Eur. J. Biochem. **270**, 570–578 (2003) © FEBS 2003 doi:10.1046/j.1432-1033.2003.03428.x

MINIREVIEW

Affinity purification-mass spectrometry

Powerful tools for the characterization of protein complexes

Andreas Bauer and Bernhard Kuster

Cellzome AG, Heidelberg, Germany

Multi-protein complexes are emerging as important entities of biological activity inside cells that serve to create functional diversity by contextual combination of gene products and, at the same time, organize the large number of different proteins into functional units. Many a time, when studying protein complexes rather than individual proteins, the biological insight gained has been fundamental, particularly in cases in which proteins with no previous functional annotation could be placed into a functional context derived from their 'molecular environment'. In this minireview, we summarize the current state of the art for the retrieval of multi-protein complexes by affinity purification and their analysis by mass spectrometry. The advances in technology made over the past few years now enable the study of protein complexes on a proteomic scale and it can be anticipated that the knowledge gathered from such projects will fuel drug target discovery and validation pipelines and that the technology is also going to prove valuable in the emerging field of systems biology.

Keyword: TAP (tandem affinity purification).

Protein complexes

In the postgenomic era, proteins are coming back into focus because it has been realized again that whole genome sequence information alone is not sufficient to explain and predict cellular phenomena, as it is largely the proteins that execute and control the majority of cellular activities. While the human genome is estimated to contain approximately 30 000–40 000 genes, the corresponding proteome is much more complex. Events such as alternative splicing of genes and post-translational modifications generate a highly diverse set of proteins that could exceed a million distinct molecular species within a given cell. This molecular diversity could contribute to explaining many of the differences between evolutionary distant species that do not differ substantially in the total number of genes encoded in their respective genomes. However, within this diverse set of molecules, it is critical to maintain functional organization. It is becoming increasingly clear that an important level of organization is provided by multi protein complexes because instead of proteins and substrates colliding in a diffusion-dependent manner, proteins generally interact with each other and form larger assemblages in a time- and space-dependent manner [1]. At the same time, protein complexes provide functional diversity that is coded by the contextual combination of gene products. Within a protein complex, each individual protein may have a particular specialized function that contributes to the overall function of the complex. In turn, this specialized function may well be dependent on the interaction with neighboring protein surfaces that may lead to e.g. modulation of protein activity through conformational changes or post-translational modifications. Prominent examples of protein complexes are the ribosome, the spliceosome and the nuclear pore complex but many more (and less macroscopic) protein complexes with more subtle or diverse functions and exemplified by the many signal transduction pathways have been described [2,3].

The beauty of studying complexes is that it allows to place proteins with hitherto unknown roles into a functional context that is provided by their associated partners, some of which may have a known function. Even when analyzing proteins of known function, novel insight can be gained from describing their molecular environment. Quite often, proteins participate to more than one complex or do so in different subcellular compartments which can help to understand cross-talk between seemingly unconnected cellular activities. The extent to which such functional connectivities are operating in cells may be best appreciated by large-scale functional proteomics projects that build comprehensive interaction maps for proteins and protein complexes [4,5]. From a simplistic pharmacological point of view, functional proteomics via the analysis of protein complexes would contribute to the identification of novel drug targets, the reconstruction of pathways and help to understand the mechanism of action and side-effects of therapeutic compounds.

Correspondence to B. Kuster, Cellzome AG, Meyerhofstrasse 1, 69117 Heidelberg, Germany. Fax: + 49 6221 13757202, E-mail: Bernhard.kuster@cellzome.com or Bernhard.kuster@cellzome.de

Abbreviations: PrP(C), cellular prion protein; N-CAM, neural cell adhesion molecule; TAP, tandem affinity purification; TEV, Tobacco etch virus; CBP, calmodulin binding peptide; PMF, peptide mass fingerprinting; MS, mass spectrometry.

Note: web page available at http://www.cellzome.com

(Received 13 September 2002, accepted 12 December 2002)

Analysis of protein complexes

Within the scope of this review, we will focus on summarizing the current technical state of the art for the biochemical retrieval of associated proteins and protein complexes from cells and tissues as well as their analysis by mass spectrometry, for it is this combination of biochemistry and mass spectrometry that is driving progress in the field of functional proteomics today. Special attention will be paid to one major technical aspect dominating both the discovery and analysis parts of functional proteomics which is the handling of very complex protein mixtures. A protein of interest will typically represent only a tiny fraction of the total protein present in the source material. More than 10 000 different genes might be expressed at the same time in a single cell or tissue and, as mentioned earlier, diversity on the protein level is much higher. In addition to the diversity on the level of primary protein sequence and the presence of modifications, complexity is further increased when considering the dynamic range of expression levels of individual proteins. While some proteins are present in several thousand copies per cell, others are just represented by a few molecules. On top of the pure quantities the availability of proteins for complex retrieval may be much lower as it is easy to imagine that just a minor percentage might be in a physiologically active state and possibly part of several protein complexes.

