical methods. The word proteomics has been associated traditionally with displaying a large number of proteins from a given cell line or organism on two-dimensional polyacrylamide gels[1–4]. In this sense proteomics already dates back to the late 1970s when researchers started to build databases of proteins using the then newly developed technique of two-dimensional gel electrophoresis[5] (Box 1). This resulted in extensive cataloguing of spots from two-dimensional gels to create databases of all expressed proteins. However, even when such gels could be run reproducibly between laboratories, determining the identity of the proteins was difficult because of a lack of sensitive and rapid analytical methods for protein characterization (such as the polymerase chain reaction and the automated sequencer for DNA analysis). In the 1990s, biological mass spectrometry emerged as a powerful analytical method that removed most of the limitations of protein analysis. This development, coupled with the availability of the entire human coding sequence in public databases, marks the beginning of a new era. Today, the term proteomics covers much of the functional analysis of gene products or 'functional genomics', including large-scale identification or localization studies of proteins and interaction studies using the yeast two-hybrid system. The more focused large-scale study of protein structure, however, is usually not included and designated 'structural genomics' instead[6]. Likewise, strategies that target only genes or messenger RNA, such as large-scale mutagenesis or antisense experiments, should not be considered part of proteomics.

Why is proteomics necessary?

With the accumulation of vast amounts of DNA sequences in databases, researchers are realizing that merely having complete sequences of genomes is not sufficient to elucidate biological function. A cell is normally dependent upon a multitude of metabolic and regulatory pathways for its survival. There is no strict linear relationship between genes and the protein complement or 'proteome' of a cell. Proteomics is complementary to genomics because it focuses on the gene products, which are the active agents in cells. For this reason, proteomics directly contributes to drug development as almost all drugs are directed against proteins.

The existence of an open reading frame (ORF) in genomic data does not necessarily imply the existence of a functional gene. Despite the advances in bioinformatics, it is still difficult to predict genes accurately from genomic data (see review in this issue by Eisenberg *et al.*, pages 823–826, and refs 7, 8). Although the sequencing of related organisms will ease the problem of gene prediction through comparative genomics, the success rate for correct prediction of the primary structure is still low[9,10]. This is particularly true in the case of small genes (which can be missed entirely) or genes with little or no homology to other known genes. A recent study concluded that the error rate was as least 8% in the annotations for 340 genes from the *Mycoplasma genitalium* genome[11]. If such error rates are extrapolated to the human genome, the outcome and consequences can easily be imagined. Therefore, verification of a gene product by proteomic methods is an important first step in 'annotating the genome'. Modifications of the proteins that are not apparent from the DNA sequence, such as isoforms and

post-translational modifications, can be determined only by proteomic methodologies. Furthermore, it may be necessary to determine the protein expression level directly as mRNA levels may or may not correlate with the protein level[12,13]. The localization of gene products, which is often difficult to predict from the sequence, can be determined experimentally. Mechanisms such as regulation of protein function by proteolysis, recycling and sequestration in cell compartments affect gene products and not genes. Finally, protein–protein interactions and the molecular composition of cellular structures such as organelles can be determined only at the protein level.

Identification and analysis of proteins

Protein preparation methods

One of the most crucial steps in proteomics is obtaining and handling the protein sample. Out of the entire complement of the genome of about 100,000 genes, a given cell line may express about 10,000 genes and an even higher number is expressed in tissues. Furthermore, the dynamic range of abundance of proteins in biological samples can be as high as 10^6. Because even the best two-dimensional gels can routinely resolve no more than 1,000 proteins, it is obvious that only the most abundant proteins can be visualized by gel electrophoresis if a crude protein mixture is used. The ideal solution to reduce complexity and differences in abundance is to use affinity-based protein purification strategies using the whole protein complement. For example, the erythropoeitin receptor is of medium abundance, occurring in about 1,000 copies per cell, or less than two picomoles (100 ng) in one litre of cell culture. This protein would not be visualized from whole-cell extracts but can be enriched easily by antibody-based affinity purification to yield a silver-stained band. This fact has to be borne in mind if signalling and other regulatory molecules are being studied.

After obtaining the protein fraction, the method of choice for proteomic studies is one- or two-dimensional gel electrophoresis. The advantages of one-dimensional electrophoresis as a preparation method are that virtually all proteins are soluble in SDS, the range of relative molecular mass from 10,000 to 300,000 is readily covered, and extremely acidic and basic proteins are easily visualized.

Mass spectrometric identification of proteins

The most significant breakthrough in proteomics has been the mass spectrometric identification of gel-separated proteins, which extends analysis far beyond the mere display of proteins. Mass spectrometry has essentially replaced the classical technique of Edman degradation even in traditional protein chemistry, because it is much more sensitive, can deal with protein mixtures and offers much higher throughput. It relies on digestion of gel-separated proteins into peptides by a sequence-specific protease such as trypsin. The reason for analysing peptides rather than proteins is that gel-separated proteins are difficult to elute and to analyse by mass spectrometry, and that the molecular weight of proteins is not usually sufficient for database identification. In contrast, peptides are easily eluted from gels and even a small set of peptides from a protein provides sufficient information for identification. The steps typically involved in the mass spectrometric analysis of a protein are illustrated by an example that shows analysis of a molecule involved in platelet-derived growth factor (PDGF) signalling (Fig. 1). A detailed protocol describing methods and strategies for the mass spectrometric identification of signalling molecules can be found in ref. 14.

There are two main approaches to mass spectrometric protein identification. In the 'peptide-mass mapping' approach, initially suggested by Henzel and co-workers[15], the mass spectrum of the eluted peptide mixture is acquired, which results in a 'peptide-mass fingerprint' of the protein being studied. This mass spectrum is obtained by a relatively simple mass spectrometric method — matrix-assisted laser desorption/ionization (MALDI) — which results in a time-of-flight distribution of the peptides comprising the mixture (Box 2 and Fig. 1b). Advances have been made in automation of the MALDI identification procedure whereby hundreds of protein spots can be excised, digested enzymatically, their mass spectra obtained and automatically searched against databases[16,17]. As more full-length human genes are represented in the database, the success rate of identification by MALDI will increase further.

In a two-step procedure for rapid and unambiguous protein identification, MALDI fingerprinting is the first step[18]. The second method for protein identification relies on fragmentation of individual peptides in the mixture to gain sequence information. In this method, the peptides are ionized by 'electrospray ionization' directly from the liquid phase. The peptide ions are sprayed into a 'tandem mass spectrometer' which has the ability to resolve peptides in a mixture, isolate one species at a time and dissociate it into amino- or carboxy-terminal-containing fragments (Fig. 1c). The tandem mass spectrometric method is technically more complex and less scalable than MALDI fingerprinting. Its main advantage is that sequence information derived from several peptides is much more specific for the identification of a protein than a list of peptide masses. The fragmentation data can not only be used to search protein sequence databases but also nucleotide databases such as expressed sequence

Box 1
Defining proteomics

Proteomics – the classical definition

- Two-dimensional gels of cell lysates and annotation
- Two-dimensional gels to visualize differential protein expression

Proteomics – in the post-genomics era

- Protein identification:
  - One-dimensional gels (for example, analysis after affinity purification)
  - Two-dimensional gels (for example, analysis after affinity purification, body fluids, etc.)
  - Protein chips (chips coated with, for example, proteins or antibodies)
  - Proteins/protein complexes in solution (identification without electrophoresis)
- Post-translational modifications
  - Phosphorylation
  - Glycosylation
- Determining Function
  - Assays for enzymatic activity or determining substrates[75]
  - Bioassays for cytokines, receptor/ligand-binding assays
  - Localization within the cells (GFP fusions)
  - Proteomic analysis using large-scale mouse knockouts[76] or RNA interference[77].
  - Phenotypic analysis using deletion strains[78]
- Molecular Medicine (no longer just pharmaceuticals)
  - Finding molecular (protein) drug targets
  - Disrupting protein–protein interactions using drugs
  - Large-scale animal assays for recombinant proteins, antibodies and inhibitors
- Differential display by two-dimensional gels (superseded by DNA-based array in many situations)
  - Limited applications in:
    - Body fluids (for example, serum and urine)
    - Variants resulting from post-translational modifications
- Protein–protein interaction
  - Direct DNA readout
    - Yeast two hybrid
    - Phage display
    - Ribosome display[79]
    - RNA–peptide fusions[80]
  - Protein identification
    - Affinity purification and mass spectrometry

