

## Topical Review

# Advances in protein complex analysis using mass spectrometry

Anne-Claude Gingras, Ruedi Aebersold and Brian Raught

*Institute for Systems Biology, Seattle, WA 98103, USA*

Proteins often function as components of larger complexes to perform a specific function, and formation of these complexes may be regulated. For example, intracellular signalling events often require transient and/or regulated protein–protein interactions for propagation, and protein binding to a specific DNA sequence, RNA molecule or metabolite is often regulated to modulate a particular cellular function. Thus, characterizing protein complexes can offer important insights into protein function. This review describes recent important advances in mass spectrometry (MS)-based techniques for the analysis of protein complexes. Following brief descriptions of how proteins are identified using MS, and general protein complex purification approaches, we address two of the most important issues in these types of studies: specificity and background protein contaminants. Two basic strategies for increasing specificity and decreasing background are presented: whereas (1) tandem affinity purification (TAP) of tagged proteins of interest can dramatically improve the signal-to-noise ratio via the generation of cleaner samples, (2) stable isotopic labelling of proteins may be used to discriminate between contaminants and *bona fide* binding partners using quantitative MS techniques. Examples, as well as advantages and disadvantages of each approach, are presented.

(Received 1 December 2004; accepted after revision 15 December 2004; first published online 16 December 2004)

**Corresponding author** A.-C. Gingras; [agingras@systemsbiology.org](mailto:agingras@systemsbiology.org); R. Aebersold; [aegersold@biotech.biol.ethz.ch](mailto:aegersold@biotech.biol.ethz.ch); B. Raught; [brought@systemsbiology.org](mailto:brought@systemsbiology.org)

## MS identification

In recent years, mass spectrometry (MS) has emerged as a powerful tool to quickly and efficiently identify proteins in biological samples (for an accessible, in-depth explanation of mass spectrometry of biological samples, see Steen & Mann, 2004), placing MS at the forefront of technologies to probe for protein interactions. Proteins most often interact with each other to form transient or stable complexes which carry out biological activities. In addition, some proteins specifically interact with non-protein molecules, such as DNA, RNA or metabolites, and these interactions are critical for function. Thus, defining the composition of protein complexes, as well as understanding how complexes are assembled and regulated yield invaluable insights into protein function. Coupled with an isolation technique to purify a specific protein complex of interest, MS can rapidly and reliably identify the components of complexes. In addition, quantitative MS techniques offer the possibility of studying dynamically regulated interactions (see below).

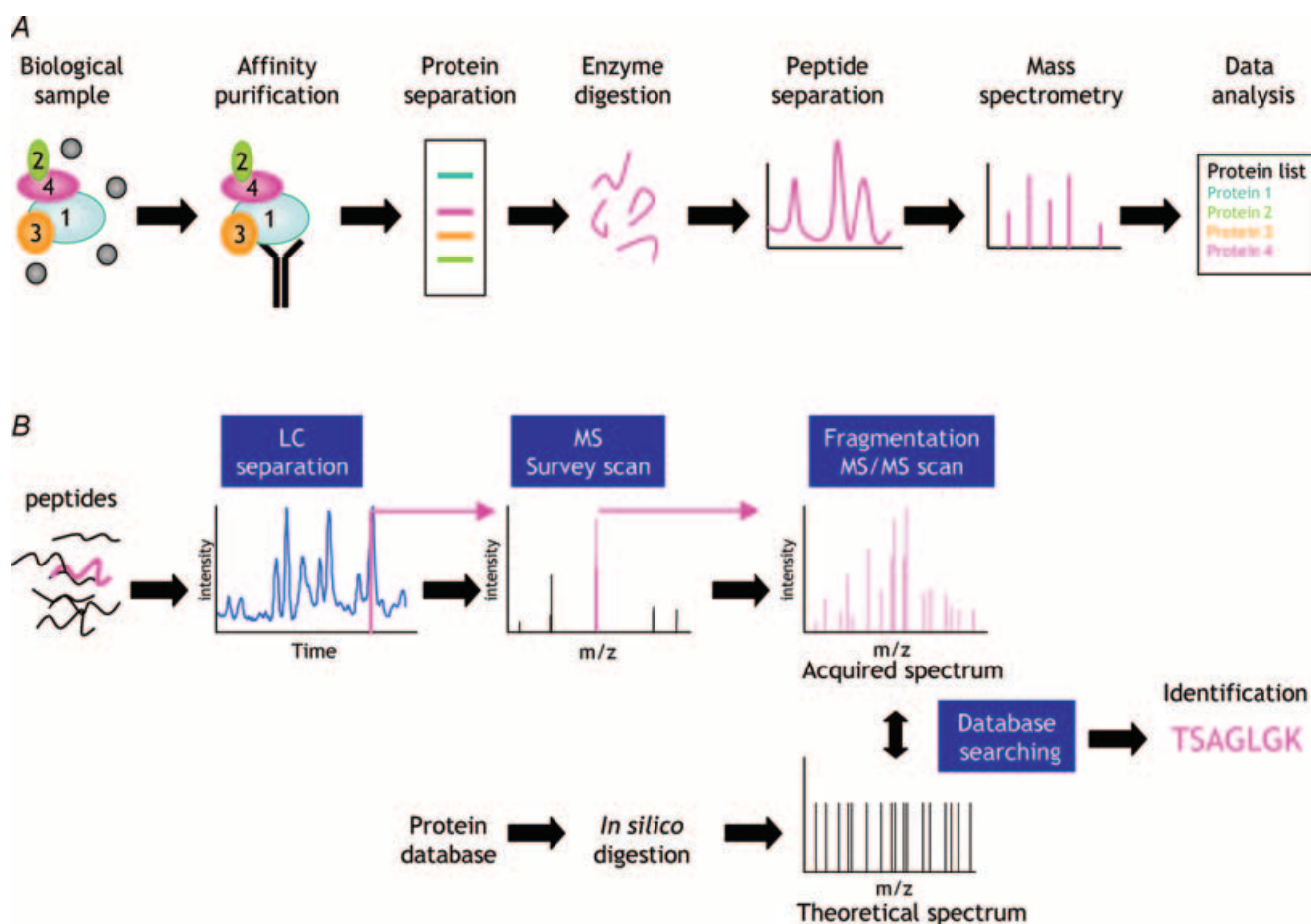
A general strategy utilized to characterize protein complex composition using MS is depicted in Fig. 1A.

