

Mass spectrometry–based functional proteomics: from molecular machines to protein networks

Thomas Köcher & Giulio Superti-Furga

The study of protein–protein interactions by mass spectrometry is an increasingly important part of post-genomics strategies to understand protein function. A variety of mass spectrometry–based approaches allow characterization of cellular protein assemblies under near-physiological conditions and subsequent assignment of individual proteins to specific molecular machines, pathways and networks, according to an increasing level of organizational complexity. An appropriate analytical strategy can be individually tailored—from an in-depth analysis of single complexes to a large-scale characterization of entire molecular pathways or even an analysis of the molecular organization of entire expressed proteomes. Here we review different options regarding protein-complex purification strategies, mass spectrometry analysis and bioinformatic methods according to the specific question that is being addressed.

The ability to characterize cellular, subcellular or organismal proteins in an unbiased fashion using mass spectrometry–based methods^{1,2} has led to important insights into cell biological processes as well as signal transduction pathways^{3–5}. We envision that together with sophisticated analytical methods targeting the genome and metabolome, ongoing advances in proteomic methodologies will eventually lead to important improvements in our understanding of pathological processes and ultimately in clinical practice^{6,7}.

Recent years have seen widely acclaimed breakthroughs in the large-scale characterization of protein complexes of entire organisms^{8,9} and pathways^{3,4}. In parallel, an ever-increasing number of highly informative studies has increased our knowledge of selected molecular machines^{10–12}. Finally, we have seen stunning cases of three-dimensional structures of molecular assemblies^{13–16}, converting the ‘seeing-is-believing’ fraction of researchers. As a result, characterization of the molecular partners of a protein has become a critical part of analyzing its biological function, next to knocking down its expression by RNA interference or studying its subcellular localization.

There are two distinct fundamental approaches in proteomics. The original concept of proteomics—expression-based proteomics—is traditionally linked to two-dimensional electrophoresis¹⁷, and can be defined as the attempt to catalog the expression of all proteins present in cells, tissues, organisms, or the differential analysis of biological systems reacting upon external stimuli or of specific disease conditions. Although this strategy has been successful in many cases^{18,19}, it suffers from the huge dynamic range of protein expression in biological systems, with regulatory proteins frequently hidden by abundant proteins. It is also limited by the difficulties associated with precise mass spectrometry–based quantitation^{20,21}, and the variability of gene and protein expression resulting from genetic and environmental variability. Additionally, most cellular processes are partly controlled by post-translational modifications, which are difficult to analyze with presently available tools in a comprehensive and quantitative manner^{22,23}. Monitoring protein expression can not be replaced by mRNA microarray technology because changes in protein abundance and mRNA abundance only moderately correlate with each

Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 19, 1090 Vienna, Austria. Correspondence should be addressed to G.S.-F. (gsuperti@cemm.oeaw.ac.at) or T.K. (tkoecher@cemm.oeaw.ac.at).

PUBLISHED ONLINE 27 SEPTEMBER 2007; DOI:10.1038/NMETH1093

other^{24,25}. Ultimately, expression-based proteomics offers only correlative relationships, requiring extensive validation.

The other approach—functional proteomics—is fundamentally and strategically different. Most cellular functions are executed by protein complexes, acting like molecular machines²⁶. The term ‘functional proteomics’ derives from the hypothesis that the association of proteins would suggest their common involvement in a biological function, analogous to the ‘guilt by association’ concept in criminal investigation. There are many different functional proteomics technologies (Fig. 1), such as those (i) based on affinity purification procedures and physical measurement of the associated protein partners from physiological fluids^{27,28}, (ii) relying on pairwise testing of two partners, based on biochemical automation or chip technologies²⁹, (iii) based on genetic readout systems such as the various two-hybrid systems and also phage-display technologies^{30–32}, and (iv) based on computational prediction methods, some of which are based on known three-dimensional structures and binding motifs^{33,34}.

The data sets from these approaches can be integrated and compared to obtain additional insights about the function and the evolution of biological systems^{35,36}. Although ‘maps’ and even compilation

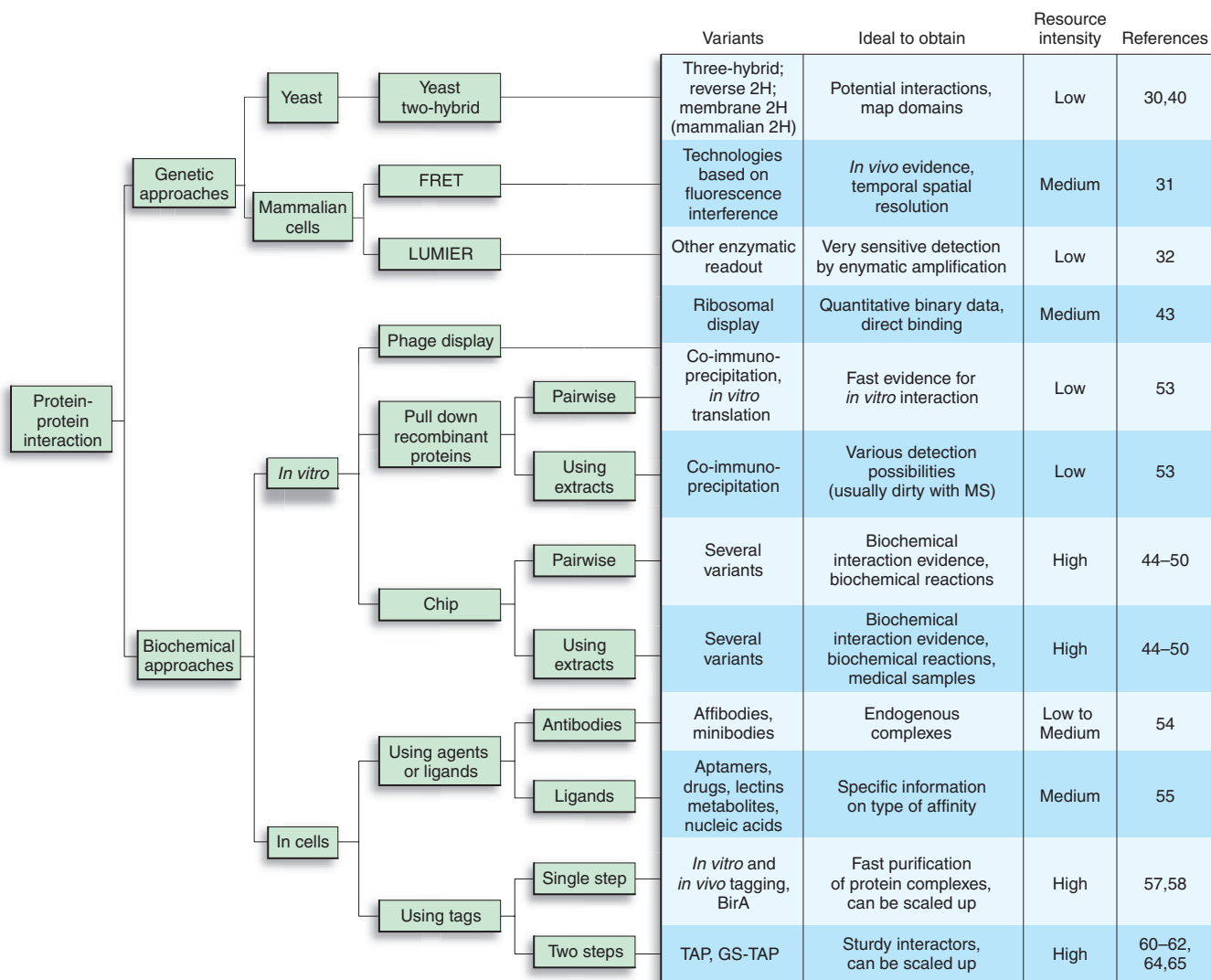
of maps into ‘atlases’ of protein-protein interaction networks are currently incomplete, future maps might describe the cellular activities in a comprehensive and potentially even quantitative fashion. Driven by the vision of quantitative biology, the quantitation of proteins is an emerging trend in proteomics. Various methods are now used to distinguish complex components from contaminating proteins^{4,37,38} and to determine the stoichiometry of the proteins present in a complex³⁹. Functional proteomics and expression-based proteomics provide complementary views onto the ensemble of proteins and their associations within a biological system.

