

Proteomics: the first decade and beyond

Scott D. Patterson^{1,3} & Ruedi H. Aebersold²

doi:10.1038/ng1106

Proteomics is the systematic study of the many and diverse properties of proteins in a parallel manner with the aim of providing detailed descriptions of the structure, function and control of biological systems in health and disease. Advances in methods and technologies have catalyzed an expansion of the scope of biological studies from the reductionist biochemical analysis of single proteins to proteome-wide measurements. Proteomics and other complementary analysis methods are essential components of the emerging 'systems biology' approach that seeks to comprehensively describe biological systems through integration of diverse types of data and, in the future, to ultimately allow computational simulations of complex biological systems.

Proteomics, like other 'discovery science' technologies (Box 1)¹, such as genomic sequencing, microarray analysis and metabolite profiling, is the direct consequence of both the results obtained from ambitious projects aimed at mapping and sequencing the complete genome of many species and the changes to our models that are catalyzed by such projects. The essence of this emerging 'systems biology' approach (Box 1) is that, for any given species, the space of possible biomolecules and their organization into pathways and processes is large but finite. In theory, therefore, the biological systems operating in a species can be described comprehensively if a sufficient density of observations on all of the elements that constitute the system can be obtained. Proteomics is a particularly rich source of biological information because proteins are involved in almost all biological activities and they also have diverse properties, which collectively contribute greatly to our understanding of biological systems. These properties are summarized in Fig. 1.

Genome sequencing, although technically complex, is conceptually simple and has a defined end point: the conclusive determination of the complete genome sequence of the species in question. Discovery science projects aimed at assaying the function and control of biological systems are based on less well-defined technologies and are devoid of clear end points. An initial goal of proteomics was the rapid identification of all of the proteins expressed by a cell or tissue—a goal that has yet to be achieved for any species. Current goals of proteomic research are more varied and directed toward the systematic determination of diverse properties of proteins. These include sequence, quantity, state of modification, interactions with other proteins, activity, subcellular distribution and structure (Fig. 1).

Many different technologies have been and are still being developed to collect the information contained in the properties of proteins. Figure 2 summarizes the current state of these technologies and how they relate to other discovery science tools.

Three characteristics of these proteomic technologies are immediately apparent: first, there is no single technology platform that can satisfy all of the desired proteomic measurements; second, the closer the measurement to protein function, the less mature the technology; and third, there is no mature, 'true' proteomic technology as yet.

In this review, we do not discuss the three-dimensional structural analysis of proteins (structural genomics), which is a large field in its own right². Instead, we discuss three phases of the emergence and maturation of proteomic concepts and technology. The first spans the emergence from protein chemistry to proteomics as a coordinated platform for discovery science. The second is the current diversification of proteomic technologies and platforms, which aim to capture the many properties of proteins. The third is a window into the future, in which the variety of data obtained by proteomics and other discovery science technologies will be integrated and collectively interpreted to achieve a comprehensive understanding of the workings of biological systems. We focus predominantly on developments in proteomics in the past decade.

From protein chemistry to proteomics

Protein chemistry was a key element of the reductionist research approaches that were a mainstay of biology in the 1980s. Also called 'forward', as in 'forward genetics', these approaches attempted to move from an observed phenotype or function to the relevant genes and their products that caused that phenotype. Together with rapidly advancing methods in molecular biology such as gene cloning, sequencing and expression analysis, protein chemistry provided the link between the observed activity or function of a biochemically isolated protein and the gene that encoded it. A key objective was therefore the development of ever-more sensitive and reliable methods for protein sequencing to make ever-smaller amounts of purified protein accessible to identification. The long-term goals of this type of research were

¹Celera Genomics Corporation, 45 West Gude Drive, Rockville, Maryland 20850, USA. ²Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103, USA. ³Present address: Farmal Biomedicines, LLC, 129 North Hill Avenue, Suite #107, Pasadena, California 91106, USA. Correspondence should be addressed to S.D.P. (e-mail: scott@farmalbiomed.com)

to reassemble *in vitro* the system under study from its isolated components and to test whether this reconstituted system would recapitulate functions observed *in vivo*³.

The advent of large-scale sequencing projects and their results⁴ catalyzed the development of 'reverse' approaches, which attempted to move from the gene sequence to function and phenotype. Such approaches included the observation of clusters of mRNA species showing coordinated expression patterns in different cellular states, either by expression arrays or by serial analysis of gene expression (SAGE)^{5,6}. The idea of defining functionally relevant patterns of gene expression by comparative pattern analysis was also applied, and in fact pioneered, in the protein science field through attempts to develop global approaches to the quantitative analysis of protein expression patterns generated by high-resolution two-dimensional gel electrophoresis or '2DE' (Box 1)^{7,8}.

Below, we describe the evolution of protein chemistry methods and their adaptation from forward to reverse research strategies.

From isolated proteins to gene sequence

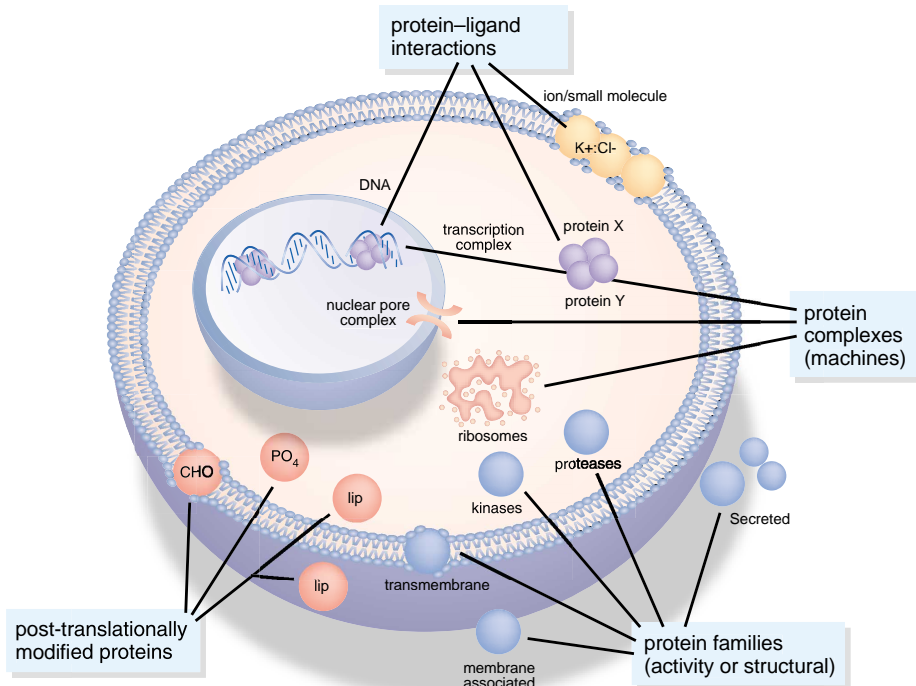
Technological developments for analytical protein chemistry in the 1980s and early 1990s were primarily directed at improving the sensitivity of techniques for identifying proteins separated by gels. Protein sequencing often provided the crucial link between the activity of a purified protein and its amino acid sequence or the sequence of the corresponding gene. Edman sequencing of the amino terminus of intact proteins or enzymatically digested fragments was the most common approach to identify gel-separated proteins^{9,10}. In the likely case that the sequence did not exist in the rather small sequence databases that existed then, degenerate oligonucleotide primers could be synthesized based on segments of the determined sequences and the corresponding gene could be cloned using methods based on PCR¹¹.

Although protein sequencing by Edman degradation was mature, reliable and automated, it also had relatively poor sensitivity and was slow. With the rapid increase in the size of sequence databases, which was fueled by systematic sequencing programs, the chance that a particular protein and/or gene sequence was

Box 1 • Glossary of experimental terms used in proteomics

2DE	two-dimensional gel electrophoresis is the separation of proteins using two orthogonal parameters, isoelectric point (charge) and relative molecular mass, which are both usually determined on the basis of protein mobility in a polyacrylamide gel matrix.
discovery science	investigation of a biological system or process by enumerating the elements of a system irrespective of any hypotheses on how the system might function.
DNA microarrays	a high-throughput differential screen of mRNA expression using complementary cDNA or oligonucleotide libraries that are printed in extremely high density on microchips; these microchips are probed with a mixture of fluorescently tagged cDNAs that are produced from two different cell populations and analyzed with a laser confocal scanner.
ESI	the electrospray ionization process is achieved by spraying a solution (such as the effluent of a HPLC column) through a charged needle at atmospheric pressure towards the inlet of the mass spectrometer; the voltage applied to the needle tip and a pressure differential result in the formation of ions for mass analysis and their transfer into the mass spectrometer.
ICAT	isotope-coded affinity tag reagent comprising a chemical modifying group linked to an affinity group through a mass-encoded linker.
ion source	mass spectrometer component designed to use the principles of an ionization method for generating ions (charged analytes) for mass analysis.
ionization	process of adding charge onto an uncharged (neutral) analyte, in other words, the formation of an ion; either ionization is conducted in a vacuum or ions formed at atmospheric pressure are transferred into the vacuum system of the mass spectrometer.
MALDI	matrix-assisted laser desorption ionization is a process by which ion formation is promoted by short laser pulses; the sample is deposited on a sample plate into the source (which is held under vacuum) and then embedded in a matrix that promotes ionization; a laser fired at the sample that is co-crystallized with the matrix results in the desorption of the analyte from the sample plate and its ionization.
mass analyzer	mass spectrometer component that can measure the mass-to-charge ratio of charged molecules (ions); ion-trap, quadrupole and time of flight (TOF) analyzers are used most often.
mass spectrometry	accurate mass measurement of charged analytes (ions); in the context of proteomics, analytes are usually peptides or less frequently protein ions; a mass spectrometer measures the mass-to-charge ratio of charged species under vacuum and comprises, broadly, an ionization source and a mass analyzer.
MS spectra	single-stage mass spectrometry spectra provide mass information on all ionizable components in a sample; these data are used, for example, for peptide mass fingerprinting.
MS/MS spectra	MS/MS spectra are generated in instruments equipped with a mass filter that can select a peptide ion from a mixture of peptide ions, a collision cell in which peptide ions are fragmented into a series of product ions (through collision of the selected precursor ion with a noble gas in a process referred to as CID), and a second mass analyzer that records the fragment ion mass spectrum; the fragment ion spectra are referred to as either MS/MS spectra or CID spectra.
protein identification	method to determine the sequence identity of a protein; two common mass spectrometry-based approaches used are peptide mass mapping and searching uninterpreted MS/MS spectra; in both methods, observed data are matched to theoretically derived peptide and/or fragment ion masses calculated from sequence databases.
systems biology	study of a biological system by the systematic and quantitative analysis of all of the components that constitute the system.
yeast two-hybrid	genetics-based method for identifying protein-protein interactions <i>in vivo</i> ; a protein fused to the DNA-binding domain (the 'bait') and a (different) protein fused to the activation domain of a transcriptional activator (the 'prey') are expressed in yeast cells; if an interaction between the bait and the prey occurs, transcription of a reporter gene is induced and detected typically by a color reaction that indicates transactivation of the reporter gene.