As an example, the cell–cell adhesion molecule beta-catenin was originally described to be associated with the plasma membrane protein E-cadherin which mediates cell–cell contact [6–8]. Unexpectedly, the protein was recently also found to bind to HMG box transcription factors (TCF, LEF-1) which drive the expression of downstream target genes of the Wnt-signaling pathway [9–11]. Both interactions are largely independent of each other and knowledge of both yields information on different aspects of beta-catenin function. While it is straightforward to purify the stable and fairly abundant cell adhesion complex, the identification of the nuclear transcription factor complex is technically complicated because the percentage of beta-catenin participating to this complex is typically very low.

Isolation of protein complexes by affinity chromatography

The purification of protein complexes has been accomplished by a multitude of different techniques ranging from classical methods such as size exclusion or ion exchange chromatography to different varieties of affinity chromatography. Following the arguments that have been made about sample complexity in the previous section, it becomes apparent that successful approaches will have to include at least one highly discriminating separation step. This is typically provided by affinity-based methods. The common theme of these is the use of an inherent interaction (affinity) of two biomolecules. If one of the molecules is immobilized on a solid support, the interacting molecule can be purified from e.g. a cell lysate along with associated proteins. There are many different such affinity reagents but we will confine ourselves to examining those that have proven useful for the retrieval of protein complexes, notably recombinant proteins, epitope-tagged proteins and antibodies (Table 1). For more specialized applications, peptides and nucleic acids have also been used.

Technically, the discovery of interacting proteins is influenced by a number of parameters. Biochemical determinants include binding affinities between components of

Table 1. Strengths and weaknesses of commonly used affinity approaches for the retrieval of protein complexes.

Approach	Pros	Cons	Typical application area
Immunoprecipitation	Independent of cloning and ectopic gene expression	Cross-reactivity of antibody	Preparative IP from tissue
	Rapid procedures (if Ab is available)	Antibody bleeding from column Not generic (availability of Ab)	Co-IP from tissue culture and tissue
Epitope-tagging	Generically applicable approach Ability to purify low abundant proteins/protein complexes	Ectopic gene expression necessary Protein-tag might influence protein function Stringent biochemical conditions necessary for affinity purification to discriminate against common contaminants	Complex retrieval from tissue culture Large-scale studies
GST-pulldown	Generically applicable approach Ability to purify low abundant proteins/protein complexes Applicable to very weak protein interactions	Complex formation *in-vitro* Competition with *in-vivo* pre-assembled complex	Protein-protein interaction studies Complex retrieval from tissue
TAP	Generically applicable approach Ability to purify low abundant proteins/protein complexes Physiological conditions throughout the biochemical purification	Ectopic gene expression necessary Protein-tag might influence protein function	Complex retrieval from tissue culture Large-scale studies

the complex as well as their stability and solubility during cell lysis and affinity purification. Biological determinants such as the cellular expression level and tissue specific expression pattern of the protein of interest (bait) can also have a major influence on the choice of approach. Last but not least, technical limitations imposed by some methods (e.g. cloning of large cDNAs) also need to be considered.

The classic approach: antibodies

The classic co-immunoprecipitation (IP) experiment using antibodies is probably the most frequently employed method for testing whether two proteins are associated *in vivo* but the method can also be successfully used for the discovery of novel interacting partners in a protein complex [12,13]. In a typical experiment, a protein complex is affinity captured from cell lysates by an immobilized antibody that specifically recognizes an epitope of one known component of the complex. The retrieved complex is washed extensively to remove unspecifically bound proteins and is subsequently eluted form the resin prior to protein identification by mass spectrometry (see below).

There are several arguments that would favor an antibody approach. Importantly, antibodies allow the retrieval of protein complexes from endogenous components of cells and tissues which is obviously closest to resembling physiological conditions as there is no need for ectopic expression of the bait protein that could lead to a variety of problems (see section on epitope tagging). In fact, there may often be no alternative to IP on endogenous protein as a particular protein along with associated partners might only be expressed in a specific tissue and the establishment of a corresponding *in vivo* tissue culture model may not always be possible. Given the available of a good quality antibody, IP experiments are also fast to perform as no cloning of complex components is involved in the process.

One limitation of antibody IPs is that an individual antibody is needed for every bait protein. Whether antibodies are commercially available or custom made, it is not easy to predict if the specificity and affinity of the produced antibodies will turn out to be useful for immunoprecipitation. There are some other, more technical, limitations to the method: Even mouse monoclonal antibodies might exhibit cross-reactivity with proteins other than the immunogen. In this case several proteins (or protein complexes for that matter) that are not related to the bait protein might be precipitated which, essentially, leads to the generation of false positives. Cross-reactivity of the antibody aside, very abundant proteins might unspecifically bind to the resin on which the antibody is immobilized. These have to be removed efficiently because otherwise the data will become more difficult to interpret. Specificity is typically increased by washing the immobilized protein complex briefly before elution using high stringency conditions (200–500 mM salt). However, components that are not tightly bound (high K_{off}) might be lost during this procedure. Although this point is raised in this section, other affinity approaches will suffer from the same limitation. Loss of particular components of a protein complex might be overcome by chemical cross-linking. Schmitt-Ulms *et al.* [14] have demonstrated that the interaction of the cellular prion protein (PrP(C)) with neural cell adhesion molecules (N-CAMs) can be maintained during an IP experiment by mild formaldehyde treatment of cells. While chemical cross-linking is an interesting approach, it is questionable how soon such methods would be applicable in a generic fashion [15,16]. Antibody bleeding from the column is another technical problem that could become serious in the later MS analysis. Large amounts of antibodies present in the eluate might mask the presence of proteins of the purified complex especially when samples are not separated by gel electrophoresis prior to MS analysis. Bleeding can be reduced by cross-linking the antibody to the resin or by specific elution (i.e. with peptides representing the epitope). In practical terms, however, cross-linking cannot fully prevent antibody bleeding and specific peptide elution is often incomplete which reduces sample recovery.