Here we review mass spectrometry-based functional proteomics approaches, including protein-complex purification strategies, mass spectrometry analysis as well as data analysis and interpretation.

Functional proteomics

Goals of protein-protein interaction studies. Modern genetics and functional genomics experiments often lead to the identification of gene products with a putative biological function but a poorly characterized biochemical mode of action. Functional proteomics experiments allow researchers to identify the interacting proteins,

© 2007 Nature Publishing Group <http://www.nature.com/naturemethods>



Katie Ris-Vicari

Figure 1 | Decision tree of options for the most common different protein-protein (or protein-ligand) interaction experimental strategies. Several permutations and variants of the individual approaches are possible. 2H, two-hybrid.

facilitating mapping of a protein to a particular biological pathway. If only individual connections to a potential pathway are sought, then the highly effective and less resource-intensive binary yeast two-hybrid approach⁴⁰ is recommended (Fig. 1).

Biochemical purification of protein complexes followed by characterization of their components by mass spectrometry⁴¹ requires a greater commitment of research capabilities, but the potential informational gain is an order of magnitude larger than the information about individual binary interactions. The researcher can discover the entire cellular machinery in which the protein of interest participates, which can seldom be recapitulated as the sum of binary interactions. Clues can be obtained about the links of the complex to various cellular signaling pathways as well as about cell biological processes governing its 'birth', 'death', stability and subcellular localization. In a characterization of protein complexes a protein of interest initially serves as the bait in cycles of purification under various physiological conditions (for example, in the presence of a stimulus) and subsequent purifications use identified preys as baits.

To characterize the proteins involved in entire biological processes and signaling pathways⁴², the approach is essentially identical to the study of a single protein complex but the scale is larger, starting with as few as five⁵ or as many as 32 (ref. 3) entry points. Although such efforts require multiplication of resources, the synergistic effect of efficiency gains and economy of labor lead to a nonlinear relationship between required effort and output.

The ultimate goal of functional proteomics is to decipher the molecular function of an entire cell by generating a construction master plan describing all molecular machines, their functions, their reactions to external stimuli and their interconnectivities. An interdisciplinary and community-wide effort will be required to realize this vision, even when limited to the characterization of a few cellular states.

Biochemical approaches to map protein-protein interactions.

Presently available protein-complex characterization methods can be grouped into methods to isolate endogenous protein complexes from cells and *in vitro* methods using recombinant proteins (Fig. 1). The simplest *in vitro* method uses immobilized recombinant proteins to capture putative interaction partners by binding. After washing, the bound proteins are eluted and typically identified by mass spectrometry.

Phage display technology⁴³ screens for interacting proteins by expressing recombinant proteins on the surface of phage particles. Interacting proteins are identified as phage to bind to selected immobilized molecules. By fusing foreign cDNA libraries into the phage genome, libraries containing billions of proteins and peptides can be screened in a single experiment. After washing, interacting phages are amplified and the DNA sequence of the putative interacting molecule is sequenced.

Protein arrays^{29,44} are constructed by spotting proteins in defined locations on a chip surface. These proteins can be recombinant proteins⁴⁵, samples from patients⁴⁶ or antibodies⁴⁷. Protein arrays have been used to detect protein-protein, protein-nucleic acid interactions and biochemical functions such as kinase activity⁴⁸. Detection can be based on fluorescent⁴⁹ or chemoluminescent probes, radioisotope labeling⁴⁸, or mass spectrometry⁵⁰.

Approaches for purifying endogenous protein complexes include antibody-based (immunochemical) methods, biochemical purification methods and affinity chromatography (Fig. 1). Subsequent characterization requires unambiguous identification of large numbers of

proteins rapidly and with high sensitivity⁴¹. This can only be achieved by using state-of-the-art mass spectrometry, which can analyze and identify thousands of proteins in a few hours.