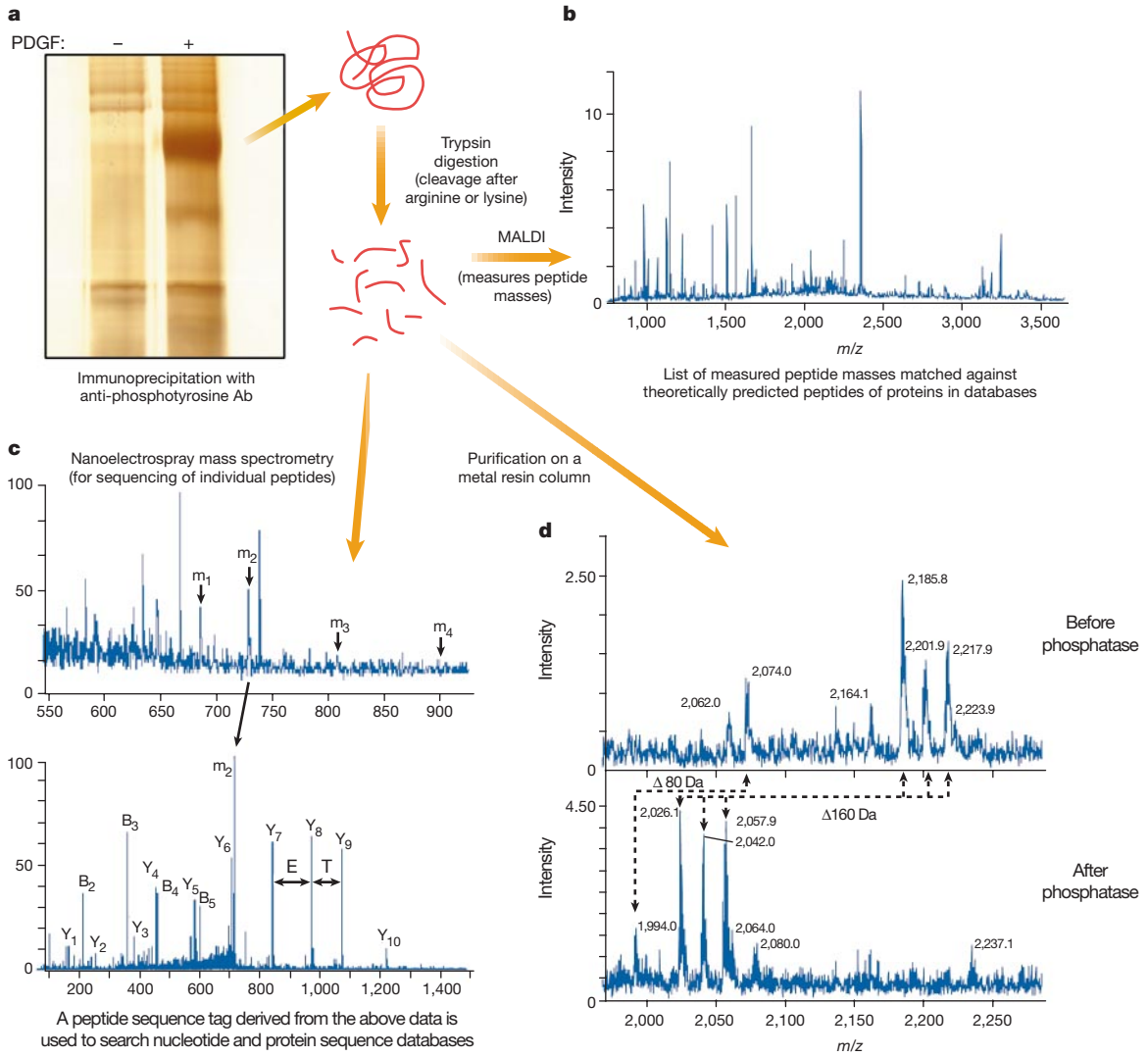

Figure 1 A strategy for mass spectrometric identification of proteins and post-translational modifications. **a**, Responsive cells such as NIH 3T3 fibroblasts are treated with PDGF followed by immunoprecipitation of cell lysates with anti-phosphotyrosine antibodies. After one-dimensional gel electrophoresis, the gel is silver stained, the protein band excised as shown and subjected to digestion with trypsin. This results in peptides with arginine or lysine at their C termini as a result of the cleavage specificity of trypsin. **b**, An aliquot of the supernatant containing tryptic peptides is analysed by MALDI, which results in a peptide-mass fingerprint of the protein. **c**, The remainder of the supernatant is desalted and analysed by nanoelectrospray tandem mass spectrometry. The top panel shows the individual peptide peaks in the mass spectrum. The bottom panel shows how sequence can be derived by fragmentation of the chosen peptide (m_2) by tandem mass spectrometry. **d**, The phosphopeptides may be enriched by purifying the peptide mixture over a metal resin microcolumn. The resulting peptides can then be analysed by MALDI as shown (and subsequently by nanoelectrospray) before and after treatment with alkaline phosphatase. The panel shows a singly phosphorylated (showing a shift of 80 Da) and a doubly phosphorylated (showing a shift of 160 Da) peptide in the MALDI spectrum. (Fig. 1d courtesy of O. N. Jensen and A. Stensballe.)

tag (EST) databases and more recently even raw genomic sequence databases (B. Küster, P. Mortensen, J. S. Andersen and M. Mann, unpublished data).

New developments in mass spectrometry

Biological mass spectrometry is still evolving rapidly owing to continued technological advances in various areas. For instance, a new type of mass spectrometer that combines a MALDI ion source with a highly efficient tandem mass spectrometer unit that can fragment the individual peptides has recently been developed[19]. If this 'MALDI quadruple time of flight' instrument proves to be sufficiently sensitive, it would combine the high throughput of the peptide mapping method with the specificity of the peptide sequencing method, allowing a one-step instead of a two-step mass spectrometric analysis strategy. In our experience, this instrument already significantly improves the analysis of small proteins and improves the throughput when analysing simple protein mixtures. There are also efforts at miniaturizing protein preparation using microfabricated 'chips', which have obtained promising results[20–22]. However, these methods have not yet yielded the sensitivity or robustness of preparations using standard tube or microtitre plate formats. There are also longstanding efforts to scan one- or two-dimensional gels directly by MALDI mass spectrometry[23,24]. A recent variation uses an intercalating membrane containing immobilized trypsin for digestion of proteins during electrophoretic transfer onto a collecting membrane. The membrane is then rasterized and analysed by MALDI yielding a peptide map for each position of the gel[25,26].

In the future, it would be desirable to analyse a protein sample directly by mass spectrometry, without gel separation or enzymatic digestion. Smith *et al.* have loaded crude protein extract into a capillary and performed capillary electrophoresis to separate the proteins

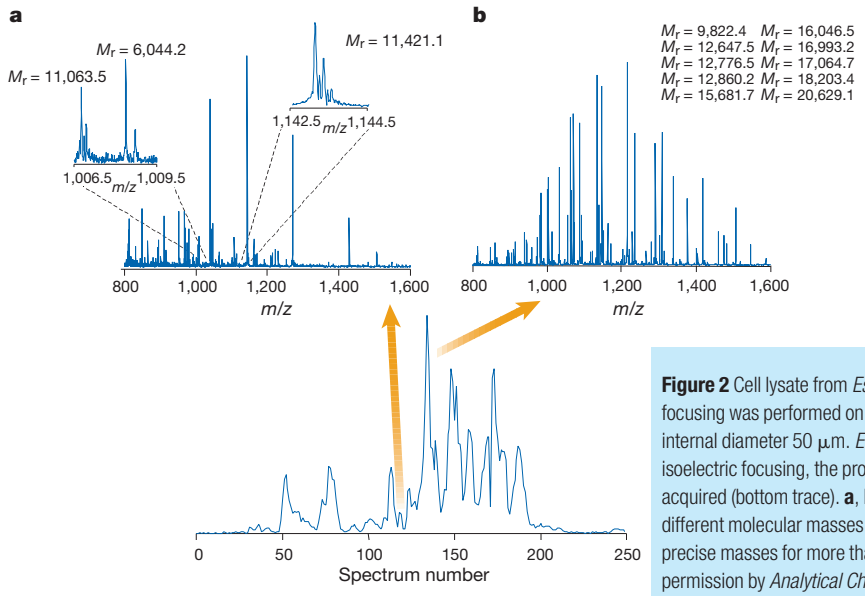

Figure 2 Cell lysate from *Escherichia coli* analysed by FTICR. Capillary isoelectric focusing was performed on ~300 ng *E. coli* total cell lysate in a coated capillary of internal diameter 50 μm. *E. coli* were grown in isotopically depleted medium. After isoelectric focusing, the proteins were eluted into the mass spectrometer and spectra acquired (bottom trace). **a**, High-resolution spectrum for charge states representing different molecular masses present in a single scan. **b**, Mass spectrum showing precise masses for more than ten co-eluting protein species. (Reprinted with permission by *Analytical Chemistry*.)