The generic approach: epitope tagging

Antibodies can be used in a more generic way for the isolation of protein complexes that circumvents the need for producing specific antibodies. For this purpose, bait proteins can be fused to an epitope-tag and an antibody directed against the tag instead of the bait protein is used for complex retrieval. As a result, many different cDNAs can be fused to the same tag in parallel and complexes retrieved using the same antibody. Often, multimers of the same tag are used to increase the affinity and accessibility to the antibody. A variety of different epitope tags (e.g. Myc, HA, Flag, KT3) have been used successfully in the past and antibodies against these tags are commercially available. The underlying advantage of the approach is the high degree of reproducibility of the results because standardized technical procedures can be used that do not need to be optimized for each individual case.

The downside of epitope-tagging is obviously the need to ectopically express proteins in cells which largely limits this approach to cell culture systems. Tight control over the expression level of bait proteins must be exercised as overexpression might adversely affect the assembly of a protein complex. In addition, overexpression can occasionally cause cytotoxicity. This might be overcome by use of sophisticated vector systems in which the expression level of the epitope tagged protein can be regulated by an inducible promotor [17]. In addition, the artificially introduced tag may interfere with protein folding, protein function, or the ability to interact with other proteins. It is therefore advisable to create N-terminal and C-terminal fusions in parallel. Despite some of these limitations, co-IP via epitope tags is used extensively on a normal lab scale to identify protein complexes [18,19] but has recently also been employed on a proteomic scale [4,5].

The generic approach: 'GST pulldown'

The standard precipitation experiment with the aid of a recombinant GST fusion protein has been widely used for the discovery and analysis of individual protein interactions and, to a lesser extent, protein complexes [20,21]. In this approach, the protein of interest is expressed in *Escherichia coli* as a recombinant fusion protein and immobilized on a

solid support. Interacting proteins can then be precipitated (or 'pulled-down') by applying a cellular lysate to the column. An obvious advantage of this method is that it is robust, easy to use and capable of retrieving even weakly interacting and low abundant proteins owing to the fact that large amounts of recombinant protein are present on the column. However, not all proteins can be easily overexpressed in a soluble form in *E. coli*. Furthermore, interactions that may be dependent on the correct post-translational processing of the bait protein may not be provided by the expression system used. Because protein complexes are formed within cells, the recombinant fusion protein is in competition with the corresponding endogenous component and therefore the complex may not be retrieved at all because all complex components are engaged in endogenous protein–protein interactions.

The tailored approach: RNA and peptides

Although antibodies are certainly the most frequently used reagents for affinity purification, other bio-molecules may be used for that purpose as well. The application of these is typically tailored to the particular system under investigation and thus can be very specific albeit not universally applicable. Neubauer *et al*. [22] described the purification of the spliceosome with the help of biotinylated RNA as the 'affinity-hook' into the complex. Spliceosomes are complex ribonucleoprotein particles that are assembled from several smaller spliceosomal complexes. The main activity of the spliceosome is the recognition of splice sites in nuclear pre-mRNA and to catalyze the accurate removal of introns and ligation of exons to yield a mature mRNA. The affinity of the splicing complexes to RNA was used in this approach for the isolation of the complex. Splicing complexes were prefractionated by gel filtration and mixed with a biotinylated pre-mRNA that was subsequently affinity captured on a streptavidin-matrix. The protein content of the purified complex was then separated on a 2D gel and analyzed by mass spectrometry.

An example for the use of peptides as affinity ligands is the study of Husi *et al*. in which the isolation of a large protein complex from the postsynaptic density fraction of brain synapses is described [23]. The NMDA receptor is part of the postsynaptic density and known to be involved in signaling processes related to the regulation of synaptic plasticity by binding to the adaptor protein PSD95. This protein contains several PDZ domains one of which binds to the few most C-terminal amino acids of the NMDA receptor. This particular feature was exploited for the retrieval of a PSD95-containing complex by immobilizing a hexapeptide corresponding to the C-terminus of the NMDA receptor and subsequent binding of PSD95 along with associated proteins from mouse brain preparations.

The large-scale approach: tandem affinity purification

While the aforementioned techniques all have particular advantages, their individual limitations may render them less applicable to studying protein complexes on a proteomic scale. Retrieval methods for such applications should meet a number of important criteria. First and foremost, the method must be highly discriminating against unspecific protein background yet retain essential components of the complex. In addition, the method must be generic in the sense that all bait proteins must be processed under the same conditions in order to yield reproducible and comparable results as well as attaining the required throughput. If an acceptable level of reproducibility can be achieved, large datasets generated in this way can be mined for information that is not necessarily provided by individual experiments and thus increase the overall insight gained into the system under investigation.