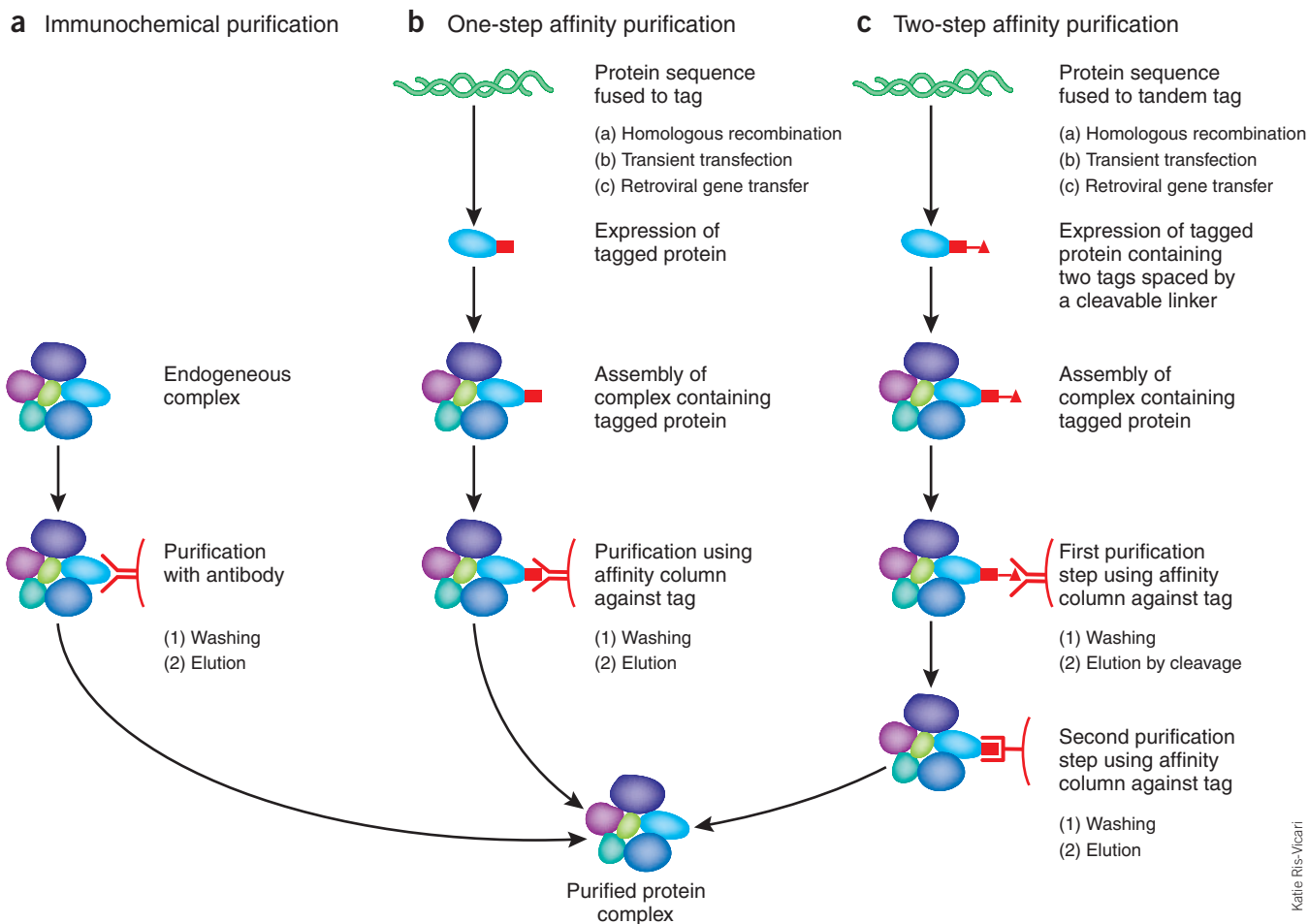
In contrast to universally applicable approaches such as immunochemical methods, biochemical purifications are now limited to rather abundant protein complexes such as the ribosome, the proteasome or the spliceosome^{51,52}. A specific purification strategy must be developed for each type of complex.

In immunochemical methods, protein complexes are precipitated from a cell lysate by using an immobilized antibody to a known component of a complex⁵³ (Fig. 2a). The protein complex is then purified by washing away nonspecific interactors. A fundamental advantage of this approach is that protein complexes can be isolated with a specific and efficient antibody from all types of biological sources, including tissue samples from patients, circumventing the need to express the target protein. Whenever feasible, this route to characterize endogenous complexes should be chosen. Initiatives are underway to generate antibodies to the entire human proteome, which should enable large-scale studies⁵⁴. Modifications of this approach use affibodies, minibodies or aptamers^{55,56}.

Affinity purification-based techniques exploit the biochemical properties of a tag attached to the bait protein to purify the other components of a protein complex (Fig. 2b,c). Using standard cloning techniques, target-protein and peptide-tag coding sequences are fused, and the resulting construct is expressed in target cells. Available tagging systems include His tags, glutathione S-transferase (GST) tags, Flag tags, the calmodulin-binding peptide, the streptavidin binding peptide or the *in vivo* biotinylation of the target tagged peptide using coexpression of the BirA ligase^{57,58}. Combinations of these tags have been also used in various configurations (Fig. 2c). Specific columns with a high specificity for a certain tag are used to enrich the protein complex. One of the first protein complexes of considerable size analyzed by tagging technology was the spliceosomal U1 small nuclear ribonucleoprotein complex⁵⁹. It was isolated by fusing a His tag to a known component of the protein complex and purifying the complex by nickel-nitrilotriacetic acid affinity chromatography.

One of the most successful tags developed to date is the tandem affinity purification (TAP) tag^{60,61}, which uses two sequential enrichment steps. Developed for yeast, the original TAP tag is composed of a protein A tag, followed by a tobacco etch virus (TEV) protease cleavage site and a calmodulin binding peptide^{60,61}. In the first purification step, the protein complex is purified from the cell lysate on an immunoglobulin gamma (IgG) affinity resin. The target protein complex is cleaved from the protein A tag with TEV protease. The eluate is then enriched in a second affinity purification step on an immobilized calmodulin column; elution yields a highly purified protein complex. Notably, all binding and elution steps are performed in mild buffer conditions, keeping the native complex as intact as possible. The TAP-tag can be fused to the N or C termini of the target protein. Fusion at one terminus might disturb the interaction of the protein with its partners; therefore, it might be necessary to test both variants. Several variants of this two-step approach, using different combinations of tags, have been described^{3,60,62-65}. The combination of protein G, a TEV protease cleavage site and streptavidin-binding peptide facilitated a tenfold improvement of recovery for complexes from mammalian cells⁶².

In yeast, bait proteins are usually expressed by replacing the endogenous gene by homologous recombination with its tagged version⁶¹. In higher eukaryotic systems such as human cell lines, the expression



Kaitie Ris-Vicari

Figure 2 | Main routes of protein-complex purification. **(a)** In immunochemical purification, the endogenous protein complex is precipitated using an antibody to the target protein, allowing protein-complex characterization without expression of a tagged protein. **(b)** In one-step affinity purification, the purified protein complex is obtained by expression of the tagged construct in the cell, followed by specific binding and elution from an affinity column. **(c)** In two-step affinity purification, two rounds of specific binding and specific elution assure a highly purified protein complex with little contaminating proteins at the cost of losing transient interactions.

of the tagged protein in the natural chromosomal context cannot be easily achieved. Consequently, other methods of expressing the fusion protein have to be used such as the stable integration of the construct by retrovirus-mediated gene transfer³ or by transient transfection⁶⁶. In our experience, the problems with transient transfection concern not only expression over endogenous levels of the protein but also the cellular shock associated with the large amount of newly translated proteins, often resulting in their association with chaperones. Therefore, whenever feasible, infection with viral vectors is preferred. Additionally, as the bait protein is overexpressed, it competes with the endogenous protein for complex formation. The recovery rate of protein complexes can be increased by abolishing competition from the endogenous protein by RNA interference-mediated knockdown⁶³.

How should the researcher judge which tag to choose? Is a one-step or a two-step procedure recommended? Specific protocols might work best for a specific protein complex, cell line or organism. Successful purification of protein complexes from eukaryotic cells has been reported using both one-step and two-step purifications²⁷. As the reader would guess, one-step purifications on average lead to a higher number of contaminating proteins and the two-step procedures tend to yield cleaner results but weak interactions can be lost⁶². Typically, the recovery yields for one-step procedures are 3–5 times

higher than for two-step purifications. Many researchers prefer to start their investigation with a shorter list of interacting proteins obtained with a two-step procedure because the validation of data is the main experimental bottleneck. A recently developed flexible tag allows one- or two-step purifications⁶².

As tagging can interfere with protein function, it is recommended to try both N-terminal and C-terminal fusion if no additional information, such as the three-dimensional structure or subcellular localization data is available. On average, however, only in 10–15% of cases does the tag interfere with the function of the protein^{28,67}.