by their isoelectric point[27]. The separated proteins were then infused directly into a specialized Fourier-transformed ion cyclotron resonance (FTICR) mass spectrometer (Fig. 2), and the precise molecular masses of hundreds of proteins were acquired during a single run. In this experiment, the mass distribution was biased towards small proteins and only the masses, not the identity of the proteins, were determined. But in the future it may become possible to use this strategy to identify proteins by on-line fragmentation of the proteins[28,29]. This would enable researchers to perform the whole proteomic analysis in a single automated experiment at least for a subset of soluble proteins of medium abundance.

Post-translational modifications

One of the unique features of proteomics studies is the ability to analyse the post-translational modifications of proteins. Phosphorylation, glycosylation and sulphation as well as many other modifications are extremely important for protein function as they can determine activity, stability, localization and turnover. These modifications are not generally apparent from genomic sequence or mRNA expression data. Whereas mass spectrometry is the proteomic method of choice to determine protein modifications, this task is much more difficult than the mere determination of protein identity. Minimal data is sufficient to identify the protein in sequence databases — often as few as one or two peptides need to be fragmented. However, for obtaining the nature and location of post-translational modifications, all the peptides that do not have the expected molecular mass need to be analysed further. Because of this and other reasons, much more material is needed to study post-translational modifications than is required for protein

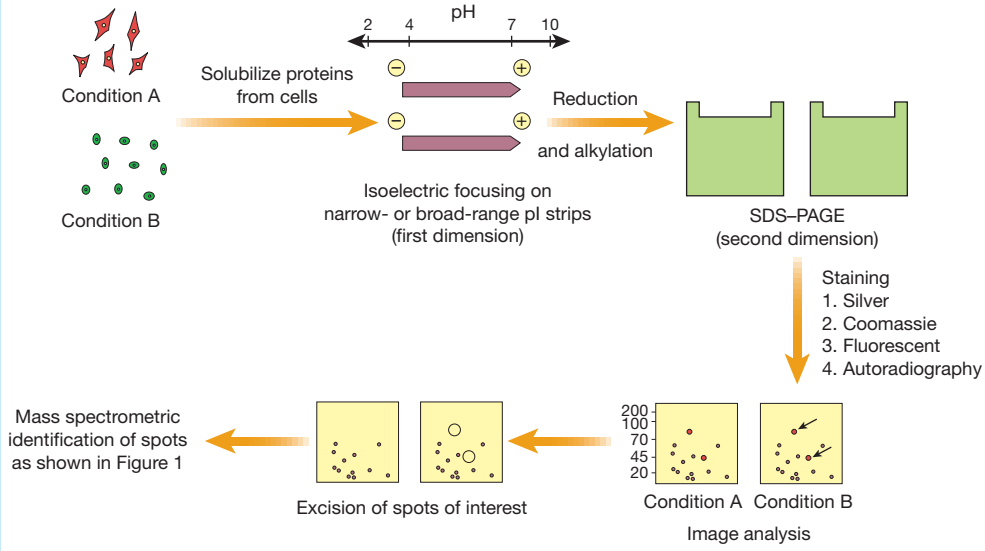

Figure 3 A schematic showing the two-dimensional gel approach. Cells (or tissue) derived from two different conditions, A and B, are harvested and the proteins solubilized. The crude protein mixture is then applied to a 'first dimension' gel strip that separates the proteins based on their isoelectric points. After this step, the strip is subjected to reduction and alkylation and applied to a 'second dimension' SDS–PAGE gel where proteins are denatured and separated on the basis of size. The gels are then fixed and the proteins visualized by silver staining. Silver staining is less quantitative than Coomassie blue but more sensitive and is also compatible with mass spectrometric analysis. After staining, the resulting protein spots are recorded and quantified. Image analysis requires sophisticated software and remains one of the most labour-intensive parts of the two-dimensional gel approach. The spots of interest are then excised and subjected to mass spectrometric analysis.

identification. Continuing progress is being made in this field, especially in the case of phosphorylation. Phosphorylation events can be studied by generic strategies, because phosphopeptides are 80 Da heavier than their unmodified counterparts, give rise to a specific fragment (PO^{3-}, mass 79), bind to metal resins, are recognized by specific antibodies and the phosphate groups can be removed by phosphatases[30–34]. As an example, Fig 1d shows the detection of phosphopeptides following metal resin-based affinity micropurification and phosphatase treatment.

Phosphorylation and signalling pathways

Several receptor-mediated signalling pathways result in tyrosine phosphorylation of a large set of substrates. To identify these substrates, the lysates from unstimulated and growth factor-stimulated cells can be prepared and resolved by two-dimensional gels. The proteins of interest can be detected by ^{32}P labelling or by western blotting with antibodies that recognize only the activated state of molecules (such as phosphotyrosine- or phosphoserine-specific antibodies). These spots can then be identified by mass spectrometry as demonstrated recently[35]. A better alternative, however, is to first enrich for these substrates by using anti-phosphotyrosine antibodies in an immunopreciptation step followed by mass spectrometric identification. Several known and new components were recently reported in one such study on the epidermal growth factor (EGF)-receptor pathway[36].

Differential-display proteomics

The two-dimensional gel approach

Until recently, proteomics was almost synonymous with two-dimensional gel electrophoresis (Fig. 3). In biomedical applications of the comparative two-dimensional gel approach, the objective is usually to identify proteins that are up- or downregulated in a disease-specific manner for use as diagnostic markers or therapeutic targets. There are several technical challenges in such experiments. First, hydrophobic and large proteins usually do not enter the second dimension of the gel. Second, the issue of dynamic range makes it difficult to visualize all but the most abundant proteins. Particularly in body fluids such as serum and cerebrospinal fluid, more than 99% of the protein complement consists of serum albumin and globulins. Third, because of the biological variation inherent in these samples, it is difficult to define normal protein-expression patterns that can be compared with the disease state. For several of these applications, methods of array-based mRNA expression profiling can not only be more comprehensive (as they provide data on all the genes applied to the chip), but also faster and more convenient, as shown by a number of studies (see review in this issue by Lockhart and Winzeler, pages 827–836, and refs 37–40).

In spite of these difficulties of comparing two-dimensional gel patterns, several applications have appeared in the literature. For example, Celis and co-workers have found a putative urinary marker, psoriasin, which can be used for the follow-up of patients with bladder squamous cell carcinomas[41]. This marker was identified when they compared the profile of secreted proteins from normal tissue with that from cancerous tissue. A similar study compared the proteome of normal human luminal and myoepithelial breast cells using immunopurified cell populations. It detected 170 protein spots that were twofold differentially expressed[42], of which 51 were identified. However, almost all of these proteins were abundant cytoskeletal proteins such as actin and keratin. A recent study compared the protein complement from different fractions of brain extracts from two different strains of mice[43], finding over 1,000 genetically variant protein spots. Such studies may be useful in other situations as well, for example, in comparing the proteome of wild-type with that of knockout mice. Toxicology studies frequently use proteomic analysis to understand the mechanism of action of a drug or to identify its targets. Aicher and colleagues discovered an association between decreased levels of a calcium-binding protein, calbindin-D 28K, and cyclosporine A-induced nephrotoxicity when kidney samples were compared from species that were either