A method that was recently developed to meet these criteria is tandem affinity purification (TAP) [24]. The basic concept of TAP is similar to the epitope tagging strategy described earlier. The main difference, however, is the sequential utilization of two tags instead of one (Fig. 1). First, the 'TAP-tagged' protein is expressed in cells to form a complex with the endogenous components. The tagged protein along with associated partners is retrieved via interaction of the ProteinA tag with immunoglobulins that are immobilized on agarose beads. In order to remove proteins that are unspecifically bound to the column, the retrieved protein complex is released by protease cleavage using the TEV (Tobacco etch virus) protease. This enzyme is a site specific protease that cleaves a seven amino-acid recognition sequence located between the first and the second tag. This sequence is only found in very few human proteins known so far thus ensuring that components of retrieved complexes are not themselves cleaved by the protease. In the second affinity step, the complex is immobilized to calmodulin coated beads via the calmodulin binding peptide (CBP) tag. This step removes the TEV protease and further contaminants that may be present. The CBP–calmodulin interaction is calcium dependent and, hence, the removal of calcium ions with chelating agents can be used as a second specific elution step that yields the final protein complex preparation.

Aside some of the limitations associated with epitope tagging mentioned earlier, the two step TAP approach has several advantages over traditional single step methods. Probably the most useful asset is the massive reduction of sample complexity through very efficient reduction of unspecific protein background. This achievement simplifies the analytical strategy chosen for protein identification by mass spectrometry and reduces the need for validating identified proteins as genuine interaction partners. Stringent purification conditions (high salt or detergent concentrations) which tend to result in the loss of associated proteins can be avoided and the complex can be kept under close to physiological conditions all along the purification procedure. Owing to the generic structure of the approach, the results of TAP purifications are highly reproducible and comparable for different bait proteins which renders the approach applicable to small- and large-scale studies alike.

Gavin *et al*. recently reported results of a large-scale proteomic study using the TAP/MS method in which 232 distinct protein complexes of the baker's yeast *Saccharomyces cerevisiae* were identified. These complexes, in turn, formed a massive network by sharing of protein components providing an unprecedented view on the level of