One new area of development is the characterization of complete protein complexes in their native form with mass spectrometry⁶⁸. Similarly to other top-down mass spectrometry techniques⁶⁹, these experiments are now limited to abundant complexes such as ribosomes⁷⁰, proteasomes⁷¹ or exosomes⁷² because large amounts of the purified protein complex are required.

Cross-linking techniques can also be applied to the study of protein complexes⁷³ and can be combined with biochemical approaches to purify protein complexes⁷⁴. In general, these approaches aim for two major improvements over conventional techniques. They either attempt to maintain the interaction of loosely bound interacting proteins by covalently linking them, or they attempt to map the topology

of the purified complex. In the latter approach, cross-linked peptides originating from different proteins prove that there is a direct interaction between two proteins and give clues about their binding interfaces. Although several successful experiments have been reported^{73,74}, the need for large amounts of protein and additional sample preparation steps have limited widespread application of these techniques.

Mass-spectrometric protein identification

Sample preparation options. In the majority of proteomics experiments the purified proteins are separated by one-dimensional SDS-PAGE and stained with a mass-spectrometry compatible dye such as silver, fluorescence-based dyes such as SYPRO ruby or Coomassie (Fig. 3). SDS-PAGE separation removes unwanted contaminants such as buffer components from the protein sample, and the sample complexity is decreased by separating the proteins according to molecular weight. Additionally, the staining pattern of the gel can be used as semiquantitative assessment of the experiment, for example, to evaluate the quantity of the bait protein versus the interacting components or compare an experiment with a control sample.

Individual protein bands of interest are excised or the entire lane is cut into slices, followed by in-gel digestion with a specific protease such as trypsin to produce peptides for mass spectrometry analysis⁷⁵. Unlike intact proteins, peptides have lower detection limits, they can be extracted from gels, and their cleavage pattern by specific proteases can provide additional information. Note that the extraction efficiency of peptides from a gel is only about 20% and is dependent on the primary structure of the peptide⁷⁶.

As an alternative to gel-based protocols, protein mixtures can be digested in solution without prior separation of individual components and analyzed by mass spectrometry⁷⁷ (Fig. 3). In many cases buffer components such as detergents prohibit direct analysis as they can interfere with the mass spectrometry ionization process. In such cases protein samples can be precipitated with trichloroacetic acid, washed and dissolved in a digestion buffer containing the appropriate protease. Although successful, the procedure requires a high-pressure liquid chromatography (HPLC) system that can readily resolve peptides in a complex mixture. An option to improve the peak capacity is two-dimensional liquid chromatography separation^{77,78}. The main advantages of in-solution digestion protocols are the reduction of the

time required for analysis⁷⁷ and a higher recovery of peptides compared to in-gel digestion.

Peptide separation options. The peptide mixture can be directly analyzed by mass spectrometry^{79,80}, or separated by HPLC before mass-spectrometric analysis (Fig. 3). Direct analysis of a peptide sample is rapid compared to liquid chromatography–mass spectrometry (LC-MS) because the loading of a sample onto a liquid chromatography system and subsequent separation is a time-consuming procedure (45–90 min per experiment). In contrast, the use of HPLC systems coupled to a mass spectrometer not only results in a dramatically higher number of detected peptides but also facilitates automation. Typically, an even greater number of proteins can be identified using two-dimensional peptide separations such as combining strong ion exchange with conventional reversed phase separation⁷⁸.

Good chromatographic resolution is important because if peptides elute in a relatively short period, the high concentration will yield high ion counts in the mass spectrometer. More peptides can be sequenced by the mass spectrometer if they are well-separated over time. In practice, most laboratories aim to achieve chromatographic resolution in the range of 10–15 s full width half maximum (FWHM). Another important aspect of on-line coupled LC-MS instrumentation is the flow rate, which inversely affects the ionization efficiency⁸⁰.

In general, low-complexity samples can be analyzed simply and rapidly without peptide separation. Complex samples such as the peptide mixtures generated from a complete pulldown require chromatographic separation before mass-spectrometric analysis.

Mass-spectrometric options. Two soft ionization techniques have been developed for the ionization of macromolecules (including proteins and peptides) into the gas phase, electrospray ionization (ESI)⁸¹ and matrix-assisted laser desorption/ionization (MALDI)⁸² (Fig. 3). In the most common instrumental designs, ESI is performed with mass spectrometers capable of tandem mass spectrometry (MS/MS) experiments. Ion traps, quadrupole time-of-flight instruments (Q-TOF), Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometers (FTMS) and the Orbitrap are the most common types of instrumentation now used in high-end protein analysis. A critical factor for the unambiguous identification of proteins is high mass

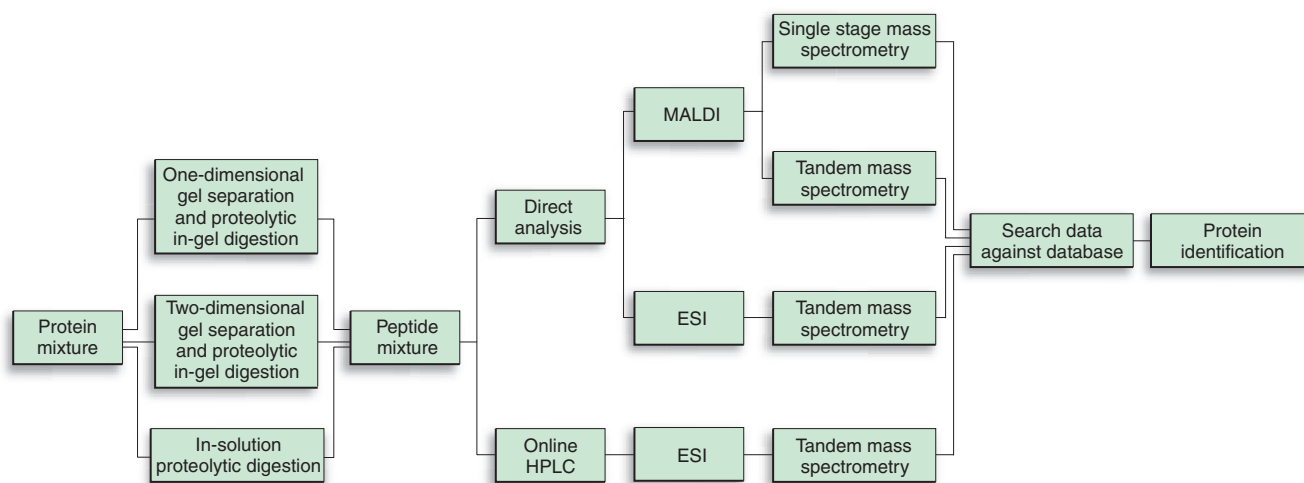


Figure 3 | Flowchart of different options for sample preparation, protein separation and mass-spectrometric analysis. The main routes to protein identification are shown.

accuracy of the mass spectrometer, best realized with FTMS⁸³. But high sensitivity and high sequencing speed might compensate for lower mass accuracy. Therefore, although all of the above mentioned mass spectrometers can be used for protein identification, a specific analytical question might be answered optimally using specific mass-spectrometric instrumentation. Q-TOF instruments now have mass accuracies for typical tryptic peptides in the range of 10–20 p.p.m., the Orbitrap can achieve mass accuracies of approximately 2 p.p.m. and FT-ICR instrumentation can be optimized for sub-p.p.m. mass accuracy.