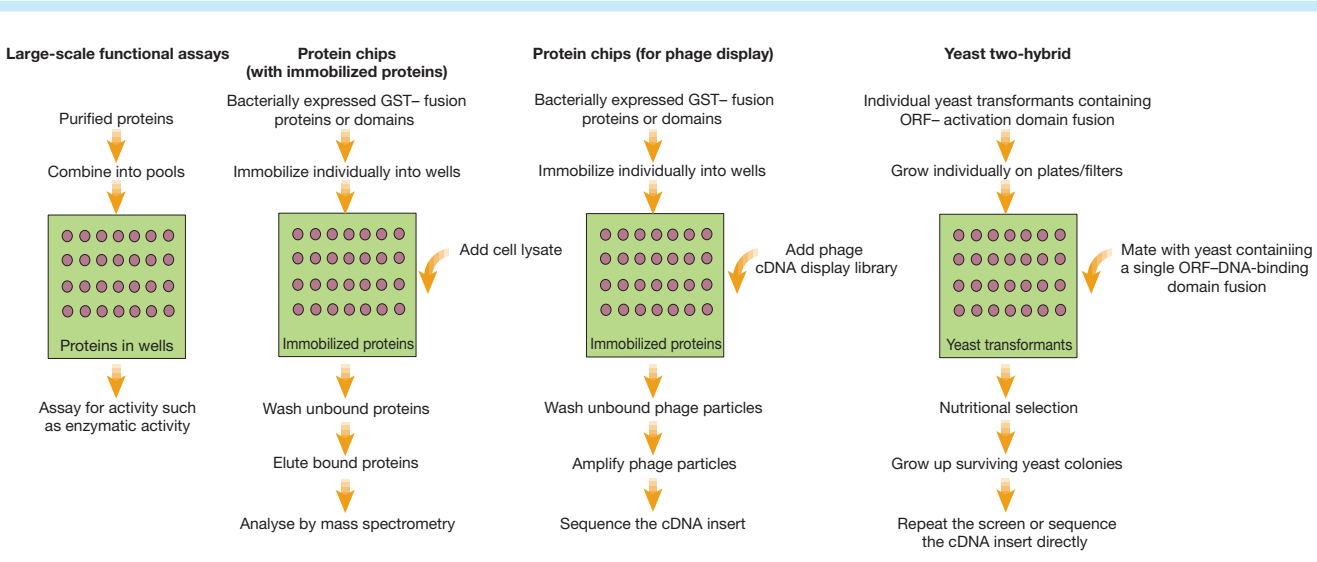

Figure 4 A schematic showing use of arrays for proteomic analysis. Recombinant proteins can be expressed and purified in a large-scale format. These proteins are pooled into wells as shown and assayed for functions such as enzymatic activity. This approach has been termed biochemical genomics. A protein chip can be prepared in several ways. The surface can be immobilized with recombinant proteins or their domains (such as bacterially expressed GST–fusion proteins) and then cell lysates containing putative interaction partners are applied to the chip followed by washing to remove unbound material. The bound proteins can then be eluted and identified by mass spectrometry. Alternatively, instead of cell lysates, a phage cDNA display library can be applied to the chip followed by washing and amplification steps to isolate individual interacting phage particles. The inserts in these phage particles can then be sequenced to determine the identity of the interacting partners. The yeast two-hybrid system is also amenable to an array-based analysis. First, yeast cells can be transformed with individual ORF–activation domain fusions. These cells can be grown in an array format on plates or filters such that each element of the array contains a yeast clone with a unique ORF. Such an array can be probed in a mating assay with yeast cells containing a single ORF–DNA-binding domain fusion, one at a time. The nutritional selection ensures that only the yeast cells containing interacting partners survive. These interacting clones can be re-screened to reduce false positives or be sequenced directly.

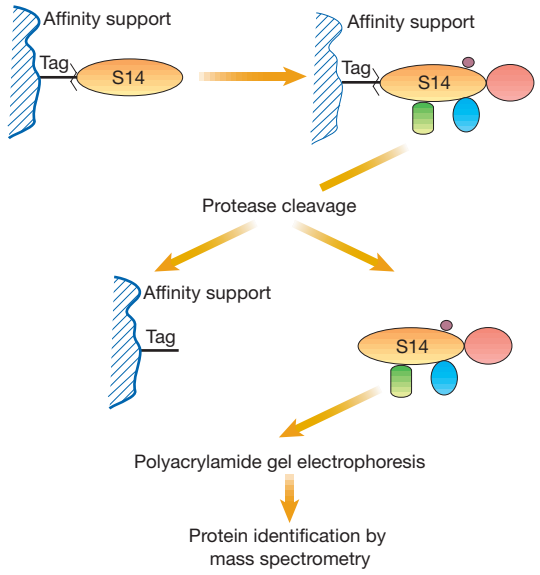

Figure 5 A generic strategy to isolate interacting proteins. The protein of interest is expressed as a fusion protein with a cleavable affinity tag to identify interacting proteins. In this case, S14 protein (spot S14 identified from gel shown in Fig. 6a) is immobilized onto agarose beads using a GST tag. Nuclear cell extracts are incubated with the beads and the beads washed extensively. Thrombin is used to cleave between the GST and the S14 protein, which results in elution of all proteins that are specifically bound to S14. The advantage of this method is that the proteins that are nonspecifically bound to the matrix or the tag itself are not eluted. The eluted proteins are resolved by one- or two-dimensional gel electrophoresis and compared to GST alone. The bands or spots corresponding to proteins specifically bound to the tagged proteins are excised and analysed by mass spectrometry. (Figure courtesy of A. King)

susceptible or resistant to nephrotoxicity[44].

When two-dimensional gels are used as a method of separating a qualitative subset of proteins, as opposed to comparing whole-cell preparations, or when immunological methods are used to highlight a subset of proteins, biologically relevant answers can be more readily obtained. For example, many secreted proteins can be identified by two-dimensional gel analysis of supernatants of cell lines and explants from tumour tissues[45]. Several groups have probed two-dimensional gels of proteins from allergy-causing organisms using antibodies derived from allergic patients[46,47]. Identification of the responsible allergen by mass spectrometry can be exploited in the rational design of preventive and therapeutic strategies.

We predict that protein expression analysis will be most useful in well-defined areas such as (1) analysis of samples that do not contain mRNA such as some body fluids; (2) cases where the protein abundance does not correlate with the mRNA abundance; (3) cases where the critical changes involve post-translational modifications of proteins such as glycosylation or phosphorylation, rather than changes in protein abundance; (4) cases where an overview of the most abundant proteins in a specialized source is itself of importance; and (5) cases where two-dimensional gels allow a relatively comprehensive overview of a simple proteome such as that of a microbe.

Protein chips

In the protein chip approach, a variety of 'bait' proteins such as antibodies can be immobilized in an array format onto specially treated surfaces (Fig. 4). The surface is then probed with the sample of interest and only the proteins that bind to the relevant antibodies remain bound to the chip[48]. Such an approach is essentially a large-scale version of enzyme-linked immunosorbent assays that are already used in clinical diagnostics. In one version, the protein chip is probed with fluorescently labelled proteins from two different cell states. Cell lysates are labelled by different fluorophores and mixed such that the colour acts as a readout for the change in abundance of the protein bound to the antibody. This system depends on reasonably specific and well-characterized antibodies and a number of technical problems would still need to be overcome. However, once developed it could provide convenient proteome analysis. In

Box 2

Mass spectrometric techniques in proteomics

MALDI and peptide-mass mapping

In this approach, a portion of the tryptic peptide mixture is analysed by MALDI mass spectrometry. Because trypsin cleaves the protein backbone at the amino acids arginine and lysine, the masses of tryptic peptides can be predicted theoretically for any entry in a protein sequence database. These predicted peptide masses are compared with those obtained experimentally by MALDI analysis. The protein can be identified correctly if there are a sufficient number of peptide matches with a protein in the database, resulting in a high score. High mass accuracy is critical for unambiguous identification and serves mainly to eliminate the false positives. MALDI identification by peptide-mass fingerprints requires that the full-length gene be present in the databases. Therefore, the success rate of this method will receive an additional boost with the availability of all predicted genes in sequence databases.