Fig. 1 Representation of a eukaryotic cell. A section through a eukaryotic cell is shown, highlighting the diverse properties of proteins. The systematic investigation of these properties constitutes the field of proteomics. The subcellular distribution, quantity, modification and interaction state, catalytic activity and structure are particularly informative for describing biological systems. Representative examples of protein properties are shown, including the subcellular distribution of proteins to specific compartments and organelles; the interaction of proteins with DNA, other proteins or small molecules to form functional complexes (or 'machines') with diverse functions; and protein modifications such as carbohydrate (CHO), phosphate (PO₄) or lipid (lip).



already represented in a sequence database, either in total or in part, also increased. This greatly facilitated gene cloning and sometimes the complete sequence of a gene of interest could be found by database searching, without the need for further experimentation.

Thus, the idea that protein sequences did not always have to be determined *de novo* began to take shape.

DNA sequencing rules the day

In the early 1990s, mass DNA sequencing of cDNAs derived from pools of mRNA generated large numbers of expressed sequence tags (ESTs)¹². These stretches of sequence provided an unprecedented window into the transcripts present in particular types of cell and tissue, and therefore a powerful tool for gene discovery⁴. Gene sequences (ESTs and others) also provided a resource that could greatly accelerate protein identification by correlating experimentally derived sequence segments with sequences in databases. At the time, it was expected that eventually every gene of a species would be represented in sequence databases and it would be possible to identify proteins and/or genes simply by looking up the gene of interest using experimentally derived data.

Large-scale EST sequencing represented the first real approach to the systematic sequencing of expressed genes¹³. Although ESTs did provide much useful data, they did not produce the depth of analysis that had been anticipated¹⁴. In part, this was due to the dynamic range of transcript numbers expressed in cells, which complicated the detection of low-abundance species. This problem could be alleviated partially through the use of normalized libraries¹⁵.

The ultimate normalized sequence libraries—that is, the complete genomic sequences—were established only a few years later for yeast¹⁶ and a decade later for human^{17,18}. With such complete libraries on hand, the rapid identification of proteins was limited only by our capacity to extract sequence information from proteins and peptides, and to correlate this information with the sequence databases. 'Mass spectrometry' (Box 1) and database search algorithms rapidly filled this gap.

Protein and peptide mass spectrometry

For years, mass spectrometry has been the analytical chemist's workhorse for analyzing small molecules. The high precision of

mass spectrometric measurements can distinguish closely related species, and tandem mass spectrometry or 'MS/MS' can provide structural information on molecular ions that can be isolated and fragmented within the instrument (Box 1).

To measure the mass or, more specifically, the mass-to-charge ratio (m/z) of a molecule in a mass spectrometer, the analyte must first be ionized and transferred into the high vacuum system of the instrument. Peptides and proteins, like other large molecules, proved difficult to ionize under conditions that did not destroy the molecule. In the late 1980s, two methods were developed that allowed the 'ionization' of peptides and proteins at high sensitivity and without excessive fragmentation. These breakthroughs were electrospray ionization ('ESI')¹⁹ and matrix-assisted laser desorption ionization ('MALDI')²⁰, which had closely followed the development of laser desorption^{21,22} (Box 1). The success of these ionization methods in analytical protein chemistry led to the development of commercial mass spectrometers equipped with robust ESI or MALDI 'ion source' instruments, which rapidly penetrated the protein chemistry community.

MALDI ion sources were most commonly coupled with time-of-flight (TOF) 'mass analyzers', whereas ESI was most often coupled with ion-trap or triple-quadrupole MS/MS spectrometers (Box 1). Although MALDI-TOF mass spectrometers can determine the mass of a protein or peptide with a high degree of accuracy, the intrinsic mass of a eukaryotic protein is not a uniquely identifying feature. It was quickly recognized, however, that the masses of the various peptides generated by fragmentation of an isolated protein with an enzyme of known cleavage specificity could uniquely identify a protein.

In 1993, five independent reports were published that described the implementation of this insight in database search algorithms^{23–27}. These algorithms, together with MALDI-TOF mass spectrometry peptide analysis, constituted a 'protein identification' method that is now known as peptide mass mapping (or peptide mass fingerprinting). In this type of analysis, the collected 'MS spectra' (Box 1) are used to generate a list of proteolytic (peptide) fragment masses, which are then matched to the

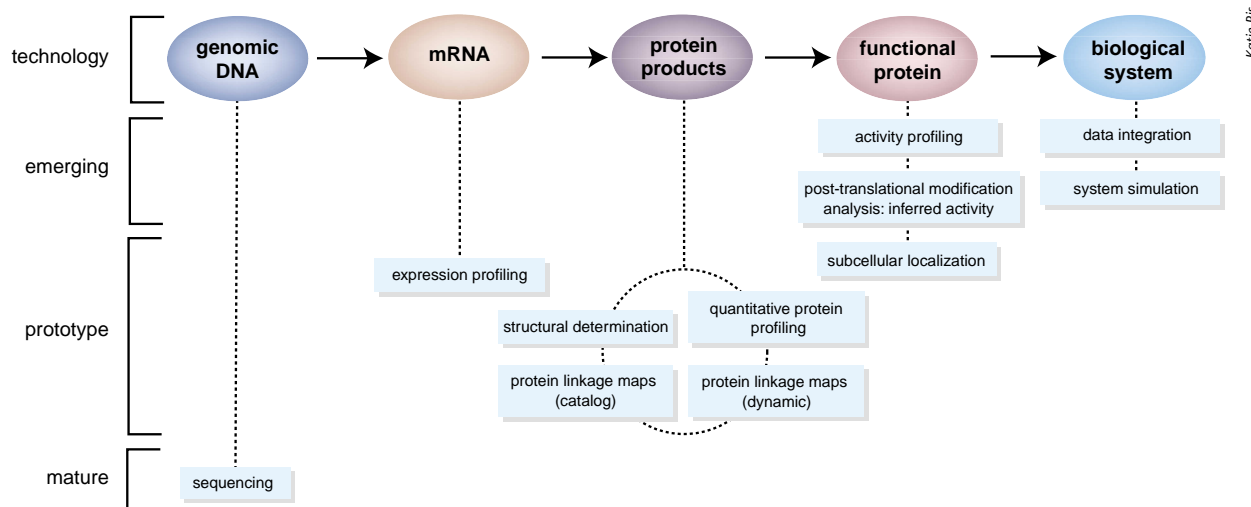


Fig. 2 The current status of proteomic technologies. The different data typically collected in proteomic research and the available technologies are listed. The relative maturity of the proteomic technologies and other key discovery science tools is apparent from the position of the respective technology on the graph.

masses calculated from the same proteo-lytic digestion of each entry in a sequence database, resulting in identification of the target protein. The success of this type of analysis is dependent on the specificity of the enzyme used (most frequently trypsin), the number of peptides identified from each protein species, and the mass accuracy of the mass spectrometer²⁸. Owing to its increasing sensitivity and ease of use, MALDI-TOF mass spectrometry has become the method of choice for protein identification by peptide mass mapping and is commonly used for identifying proteins separated by 2DE.

As discussed above, ESI ion sources were originally coupled mostly with triple-quadrupole or ion-trap instruments. More recently, hybrid-quadrupole TOF MS/MS spectrometers have become available and are also used frequently with ESI. In addition to measuring peptide mass, all of these instruments can isolate specific ions from a mixture on the basis of their *m/z* ratio and fragment these ions in the gas phase within the instrument, allowing the recording of MS/MS spectra. Because peptide ions fragment in a sequence-dependent manner, the MS/MS spectrum of a peptide, in principle, represents its amino acid sequence.

Algorithms that match MS/MS spectra to sequence databases^{29,30} have greatly facilitated mass spectrometric protein identification by this approach³¹. Because a peptide sequence, and thus the MS/MS spectrum of a peptide, can uniquely identify a protein, the specificity of MS/MS-based protein identifications is often much higher than that of peptide mass mapping. MS/MS spectra are also ideally suited to search translated EST and other sequence databases containing incomplete sequences. ESI-MS/MS is popular because it can be combined easily with standard peptide separation tools, such as chromatography, and because it is directly compatible with the solvents that are used to solubilize peptides.

The method of nanospray-ESI, in which unseparated peptide mixtures were sprayed into the mass spectrometer at very low flow rates and detected at sensitivities not previously achieved by ESI-MS, was developed subsequently^{32,33}. The very slow sample consumption afforded by the low flow rate provided the opportunity to generate fragment ion spectra of several of the observed precursor ions. This was achieved by the operator, who manually selected precursor ions. Subsequent developments in instrument control software facilitated computer-controlled ion selection,

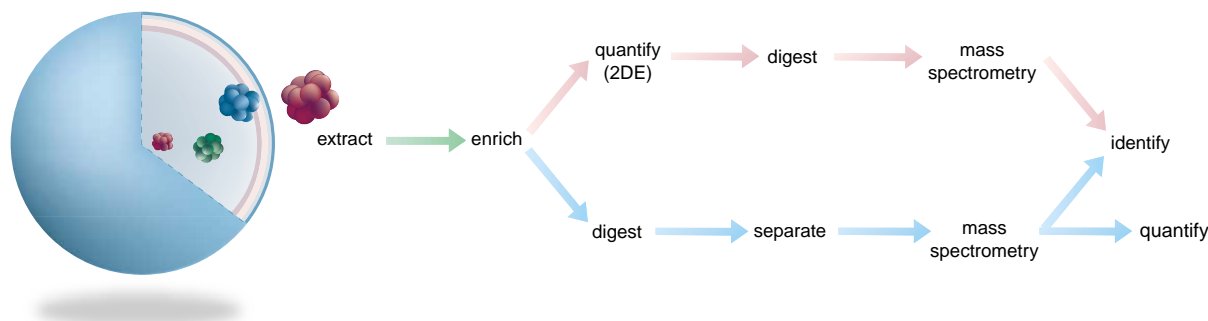
such that MS/MS spectra could be generated from many peptide ions in a given sample without the need for operator intervention, effectively automating the process. This was developed mostly for analyses in which mixtures of peptides were supplied to the mass spectrometer from an online, capillary, high-pressure liquid chromatography (HPLC) system—an approach referred to as LC-MS/MS.

As summarized in Fig. 3, peptide and protein separation technologies, advanced mass spectrometry and MS/MS instrumentation and algorithms for searching mass spectrometric data against sequence databases have been combined in different ways to create a set of technologies for protein and proteome analyses (Fig. 4).