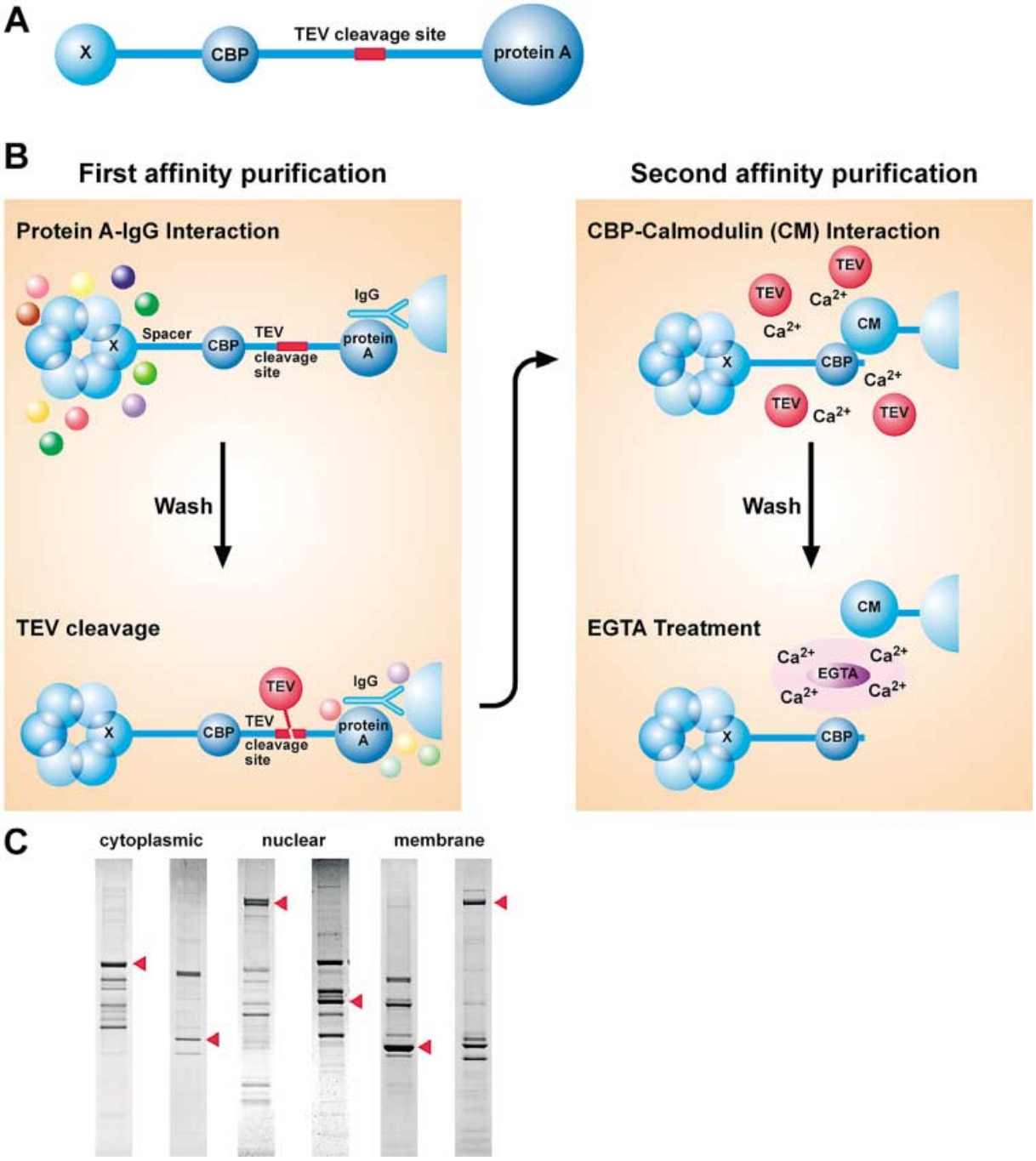

Fig. 1. Schematic representation of the tandem affinity purification method. (A) Structure of the TAP-tag. (B) TAP strategy. A protein complex containing the TAP-tagged protein is purified sequentially by two independent affinity steps on IgG- and calmodulin-containing resins, respectively. The immobilized protein is specifically eluted in the first instance by protease cleavage (TEV) and in the second step by lowering the calcium-dependent affinity of CBP to calmodulin. (C) Examples of TAP purified protein complexes from different subcellular localizations of cultured human cell lines. Bait proteins are marked with an arrow (CBP, calmodulin binding peptide; TEV protease, tobacco etch virus protease).

functional diversity and organization of a eukaryotic cell. Although most applications of the method have thus far been described for yeast complexes, the TAP/MS approach is equally applicable for the retrieval of protein complexes from higher eukaryotes such as human [4,25] and Drosophila (A. Veraksa, Harvard Medical School, personal communication).

Protein identification by mass spectrometry

Mass spectrometry is the method of choice for the identification of proteins in proteomics projects because of its superior speed, sensitivity and versatility compared to traditional protein sequencing by Edman degradation. To date, hundreds to several thousands of proteins may be

Table 2. Merits of mass spectrometry based protein identification strategies used for the analysis of protein complexes.

	MALDI-TOF MS	NanoES MS/MS	nLC-MS/MS	LC/LC-MS/MS
No use of gels	Very poor	Poor	Good	Very good
Use with 1D gels	Good	Good	Very good	Fair
Use with 2D gels	Very good	Very good	Good	Poor
Throughput	Very high	Poor	High	High
Sensitivity	Very high	Very high	High	Fair
Dynamic range	Poor	Fair	Good	Very good
Mixture analysis	Fair	Good	Very good	Very good
Robustness	Very high	Fair	High	Fair
Special expertise	Low	Fair	High	High

identified within one day from sample quantities in the subpicomol range whether or not proteins are N-terminally or otherwise modified. This advance in sensitivity and scope together with the extensive availability of protein and nucleotide sequence information has made mass spectrometry and biology much more compatible than in the past. The 'joined forces' of sophisticated biochemical and mass spectrometry approaches are the real driving forces in the entire field of proteomics today and have enabled researchers to embark on studies of much larger and more complicated systems than previously possible.

Just as much as complexity is an important aspect to consider when designing a biological experiment, this factor is also, to a large extent, governing the choice of which analytical strategy is taken. Gels, liquid chromatography (LC) and mass spectrometers are all separation devices (protein, peptide, mass, respectively) that can be used to break down sample complexity. Combination of two or more of these methods can attain enormous separation power, but only the combination with mass spectrometry generates data that is sufficient for the identification of a protein. Questions around whether or not a separation step prior to mass spectrometric analysis is required or which flavor of MS-based protein identification is appropriate to use are sometimes not trivial to answer as they largely depend on the analytical problem to which the technology is applied. Table 2 may serve as a rough guideline for assessing the merits of the more widely available MS strategies. In short, protein separation methods with high resolving power such as 2D gel electrophoresis are generally well compatible with MS approaches that have only limited capabilities for mixture analysis. Conversely, when the discrimination power of the analytical system is very high, there is less need for protein separation prior to MS analysis.