Electrospray and tandem mass spectrometry

There are two major mass spectrometric strategies that use electrospray ionization. In one of them the unseparated mixture of peptides is applied to a low-flow device called nanoelectrospray[81,82]. The peptide mixture is then electrosprayed from a very fine needle (tip internal diameter of 1 μm) into the mass spectrometer. Individual peptides from the mixture are isolated in the first step and fragmented during the second step to sequence the peptide (hence tandem mass spectrometry). The fragments obtained by this method are derived from the N or C terminus of the protein and are designated 'b' or 'y' ions, respectively[83]. The other strategy uses liquid chromatography for initial separation of peptides followed by sequencing as they elute into the electrospray ion source. This method can also be used without gel electrophoresis where a mixture of proteins is digested in solution and the 'scrambled' set of peptides are sequenced, ideally resulting in several peptide hits for each of the proteins that was initially present in the mixture[54,84]. A great deal of data can be obtained from a single run in an automated fashion.

Using mass spectrometry data to search databases

In tandem mass spectrometry, peptides are fragmented by collision with gas molecules in the mass spectrometer. Spacing of these fragments by the molecular mass of one amino acid reveals the identity and location of that amino acid in the peptide. Only two such amino acids, combined with the knowledge of their location in the peptide — a 'peptide sequence tag' — is sufficient to locate the peptide in large sequence databases (Fig. 1c)[85,86]. Alternatively, the theoretical fragmentation spectra of all possible peptides can be compared with the experimental spectrum to find the sequence that most likely gave rise to it[87,88]. As a result, more complex mixtures of proteins can be analysed and the corresponding peptides found in EST databases or even directly in genomic databases. Routine sensitivities achieved by many laboratories are in the low picomole range (50–100 ng for most proteins), but specialized laboratories achieve higher sensitivities down to the low femtomole range of protein applied to the gel. The overall sensitivity of detection is determined mainly by the protein preparation methods as the mass spectrometer itself is capable of detecting sub-femtomole amounts of peptides under optimized conditions.

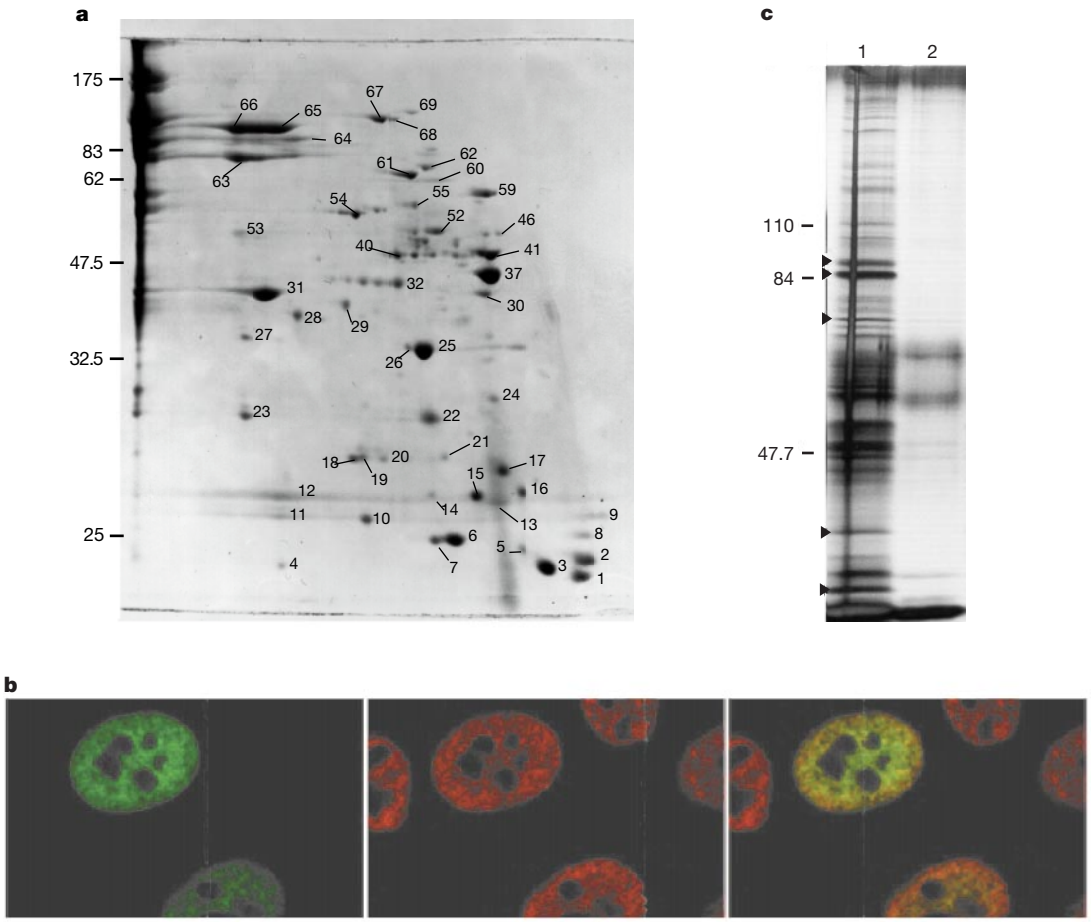

Figure 6 Characterization of the multi-protein spliceosome complex. **a**, A two-dimensional gel of spliceosome-associated factors. **b**, Expression of a green fluorescent protein (GFP)-tagged version of a protein, SPF45 (spot S28), identified from the gel shown in panel **a**. HeLa cells were transiently transfected with a plasmid encoding SPF45, which was tagged with GFP at its N terminus. The green fluorescence observed is due to localization of the GFP-tagged protein to the nucleus. Immunofluorescence using an antibody against a known nuclear protein, U1-specific snRNP protein or U1 (red signal), shows similar sub-nuclear localization as shown by the overlay (yellow signal). **c**, The strategy shown in Fig. 5 was used to isolate molecules interacting with S14. A one-dimensional gel showing proteins eluted from GST beads alone or GST–S14 is shown. The gel was silver stained and the bands indicated by arrowheads were excised and identified by mass spectrometry. These were again found to be proteins in the spliceosome complex, confirming the presence of S14 in the complex and providing insight to S14's role. (Fig. 6c courtesy of A. King.)

other modifications, peptides, protein fragments or proteins may also be immobilized onto chips and samples (for example, phage library or patient serum) applied onto the chip followed by detection of binding. One approach using protein chips couples the above techniques with a direct MALDI readout of the bound material[49,50].

Quantification by mass spectrometry

In addition to the above methods, differential-display proteomics can also be done using limited or no protein separation followed by mass spectrometric quantification. Because the intensity of a peptide peak in the mass spectrum cannot be predicted, quantification is achieved by labelling one of the two states by stable isotopes. Such methods have been used traditionally in mass spectrometry of small molecules but have only recently been applied to proteomics. Microbes can, for example, be grown in one state in normal medium and in another state in medium containing only N^{15} instead of N^{14}. Protein preparations from the two states are then mixed, separated and analysed by mass spectrometry. Two versions of any peptide can now be detected where one is greater in mass by its number of nitrogen atoms and the ratio of peak heights accurately quantifies the relative amounts of the corresponding proteins. As an alternative, Aebersold and colleagues introduced an isotopic non-radioactive label on cysteines after cell lysis before quantifying the samples by mass spectrometry[51]. This strategy enables quantification of peptides from the most abundant components of very crude protein mixtures without gel electrophoresis.

Protein–protein interactions

A key question about a protein, in addition to when and where it is expressed, is with which other proteins does it interact. Interaction partners are an immediate lead into biological function and can potentially be exploited for therapeutic purposes. Creation of a protein–protein interaction map of the cell would be of immense value to understanding the biology of the cell.

Purification of protein complexes

Proteomics can make a key contribution to the study of protein–protein interactions[52–55]. An attractive way to study protein–protein interactions is to purify the entire multi-protein complex by affinity-based methods. This can be achieved in a variety of ways such as by using glutathione *S*-transferase (GST)–fusion proteins, antibodies, peptides, DNA, RNA or a small molecule binding specifically to a cellular target. One of the generic ways of identifying the interaction partners of a new protein is to tag it with an epitope. This protein can then be overexpressed in cells and — together with its interaction partners — immunoprecipitated by an antibody against the epitope.