The 'proteome' is born

Long before global differential analysis of mRNA expression was possible, protein science relied on 2DE for generating reproducible protein arrays, displaying large numbers of separated features and indicating their quantities^{34–36}. On the strength of such 2DE protein profiles, in the 1970s and 1980s ideas were proposed to build protein databases (such as the human protein index³⁷) and to apply reverse strategies based on subtractive pattern analysis^{38–40}, similar to current popular strategies for analyzing data obtained from gene expression microarrays. In fact, at that time many of the principles now commonly used for the global, quantitative analysis of gene expression patterns, such as the use of clustering algorithms and multivariate statistics, were developed in the context of 2DE^{7,8}. These ideas were not substantially implemented then, however, mainly because 2DE by itself is an essentially descriptive technique that does not indicate the identity of the separated proteins and because the technique had been plagued by reproducibility and other technical problems.

With the rapid advances in protein analytical technologies, fueled by the addition of mass spectrometry, sequence databases and database search tools, in the early 1990s it became possible for protein chemists to identify and to examine the expression of many, if not all, of the proteins resolvable by 2DE, and the possibility for large-scale protein studies seemed attainable⁴¹. It was in this context that in 1994, at the first 2DE meeting in Siena, Italy, the term 'proteome' was coined⁴². The proteome was defined as the protein complement of the genome, and the process of studying the proteome became known as 'proteomics'.



Karlis Rits

Fig. 3 Quantitative protein analysis from the cell to the identified protein. The two most common processes for quantitative proteome analysis are shown. In the first (top), 2DE is used to separate and to quantify proteins, and selected proteins are then isolated and identified by mass spectrometry. In the second (bottom), LC-MS/MS is used to analyze enzyme digests of unseparated protein mixtures, and accurate quantification is achieved by labeling the peptides with stable isotope. Both processes are compatible with protein fractionation or separation methods, such as subcellular fractionation, protein complex isolation and electrophoresis and chromatography, thereby providing additional biological context to the protein samples being analyzed.

In comparison to its nucleic acid-based counterpart, genomics, the experimental complexity of proteomics is far greater. The technology is also not as mature and, owing to the lack of amplification schemes akin to PCR, only proteins isolated from a natural source can be analyzed. Proteomic analyses are therefore generally limited by substrate. The complexities of the proteome arise because most proteins seem to be processed and modified in complex ways and can be the products of differential splicing; in addition, protein abundance spans a range estimated at five to six orders of magnitude for yeast cells⁴³ and more than ten orders of magnitude for human blood serum—for example, from interleukin-6 at ~2 pg/ml (ref. 44) to albumin at 50 mg/ml (ref. 45). Thus, the relatively low number of human genes predicted from the genome sequence^{17,18} has the potential to generate a proteome of enormous and as yet undetermined complexity.

It became rapidly apparent that a proteomic technology based strictly on 2DE was technically complex, labor- and therefore cost-intensive, and fundamentally limited. The increased use of MALDI-MS and ESI-MS/MS in the identification of 2DE-separated proteins showed that the incidence of comigrating proteins even in this, the highest resolving protein method known, was more prevalent than had been originally thought⁴⁶. Because quantification in 2DE relies on the assumption that one protein is present in each spot, comigration compromises such analyses. It was also observed that with conventional protein staining methods only a relatively small subset of a cellular proteome was apparent when unfractionated cell lysates were separated by 2DE^{46,47}. Thus, despite the maturity and unmatched performance of 2DE for separating intricate patterns of differentially modified and processed proteins⁴⁸, and despite the continuing evolution of 2DE separation^{49–52} and detection technology⁵³, alternative methods for large-scale protein expression analysis began to be investigated more vigorously.

Hunt and co-workers⁵⁴ laid the groundwork for a gel-independent approach to proteomics by demonstrating the ability of LC-MS/MS systems to handle extremely complex peptide mixtures. Antigen-presenting lymphocytes continually digest proteins and present some of the resulting peptides bound to major histocompatibility complex (MHC) proteins for immune surveillance. Hunt and co-workers⁵⁴ used immunoprecipitation to isolate the peptide-MHC complexes, extracted the antigenic peptides and subjected the complex peptide mixtures to successive LC-MS/MS analyses. They also used a specific cytotoxic T cell response as a bioassay to confirm the presence of antigenic peptides in each fraction and cor-

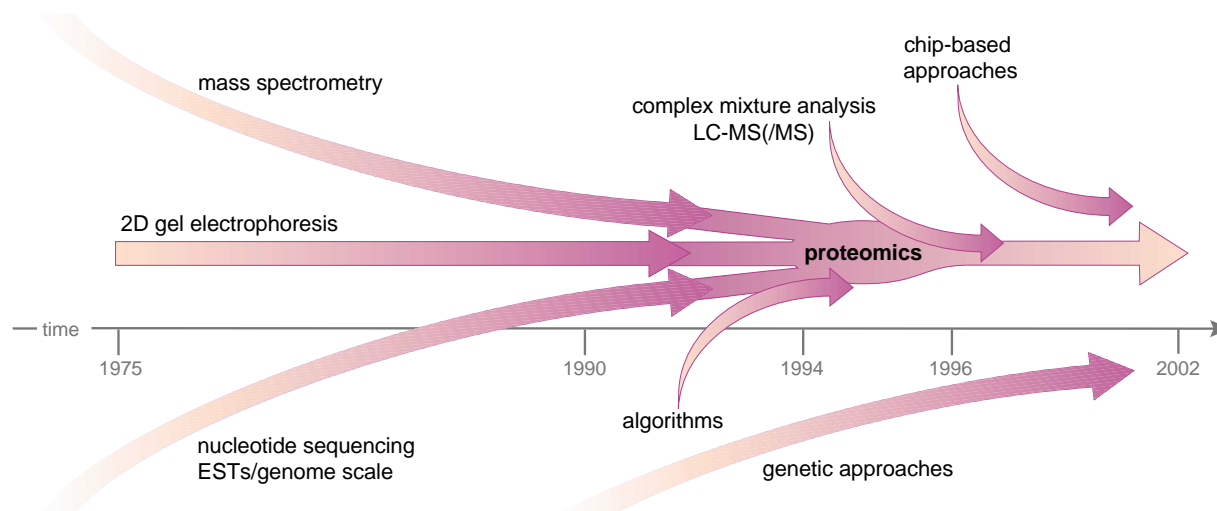
related this functional data with the mass spectrometric data, thereby identifying the sequence of the antigenic peptides^{55–57}. This approach clearly established LC-MS/MS as a powerful tool for analyzing complex peptide mixtures, and the application of this method to the analysis of peptide mixtures generated by the proteolysis of complex protein samples was a considerable step toward gel-independent proteomic technologies⁵⁸.

Gel-independent quantitative profiling of the proteome

The combination of LC-MS/MS and sequence database searching has been widely adopted for the analysis of complex peptide mixtures generated from the proteolysis of samples containing several proteins. This approach is often referred to as ‘shotgun’ proteomics^{59,60} and has the ability to catalog hundreds, or even thousands, of components contained in samples isolated from very different sources. Specific examples include the identification of proteins in the periplasmic space of bacteria⁶¹, yeast ribosomal complexes⁶², murine nuclear interchromatin granule clusters (nuclear speckles)⁶³, murine mitochondrial soluble intermembrane proteins⁶⁴, human urinary proteins⁶⁵, yeast TFIID-associated proteins⁶⁶, proteasomal proteins⁶⁷, human microsomal proteins⁶⁸, human membrane proteins⁶⁹ and yeast nuclear pore proteins⁷⁰ pre-fractionated by SDS polyacrylamide gel electrophoresis, and proteins from yeast lysates⁷¹.

Such studies have also highlighted the limitations of shotgun proteomics, including the difficulty of detecting and analyzing by collision-induced dissociation (CID) mass spectrometry all of the peptides in a sample, the qualitative nature of data-dependent experiments, and the challenge of processing the tens of thousands of CID spectra generated in a typical experiment—one of the many informatics challenges that still faces scientists in this field. On average, a protein digested with trypsin will generate 30–50 different peptides. A tryptic digest of the proteome of a typical human cell will therefore generate a peptide mixture containing at least hundreds of thousands of peptides. Even the most advanced LC-MS/MS systems cannot resolve and analyze such complexity in a reasonable amount of time.

To use LC-MS/MS for the analysis of most proteomes, therefore, a form of complexity reduction (fractionation) is required. Two approaches have been developed to tackle this problem. The first approach is the selective enrichment of a subset of peptides from a complex mixture. This has been mostly achieved by specifically targeting peptides that contain a distinguishing feature, such as chemically reactive sulfhydryl groups (cysteine



Karlis Rits

Fig. 4 Time line indicating the convergence of different technologies and resources into a proteomic process. Advances in mass spectrometry and the generation of large quantities of nucleotide sequence information, combined with computational algorithms that could correlate the two, led to the emergence of proteomics as a field.

residues)^{59,72} or residues that have been modified post-translationally with phosphate^{73,74} or carbohydrate^{75,76}. Ideally, such strategies would select precisely one idiotypic peptide per protein. Although this has not been realized, substantial reductions in sample complexity have been achieved. But the trade-off is the loss of those proteins that do not contain the selected feature.

The second approach relies on extended upstream fractionation of the complex peptide mixtures, with the aim of increasing the potential of the mass spectrometer to detect and to sequence all of the components of the sample. This has been implemented by using two or three orthogonal peptide separation methods in sequence, the most successful of which have involved cation exchange and capillary reverse-phase chromatography^{68,71,77}. Although both approaches—either by themselves or in combination—have produced evidence that proteome cataloguing using such shotgun proteomic strategies is feasible^{71,78,79}, a complete map of the proteome of any species has yet to be produced by any method.

For proteomic studies applying a forward (function to sequence) approach, determination of the sequence of the target proteins is usually a defined end point, because detailed functional analyses of the isolated species precede sequence analysis. For studies that apply reverse (sequence to function) approaches, knowing the sequence of the proteins in a sample is necessary but not sufficient. Reverse approaches, which are used in many proteomic studies, typically involve quantitative comparison of the protein profiles expressed by cells or tissues in different states.

The most valuable information on the system being studied is obtained from those proteins that are expressed differentially in a matrix of proteins of unchanged expression; therefore, proteomic technologies detecting differences in protein profiles need to be quantitative. Unfortunately, peptides analyzed in a mass spectrometer will produce different specific signal intensities depending on their chemical composition, on the matrix in which they are present and on other poorly understood variables. Thus, the intensity of a peptide ion signal does not accurately reflect the amount of peptide in a sample; in other words, mass spectrometry is inherently not a quantitative technique. However, two peptides of identical chemical structure that differ in

mass because they differ in isotopic composition are expected, according to stable isotope dilution theory⁸⁰, to generate identical specific signals in a mass spectrometer. To turn shotgun proteomics into a quantitative protein profiling method, therefore, stable isotope dilution has been combined with the complexity reduction techniques described above.