But even mass windows in the range of 1 Da (typical for ion traps or triple quadrupole instruments) allow the unambiguous identification of proteins. In such instances, improved fragment ion data are necessary to ensure correct protein identifications. In most cases, ESI experiments are performed by directly coupling a HPLC system to the mass spectrometer. The experiment is performed in a data-dependent acquisition mode whereby the MS/MS experiments are automated based on the eluting peptides¹. The *m/z* values of the peptides eluting at a given time from the column are recorded, and the peptide with the most intense signal is automatically selected, fragmented and the fragment ion spectrum is recorded. This procedure can be repeated with the other eluting peptides, but is limited by the duty cycle of the mass spectrometer, detection limits and the *m/z* value of the peptides.

An alternative method to ESI-MS is MALDI-MS, usually performed with time-of-flight (TOF) instruments recording only the mass spectrum of the peptide ions; no fragmentation information is collected. This analytical strategy is rapid but protein identification based on MS¹ data alone uses less statistical information as compared to MS/MS data. This can be problematic when analyzing protein mixtures or samples containing modified proteins, common in samples from higher eukaryotes. Instruments such as MALDI-Q-TOFs or MALDI-TOF/TOFs can operate in an MS/MS mode. It is common practice to separate complex peptide samples off-line with an HPLC system. A limitation of MALDI-MS/MS, however, is that as mostly singly charged ions are generated, and thus weaker fragment ion spectra are produced compared to those generated by multiply charged electrospray ions.

If analysis time is the critical factor in the analytical strategy, a simple MALDI-TOF analysis without HPLC separation (with a time frame of seconds) is recommended. If high sample complexity is expected and identification of numerous proteins covering a substantial dynamic range is required, then LC-MS/MS is the method now chosen by most proteomics researchers. Given the importance of unambiguous identification of proteins, most proteomic studies published today are based on LC-MS/MS data acquired from peptides ionized via ESI.

Approaches to assess and increase confidence of data sets. In contrast to the plethora of routes to analyze peptide mixtures, approaches to analyze the resulting data sets are fairly similar. In most cases, the raw data files are first processed by the software controlling the respective mass spectrometry instrument. Typical processing steps include smoothing, centroiding and charge-state deconvolution of the acquired spectra. The generated data sets are then searched against a protein database. The two most commonly used algorithms use either a probabilistic approximation such as the search engine MASCOT⁸⁴ or a mathematical correlation method such as SEQUEST⁸⁵. Although these and other search engines differ in their mathematical approach and exact statistical methods, the most crucial factor affecting

false positive and false negative rates is the applied mass accuracy. Consequently, the greatest care should be taken in defining the thresholds of the minimum scores and the allowed mass tolerances for the precursor ion and the fragment ions. A valid approach for validation of the chosen parameters is to search the obtained data sets against a decoy protein database⁸⁶.

After protein identification by one of the available search engines, the data might be further filtered by setting specific thresholds such as a minimum peptide length or a specific number of peptides to consider a protein identification to be correct. There is currently a controversy about whether protein identifications should be based on a minimum of two tryptic peptides or if protein identifications based on a single tandem mass spectrum with very high statistical significance or MS/MS/MS data of one peptide are sufficient⁸⁷.

Often a set of identified peptides will match to more than one protein, which are often homologous proteins or different isoforms of the same reading frame⁸⁸. Commercial algorithms have been developed to group these proteins, facilitating assessment and interpretation of data.

Data standardization, interpretation and validation

Data standardization. The high-throughput nature of proteomics has generated a rapidly increasing flow of progressively complex protein identification data. Given the different types of mass spectrometric instrumentation, ionization processes and software platforms, the assessment of published data becomes increasingly difficult. To facilitate sharing experimental data, common standards in data acquisition, data interpretation and data storage are required. The proteomics community has begun a process toward defining the minimal standards for generating and publishing mass-spectrometric data and proteomics experiments. Although this process is far from complete and many different groups of researchers have defined their own standardization protocols⁸⁹, it is foreseeable that in the future a common format of publishing and storing of proteomic data will exist⁹⁰. It is evident that fundamental experimental information such as the mass-spectrometric instrumentation, resolution, mass accuracy, software for data interpretation, HPLC flow rates and composition or type of MALDI sample plates must be reported and stored in a specific format.

Related standardization initiatives tackle the protein identification process. Descriptions of the software used, input parameters such as the database queried, the restrictions applied to the search, the cleavage agents and the mass tolerances used should be reported in proteomics publications. Specific requirements for publishing the output of the identification process, such as the accession codes of the identified proteins, the protein scores or the obtained sequence coverage have also been defined. In addition, any publication reporting proteomics data should contain a statistical analysis of the data such as the determination of the false positive rate.

Similar trends have emerged for the standardization of proteomic protein-protein interaction data⁹¹. In recent years, several public databases have been created for storing functional proteomics data sets. Now the data are manually curated, and are often extracted from publications. The main aim of standardization in reporting protein-protein interaction data is to define common standards, similar to those of nucleotide databases. The use of ambiguous protein identifiers and unclear descriptions of experimental conditions seriously hinder the development of interaction databases and the exchange of information between them. Key data such as exact accession

numbers and classification of the molecular role of any published proteins should be included in each publication.

Data validation and interpretation. Because mass spectrometry is such a sensitive technique, an undesired side effect is that contaminating proteins such as keratins and highly abundant proteins are also identified in purification experiments. In the case of protein-protein interaction experiments at a small scale, these contaminants do not impose substantial problems upon data validation, but can be removed from the data set based on biological knowledge and experimental data sets. Putative interacting proteins can be evaluated by various orthogonal methods such as colocalization studies, gain-of-function or loss-of-function experiments²⁸.

Quantitative methods in mass spectrometry can also facilitate this process (for reviews see refs. 20,21,38.). By applying relative quantification methods such as the isotope-coded affinity tag (ICAT), the genuine components of a protein complex can be distinguished from contaminant proteins by comparing the relative abundances of the differently isotope-labeled peptides derived from a control sample and the purified specific complex³⁷. Absolute quantitation of proteins can be achieved by spiking a single sample with isotopically labeled standard peptides⁹². This technique can be used to define the stoichiometry of a protein complex.

One critical general limitation encountered in the interpretation of the data obtained from the purification of a protein complex is a lack of information about binary interactions. At least in theory it is possible that several proteins copurify, each binding like beads on a string to maximally two components. From biophysical, structural or genetic methods in combination with biochemistry, we know that complexes assemble according to precise assembly steps. The order of complex assembly, post-translational modifications, allosteric effects and cooperative binding are fundamental parts of the biological integration of the machine in the cellular orchestration. Bioinformatics tools can use existing protein-protein interaction data sets from the literature, binding prediction and structural considerations to compose a possible interaction diagram for each complex^{33,93}. Thus binary interaction data, as provided by yeast two-hybrid screens or pairwise biochemical interaction experiments (Fig. 4a), is perfectly complementary to mass spectrometry-based approaches.