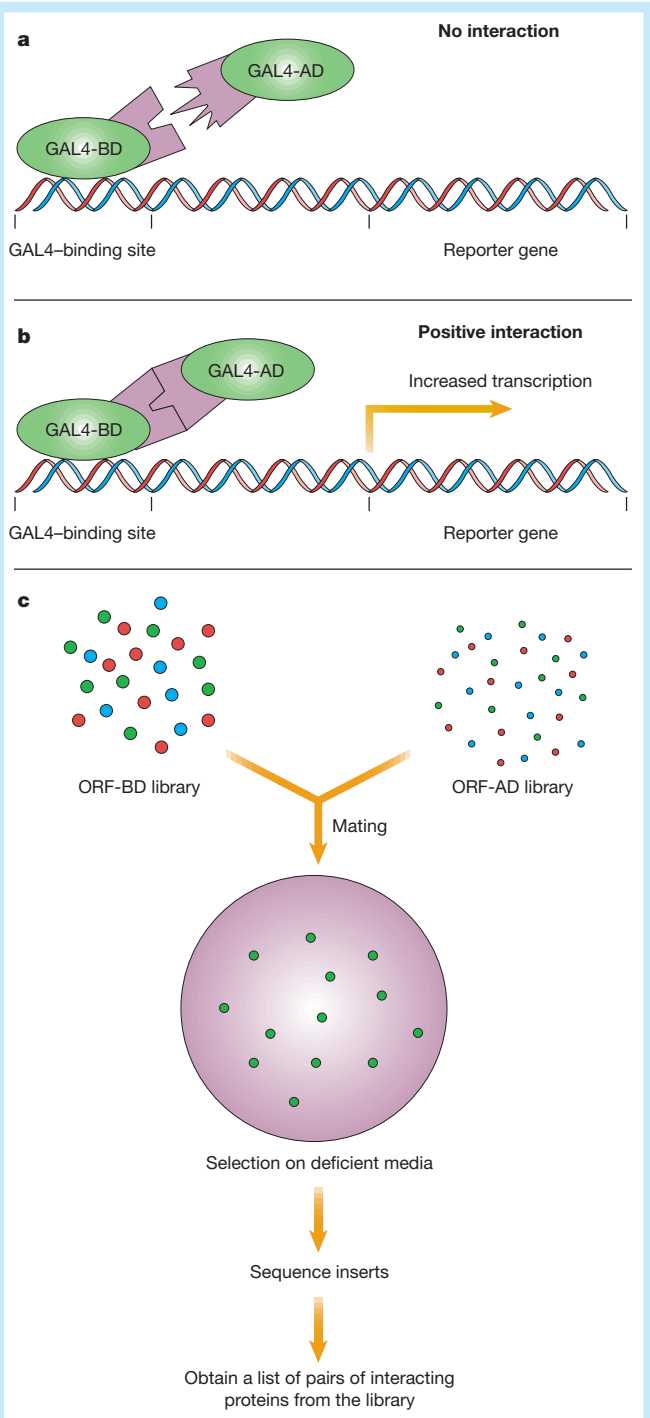

Figure 7 The yeast two-hybrid system. **a**, Different ORFs are expressed as fusion proteins to either the GAL4 DNA-binding domain (GAL4-BD) or its activation domain (GAL4-AD). If the proteins encoded by the ORFs do not interact with each other, the fusion proteins are not brought into close proximity and there is no activation of transcription of the reporter gene containing the upstream GAL4-binding sites. **b**, If the ORFs encode proteins that interact with each other, the fusion proteins are assembled at the GAL4-binding site of the reporter gene, which leads to activation of transcription. **c**, Library-based yeast two-hybrid screening method. In this strategy, two different yeast strains containing two different cDNA libraries are prepared. In one case, the ORFs are expressed as GAL4-BD fusions and in the other case, they are expressed as GAL4-AD fusions. The two yeast strains are then mated and diploids selected on deficient media. Thus, only the yeast cells expressing interacting proteins survive. The inserts from both the plasmids are then sequenced to obtain a pair of interacting genes.

This requires only the full-length complementary DNA clone of the gene and no time is spent in generating a precipitating antibody against the gene of interest. Because full-length cDNAs may soon be available for most human genes[56], large-scale interaction studies will become possible. Making fusion proteins such as GST–fusions is another generic way to obtain interaction partners (Fig. 5). The multi-protein complex associates with the 'bait', which is immobilized on a solid support. After washing away the proteins that interact nonspecifically, the protein complex is eluted, separated by gel electrophoresis and analysed by mass spectrometry. Thus, in a single experiment, the components of an entire multi-protein complex can be identified. As an example, the human spliceosome has been purified using biotinylated RNA as the 'bait' on which the complex assembled[57]. Its protein components were then displayed by two-dimensional gel electrophoresis (Fig. 6a). From a single two-dimensional gel, 19 new factors were obtained (mostly in EST databases) and several of them were cloned and analysed further. Co-localization using immunofluorescence of the new protein with other members of the complex served to establish that they are bona fide members of the complex (Fig. 6b). Several of the new factors identified from this study were cloned and GST–fusion proteins generated. Using the strategy shown in Fig. 5, one of these proteins, designated S14, precipitated a subset of the spliceosome proteins (Fig. 6c), which, together with other experiments and bioinformatics analysis of the sequence, indicated a function of this protein. Many protein complexes have now been characterized using the strategy outlined above. Some of these complexes include the yeast Arp2/3 complex[58], proteins found in the yeast nuclear-pore complex[59] and proteins bound to the chaperonin GroEL[60].

These studies provide insight into mechanisms and open up new lines of investigations. Because no assumptions are made about the complex, unsuspected connections between cellular processes routinely emerge. For example, a study of profilin-I and -II binding proteins in mouse brain resulted in the discovery of two sets of proteins, one consisted of signalling molecules that regulate actin cytoskeleton and the other was involved in endocytosis. This indicated a link between signal transduction pathways and microfilament assembly involving profilin[61].

Once members of a multi-protein complex have been identified by mass spectrometry, their function is studied by pertinent assays. At this stage, proteomics can be used in an iterative fashion to define either direct interaction partners of a new protein in the complex and/or to connect to other complexes in the cell[62].

The success of the above-mentioned strategies relies on sufficient affinity of the protein complex to the bait and on optimized conditions for purification steps. For example, use of a double-tagging strategy improves complex recovery and reduces nonspecific protein binding[63]. Lower-affinity interactions can potentially be captured by chemically crosslinking the protein complex before affinity purification because it relies on spatial proximity rather than affinity. Crosslinking can also help in elucidating the topological structure of a protein complex by the determination of nearest neighbours[64].

Components of specific organelles have also begun to be analysed. The yeast Golgi apparatus has been catalogued and the components of the chloroplast of garden pea have been similarly investigated to identify proteins involved in the processing, targeting, insertion and assembly of photosynthetic complexes[65,66]. The interchromatin granules have been examined by the analysis of the crude peptide mixture obtained after digestion in solution of the entire sample[67].

Yeast two-hybrid system

The yeast two-hybrid system has emerged as a powerful tool to study protein–protein interactions[68]. It is a genetic method based on the modular structure of transcription factors wherein close proximity of the DNA-binding domain to the activation domain induces increased transcription of a set of genes. The yeast hybrid system uses ORFs fused to the DNA-binding or -activation domain of GAL4 such that increased transcription of a reporter gene results when the

proteins encoded by two ORFs interact in the nucleus of the yeast cell (Fig. 7a, b). One of the main consequences of this is that once a positive interaction is detected, the ORF is identified simply by sequencing the relevant clones. For these reasons it is a generic method that is simple and amenable to high-throughput screening of protein–protein interactions.

On a large scale, this strategy has been used in two formats. In the array method, yeast clones containing ORFs as fusions to DNA or activation domains are arrayed onto a grid and the ORFs to be tested (as reciprocal fusions) are screened against the entire grid to identify interacting clones (Fig. 4). In the library screening method, one set of ORFs are first pooled to generate a library and then the reciprocal ORF–fusions are mated with the library one by one or several at a time (Fig. 7c).

Such analyses on a genome-wide scale have already been reported in *Saccharomyces cerevisiae* and to a more limited extent in *Caenorhabditis elegans*[69–71]. In yeast, the array method was performed on 192 ORFs and the library screening method for 87% of the yeast genome. Together, this experiment resulted in 957 putative interactions[70]. Another group analysed the results of 10% of an exhaustive library screen in yeast, resulting in 183 putative interactions[71]. The vast majority of the interactions found in these two large-scale studies were new. Several of these interactions seem plausible based on previous genetic or biochemical studies, whereas the relevance of most others cannot easily be determined. Therefore, such studies provide only potential interactions that have to be confirmed or eliminated by further biological experimentation. The main advantage of these methods is that they can be performed with a high throughput and in an automated manner. A recently described modification of the yeast two-hybrid method, termed 'reverse' two hybrid, can be used for identification of compounds and peptides that disrupt protein–protein interactions[72]. This can lead to development of drugs that have activities *in vivo* as opposed to drug screens that are conventionally done *in vitro*.