The first such approach was based on a class of reagents termed isotope-coded affinity tags or 'ICAT' (Box 1), LC-MS/MS and sequence database searching⁸¹. The reagents consist of an alkylating group (iodoacetic acid) that covalently attaches the reagent to reduced cysteine residues, a polyether mass-encoded linker containing either eight hydrogens (d0) or eight deuteriums (d8) that represents the isotope dilution and a biotin affinity tag through which tagged peptides are selectively isolated. For quantitative proteomics, the ICAT reagent approach, or similar techniques, now provides an alternative method to the subtractive 2DE-based approaches discussed above.

As shown in Fig. 5, the ICAT reagent approach involves labeling the cysteine residues in one sample with d0-ICAT reagent and the cysteine residues in a second sample with the d8-ICAT reagent. The samples are then combined. After optional protein enrichment and enzymatic digestion of the combined samples, the biotinylated ICAT-labeled peptides are enriched by means of avidin affinity chromatography and analyzed by LC-MS/MS. Each cysteinyl peptide appears as a pair of signals differing by the mass differential encoded in the mass tag. The ratio of these signal intensities precisely indicates the ratio of abundance of the protein from which the peptide originates, and the MS/MS spectrum of either isotopic form of the peptide allows the protein to be identified. Thus, in a single, automated operation this method identifies the proteins present in two related samples and determines the ratio of relative abundance.

Variations of this approach have been described. For example, alternative labeling chemistries have been explored^{82,83}; stable isotope labeling has been achieved by a solid-phase isotope tag transfer method⁸⁴; ¹⁶O or ¹⁸O has been incorporated from H₂¹⁶O or H₂¹⁸O, respectively, at the carboxy terminus of peptides during proteolytic cleavage by trypsin^{85–87}; and stable-isotope metabolic protein labeling has been attempted before mass

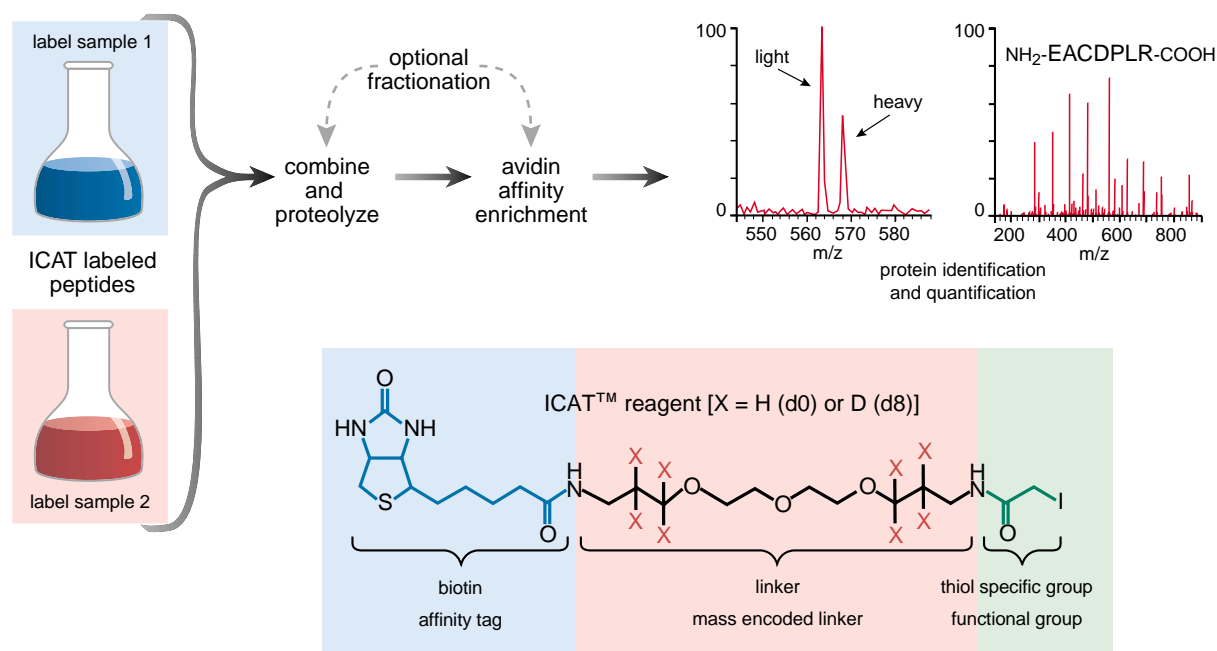


Fig. 5 Quantitative proteomics using ICAT reagents. The ICAT reagent comprises a protein reactive group (such as a sulfhydryl-specific reactivity), a mass-encoded linker and an affinity tag (such as biotin). Variations of any of the three can be used to facilitate the quantification of many different modifications or activities. The general scheme used for this reagent is shown: first, the protein reactive groups (such as cysteine residues) are labeled separately with either light or heavy reagent, and then the proteins are mixed and digested by enzymes; second, the labeled peptides are captured and then quantified and identified by LC-MS/MS.

spectrometric analysis of intact proteins or proteolytic peptides in a high LC-MS/MS system^{88–91}. Although isotope-tagging methods based on chemical labeling after isolation are compatible with essentially any protein sample, including organelles, body fluids or subcellular and biochemical fractions, the application of metabolic labeling is limited to those situations in which cells can be cultured in isotopically defined media.

Contemporary proteomics: a bewildering array of tools

From its inception to the present day, proteomics has evolved substantially (Fig. 4). Conceptually, proteomics has become a biological assay for the quantitative, subtractive analysis of complex protein samples. Technologically, proteomics has become a suite of relatively mature tools that support protein cataloguing and quantitative proteome measurements reliably, sensitively and at high throughput. The impressive successes of gene expression profiling ('DNA microarrays'; Box 1) for classifying cell types and cellular states in health and disease^{92–95} and for dissecting cellular pathways^{96,97} have illustrated the value of the information obtained by systematic, quantitative expression profiling experiments, and several insights have become apparent from such studies.

First, global data sets are rich in information but difficult to analyze using traditional knowledge-based interpretation. Second, the more data the better: in comparison to interpretations of one or a few expression profiles, it has been much more informative to collect several global data sets on the same, differentially perturbed system, and to use mathematical tools such as cluster analysis or singular value decomposition^{98,99} to extract biological insights or to formulate hypotheses. Third, proteomic and genomic measurements done on the same system provide complementary information¹⁰⁰ because neither the steady-state quantity of^{43,101,102} nor the response to perturbation-induced changes in mRNA or protein is mutually predictable^{103,104}.

Extrapolating from insights gained from the comparison of mRNA and protein expression profiles, it is expected that additional systematic proteomic data, including activity profiles, interaction maps and profiles of (regulatory) modifications, will provide yet further insights into the structure, function and control of biological systems.

Much of the effort in contemporary proteomics is therefore directed toward the development of suitable platforms on which to generate these data. These efforts can be broadly grouped into three categories based on mass spectrometry, microchips and genetics (Fig. 4).

Mass spectrometry-based methods for studying the proteome

The success of combining the selective chemical labeling of proteins, stable isotope tagging, LC-MS/MS and database searching as the basis for quantitative protein profiling suggested that the same general approach might be adapted to other types of quantitative proteomic measurement. Over the years, protein chemistry has provided many reagents for selectively capturing classes of protein with specific biochemical properties, including lectins to capture glycoproteins, immobilized metal ions to affinity-capture phosphorylated peptides, and suicide enzyme substrates to capture specific enzymes. If such reagents could be adapted to interact with their targets specifically and tightly (ideally covalently), and if they could be made compatible with stable isotope tagging methods, they might be used to profile quantitatively the targeted functional group, activity or other property on a proteome-wide scale.

Activity-based reagents covalently label only enzymatically active forms of proteins and can thus directly determine activity as opposed to total protein^{105–107}. For example, the conversion of enzyme substrates into enzyme suicide inhibitors has been achieved for serine hydrolases and cysteine proteases. In fact, the

study of proteases and their substrates and inhibitors in an organism has been called “degradomics”¹⁰⁸, and several studies on the cysteine and serine protease families have provided ‘proof of principle’ for the idea. These enzymes form covalently bound acyl intermediates during their catalytic cycle. This property can be used to introduce tagged chemical probes by means of a structure that mimics a general inhibitor and captures the active form of the proteases covalently. Fluorescently tagged probes^{107,109–111} or biotinylated probes^{112,113} have been used successfully, and isotope-tagged versions of these reagents are being developed.

Although the range of enzyme families to which this strategy has been applied so far is limited, additional enzyme families such as glycosidases also seem amenable to activity profiling¹¹⁴. In addition, drug screening projects that have been aborted because of poor compound specificity provide a rich source of potential leads for the development of class-specific reagents. For example, methods for capturing ATP-binding proteins have been developed to enrich a range of proteins, including protein kinases, with the potential for subsequent (activity) profiling of the enriched proteins by quantitative proteomic strategies¹¹⁵.

When enzyme activities cannot be measured directly, they may be inferred from the analysis of converted substrates. This has been achieved by the quantitative measurement of substrate conversion using isotope-tagged substrates and mass spectrometry¹¹⁶, or by trapping enzyme substrates on chemically or genetically altered enzymes and then analyzing the isolated enzyme-substrate conjugates by mass spectrometry. For example, chemically modified trypsin¹¹⁷ has been used to identify incompletely processed neuropeptides^{118,119}, and protein tyrosine phosphatases in which the invariant catalytic amino acid aspartic acid is mutated to alanine have been used as ‘substrate traps’, allowing the identification of their physiological substrates^{120,121}.

Selective reagents or chemical reactions have been used to extract phosphorylated peptides from complex mixtures. Immobilized metal affinity chromatography^{73,122,123} has been used successfully to enrich phosphopeptides from relatively simple mixtures, but it lacks the specificity to be effective with more-complex peptide samples. To increase the specificity of metal affinity chromatography for phosphopeptides, Ficarro *et al.*⁷⁴ eliminated interactions between the resin and carboxyl groups by capping the carboxyl groups with methyl ester groups. This allowed them to extract phosphopeptides selectively from the tryptic digest of a yeast lysate. Subsequent LC-MS/MS analysis of the selected peptides indicated the activation state of some yeast kinases from their phosphorylation state—information that is not accessible by most conventional proteomic approaches. Two covalent chemistries, one based on β -elimination reactions^{124,125} and one based on the formation of phosphoramidates¹²⁶, have been developed with the same objective. Although none of these methods is currently capable of quantitative phosphoprotein profiling on a proteome-wide scale, these early studies represent a path to such a technology.

Mass spectrometry-based proteomics is also rapidly becoming the method of choice for analyzing functional protein complexes^{62,66,70,127,128}. By providing a means to identify multiple members of complexes, this approach complements the view recently emphasized by Hartwell *et al.*¹²⁹ that a cell is a collection of interconnected modules (Fig. 1)—that is, groups of proteins with many network interconnections that act synergistically to execute a particular cellular function.