As far as protein complexes are concerned, the analysis typically starts from 1D gels that are used to separate the components of a protein complex (Fig. 2). 1D gels are primarily used for this purpose because the complexity of the protein mixture is normally not extremely high after affinity purification and the fact that running good quality 1D gels is technically trivial compared to 2D gels. Some details regarding protein isoforms and modifications may be lost but that loss can often be compensated for by the excellent extra separation dimension offered by the mass spectrometer. Following protein separation, the protein to be identified is cut from the gel and cleaved into peptides. Trypsin and Lys-C are most frequently employed for this purpose because they give rise to fragments that have favorable physico-chemical properties for peptide detection and sequencing in an MS experiment, notably the presence of a basic amino acid at the C-terminus of a peptide. There is a new trend that aims at avoiding protein separation by

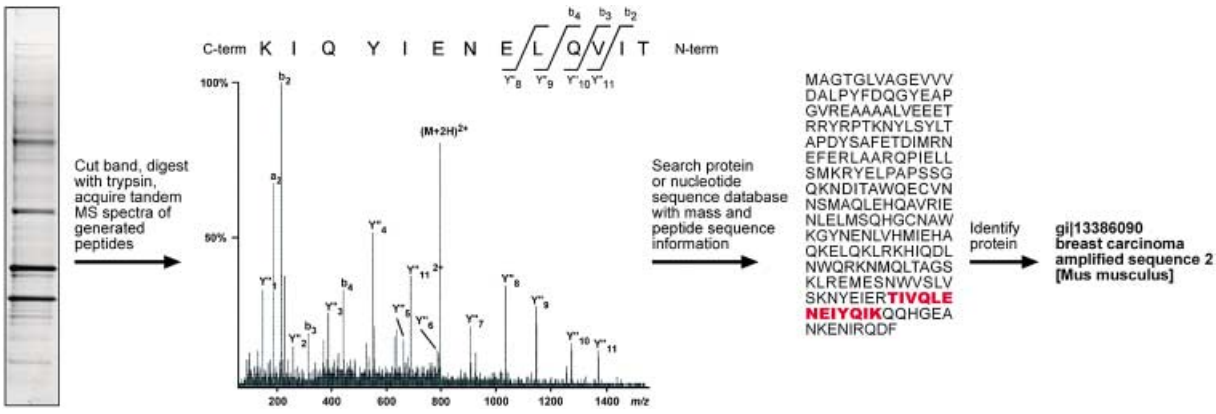

Fig. 2. Protein identification by tandem mass spectrometry. A protein complex is separated by 1D PAGE and bands of interest are cut from the gel, digested with trypsin and the generated peptides are partially sequenced by tandem mass spectrometry. Series of signals marked with Y''_n allow the corresponding peptide sequence to be determined from the tandem mass spectrum. Mass and partial amino-acid sequence data are simultaneously searched against protein or nucleotide sequence databases for protein identification.

gels altogether both for the profiling of protein complexes [26] as well as entire proteomes [27]. In these methods, all proteins of a preparation are digested together to generate a vast mixture of peptides. Subsequently, two orthogonal LC peptide separation methods in combination with tandem mass spectrometry (LC/LC-MS/MS) are employed to identify the proteins that were originally part of the mixture. This so-called shotgun sequencing approach (or multi dimensional protein identification technology, MudPit) is currently attracting a lot of interest but it is not clear yet if such approaches offer generic advantages for the analysis of protein complexes both in terms of dynamic range (i.e. how little of one protein can be identified in the presence of how much of a mixture of other proteins) and absolute sensitivity as the MudPit approach compromises sensitivity for the ability to cope with complexity.

Protein identification by peptide mass fingerprinting

Two fundamentally different approaches to protein identification using mass spectrometric information can be differentiated. The first of these methods is called peptide mass fingerprinting (PMF) [28]. In this technique, the masses of peptides from a tryptic protein digest are determined using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI TOF MS). It is important to note that no peptide sequence is generated in the experiment but that the set of measured peptides for that protein is characteristic and can serve as a fingerprint that enables its identification. Technically, this takes the form of searching the list of determined peptide masses against a sequence database in which every protein has been digested *in silico* using the same enzyme. Proteins are identified by a statistically significant overlap between the experimentally determined and theoretically predicted peptide masses. Peptide mass fingerprinting works by the statistical rational that although a single peptide mass might correspond to many different peptides in many different proteins, it is extremely unlikely that the same set of peptide masses would be found in a number of different (random) proteins by chance. The strengths of the technique are that it is experimentally simple to perform, very sensitive, fast and that the results are usually straightforward to interpret. The downsides are the statistical limitations imposed by the method. These include the requirement that the majority of the coding sequence ($> 80\%$) of a protein has to be present in a database and that a sufficiently high number of peptides must be detected in the experiment. Sequence information as contained in isolated ESTs and pieces of raw genomic DNA is generally not suitable for interrogation using PMF data because these sequences are generally too short to accommodate a sufficiently high number of peptides of the complete protein that would be required for an unambiguous identification. Clustering of ESTs and gene prediction can make this information available for searching but it has to be kept in mind that even small errors in EST clustering or gene prediction can lead to substantial effects on the mass values of the predicted peptides. Very small proteins (< 15 kDa) sometimes present problems too as they may not produce a high enough number of tryptic peptides to warrant identification. Although PMF is capable of identifying proteins in mixtures, a practical limit appears to be 2–5 proteins that are present in roughly equal molar amounts (low dynamic range). For the analysis of protein complexes that means that PMF can only be used in conjunction with 1D or 2D gels as a protein separation step.