In contrast, the nature of systematic large-scale experiments does not allow for the subjective and individual evaluation of their results. In these cases, the removal of potential contaminating proteins can not be based on judging individual purifications. One possible approach to highlight potentially contaminating proteins in high-throughput data is to quantitatively compare them³⁷ against core proteomes⁹⁴, defined as the subset of highly expressed proteins in a cell. Another approach is to subtract the proteins identified in a larger number of pull-down assays than a specific cut-off value⁹⁵ or present in pull-downs of unrelated proteins³. In various yeast studies^{8,95,96} using the TAP tagging

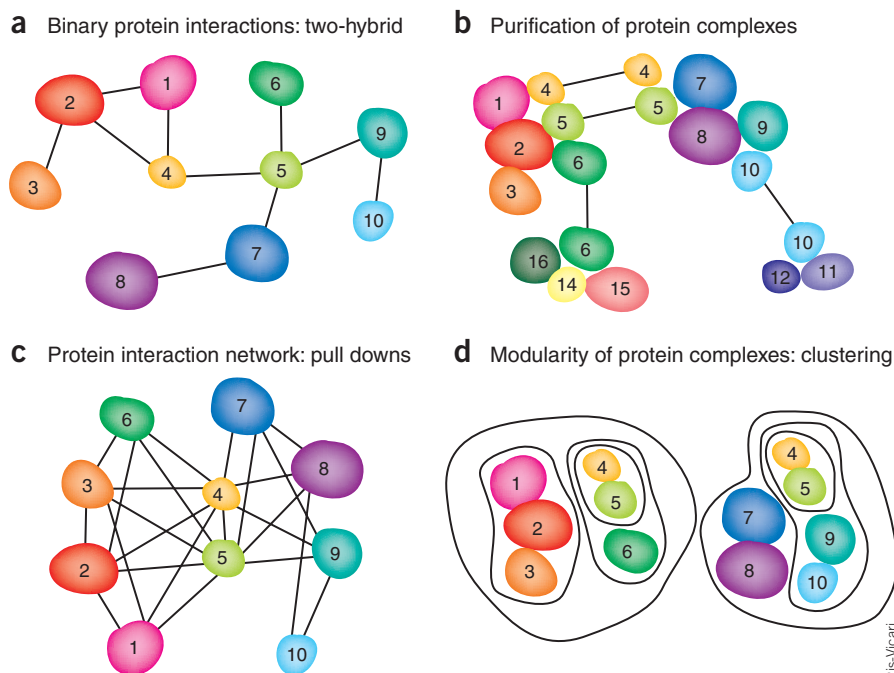


Figure 4 | Illustration of the types of protein networks that can be elucidated with different experimental approaches. (a) Binary protein interactions are typically obtained from two-hybrid assays. (b) The purification of protein complexes leads to a corresponding protein network where two protein complexes are connected by sharing one or more proteins, indicated by lines. (c) Protein interaction networks can be generated from protein pulldown assays. The matrix model represents purified protein assemblies as if interacting all with each other. (d) Statistical analysis and clustering demonstrates the modularity of protein complexes. Core components and alternative attachments or modules of one or more other proteins are depicted.

technique⁶¹, the problems associated with sticky proteins appear to be relatively moderate, also because the redundancy of repeated analysis of a limited number of complexes using different baits allows for robust statistical and probabilistic analyses^{95,96}. In experiments using cell lines derived from higher organisms, contaminating proteins can be a critical issue, particularly if too few repetitions are performed to allow for statistical analysis²⁸.

Medium- and large-scale experiments require a completely automated interpretation routine, and several bioinformatics approaches to this problem have been developed. Complexes can be represented as static entities, leading to the creation of networks in which the edges represent proteins shared between complexes (that is, present in different assemblies, Fig. 4b)⁸. This however only poorly captures the dynamics of complex composition and is very sensitive to any abundant false positive result that survives the statistical filtering for significance. If one assumes that each protein in a purified complex interacts with each other protein in the same complex (Fig. 4c), then one can easily generate large networks in which each protein is represented as an entity connected to various degrees with other proteins, in a spiderweb of binary interactions similar to those generated by two-hybrid screens^{97,98} (Fig. 4a). The comprehensive analyses of protein interactions in the yeast proteome have been performed with a high level of redundancy (via repurification using different baits), facilitating the calculation of how often a protein was retrieved in a reciprocal manner (spike model)⁹⁶. Building upon the matrix model, a conceptual 'affinity-index' can be assigned to each protein pair. The data can be further processed by clustering analysis and

applying a cutoff that corresponds to the ideal representation of very well characterized examples of protein complexes from the literature. Complexes are thus represented as modular entities consisting of core components that are always together and alternative 'attachments' of one or more other proteins⁹⁶ (Fig. 4d). Decomposed this way, the modular organization of the proteome allows for the prediction of evolutionarily more conserved elements (cores) and less conserved connections (attachments) as well as for diversification of function based on a combination of limited sets of components⁹⁶.

Future perspectives

Recent major technical developments in protein-complex purification, mass spectrometry and bioinformatics will facilitate analysis of protein interactions. Several large-scale studies of protein complexes in yeast^{95–98} and animals have been reported, and there are plans to tackle the human proteome.

What are the major challenges and future goals? The major ones include (i) the ever-changing nature of the proteome composition, (ii) the dynamics of protein complex assembly and disassembly, (iii) the absolute and relative quantitation of proteins, (iv) the capturing of endogenous complexes from native cells and tissues, (v) the integration of mass spectrometry-based data sets with the data sets from binary interaction screens, localization studies, genetic or structural information and known interactions with metabolites, (vi) and the generation of a 'molecular anatomy of the cell view' bridging the structure of the molecular machine to the organelle substructural level^{93,99}. Current experimental limitations hinder the mapping of transient or weak interactions between proteins and the comprehensive characterization of their post-translational modifications. We are confident, however, that all these experimental limitations will be overcome as an increasing number of scientists apply proteomic methods to biological as well as to clinically relevant questions in biomedical research. We predict that medicine will also profit enormously from these emerging trends¹⁰⁰, especially from the development of technologies for the identification and characterization of multiparameter diagnostic and prognostic tools as well as from the discovery of new targets for therapeutics.

ACKNOWLEDGMENTS

We thank the members of the Center for Molecular Medicine of the Austrian Academy of Sciences (CeMM) for fruitful discussions, and K. Bennett, J. Colinge and T. Buerckstuegger for critical reading of the manuscript. Work in our laboratory is supported by the Austrian Academy of Sciences, the Austrian Federal Ministry for Science and Research with the DRAGON and APP-II projects of the GEN-AU program, by the Austrian Science Fund FWF and the Austrian National Bank. We apologize to colleagues if due to space limitations we omitted important original research papers.