Indeed, two large-scale ‘protein complex’ studies recently used affinity-tagged bait proteins expressed in yeast cells to isolate the bait protein together with its associated proteins^{130,131}. The composition of the isolated complexes was analyzed by gel electrophoresis and mass spectrometry, and thousands of protein

interactions—many previously unknown—were identified. But the poor overlap between the two data sets when the same bait proteins were used, and between either data set and results obtained by a genetic yeast two-hybrid screen (see below)¹³², suggests that the conditions and parameters for such experiments need further optimization and standardization.

In an alternative approach using ICAT reagents to label the components of a target protein complex and a suitable control isolate before quantitative LC-MS/MS analysis, specific components of the complex could be distinguished from nonspecifically associated proteins and changes in the composition of protein complexes isolated from cells at different states could be observed readily¹³³.

In the near future, it can be expected that quantitative proteomic technologies will mature to become the obvious choice for systematically determining the important, diverse properties of proteins, including their activity and state of modification, and the composition and dynamics of their functional modules.

Chip-based methods for studying the proteome

Complementary DNA or oligonucleotide microarrays have proved invaluable for analyzing transcript levels in several biological systems^{5,134,135}, and technological improvements continue to increase their utility^{136,137}. Array-based profiling techniques are conceptually simple. A probe that is specific for a particular analyte is placed at a defined position in a two-dimensional array, and the interaction of the probe with its target molecule is detected. Signals indicate that interactions have occurred, their intensity and position on the array are recorded, and thus the probed molecules and their quantity can be identified.

It therefore seemed obvious to apply similar principles to proteomic analyses—in particular to protein expression profiling—by generating ordered arrays of specific protein-binding modules^{138–140}. Such modules have included phage library-selected scFv antibodies¹⁴¹, minibodies^{142,143}, cyclic peptides¹⁴⁴, reagents resulting from the scaffold engineering of various proteins^{145–147}, aptamers¹⁴⁸, antibodies¹⁴⁹ and antibody mimics¹⁵⁰. In practice, however, the translation of array-based profiling from nucleic acids to proteins has faced many difficulties¹⁵¹.

First, unlike cDNAs, proteins usually need to be captured in their native conformation. This restricts the range of conditions that can be applied to maintain solubility, to optimize interactions and to remove nonspecific contaminants. Second, because at present proteins cannot be amplified before analysis, detection methods must be very sensitive. Third, proteins have no inherent properties that make them measurable at high sensitivity, and the attachment of a detectable tag is prone to interfere with a protein’s interactions. Fourth, the interactions between proteins and their binding reagents are less specific and of lower and more variable affinity than those between Watson-Crick base-paired nucleic acids. This increases the potential for crossreactivity and complicates quantification, because substantial differences in the dissociation constants for each protein-protein interaction must be taken into account.

The challenge of obtaining antibodies of sufficient specificity to make protein expression arrays possible was recently demonstrated in studies in which monoclonal antibodies that were probed against numerous expressed proteins showed a considerable propensity to crossreact with proteins other than the intended target protein^{152,153}. In spite of advances such as the development of high-sensitivity detection methods¹⁵⁴ and chip surface engineering¹⁵⁵, it is therefore unsurprising that the most successful protein microarrays for measuring an analyte in a complex protein mixture use relatively simple arrays of well-characterized antibodies¹³⁸, most notably those directed towards cytokines¹⁵⁴.

At present, the use of 'protein chips' has been more successful for systematically measuring or inferring functional properties of proteins. Direct functional measurements include the activities of diverse enzymes¹⁵⁶ such as protein kinases^{157,158}, protein-DNA interactions^{159,160}, profiles of reaction antibodies (autoantibodies) in blood serum and other clinical samples¹⁵², and the interactions of proteins with small molecules¹⁶¹. Using cDNA arrays assayed with probes generated from mRNAs extracted from membrane-associated polysomes, Diehn *et al.*¹⁶² could infer the identity of secreted and membrane-associated proteins. Similarly, in a technique that they call 'translation state array analysis', Morris and co-workers¹⁶³ have profiled polysome-associated mRNAs using conventional cDNA array technology to distinguish actively translated from non-translated mRNAs. As the chemistry of protein attachment continues to improve, the size of the arrays decreases¹⁶⁴; and, as different types of array are developed, additional uses for this technology will undoubtedly emerge.

Genetic methods for studying the proteome

Genetic methods for interrogating the proteome are generally based on recombinant DNA techniques that introduce different tags into all or selected proteins expressed by a cell. These tags can be used to observe directly or to infer specific properties of the tagged proteins.

The strengths of genetic methods are their ability to target potentially every protein, to use selection for inferring function, to assay the protein in a cellular environment and to engineer cell strains with specific properties, as well as the ease of assay automation¹⁶⁵. The weaknesses of genetic methods are that tagging itself can potentially interfere with the observed function and that the range of species amenable to rigorous genetic engineering is limited. This means that most observations—particularly those on mammalian proteins—are made in a heterologous environment and that most readouts, with the exception of certain fluorescence measurements, are indirect.

The prototypical genetic proteomic assay is the 'yeast two-hybrid system' (Box 1), which was initially developed for detecting protein-protein interactions¹⁶⁶ and has been reviewed recently¹⁶⁷. This method has been used for large-scale protein interaction screens in, for example, bacteriophage¹⁶⁸, vaccinia¹⁶⁹, yeast^{170–172}, *Helicobacter pylori*¹⁷³ and *Caenorhabditis elegans*^{174,175}. The relative merits of the systematic protein interaction maps generated by two-hybrid assay and those generated by mass spectrometry^{130,131}, together with the matter of which one, if any, best represents the true space of protein interactions in a cell, is an important and currently unresolved issue¹³³.

This type of method has also been extended to facilitate the determination of protein-mRNA interactions¹⁷⁶ and for sensing protein-ligand binding¹⁷⁷. Related approaches in which proteins are expressed as fusions of green fluorescent protein or one of its variants^{178,179} are proving invaluable for probing several functional properties of proteins by direct *in vivo* observation¹⁸⁰. Although such observations have been made mostly during specific research projects, an initial compendium of the subcellular localization of the yeast proteome¹⁸¹ clearly shows the potential of this approach as a proteomic assay.

A window to the future

The scope of proteomic investigations has considerably broadened in the past few years. Whereas initial efforts were focused on determining the identity and quantity of proteins using a narrow selection of methods, many emerging technologies now attempt to measure systematically all of the biologically

important properties of proteins. Although few, if any, of these methods have reached the status of validated proteomic tools, the rapid pace at which they are developing suggests that the rich and varied sources of information contained in the proteome will become increasingly accessible. The main challenges of the future will be the validation, visualization, integration and interpretation, in a biological context, of the vast amounts of diverse data generated by the application of proteomic and genomic discovery science tools.

In 1996, the first complete genomic sequence of a eukaryotic species, *Saccharomyces cerevisiae*, was published¹⁶. The early availability of this resource, the richness of knowledge already acquired through decades of hypothesis-driven research and the ease with which it could be experimentally manipulated made yeast the model organism in which to test and to validate most of the technologies and approaches discussed above. Because in many respects the analysis of mammalian cells is more complex and technically challenging, yeast can be viewed as a window into the future of genomics-based and proteomics-based biological research. What insights might this view provide?

First, there will be a convergence of discovery science and hypothesis-driven research. The beginning of such a convergence is already apparent in information resources such as the yeast protein database¹⁸² and the *Saccharomyces* Genome Database¹⁸³, in which systematically collected data and the results from hypothesis-driven research published in the scientific literature have been combined into highly useful resources for experimental biologists. In addition, a recent study has shown the value of this union of data as an experimental strategy to gain insights into the physiology of a cell¹⁰⁰. In this study, both genomic and proteomic data were collected from yeast cells in which all of the known components of the galactose induction pathway had been perturbed systematically. The different data were integrated into a mathematical model consistent with the available information. This model was then used to predict previously unknown interactions within the pathway and between the galactose induction pathway and other cellular processes. Some of these predictions were then verified experimentally¹⁰⁰.

Second, systems biology approaches will detect connections between broad cellular functions and pathways that were neither apparent nor predictable despite decades of biochemical and genetic analysis of the biological system in question. This has been validated broadly by numerous mRNA expression studies, and also by studies based on quantitative protein profiling (refs. 184, 185). The rich and diverse information represented in large proteomic data sets is expected to accelerate our understanding of the interdependence of cellular processes.

Third, our ability to collect large proteomic data sets already outstrips our ability to validate, to interpret and to integrate such data for the purpose of creating biological knowledge. Therefore, software tools will be developed to help manage, interpret, integrate and understand proteomic data. The lack of suitable software tools currently limits essentially all areas of proteomic data analysis, from database searching using MS/MS spectra to the assembly of large data sets containing different types of data in relational databases (Fig. 6). To derive value from the data that goes beyond an initial scan for 'interesting observations' and to make data portable and comparable, it will be necessary to develop algorithms that assign a score to each data point that estimates the probability that the observation is correct. Just as the assignment of quality scores to each base in DNA sequencing using the algorithm Phred^{186,187} was essential for the success of genome sequencing programs, it can be expected that probability-based scores calculated for proteomic data¹⁸⁸ will have a similar impact on proteomics.

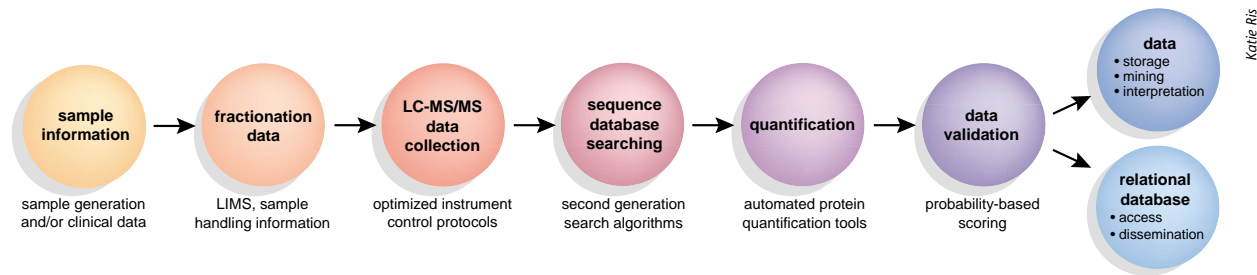


Fig. 6 Quantitative proteomics and informatics. Brief descriptions of the informatics requirements for each of the processes of biological analysis are listed. Handling these data requires significant computational infrastructure if it is to be carried out repeatedly on a large scale. Many of the algorithms used in the process are still not mature.

It may seem that these trends might be realized with incremental improvements of current proteomic technologies, and this may be largely true for unicellular organisms. But multicellular organisms contain levels of organization, such as the arrangement of cells in tissues, and carry out processes, such as memory and immunity, for which there are no counterparts in yeast or other unicellular organisms. The transition from the application of proteomic strategies in unicellular organisms to their application in higher eukaryotes is therefore much more complex than one could estimate from a simple comparison of the numbers of genes in the respective genomes.