Protein identification by tandem mass spectrometry

An alternative approach that overcomes some of the limitations of peptide mass fingerprinting uses a combination of partial peptide sequence and mass information for protein identification (tandem MS or MS/MS). In this technique, a particular peptide is first mass measured, then isolated from the mixture (within the mass spectrometer) and subjected to collisions with inert gas molecules. These collisions result in cleavage of the peptide along the peptide backbone and creates a set of fragments that differ in length by one amino acid each. The masses of the fragments can again be measured within the mass spectrometer to produce a series of signals which correspond in mass to adjacent amino-acid residues in the sequence (Fig. 2). Quite often, only a part of the sequence can be read from the sequence. However, this stretch of consecutive sequence is 'locked' within the peptide by the masses of the fragments that define the beginning and the end of the determined sequence. Information on peptide sequence, peptide mass and fragment mass can be queried simultaneously against a database in which the fragmentation patterns of all peptides derived from all proteins in that database are computed and compared to the experimentally determined spectrum in order to identify the underlying protein [29–31]. Each analyzed peptide independently identifies a given protein provided that this peptide sequence is unique. Analysis of many peptides of the digest can confirm the identification of a protein or identify a different protein that happens to be part of the mixture. It can be shown that even a consecutive sequence read of three or four amino acids from a single partially sequenced peptide is sufficient for protein identification. As a result, protein, EST and genome sequence databases can be made available for protein identification [32]. The latter aspect is particularly useful for the study of model organisms for which only limited sequence information on the protein level is available. Even in cases where no sequence information is available, tandem mass spectrometry has often been used to generate peptide sequence information for a protein that can be used to attempt cloning of the corresponding cDNA.

Tandem mass spectrometry is generally combined with nanoelectrospray ionization (nanoES) [33] or on-line nanoLC peptide separation [34]. Nanoelectrospray-MS allows for extended measurement time which is advantageous when available sample quantities are very low (< 100 fmol) whereas nanoLC-MS is much more amenable to automation, provides enhanced sequence coverage of a protein (for analysis of post-translational modifications) and generally allows more efficient handling of complex mixtures especially when the relative quantities of proteins in the sample are very different. On-line nanoLC-MS provides sufficient dynamic range to allow the identification of proteins that constitute as little 2–5% of the total protein mixture which is an important asset when samples contain a

large excess of contaminating proteins (e.g. antibodies in IP experiments).

Conclusions

The combination of affinity methods for the purification of protein complexes and the identification of their components by mass spectrometry is continuing to prove extremely successful. Examples from the literature cover all compartments of a cell from nuclear, cytoplasmic, endo-membrane to plasma membrane localization and cellular functions ranging from cell cycle control to metabolic and signal transduction pathways. The analysis of protein interactions and protein complexes are currently driving significant efforts in academic and commercial settings to elucidate systematically pathways and the functional context in which proteins operate in a variety of organisms and cell types. The specificity of affinity chromatography and the sensitivity and certainty of MS based protein identification now allows to access even proteins that are present in rather few copies per cell including many of the hitherto functionally unassigned proteins. By putting these proteins into a physiological context, it is possible to discover new biological phenomena, set up and test new biological hypothesis and design experiments to either prove or disprove a particular role of a protein under investigation. Past experience shows that many times that protein complexes were studied using approaches such as the ones described above, the biological insight gained has been fundamental for the understanding and solving of biological puzzles. This will continue to be the case particularly as the technology is now in place to study protein complexes on a proteomic scale. It can be anticipated that the knowledge gathered from such projects will fuel drug target discovery and validation pipelines and that the technology is going to prove valuable in the emerging field of systems biology [35].

Acknowledgements

The authors wish to thank Paola Grandi, Gitte Neubauer and Giulio Superti-Furga for critically reading the manuscript and Frank Weisbrodt for help with the graphics.