Phage display

Phage display is a method where bacteriophage particles are made to express either a peptide or protein of interest fused to a capsid or coat protein. It can be used to screen for peptide epitopes, peptide ligands, enzyme substrates or single-chain antibody fragments. Although combinatorial peptide libraries have generally been used in most phage display-based studies, more informative large-scale protein interaction studies can now be done if the products of cDNA libraries are displayed on phage particles. Any 'bait' protein can then be immobilized to capture phage particles displaying interacting proteins. This method is similar to the yeast two-hybrid system in that it is simple and can be performed with high throughput. Depending on the particular class of proteins being studied (such as cytoplasmic versus cell surface proteins), this method may be superior or inferior to the two-hybrid system because the interactions take place in solution as opposed to the nucleus of the yeast cell. Furthermore, this method is applicable in principle to transcription factors, which are not amenable to the yeast two-hybrid system. Methods have recently been optimized to display cDNA libraries on phages to isolate signalling molecules in the EGF-receptor signalling pathway as well as to identify antigens that react with certain antibodies[73,74].

Conclusions

Proteomics provides a powerful set of tools for the large-scale study of gene function directly at the protein level. In particular, the mass spectrometric study of gel-separated proteins is leading to a renaissance in biochemical approaches to protein function. Protein characterization will continue to improve in throughput, sensitivity and completeness. Post-translational modifications cannot currently be studied at high throughput but certain categories such as phosphorylation are beginning to be amenable to generic approaches. We predict that proteomics will move away from the monitoring of protein expression using two-dimensional gels. Mass spectrometry-based methods that use affinity purification followed by only one-dimensional electrophoresis will continue to gain in importance. In the near future, proteomics will provide a wealth of protein–protein interaction data, which will probably be its most important and immediate impact on biological science. Because proteins are one step closer to function than are genes, these studies frequently lead directly to biological discoveries or hypotheses. The ready availability of many human genes as full-length clones is itself an extremely important extension of the genome projects that will make possible several proteomic strategies. Assays to determine protein function using purified proteins will be automated and performed in miniaturized grid formats in parallel for thousands of proteins. Finally, advances in genomics will directly fuel large-scale protein assays that use genetics as a readout, such as the two-hybrid screen. □

1. Wilkins, M. R., Williams, K. L., Apple, R. D. & Hochstrasser, D. F. *Proteome Research: New Frontiers in Functional Genomics* 1–243 (Springer, Berlin, 1997).
2. Wilkins, M. R. *et al.* From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *BioTechnology* **14,** 61–65 (1996).
3. Celis, J. *et al.* Human 2-D PAGE databases for proteome analysis in health and disease: http://biobase.dk/cgi-bin/celis. *FEBS Lett.* **398,** 129–134 (1996).
4. Anderson, N. G. & Anderson, N. L. Twenty years of two-dimensional electrophoresis: past, present and future. *Electrophoresis* **17,** 443–453 (1996).
5. O'Farrell, P. H. High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250,** 4007–4021 (1975).
6. Burley, S. K. *et al.* Structural genomics: beyond the human genome project. *Nature Genet.* **23,** 151–157 (1999).
7. Krogh, A. in *Guide to Human Genome Computing* (ed. Bishop, M. J.) 261–274 (Academic, San Diego, 1998).
8. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).
9. Claverie, J. M. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6,** 1735–1744 (1997).
10. Pandey, A. & Lewitter, F. Nucleotide sequence databases: a gold mine for biologists. *Trends Biochem. Sci.* **24,** 276–280 (1999).
11. Brenner, S. E. Errors in genome annotation. *Trends Genet.* **15,** 132–133 (1999).
12. Gygi, S., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19,** 1720–1730 (1999).
13. Futcher, B. *et al.* A sampling of the yeast proteome. *Mol. Cell. Biol.* **19,** 7357–7368 (1999).
14. Pandey, A., Andersen, J. S., & Mann, M. Use of mass spectrometry to study signaling pathways. *Science's STKE* (in the press).
15. Henzel, W. J., Billeci, T. M., Stults, J. T. & Wong, S. C. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl Acad. Sci. USA* **90,** 5011–5015 (1993).
16. Jensen, O. N., Mortensen, P., Vorm, O. & Mann, M. Automation of matrix assisted laser desorption/ionization mass spectrometry using fuzzy logic feedback control. *Anal. Chem.* **69,** 1706–1714 (1997).
17. Berndt, P., Hobohm, U. & Langen, H. Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. *Electrophoresis* **20,** 3521–3526 (1999).
18. Shevchenko, A. *et al.* Linking genome and proteome by mass spectrometry: large scale identification of yeast proteins from two dimensional gels. *Proc. Natl Acad. Sci. USA* **93,** 14440–14445 (1996).
19. Shevchenko, A. *et al.* MALDI quadruple time-of-flight mass spectrometry: powerful tool for proteomic research. *Anal. Chem.* **72,** 2132–2141 (2000).
20. Zhang, B., Liu, H., Karger, B. L. & Foret, F. Microfabricated devices for capillary electrophoresis-electrospray mass spectrometry. *Anal. Chem.* **71,** 3258–3264 (1999).
21. Figeys, D., Gygi, S. P., McKinnon, G. & Aebersold, R. An integrated microfluidics-tandem mass spectrometry system for automated protein analysis. *Anal. Chem.* **70,** 3728–3734 (1998).
22. Li, J. *et al.* Integration of microfabricated devices to capillary electrophoresis—electrospray mass spectrometry using a low dead volume connection: application to rapid analyses of proteolytic digests. *Anal. Chem.* **71,** 3036–3045 (1999).
23. Eckerkorn, C. *et al.* Mass spectrometric analysis of blotted proteins after gel electrophoresis separation by matrix-assisted laser desorption/ionization. *Electrophoresis* **13,** 664–665 (1992).
24. Strupat, K. *et al.* Matrix-assisted laser desorption ionization mass spectrometry of proteins electroblotted after polyacrylamide gel electrophoresis. *Anal. Chem.* **66,** 464–470 (1994).
25. Bienvenut, W. V. *et al.* Toward a clinical molecular scanner for proteome research: parallel protein chemical processing before and during western blot. *Anal. Chem.* **71,** 4800–4807 (1999).
26. Binz, P. A. *et al.* A molecular scanner to automate proteomic research and to display proteome images. *Anal. Chem.* **71,** 4981–4988 (1999).
27. Jensen, P. K. *et al.* Probing proteomes using capillary isoelectric focusing-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **71,** 2076–2084 (1999).
28. Mørtz, E. *et al.* Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. *Proc. Natl. Acad. Sci. USA* **93,** 8264–8267 (1996).
29. Li, W., Hendrickson, C. L., Emmett, M. R. & Marshall, A. G. Identification of intact proteins in mixtures by alternated capillary liquid chromatography electrospray ionization and LC ESI infrared multiphoton dissociation Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **71,** 4397–4402 (1999).
30. Nuwaysir, L. & Stults, J. T. ESI mass spectrometry of phosphopeptides isolated by on-line immobilized metal affinity chromatography. *J. Am. Soc. Mass Spectrom.* **4,** 662–669 (1993).
31. Betts, J. C., Blackstock, W. P., Ward, M. A. & Anderton, B. H. Identification of phosphorylation sites on neurofilament proteins by nanoelectrospray mass spectrometry. *J. Biol. Chem.* **272,** 12922–12927 (1997).
32. Neubauer, G. & Mann, M. Mapping of phosphorylation sites of gel-isolated proteins by nanoelectrospray