Some of these complexities, in particular the organization of cells into tissues and to some extent the topology of the proteins contained in them, are being investigated by an innovative emerging technique called imaging mass spectrometry^{189,190}. Although at present this technique does not afford the sensitivity and resolution to study biological processes, it is immediately useful for generating diagnostic patterns. Despite these advances and substantial increases in the sensitivity, resolution and mass accuracy of new types of mass spectrometer such as linear ion traps¹⁹¹, MALDI-TOF-TOF¹⁹² and FT-ICR-MS^{79,193}, proteomics—especially when applied to higher eukaryotic species—will remain limited by technology for the foreseeable future.

We see four main challenges to be addressed in order for proteomics to have a substantial impact on eukaryotic biology within the systems biology model. The first challenge is the enormous complexity of the proteome. For some proteins, in excess of 1,000 variants (splice and translation isoforms, differentially modified and processed species) have been described¹⁹⁴. The detection, and particularly the molecular analysis of this complexity, remains an unmatched task. The second challenge is the need for a general technology for the targeted manipulation of gene expression in eukaryotic cells. An approach that has proved successful for the systematic analysis of biological systems relies on iterative cycles of targeted perturbations of the system under study and the systematic analysis of the consequences of each perturbation¹⁰⁰. Although recent advances in using RNA interference in higher eukaryotic cells open up exciting possibilities, the general targeted manipulation of biological systems in these species remains unsolved. The third challenge is the limited throughput of today's proteomic platforms: iterative, systematic measurements on differentially perturbed systems demand a sample throughput that is not matched by current proteomic platforms. The fourth challenge is the lack of a general technique for the absolute quantification of proteins. The ability to quantify proteins absolutely, thereby eliminating the need for a reference sample, would have far-reaching implications for proteomics—from the determination of the stoichiometry of protein complexes to the design of clinical studies aimed at discovering diagnostic markers.

Fortunately, proteomics will have an impact on clinical and biological research well before these challenges are met. We expect that precise clinical diagnosis based on highly discriminating patterns of proteins in easily accessible samples, particularly body fluids, may be the area in which proteomics will make its first significant contribution^{195,196}. In the short term, proteomics also can be expected to provide partial data sets of sufficient quality, density and information content to provide the basis for generating sophisticated mathematical models of biological processes that will be able to simulate system properties such as adaptation or robustness^{197,198}, which may not be apparent from the analysis of isolated elements of a system.

In its first decade, the field of proteomics has grown rapidly to encompass numerous advanced technologies that strive to provide the molecular data necessary for a comprehensive understanding of biological processes. Although much ground has been covered, continued advances in methods, instrumentation and computational analysis will be needed to get closer to the workings of biology through the analysis of these systems.

Acknowledgments

We would like to thank L. Feltz for administrative assistance and J. Watts for review of the manuscript.

1. Aebersold, R., Hood, L.E. & Watts, J.D. Equipping scientists for the new biology. *Nat. Biotechnol.* **18**, 359 (2000).
2. Thornton, J. Structural genomics takes off. *Trends Biochem. Sci.* **26**, 88–99 (2001).
3. Aebersold, R. & Patterson, S.D. Current problems and technical solutions in protein biochemistry. In *PROTEINS: Analysis & Design* (ed. Angeletti, R.H.) 3–120 (Academic, San Diego, 1998).
4. Adams, M.D. et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377**, 3–174 (1995).
5. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
6. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
7. Anderson, N.L., Hofmann, J.P., Gemmill, A. & Taylor, J. Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. *Clin. Chem.* **30**, 2031–2036 (1984).
8. Tarroux, P., Vincens, P. & Rabilloud, T. HERMeS: A second generation approach to the automatic analysis of two-dimensional electrophoresis gels. Part V: Data analysis. *Electrophoresis* **8**, 187–199 (1987).
9. Aebersold, R.H., Leavitt, J., Saavedra, R.A., Hood, L.E. & Kent, S.B. Internal amino acid sequence analysis of proteins separated by one- or two-dimensional gel electrophoresis after *in situ* protease digestion on nitrocellulose. *Proc. Natl. Acad. Sci.* **84**, 6970–6974 (1987).
10. Vandekerckhove, J., Bauw, G., Puype, M., Van Damme, J. & Van Montagu, M. Protein-blotting on polybrene-coated glass-fiber sheets. *Eur. J. Biochem.* **152**, 9–19 (1985).
11. Tempst, P., Link, A.J., Riviere, L.R., Fleming, M. & Elicone, C. Internal sequence analysis of proteins separated on polyacrylamide gels at the submicrogram level: improved methods, applications and gene cloning strategies. *Electrophoresis* **11**, 537–553 (1990).
12. Adams, M.D. et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651–1656 (1991).
13. Adams, M.D., Kerlavage, A.R., Fields, C. & Venter, J.C. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat. Genet.* **4**, 256–267 (1993).

14. Zhang, L. *et al.* Gene expression profiles in normal and cancer cells. *Science* **276**, 1268–1272 (1997).
15. Bonaldo, M.F., Lennon, G. & Soares, M.B. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* **6**, 791–806 (1996).
16. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 563–567 (1996).
17. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
18. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
19. Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. & Whitehouse, C.M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71 (1989).
20. Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10000 daltons. *Anal. Chem.* **60**, 2299–2301 (1988).
21. Tanaka, K., Ido, Y., Akita, S., Yoshida, Y. & Yoshida, T. Detection of high mass molecules by laser desorption time-of-flight mass spectrometry. In *Proc. 2nd Japan-China Joint Symp. Mass Spectrom.* (eds. Matsuda, H. & Xiao-tian, L.) 185–188 (Osaka, Japan, 1987).
22. Tanaka, K. *et al.* Protein and polymer analyses up to m/z 100,000 by laser ionization TOF-MS. *Rapid Commun. Mass Spectrom.* **2**, 151–153 (1988).
23. Henzel, W.J. *et al.* 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).
24. Mann, M., Hojrup, P. & Roepstorff, P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **22**, 338–345 (1993).
25. Pappin, D.J.C., Hojrup, P. & Bleasby, A.J. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3**, 327–332 (1993).
26. James, P., Quadroni, M., Carafoli, E. & Gonnet, G. Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* **195**, 58–64 (1993).
27. Yates, J.R., III, Speicher, S., Griffin, P.R. & Hunkapiller, T. Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* **214**, 397–408 (1993).
28. Patterson, S.D. & Aebersold, R. Mass spectrometric approaches for the identification of gel-separated proteins. *Electrophoresis* **16**, 1791–1814 (1995).
29. Eng, J.K., McCormack, A.L. & Yates, J.R., III. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
30. Mann, M. Sequence database searching by mass spectrometric data. In *Microcharacterization of Proteins* (eds. Kellner, R., Lottspeich, F. & Meyer, H.E.) 223–245 (VCH, Weinheim, 1994).
31. Gras, R. & Muller, M. Computational aspects of protein identification by mass spectrometry. *Curr. Opin. Mol. Ther.* **3**, 526–532 (2001).
32. Wilm, M.S. & Mann, M. Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? *Int. J. Mass Spectrom. Ion Proc.* **136**, 167–180 (1994).
33. Wilm, M. *et al.* Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass-spectrometry. *Nature* **379**, 466–469 (1996).
34. Scheele, G.A. Two-dimensional gel analysis of soluble proteins. Characterization of guinea pig exocrine pancreatic proteins. *J. Biol. Chem.* **250**, 5375–5385 (1975).
35. Klose, J. Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues: a novel approach to testing for induced point mutations in mammals. *Humangenetik* **26**, 231–243 (1975).
36. O'Farrell, P.H. High resolution two-dimensional gel electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021 (1975).
37. Anderson, N.G. & Anderson, L. The human protein index. *Clin. Chem.* **28**, 739–748 (1982).
38. Garrels, J.I. The QUEST system for quantitative analysis of two-dimensional gels. *J. Biol. Chem.* **264**, 5269–5282 (1989).
39. Garrels, J.I. & Franza, B.R., Jr. Transformation-sensitive and growth-related changes of protein synthesis in REF52 cells. A two-dimensional gel analysis of SV40-, adenovirus-, and Kirsten murine sarcoma virus-transformed rat cells using the REF52 protein database. *J. Biol. Chem.* **264**, 5299–5312 (1989).
40. Garrels, J.I. & Franza, B.R., Jr. The REF52 protein database. Methods of database construction and analysis using the QUEST system and characterizations of protein patterns from proliferating and quiescent REF52 cells. *J. Biol. Chem.* **264**, 5283–5298 (1989).
41. Anderson, N.L., Matheson, A.D. & Steiner, S. Proteomics: applications in basic and applied biology. *Curr. Opin. Biotechnol.* **11**, 408–412 (2000).
42. Wilkins, M.R. *et al.* Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotech. Gen. Eng. Rev.* **13**, 19–50 (1995).
43. Gygi, S.P., Rochon, Y., Franza, B.R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
44. Lai, R. *et al.* Prognostic value of plasma interleukin-6 levels in patients with chronic lymphocytic leukemia. *Cancer* **95**, 1071–1075 (2002).
45. Ritchie, R.F., Palomaki, G.E., Neveux, L.M. & Navolotskaia, O. Reference distributions for the negative acute-phase proteins, albumin, transferrin, and transthyretin: a comparison of a large cohort to the world's literature. *J. Clin. Lab. Anal.* **13**, 280–286 (1999).
46. Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y. & Aebersold, R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. USA* **97**, 9390–9395 (2000).
47. Corthals, G.L., Wasinger, V.C., Hochstrasser, D.F. & Sanchez, J.C. The dynamic range of protein expression: a challenge for proteomic research. *Electrophoresis* **21**, 1104–1115 (2000).
48. Gaus, C., Kalkum, M., Lowe, M., Lehrach, H. & Klose, J. 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).
49. Rabilloud, T. Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. *Proteomics* **2**, 3–10 (2002).
50. Herbert, B. Advances in protein solubilisation for two-dimensional electrophoresis. *Electrophoresis* **20**, 660–663 (1999).
51. Gorg, A. *et al.* The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **21**, 1037–1053 (2000).
52. Unlu, M., Morgan, M.E. & Minden, J.S. Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis* **18**, 2071–2077 (1997).
53. Rabilloud, T., Strub, J.M., Luche, S., van Dorsselaer, A. & Lunardi, J. A comparison between Sypro Ruby and ruthenium II tris (bathophenanthroline disulfonate) as fluorescent stains for protein detection in gels. *Proteomics* **1**, 699–704 (2001).
54. Appella, E., Padlan, E.A. & Hunt, D.F. Analysis of the structure of naturally processed peptides bound by class I and class II major histocompatibility complex molecules. *EXS* **73**, 105–119 (1995).
55. Hunt, D.F. *et al.* Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255**, 1261–1263 (1992).
56. Henderson, R.A. *et al.* HLA-A2.1-associated peptides from a mutant cell line: a second pathway of antigen presentation. *Science* **255**, 1264–1266 (1992).
57. Hunt, D.F. *et al.* Peptides presented to the immune system by the murine class II major histocompatibility complex molecule I-Ad. *Science* **256**, 1817–1820 (1992).
58. Yates, J.R., III, McCormack, A.L., Schieltz, D., Carmack, E. & Link, A. Direct analysis of protein mixtures by tandem mass spectrometry. *J. Prot. Chem.* **16**, 495–497 (1997).
59. Spahr, C.S. *et al.* Simplification of complex peptide mixtures for proteomic analysis: reversible biotinylation of cysteinyl peptides. *Electrophoresis* **21**, 1635–1650 (2000).
60. Wolters, D.A., Washburn, M.P. & Yates, J.R., III. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73**, 5683–5690 (2001).
61. Link, A.J., Carmack, E. & Yates, J.R., III. A strategy for the identification of proteins localized to subcellular spaces: Application to *E. coli* periplasmic proteins. *Int. J. Mass Spectrom. Ion Proc.* **160**, 303–316 (1997).
62. Link, A.J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682 (1999).
63. Mintz, P.J., Patterson, S.D., Neuwald, A.F., Spahr, C.S. & Spector, D.L. Purification and biochemical characterization of interchromatin granule clusters. *EMBO J.* **18**, 4308–4320 (1999).
64. Patterson, S.D. *et al.* Mass spectrometric identification of proteins released from mitochondria undergoing permeability transition. *Cell Death Diff.* **7**, 137–144 (2000).
65. Spahr, C.S. *et al.* Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry I. Profiling an unfractionated tryptic digest. *Proteomics* **1**, 93–107 (2001).
66. Sanders, S.L., Jennings, J., Canutescu, A., Link, A.J. & Weil, P.A. Proteomics of the eukaryotic transcription machinery: identification of proteins associated with components of yeast TFIID by multidimensional mass spectrometry. *Mol. Cell. Biol.* **22**, 4723–4738 (2002).
67. Verma, R. *et al.* Proteasomal proteomics: identification of nucleotide-sensitive proteasome-interacting proteins by mass spectrometric analysis of affinity-purified proteasomes. *Mol. Biol. Cell* **11**, 3425–3439 (2000).
68. Han, D.K., Eng, J., Zhou, H. & Aebersold, R. Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* **19**, 946–951 (2001).
69. Simpson, R.J. *et al.* Proteomic analysis of the human colon carcinoma cell line (LIM 1215): Development of a membrane protein database. *Electrophoresis* **21**, 1707–1732 (2000).
70. Rout, M.P. *et al.* The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148**, 635–651 (2000).
71. Washburn, M.P., Wolters, D. & Yates, J.R., III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).
72. Patterson, S.D. Using MS fragment-ion data to identify proteins from large sequence databases. In *Proteomics, Integrating Protein-based Tools and Applications for Drug Discovery* (ed. Savage, L.M.) 127–135 (International Business Communications, Southborough, 1998).
73. Nuwaysir, L.M. & Stults, J.T. Electrospray ionization mass spectrometry of phosphopeptides isolated by on-line immobilized metal-ion affinity chromatography. *J. Am. Soc. Mass Spectrom.* **4**, 662–669 (1993).
74. Ficarro, S.B. *et al.* Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **20**, 301–305 (2002).
75. Hayes, B.K., Greis, K.D. & Hart, G.W. Specific isolation of O-Linked N-acetylglucosamine glycopeptides from complex mixtures. *Anal. Biochem.* **228**, 115–122 (1995).
76. Greis, K.D. *et al.* Selective detection and site-analysis of O-GlcNAc-modified glycopeptides by β -elimination and tandem electrospray mass spectrometry. *Anal. Biochem.* **234**, 38–49 (1996).
77. Davis, M.T. *et al.* Automated LC-LC-MS-MS platform using binary ion-exchange and gradient reversed-phase chromatography for improved proteomic analyses. *J. Chromatogr. B* **752**, 281–291 (2001).
78. Gygi, S.P., Rist, B., Griffin, T.J., Eng, J. & Aebersold, R. Proteome analysis of low-abundance proteins using multidimensional chromatography and isotope-coded affinity tags. *J. Proteome Res.* **1**, 47–54 (2002).
79. Smith, R.D. *et al.* An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2**, 513–523 (2002).
80. De Leenheer, A.P. & Thienpont, L.M. Application of isotope dilution-mass spectrometry in clinical chemistry, pharmacokinetics, and toxicology. *Mass Spectrom. Rev.* **11**, 249–307 (1992).
81. Gygi, S.P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999 (1999).
82. Munchbach, M., Quadroni, M., Miotto, G. & James, P. Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. *Anal. Chem.* **72**, 4047–4057 (2000).
83. Cagney, G. & Emili, A. De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nat. Biotechnol.* **20**, 163–170 (2002).
84. Zhou, H., Ranish, J.A., Watts, J.D. & Aebersold, R. Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. *Nat. Biotechnol.* **20**, 512–515 (2002).
85. Mirgorodskaya, O.A. *et al.* Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using ^{18}O -labeled internal standards. *Rapid Commun. Mass Spectrom.* **14**, 1226–1232 (2000).