References

1. Alberts, B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291–294.
2. Garrels, J.I. (1996) YPD-A database for the proteins of *Saccharomyces cerevisiae*. *Nucleic Acids Res*. **24**, 46–49.
3. Bader, G.D & Hogue, C.W. (2000) BIND – a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **16**, 465–477.
4. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. & Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147.
5. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D. & Tyers, M. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183.
6. Ozawa, M., Baribault, H. & Kemler, R. (1989) The cytoplasmic domain of the cell adhesion molecule uvomorulin associates with three independent proteins structurally related in different species. *EMBO J*. **8**, 1711–1717.
7. Aberle, H., Butz, S., Stappert, J., Weissig, H., Kemler, R. & Hoschuetzky, H. (1994) Assembly of the cadherin-catenin complex *in vitro* with recombinant proteins. *J. Cell Sci*. **107**, 3655–3663.
8. Rimm, D.L., Koslov, E.R., Kebriaei, P., Cianci, C.D. & Morrow, J.S. (1995) Alpha 1 (E) -catenin is an actin-binding and -bundling protein mediating the attachment of F-actin to the membrane adhesion complex. *Proc. Natl Acad. Sci. USA* **92**, 8813–8817.
9. Molenaar, M., van de Wetering, M., Oosterwegel, M., Peterson-Maduro, J., Godsave, S., Korinek, V., Roose, J., Destree, O. & Clevers, H. (1996) XTcf-3 transcription factor mediates beta-catenin-induced axis formation in *Xenopus* embryos. *Cell* **86**, 391–399.
10. Behrens, J., von Kries, J.P., Kuhl, M., Bruhn, L., Wedlich, D., Grosschedl, R & Birchmeier, W. (1996) Functional interaction of beta-catenin with the transcription factor LEF-1. *Nature* **382**, 638–642.
11. Huber, O., Korn, R., McLaughlin, J., Ohsugi, M., Herrmann, B.G & Kemler, R. (1996) Nuclear localization of beta–catenin by interaction with transcription factor LEF-1. *Mech. Dev*. **59**, 3–10.
12. Ajuh, P., Kuster, B., Panov, K., Zomerdijk, J.C., Mann, M & Lamond, A.I. (2000) Functional analysis of the human CDC5L complex and identification of its components by mass spectrometry. *EMBO J*. **19**, 6569–6581.
13. Wang, Y., Cortez, D., Yazdi, P., Neff, N., Elledge, S.J & Qin, J. (2000) BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures. *Genes Dev*. **14**, 927–939.
14. Schmitt-Ulms, G., Legname, G., Baldwin, M.A., Ball, H.L., Bradon, N., Bosque, P.J., Crossin, K.L., Edelman, G.M., DeArmond, S.J., Cohen, F.E & Prusiner, S.B. (2001) Binding of neural cell adhesion molecules (N-CAMs) to the cellular prion protein. *J. Mol. Biol*. **314**, 1209–1225.
15. Jackson, V. (1999) Formaldehyde cross-linking for studying nucleosomal dynamics. *Methods* **17**, 125–139.
16. Fancy, D.A. (2000) Elucidation of protein–protein interactions using chemical cross- linking or label transfer techniques. *Curr. Opin. Chem. Biol*. **4**, 28–33.
17. Medina, D., Moskowitz, N., Khan, S., Christopher, S & Germino, J. (2000) Rapid purification of protein complexes from mammalian cells. *Nucleic Acids Res*. **28**, E61.
18. Zachariae, W., Shevchenko, A., Andrews, P.D., Ciosk, R., Galova, M., Stark, M.J., Mann, M & Nasmyth, K. (1998) Mass spectrometric analysis of the anaphase-promoting complex from yeast: identification of a subunit related to cullins. *Science* **279**, 1216–1219.
19. Ikura, T., Ogryzko, V.V., Grigoriev, M., Groisman, R., Wang, J., Horikoshi, M., Scully, R., Qin, J & Nakatani, Y. (2000)

Involvement of the TIP60 histone acetylase complex in DNA repair and apoptosis. *Cell* **102**, 463–473.
20. Becamel, C., Alonso, G., Galeotti, N., Demey, E., Jouin, P., Ullmer, C., Dumuis, A., Bockaert, J & Marin, P. (2002) Synaptic multiprotein complexes associated with 5-HT (2C) receptors: a proteomic approach. *EMBO J.* **21**, 2332–2342.
21. Rappsilber, J., Ajuh, P., Lamond, A.I & Mann, M. (2001) SPF30 is an essential human splicing factor required for assembly of the U4/U5/U6 tri-small nuclear ribonucleoprotein into the spliceosome. *J. Biol. Chem.* **276**, 31142–31150.
22. Neubauer, G., King, A., Rappsilber, J., Calvio, C., Watson, M., Ajuh, P., Sleeman, J., Lamond, A & Mann, M. (1998) Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat. Genet.* **20**, 46–50.
23. Husi, H., Ward, M.A., Choudhary, J.S., Blackstock, W.P & Grant, S.G. (2000) Proteomic analysis of NMDA receptor-adhesion protein signaling complexes. *Nat Neurosci.* **3**, 661–669.
24. Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M & Seraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030–1032.
25. Westermarck, J., Weiss, C., Saffrich, R., Kast, J., Musti, A.M., Wessely, M., Ansorge, W., Seraphin, B., Wilm, M., Valdez, B.C & Bohmann, D. (2002) The DEXD/H-box RNA helicase RHII/Gu is a co-factor for c-Jun-activated transcription. *EMBO J.* **21**, 451–460.
26. Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., Yates, J.R & 3rd. (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682.
27. Washburn, M.P., Wolters, D. & Yates, J.R. 3rd. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol.* **19**, 242–247.
28. Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C & Watanabe, C. (1993) 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.
29. Mann, M & Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal Chem.* **66**, 4390–4399.
30. Eng, J., McCorrmack, A & Yates, J. (1994) 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.
31. Perkins, D.N., Pappin, D.J., Creasy, D.M & Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567.
32. Kuster, B., Mortensen, P., Andersen, J.S & Mann, M. (2001) Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **1**, 641–650.
33. Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T & Mann, M. (1996) Femtomole sequencing of proteins from polyacrylamide gels by nano- electrospray mass spectrometry. *Nature* **379**, 466–469.
34. McCormack, A.L., Schieltz, D.M., Goode, B., Yang, S., Barnes, G., Drubin, D & Yates, J.R.R. (1997) Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal Chem.* **69**, 767–776.
35. Ideker, T., Galitski, T & Hood, L. (2001) A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372.