Published online at <http://www.nature.com/naturemethods/>
Reprints and permissions information is available online at
<http://npg.nature.com/reprintsandpermissions>

1. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711 (2004).
2. Ferguson, P.L. & Smith, R.D. Proteome analysis by mass spectrometry. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 399–424 (2003).
3. Bouwmeester, T. *et al.* A physical and functional map of the human TNF- α NF- κ B signal transduction pathway. *Nat. Cell Biol.* **6**, 97–105 (2004).
4. Blagoev, B. *et al.* A proteomics strategy to elucidate functional protein-protein interactions applied to EGF signaling. *Nat. Biotechnol.* **21**, 315–318 (2003).
5. Major, M.B. *et al.* Wilms tumor suppressor WTX negatively regulates WNT/ β -catenin signaling. *Science* **316**, 1043–1046 (2007).
6. Weston, A.D. & Hood, L. Systems biology, proteomics, and the future of health care: Toward predictive, preventative, and personalized medicine. *J. Proteome Res.* **3**, 179–196 (2004).
7. Anderson, N.L. & Anderson, N.G. The human plasma proteome — history,

8. character, and diagnostic prospects. *Mol. Cell. Proteomics* **1**, 845–867 (2002).
9. Gavin, A.C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
10. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
11. Riedel, C.G. *et al.* Protein phosphatase 2A protects centromeric sister chromatid cohesion during meiosis I. *Nature* **441**, 53–61 (2006).
12. Vanacova, S. *et al.* A new yeast poly(A) polymerase complex involved in RNA quality control. *PLoS Biol.* **3**, 986–997 (2005).
13. Bertwistle, D., Sugimoto, M. & Sherr, C.J. Physical and functional interactions of the Arf tumor suppressor protein with nucleophosmin/B23. *Mol. Cell. Biol.* **24**, 985–996 (2004).
14. Lorentzen, E. *et al.* The archaeal exosome core is a hexameric ring structure with three catalytic subunits. *Nat. Struct. Mol. Biol.* **12**, 575–581 (2005).
15. Hao, B., Oehlmann, S., Sowa, M.E., Harper, J.W. & Pavletich, N.P. Structure of a Fbw7-Skp1-cyclin E complex: multisite-phosphorylated substrate recognition by SCF ubiquitin ligases. *Mol. Cell* **26**, 131–143 (2007).
16. Nickell, S., Kofler, C., Leis, A.P. & Baumeister, W. A visual approach to proteomics. *Nat. Rev. Mol. Cell Biol.* **7**, 225–230 (2006).
17. Groll, M., Bochtler, M., Brandstetter, H., Clausen, T. & Huber, R. Molecular machines for protein degradation. *ChemBioChem* **6**, 222–256 (2005).
18. Gorg, A., Weiss, W. & Dunn, M.J. Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **4**, 3665–3685 (2004).
19. Wulfsberg, J.D. *et al.* Proteomics of human breast ductal carcinoma *in situ*. *Cancer Res.* **62**, 6740–6749 (2002).
20. Le Naour, F. *et al.* Profiling changes in gene expression during differentiation and maturation of monocyte-derived dendritic cells using both oligonucleotide microarrays and proteomics. *J. Biol. Chem.* **276**, 17920–17931 (2001).
21. Ong, S.E. & Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nat. Chem. Biol.* **1**, 252–262 (2005).
22. Zhang, H., Yan, W. & Aebersold, R. Chemical probes and tandem mass spectrometry: a strategy for the quantitative analysis of proteomes and subproteomes. *Curr. Opin. Chem. Biol.* **8**, 66–75 (2004).
23. Jensen, O.N. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol.* **8**, 33–41 (2004).
24. Mann, M. & Jensen, O.N. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.* **21**, 255–261 (2003).
25. 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).
26. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
27. Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291–294 (1998).
28. Terpe, K. Overview of tag protein fusions: from molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* **60**, 523–533 (2003).
29. Bauch, A. & Superti-Furga, G. Charting protein complexes, signaling pathways, and networks in the immune system. *Immunol. Rev.* **210**, 187–207 (2006).
30. Zhu, H. & Snyder, M. Protein chip technology. *Curr. Opin. Chem. Biol.* **7**, 55–63 (2003).
31. Chien, C.T., Bartel, P.L., Sternglanz, R. & Fields, S. The 2-hybrid system — a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl. Acad. Sci. USA* **88**, 9578–9582 (1991).
32. Giepmans, B.N.G., Adams, S.R., Ellisman, M.H. & Tsien, R.Y. The fluorescent toolbox for assessing protein location and function. *Science* **312**, 217–224 (2006).
33. Barrios-Rodiles, M. *et al.* High-throughput mapping of a dynamic signaling network in mammalian cells. *Science* **307**, 1621–1625 (2005).
34. Aloy, P. & Russell, R.B. Structural systems biology: modelling protein interactions. *Nat. Rev. Mol. Cell Biol.* **7**, 188–197 (2006).
35. Marcotte, E.M. *et al.* Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
36. Tornow, S. & Mewes, H.W. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res.* **31**, 6283–6289 (2003).
37. Sharan, R. & Ideker, T. Modeling cellular machinery through biological network comparison. *Nat. Biotechnol.* **24**, 427–433 (2006).
38. Ranish, J.A. *et al.* The study of macromolecular complexes by quantitative proteomics. *Nat. Genet.* **33**, 349–355 (2003).
39. Mann, M. Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell Biol.* **7**, 952–958 (2006).
40. Hochleitner, E.O. *et al.* Protein stoichiometry of a multiprotein complex, the human spliceosomal U1 small nuclear ribonucleoprotein — absolute quantification using isotope-coded tags and mass spectrometry. *J. Biol. Chem.* **280**, 2536–2542 (2005).
41. Fields, S. & Song, O.K. A novel genetic system to detect protein-protein