86. Yao, X., Freas, A., Ramirez, J., Demirev, P.A. & Fenselau, C. Proteolytic ¹⁸O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal. Chem.* **73**, 2836–2842 (2001).
87. Uttenweiler-Joseph, S., Neubauer, G., Christoforidis, S., Zerial, M. & Wilm, M. Automated *de novo* sequencing of proteins using the differential scanning technique. *Proteomics* **1**, 668–682 (2001).
88. Oda, Y., Huang, K., Cross, F.R., Cowburn, D. & Chait, B.T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96**, 6591–6596 (1999).
89. Conrads, T.P. et al. Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and ¹⁵N-metabolic labeling. *Anal. Chem.* **73**, 2132–2139 (2001).
90. Smith, R.D. et al. Rapid quantitative measurements of proteomes by Fourier transform ion cyclotron resonance mass spectrometry. *Electrophoresis* **22**, 1652–1668 (2001).
91. Washburn, M.P., Ulaszek, R., Deciu, C., Schieltz, D.M. & Yates, J.R. III. Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Anal. Chem.* **74**, 1650–1657 (2002).
92. Siebert, R., Rosenwald, A., Staudt, L.M. & Morris, S.W. Molecular features of B-cell lymphoma. *Curr. Opin. Oncol.* **13**, 316–324 (2001).
93. Bangur, C.S. et al. Identification of genes over-expressed in small cell lung carcinoma using suppression subtractive hybridization and cDNA microarray expression analysis. *Oncogene* **21**, 3814–3825 (2002).
94. van't Veer, L.J. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
95. Spellman, P.T. et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
96. Hughes, T.R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
97. 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).
98. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
99. Alter, O., Brown, P.O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97**, 10101–10106 (2000).
100. Ideker, T. et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
101. Anderson, L. & Seilhamer, J. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18**, 533–537 (1997).
102. Futcher, B., Latter, G.I., Monardo, P., McLaughlin, C.S. & Garrels, J.I. A sampling of the yeast proteome. *Mol. Cell. Biol.* **19**, 7357–7368 (1999).
103. Betts, J.C., Lukey, P.T., Robb, L.C., McAdam, R.A. & Duncan, K. Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. *Mol. Microbiol.* **43**, 717–731 (2002).
104. Griffin, T.J. et al. Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **1**, 323–333 (2002).
105. Adam, G.C., Cravatt, B.F. & Sorensen, E.J. Profiling the specific reactivity of the proteome with non-directed activity-based probes. *Chem. Biol.* **8**, 81–95 (2001).
106. Adam, G.C., Sorensen, E.J. & Cravatt, B.F. Proteomic profiling of mechanistically distinct enzyme classes using a common chemotype. *Nat. Biotechnol.* **20**, 805–809 (2002).
107. Greenbaum, D., Medzihradsky, K.F., Burlingame, A. & Bogoy, M. Epoxide electrophiles as activity-dependent cysteine protease profiling and discovery tools. *Chem. Biol.* **7**, 569–581 (2000).
108. Lopez-Otin, C. & Overall, C.M. Protease degradomics: a new challenge for proteomics. *Nat. Rev. Mol. Cell Biol.* **3**, 509–519 (2002).
109. Bogoy, M., Shin, S., McMaster, J.S. & Ploegh, H.L. Substrate binding and sequence preference of the proteasome revealed by active-site-directed affinity probes. *Chem. Biol.* **5**, 307–320 (1998).
110. Cravatt, B.F. & Sorensen, E.J. Chemical strategies for the global analysis of protein function. *Curr. Opin. Chem. Biol.* **4**, 663–668 (2000).
111. Greenbaum, D. et al. Chemical approaches for functionally probing the proteome. *Mol. Cell. Proteomics* **1**, 60–68 (2002).
112. Kidd, D., Liu, Y. & Cravatt, B.F. Profiling serine hydrolase activities in complex proteomes. *Biochemistry* **40**, 4005–4015 (2001).
113. Liu, Y., Patricelli, M.P. & Cravatt, B.F. Activity-based protein profiling: the serine hydrolases. *Proc. Natl. Acad. Sci. USA* **96**, 14694–14699 (1999).
114. Withers, S.G. & Aebersold, R. Approaches to labeling and identification of active-site residues in glycosidases. *Protein Sci.* **4**, 361–372 (1995).
115. Haystead, C.M., Gregory, P., Sturgill, T.W. & Haystead, T.A. γ -Phosphate-linked ATP-sepharose for the affinity purification of protein kinases. Rapid purification to homogeneity of skeletal muscle mitogen-activated protein kinase kinase. *Eur. J. Biochem.* **214**, 459–467 (1993).
116. Turecek, F. Mass spectrometry in coupling with affinity capture-release and isotope-coded affinity tags for quantitative protein analysis. *J. Mass Spectrom.* **37**, 1–14 (2002).
117. Kumazaki, T., Terasawa, K. & Ishii, S. Affinity chromatography on immobilized anhydrotrypsin: general utility for selective isolation of C-terminal peptides from protease digests of proteins. *J. Biochem.* **102**, 1539–1546 (1987).
118. Fricker, L.D. et al. Identification and characterization of proSAAS, a granin-like neuroendocrine peptide precursor that inhibits prohormone processing. *J. Neurosci.* **20**, 639–648 (2000).
119. Bures, E.J. et al. Identification of incompletely processed potential carboxypeptidase E substrates from CpE^{fat}/CpE^{fat} mice. *Proteomics* **1**, 79–92 (2001).
120. Flint, A.J., Tiganis, T., Barford, D. & Tonks, N.K. Development of 'substrate-trapping' mutants to identify physiological substrates of protein tyrosine phosphatases. *Proc. Natl. Acad. Sci. USA* **94**, 1680–1685 (1997).
121. Zhang, S.H., Liu, J., Kobayashi, R. & Tonks, N.K. Identification of the cell cycle regulator VCP (p97/CDC48) as a substrate of the band 4.1-related protein-tyrosine phosphatase PTPH1. *J. Biol. Chem.* **274**, 17806–17812 (1999).
122. Belew, M. & Porath, J. Immobilized metal ion affinity chromatography. Effect of solute structure, ligand density and salt concentration on the retention of peptides. *J. Chromatogr.* **516**, 333–354 (1990).
123. Posewitz, M.C. & Tempst, P. Immobilized gallium(III) affinity chromatography of phosphopeptides. *Anal. Chem.* **71**, 2883–2892 (1999).
124. Oda, Y., Nagasu, T. & Chait, B.T. Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nat. Biotechnol.* **19**, 379–382 (2001).
125. Goshe, M.B. et al. Phosphoprotein isotope-coded affinity tag approach for isolating and quantitating phosphopeptides in proteome-wide analyses. *Anal. Chem.* **73**, 2578–2586 (2001).
126. Zhou, H., Watts, J.D. & Aebersold, R. A systematic approach to the analysis of protein phosphorylation. *Nat. Biotechnol.* **19**, 375–378 (2001).
127. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405**, 837–846 (2000).
128. Deshaies, R.J. et al. Charting the protein complexome in yeast by mass spectrometry. *Mol. Cell. Proteomics* **1**, 3–10 (2002).
129. Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
130. Gavin, A.C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
131. Ho, Y. et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
132. von Mering, C. et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
133. Ranish, J.A. et al. The study of macromolecular complexes by quantitative proteomics. *Nat. Genet.* (in press).
134. Pease, A.C. et al. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. USA* **91**, 5022–5026 (1994).
135. Hughes, T.R. et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* **19**, 342–347 (2001).
136. van Berkum, N.L. & Holstege, F.C. DNA microarrays: raising the profile. *Curr. Opin. Biotechnol.* **12**, 48–52 (2001).
137. Nadon, R. & Shoemaker, J. Statistical issues with microarrays: processing and analysis. *Trends Genet.* **18**, 265–271 (2002).
138. 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 (2001).
139. Jenkins, R.E. & Pennington, S.R. Arrays for protein expression profiling: towards a viable alternative to two-dimensional gel electrophoresis? *Proteomics* **1**, 13–29 (2001).
140. Zhou, H., Roy, S., Schulman, H. & Natan, M.J. Solution and chip arrays in protein profiling. *Trends Biotechnol.* **19**, 534–539 (2001).
141. McCafferty, J., Griffiths, A.D., Winter, G. & Chiswell, D.J. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* **348**, 552–554 (1990).
142. Tramontano, A. et al. The making of the minibody: an engineered β -protein for the display of conformationally constrained peptides. *J. Mol. Recognit.* **7**, 9–24 (1994).
143. Martin, F. et al. Coupling protein design and *in vitro* selection strategies: improving specificity and affinity of a designed β -protein IL-6 antagonist. *J. Mol. Biol.* **255**, 86–97 (1996).
144. Koivunen, E., Wang, B. & Ruoslahti, E. Phage libraries displaying cyclic peptides with different ring sizes: ligand specificities of the RGD-directed integrins. *Biotechnology* **13**, 265–270 (1995).
145. McConnell, S.J. & Hoess, R.H. Tandemistat as a scaffold for conformationally constrained phage peptide libraries. *J. Mol. Biol.* **250**, 460–470 (1995).
146. Nord, K., Nilsson, J., Nilsson, B., Uhlen, M. & Nygren, P.A. A combinatorial library of an α -helical bacterial receptor domain. *Protein Eng.* **8**, 601–608 (1995).
147. Choo, Y. & Klug, A. Designing DNA-binding proteins on the surface of filamentous phage. *Curr. Opin. Biotechnol.* **6**, 431–436 (1995).
148. Brody, E.N. et al. The use of aptamers in large arrays for molecular diagnostics. *Mol. Diagn.* **4**, 381–388 (1999).
149. Holt, L.J., Enever, C., de Wildt, R.M. & Tomlinson, I.M. The use of recombinant antibodies in proteomics. *Curr. Opin. Biotechnol.* **11**, 445–449 (2000).
150. Eklund, M., Axelsson, L., Uhlen, M. & Nygren, P.A. Anti-idiotypic protein domains selected from protein A-based antibody libraries. *Proteins* **48**, 454–462 (2002).
151. Mitchell, P. A perspective on protein microarrays. *Nat. Biotechnol.* **20**, 225–229 (2002).
152. Cahill, D.J. Protein and antibody arrays and their medical applications. *J. Immunol. Methods* **250**, 81–91 (2001).
153. Walter, G., Bussow, K., Cahill, D., Lueking, A. & Lehrach, H. Protein arrays for gene expression and molecular interaction screening. *Curr. Opin. Microbiol.* **3**, 298–302 (2000).
154. Schweitzer, B. et al. Multiplexed protein profiling on microarrays by rolling-circle amplification. *Nat. Biotechnol.* **20**, 359–365 (2002).
155. Houseman, B.T. & Mrksich, M. Towards quantitative assays with peptide chips: a surface engineering approach. *Trends Biotechnol.* **20**, 279–281 (2002).
156. Martzen, M.R. et al. A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286**, 1153–1155 (1999).
157. Houseman, B.T., Huh, J.H., Kron, S.J. & Mrksich, M. Peptide chips for the quantitative evaluation of protein kinase activity. *Nat. Biotechnol.* **20**, 270–274 (2002).
158. Zhu, H. et al. Analysis of yeast protein kinases using protein chips. *Nat. Genet.* **26**, 283–289 (2000).
159. Ren, B. et al. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
160. Krylov, A.S., Zasedateleva, O.A., Prokopenko, D.V., Rouviere-Yaniv, J. & Mirzabekov, A.D. Massive parallel analysis of the binding specificity of histone-like protein HU to single- and double-stranded DNA with generic oligodeoxyribonucleotide microchips. *Nucleic Acids Res.* **29**, 2654–2660 (2001).
161. MacBeath, G. & Schreiber, S.L. Printing proteins as microarrays for high-throughput function determination. *Science* **289**, 1760–1763 (2000).
162. Diehn, M., Eisen, M.B., Botstein, D. & Brown, P.O. Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. *Nat. Genet.* **25**, 58–62 (2000).
163. Zong, Q., Schummer, M., Hood, L. & Morris, D.R. Messenger RNA translation state: the second dimension of high-throughput expression screening. *Proc. Natl. Acad. Sci. USA* **96**, 10632–10636 (1999).