tandem mass spectrometry: potentials and limitations. *Anal. Chem.* **71,** 235–242 (1999).
33. Zhang, X. *et al.* Identification of phosphorylation sites in proteins separated by polyacrylamide gel electrophoresis. *Anal. Chem.* **70,** 2050–2059 (1998).
34. Cortez, D., Wang, Y., Qin, J. & Elledge, S. J. Requirement of ATM-dependent phosphorylation of brca1 in the DNA damage response to double-strand breaks. *Science* **286,** 1162–1166 (1999).
35. Soskic, V. *et al.* Functional proteomics analysis of signal transduction pathways of the platelet-derived growth factor beta receptor. *Biochemistry* **38,** 1757–1764 (1999).
36. Pandey, A. *et al.* Analysis of receptor signaling pathways by mass spectrometry: identification of Vav-2 as a substrate of the epidermal and platelet-derived growth factor receptors. *Proc. Natl Acad. Sci. USA* **97,** 179–184 (2000).
37. DeRisis, J., Iyer, V. R. & Brown, O. P. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278,** 680–686 (1997).
38. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537 (1999).
39. Roberts, C. J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287,** 873–880 (2000).
40. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403,** 503–511 (2000).
41. Ostergaard, M., Wolf, H., Orntoft, T. F. & Celis, J. E. Psoriasin (S100A7): a putative urinary marker for the follow-up of patients with bladder squamous cell carcinomas. *Electrophoresis* **20,** 349–354 (1999).
42. Page, M. J. *et al.* Proteomic definition of normal human luminal and myoepithelial breast cells purified from reduction mammoplasties. *Proc. Natl Acad. Sci. USA* **96,** 12589–12594 (1999).
43. Gauss, C. *et al.* Analysis of the mouse proteome. (I) Brain proteins: separation by two-dimensional electrophoresis and identification by mass spectrometry and genetic variation. *Electrophoresis* **20,** 575–600 (1999).
44. Aicher, L. *et al.* New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis* **19,** 1998–2003 (1998).
45. Celis, J. E. *et al.* A comprehensive protein resource for the study of bladder cancer: http://biobase.dk/cgi-bin/celis. *Electrophoresis* **20,** 300–309 (1999).
46. Breitenbach, M. *et al.* Biological and immunological importance of Bet v 1 isoforms. *Adv. Exp. Med. Biol.* **409,** 117–126 (1996).
47. Sander, I. *et al.* Allergy to aspergillus-derived enzymes in the baking industry: identification of beta-xylosidase from aspergillus niger as a new allergen (Asp n 14). *J. Allergy Clin. Immunol.* **102,** 256–264 (1998).
48. Lueking, A., Horn, M., Eickhoff, H., Lehrach, H. & Walter, G. Protein microarrays for gene expression and antibody screening. *Anal. Biochem.* **270,** 103–111 (1999).
49. Davies, H., Lomas, L. & Austen, B. Profiling of amyloid beta peptide variants using SELDI Protein Chip arrays. *Biotechniques* **27,** 1258–1261 (1999).
50. Nelson, R. W. The use of bioreactive probes in protein characterization. *Mass Spectrom. Rev.* **16,** 353–376 (1997).
51. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.* **17,** 994–999 (1999).
52. Neubauer, G. *et al.* Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl Acad. Sci. USA* **94,** 385–390 (1997).
53. Lamond, A. I. & Mann, M. Cell biology and the genome projects—a concerted strategy for characterizing multi-protein complexes using mass spectrometry. *Trends Cell Biol.* **7,** 139–142 (1997).
54. Link, A. J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnol.* **17,** 676–682 (1999).
55. Blackstock, W. P. & Weir, M. P. Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* **17,** 121–127 (1999).
56. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. The mammalian gene collection. *Science* **286,** 455–457 (1999).
57. Neubauer, G. *et al.* Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nature Genet.* **20,** 46–50 (1998).
58. Winter, D., Podtelejnikiov, A. V., Mann, M. & Li, R. The complex containing actin-related proteins Arp2 and Arp3 is required for the motility and integrity of yeast actin patches. *Curr. Biol.* **7,** 519–529 (1997).
59. Rout, M. P. *et al.* The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell. Biol.* **148,** 635–651 (2000).
60. Houry, W. A. *et al.* Identification of in vivo substrates of the chaperonin GroEL. *Nature* **402,** 147–154 (1999).
61. Witke, W. *et al.* In mouse brain profilin I and profilin II associate with regulators of the endocytic pathway and actin assembly. *EMBO J.* **17,** 967–976 (1998).
62. Shevchenko, A. & Mann, M. in *Mass Spectrometry in Biology and Medicine* (eds Burlingame, A., Carr, C. A. & Baldwin, M. A.) 237–269 (Humana, Totowa, 1999).
63. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17,** 1030–1032 (1999).
64. Rappsilber, J., Siniossoglou, S., Hurt, E. C. & Mann, M. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal. Chem.* **72,** 267–275 (2000).
65. Rowley, A. *et al.* Applications of protein mass spectrometry in cell biology. *Methods* **20,** 383–397 (2000).
66. Peltier, J. B. *et al.* Proteomics of the chloroplast. Systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. *Plant Cell* **12,** 319–342 (2000).
67. Mintz, P. J. *et al.* Purification and biochemical characterization of interchromatin granule clusters. *EMBO J.* **18,** 4308–4320 (1999).
68. Fields, S. & Song, O. K. A novel genetic system to detect protein-protein interactions. *Nature* **340,** 245–246 (1989).
69. Walhout, A. J. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287,** 116–122 (2000).
70. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403,** 623–627 (2000).
71. Ito, T. *et al.* Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA* **97,** 1143–1147 (2000).
72. Vidal, M. & Endoh, H. Prospects for drug screening using the reverse two-hybrid system. *Trends Biotechnol.* **17,** 374–381 (1999).
73. Zozulya, S. *et al.* Mapping signal transduction pathways by phage display. *Nature Biotechnol.* **17,** 1193–1198 (1999).
74. Hufton, S. E. *et al.* Phage display of cDNA repertoires: the pVI display system and its applications for the selection of immunogenic ligands. *J. Immunol. Methods* **231,** 39–51 (1999).
75. Martzen, M. R. *et al.* A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286,** 1153–1155 (1999).
76. Zambrowicz, B. P. *et al.* Disruption and sequence identification of 2,000 genes in mouse embryonic stem cells. *Nature* **392,** 608–611 (1998).
77. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391,** 806–811 (1998).
78. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285,** 901–906 (1999).
79. Mattheakis, L. C., Bhatt, R. R. & Dower, W. J. An in vitro polysome display system for identifying ligands from very large peptide libraries. *Proc. Natl Acad. Sci. USA* **91,** 9022–9026 (1994).
80. Roberts, R. W. & Szostak, J. W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl Acad. Sci. USA* **94,** 12297–12302 (1997).
81. Wilm, M. & Mann, M. Analytical properties of the nanoelectrospray ion source. *Anal. Chem.* **68,** 1–8 (1996).
82. Wilm, M. *et al.* Femtomole sequencing of proteins from polyacrylamide gels by nano electrospray mass spectrometry. *Nature* **379,** 466–469 (1996).
83. Roepstorff, P. & Fohlman, J. Proposed nomenclature for sequence ions. *Biomed. Mass Spectrom.* **11,** 601 (1984).
84. Yates, J. R. Mass spectrometry. From genomics to proteomics. *Trends Genet.* **16,** 5–8 (2000).
85. Mann, M. & Wilm, M. S. Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66,** 4390–4399 (1994).
86. Mann, M. A shortcut to interesting human genes: peptide sequence tags, ESTs and computers. *Trends Biochem. Sci.* **21,** 494–495 (1996).
87. Eng, J. K., McCormack, A. L. & J. R. Yates, I. An approach to correlate MS/MS data to amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989 (1994).
88. Yates, J. R. III Database searching using mass spectrometry data. *Electrophoresis* **19,** 893–900 (1998).

Acknowledgements

We thank B. Blagoev and M. Fernandez for their expert assistance with cell culture and immunoprecipitation experiments. We thank all other members of the Protein Interaction Laboratory for valuable discussions and comments on the manuscript and A. King, Protana A/S, for obtaining the data for the new spliceosomal protein. O. N. Jensen and A. Stensballe are acknowledged for their contributions in the analysis of phosphopeptides. A.P. was supported by the Howard Temin Award from the National Cancer Institute. This work was funded in part by a grant from the Danish National Research Foundation to M.M.'s laboratory (www.pil.sdu.dk) at the Center for Experimental BioInformatics (CEBI).