- interactions. *Nature* **340**, 245–246 (1989).
41. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
 42. Fishman, M.C. & Porter, J.A. Pharmaceuticals—a new grammar for drug discovery. *Nature* **437**, 491–493 (2005).
 43. Rodi, D.J. & Makowski, L. Phage-display technology—finding a needle in a vast molecular haystack. *Curr. Opin. Biotechnol.* **10**, 87–93 (1999).
 44. Cahill, D.J. Protein and antibody arrays and their medical applications. *J. Immunol. Methods* **250**, 81–91 (2001).
 45. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105 (2001).
 46. Pawelz, C.P. *et al.* Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene* **20**, 1981–1989 (2001).
 47. Huang, R.P., Huang, R.C., Fan, Y. & Lin, Y. Simultaneous detection of multiple cytokines from conditioned media and patient's sera by an antibody-based protein array system. *Anal. Biochem.* **294**, 55–62 (2001).
 48. Zhu, H. *et al.* Analysis of yeast protein kinases using protein chips. *Nat. Genet.* **26**, 283–289 (2000).
 49. Hamelinck, D. *et al.* Optimized normalization for antibody microarrays and application to serum-protein profiling. *Mol. Cell. Proteomics* **4**, 773–784 (2005).
 50. Davies, H., Lomas, L. & Austen, B. Profiling of amyloid beta peptide variants using SELDI ProteinChip (R) arrays. *Biotechniques* **27**, 1258–1261 (1999).
 51. Hua, S., To, W.Y., Nguyen, T.T., Wong, M.L. & Wang, C.C. Purification and characterization of proteasomes from *Trypanosoma brucei*. *Mol. Biochem. Parasitol.* **78**, 33–46 (1996).
 52. Huang, L. *et al.* Functional assignment of the 20S proteasome from *Trypanosoma brucei* using mass spectrometry and new bioinformatics approaches. *J. Biol. Chem.* **276**, 28327–28339 (2001).
 53. Harlow, E. & Lane, D. *Antibodies: a Laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold Springs Harbor, New York, 1988).
 54. Uhlen, M. & Ponten, F. Antibody-based proteomics for human tissue profiling. *Mol. Cell. Proteomics* **4**, 384–393 (2005).
 55. Nord, K. *et al.* Binding proteins selected from combinatorial libraries of an alpha-helical bacterial receptor domain. *Nat. Biotechnol.* **15**, 772–777 (1997).
 56. Hermann, T. & Patel, D.J. Biochemistry - Adaptive recognition by nucleic acid aptamers. *Science* **287**, 820–825 (2000).
 57. Waugh, D.S. Making the most of affinity tags. *Trends Biotechnol.* **23**, 316–320 (2005).
 58. de Boer, E. *et al.* Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice. *Proc. Natl. Acad. Sci. USA* **100**, 7480–7485 (2003).
 59. Neubauer, G. *et al.* Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl. Acad. Sci. USA* **94**, 385–390 (1997).
 60. Puig, O. *et al.* The tandem affinity purification (TAP) method: A general procedure of protein complex purification. *Methods* **24**, 218–229 (2001).
 61. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030–1032 (1999).
 62. Burckstummer, T. *et al.* An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. *Nat. Methods* **3**, 1013–1019 (2006).
 63. Forler, D. *et al.* An efficient protein complex purification method for functional proteomics in higher eukaryotes. *Nat. Biotechnol.* **21**, 89–92 (2003).
 64. Drakas, R., Prisco, M. & Baserga, R. A modified tandem affinity purification tag technique for the purification of protein complexes in mammalian cells. *Proteomics* **5**, 132–137 (2005).
 65. Knuesel, M. *et al.* Identification of novel protein-protein interactions using a versatile mammalian tandem affinity purification expression system. *Mol. Cell. Proteomics* **2**, 1225–1233 (2003).
 66. Westermarck, J. *et al.* The DEXD/H-box RNA helicase RHII/Gu is a co-factor for c-Jun-activated transcription. *EMBO J.* **21**, 451–460 (2002).
 67. Gavin, A.C. & Superti-Furga, G. Protein complexes and proteome organization from yeast to man. *Curr. Opin. Chem. Biol.* **7**, 21–27 (2003).
 68. Benesch, J.L.P., Aquilina, J.A., Ruotolo, B.T., Sobott, F. & Robinson, C.V. Tandem mass spectrometry reveals the quaternary organization of macromolecular assemblies. *Chem. Biol.* **13**, 597–605 (2006).
 69. Kelleher, N.L. *et al.* Top down versus bottom up protein characterization by tandem high-resolution mass spectrometry. *J. Am. Chem. Soc.* **121**, 806–812 (1999).
 70. Benjamin, D.R., Robinson, C.V., Hendrick, J.P., Hartl, F.U. & Dobson, C.M. Mass spectrometry of ribosomes and ribosomal subunits. *Proc. Natl. Acad. Sci. USA* **95**, 7391–7395 (1998).
 71. Sharon, M., Taverner, T., Ambroggio, X.I., Deshaies, R.J. & Robinson, C.V. Structural organization of the 19S proteasome lid: Insights from MS of intact complexes. *PLoS Biol.* **4**, 1314–1323 (2006).
 72. Hernandez, H., Dziembowski, A., Taverner, T., Seraphin, B. & Robinson, C.V. Subunit architecture of multimeric complexes isolated directly from cells. *EMBO Rep.* **7**, 605–610 (2006).
 73. Rappsilber, J., Siniosoglou, S., Hurt, E.C. & Mann, M. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal. Chem.* **72**, 267–275 (2000).
 74. Vasilescu, J., Guo, X.C. & Kast, J. Identification of protein-protein interactions using *in vivo* cross-linking and mass spectrometry. *Proteomics* **4**, 3845–3854 (2004).
 75. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J.V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1**, 2856–2860 (2006).
 76. Havlis, J. & Shevchenko, A. Absolute quantification of proteins in solutions and in polyacrylamide gels by mass spectrometry. *Anal. Chem.* **76**, 3029–3036 (2004).
 77. Link, A.J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682 (1999).
 78. Washburn, M.P., Wolters, D. & Yates, J.R. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).
 79. Wilm, M. *et al.* Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379**, 466–469 (1996).
 80. Wilm, M. & Mann, M. Analytical properties of the nanoelectrospray ion source. *Anal. Chem.* **68**, 1–8 (1996).
 81. 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).
 82. Karas, M., Bachmann, D., Bahr, U. & Hillenkamp, F. Matrix-assisted ultraviolet-laser desorption of nonvolatile compounds. *Int. J. Mass Spectrom. Ion Process.* **78**, 53–68 (1987).
 83. Clauser, K.R., Baker, P. & Burlingame, A.L. Role of accurate mass measurement (+/– 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **71**, 2871–2882 (1999).
 84. Perkins, D.N., Pappin, D.J.C., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
 85. Eng, J.K., McCormack, A.L. & Yates, J.R. 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).
 86. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).
 87. Olsen, J.V. & Mann, M. Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc. Natl. Acad. Sci. USA* **101**, 13417–13422 (2004).
 88. Nesvizhskii, A.I. & Aebersold, R. Interpretation of shotgun proteomic data—the protein inference problem. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).
 89. Brazma, A., Krestyaninova, M. & Sarkans, U. Standards for systems biology. *Nat. Rev. Genet.* **7**, 593–605 (2006).
 90. Taylor, C. F. *et al.* The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* **25**, 887–893 (2007).
 91. Orchard, S. *et al.* The minimum information required for reporting a molecular interaction experiment (MIMIX). *Nat. Biotechnol.* **25**, 894–898 (2007).
 92. Gerber, S.A., Rush, J., Stemman, O., Kirschner, M.W. & Gygi, S.P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. USA* **100**, 6940–6945 (2003).
 93. Aloy, P. *et al.* Structure-based assembly of protein complexes in yeast. *Science* **303**, 2026–2029 (2004).
 94. Schirle, M., Heurtier, M.A. & Kuster, B. Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. *Mol. Cell. Proteomics* **2**, 1297–1305 (2003).
 95. Krogan, N.J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006).
 96. Gavin, A.C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
 97. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
 98. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98**, 4569–4574 (2001).
 99. Andersen, J.S. & Mann, M. Organellar proteomics: turning inventories into insights. *EMBO Rep.* **7**, 874–879 (2006).
 100. Goh, K.I. *et al.* The human disease network. *Proc. Natl. Acad. Sci. USA* **104**, 8685–8690 (2007).