164. Lee, K.B., Park, S.J., Mirkin, C.A., Smith, J.C. & Mrksich, M. Protein nanoarrays generated by dip-pen nanolithography. *Science* **295**, 1702–1705 (2002).
165. Winzler, E.A. et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
166. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246 (1989).
167. Tucker, C.L., Gera, J.F. & Uetz, P. Towards an understanding of complex protein networks. *Trends Cell Biol.* **11**, 102–106 (2001).
168. Bartel, P.L., Roecklein, J.A., SenGupta, D. & Fields, S. A protein linkage map of *Escherichia coli* bacteriophage T7. *Nat. Genet.* **12**, 72–77 (1996).
169. McCraith, S., Holtzman, T., Moss, B. & Fields, S. Genome-wide analysis of vaccinia virus protein-protein interactions. *Proc. Natl. Acad. Sci. USA* **97**, 4879–4884 (2000).
170. Uetz, P. et al. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
171. 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).
172. Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
173. Rain, J.C. et al. The protein-protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211–215 (2001).
174. Walhout, A.J. et al. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
175. Davy, A. et al. A protein-protein interaction map of the *Caenorhabditis elegans* 26S proteasome. *EMBO Rep.* **2**, 821–828 (2001).
176. Zhang, B., Kraemer, B., SenGupta, D., Fields, S. & Wickens, M. Yeast three-hybrid system to detect and analyze RNA-protein interactions. *Methods Enzymol.* **318**, 399–419 (2000).
177. Tucker, C.L. & Fields, S. A yeast sensor of ligand binding. *Nat. Biotechnol.* **19**, 1042–1046 (2001).
178. Heim, R., Prasher, D.C. & Tsien, R.Y. Wavelength mutations and posttranslational autoxidation of green fluorescent protein. *Proc. Natl. Acad. Sci. USA* **91**, 12501–12504 (1994).
179. Tsien, R.Y. The green fluorescent protein. *Annu. Rev. Biochem.* **67**, 509–544 (1998).
180. Tsien, R.Y. & Miyawaki, A. Seeing the machinery of live cells. *Science* **280**, 1954–1955 (1998).
181. Kumar, A. et al. Subcellular localization of the yeast proteome. *Genes Dev.* **16**, 707–719 (2002).
182. Garrels, J.I. YPD—a database for the proteins of *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **24**, 46–49 (1996).
183. Cherry, J.M. et al. SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* **26**, 73–79 (1998).
184. Shiio, Y. et al. Quantitative proteomic analysis of Myc oncoprotein function. *EMBO J.* **21**, 5088–5096 (2002).
185. Guina, T. et al. Quantitative proteomic analysis of *Pseudomonas aeruginosa* indicates synthesis of quinolone signal in adaptation to cystic fibrosis airways. *Proc. Natl. Acad. Sci. USA* (in the press).
186. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
187. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
188. Keller, A. et al. Experimental protein mixture for validating tandem mass spectral analysis. *Omics* **6**, 207–212 (2002).
189. Caprioli, R.M., Farmer, T.B. & Gile, J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Anal. Chem.* **69**, 4751–4760 (1997).
190. Stoeckli, M., Chaurand, P., Hallahan, D.E. & Caprioli, R.M. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nat. Med.* **7**, 493–496 (2001).
191. Collings, B.A., Sudakov, M. & Londry, F.A. Resonance shifts in the excitation of the $n = 0$, $K = 1$ to 6 quadrupolar resonances for ions confined in a linear ion trap. *J. Am. Soc. Mass Spectrom.* **13**, 577–586 (2002).
192. Medzihradsky, K.F. et al. The characteristics of peptide collision-induced dissociation using a high-performance MALDI-TOF/TOF tandem mass spectrometer. *Anal. Chem.* **72**, 552–558 (2000).
193. Marshall, A.G., Hendrickson, C.L. & Jackson, G.S. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom. Rev.* **17**, 1–35 (1998).
194. Ullrich, B., Ushkaryov, Y.A. & Sudhof, T.C. Cartography of neuexins: more than 1000 isoforms generated by alternative splicing and expressed in distinct subsets of neurons. *Neuron* **14**, 497–507 (1995).
195. Petricoin, E.F. et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572–577 (2002).
196. Adam, B.L. et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* **62**, 3609–3614 (2002).
197. Alon, U., Surette, M.G., Barkai, N. & Leibler, S. Robustness in bacterial chemotaxis. *Nature* **397**, 168–171 (1999).
198. Guet, C.C., Elowitz, M.B., Hsing, W. & Leibler, S. Combinatorial synthesis of genetic networks. *Science* **296**, 1466–1470 (2002).