A protein complex to be analysed is first purified using an appropriate approach – most often including an affinity chromatography step. Appropriate negative controls (e.g. an immunoprecipitate using pre-immune serum) are conducted in parallel to discriminate between *bona fide* components of the complex and background contaminants. Purified protein complex components may then be separated using techniques such as SDS-PAGE, isoelectric focusing, or various two-dimensional separation methods. Individual proteins may then be visualized by staining, and recovered for analysis. Isolated proteins are proteolytically digested (most often using trypsin) to generate a mixture of peptides that can be identified by MS (see below). If the sample complexity is low, gel separation may not be necessary. Instead, the digested protein complex may be subjected directly to MS. The proteins present in the sample are deduced by recombining the identified peptides, using suitable informatics tools (Nesvizhskii & Aebersold, 2004; Eng *et al.* 2005).

A particularly reliable and robust approach for the identification of peptides in complex mixtures is

reversed-phase capillary liquid chromatography (RPLC), directly in-line with a mass spectrometer (Fig. 1B). The RPLC column is packed with fused silica particles conjugated to extended carbon chains, and thus separates peptides on the basis of hydrophobicity. Once loaded onto the column, bound peptides are subjected to increasing amounts of organic solvent (such as acetonitrile), such that more hydrophilic peptides move into the mobile phase earlier in the gradient, and more hydrophobic peptides are eluted later in the gradient. In-line RPLC columns used in most laboratories are very small (typically in the range of 50–150  $\mu\text{m}$  internal diameter), allowing for the elution of small quantities of peptides in minimal buffer volumes, and therefore at high concentrations. In a process referred to as electrospray ionization (ESI; Steen & Mann,

2004), as the positively charged peptides elute from the column, they pass through a fine tip to which a high electrical potential is applied, and are sprayed into the MS within fine droplets. As the droplets pass through a heated chamber, the buffer is evaporated, sending desolvated peptide ions to the instrument. (Other systems which uncouple liquid chromatography and mass spectrometry may also be used. Such configurations are typically based on matrix-assisted laser desorption/ionization (MALDI), and are reviewed elsewhere (Steen & Mann, 2004).) As the peptide ions enter the spectrometer, their mass/charge ratios (in the effective range of the MS, usually between  $\sim 200$ – $2000$  Da) are measured in real time. Peptides are then automatically selected for fragmentation, a process in which individual peptide populations are forced to



**Figure 1. General strategy for protein complex identification using mass spectrometry**

**A**, a biological sample is purified and separated into its constituents, which are then proteolysed and analysed by LC-MS (see text for details). **B**, mass-spectrometry-based protein identification. A mixture of peptides (the peptide of interest is highlighted in pink) is separated by reversed-phase HPLC. The chromatography column is located immediately in-line with the MS, and peptides are analysed as they elute from the column. The mass/charge ratios ( $m/z$ ) of all co-eluting peptides are first analysed in a survey (or MS) scan. Individual peptide populations are then selected (usually based on abundance) for fragmentation. Finally, the  $m/z$  of the peptide fragments are analysed to generate an MS/MS or CID spectrum. The acquired MS/MS spectrum is compared with theoretical spectra obtained via an *in silico* digest of a relevant protein database. Significant matches are reported, yielding peptide identification.

collide with gas particles, leading to the disruption of peptide bonds. The resulting peptide fragments are analysed to generate a tandem (MS/MS) mass spectrum (also known as collision induced dissociation – or CID – spectrum), which is rich in sequence information (Fig. 1B). Finally, peptide sequencing software (such as SEQUEST, Mascot, etc.) compares the acquired MS/MS spectrum to theoretical spectra generated from protein sequences in an appropriate database. The peptides identified are used to ascertain the protein population in the original sample. Many excellent database search tools, as well as accompanying analytical software, have been developed to evaluate the quality of peptide and protein assignments, and are reviewed elsewhere (Nesvizhskii & Aebersold, 2004).

### Binding partner isolation/identification

In characterizing binding partners for a molecule of interest using MS, the major challenge is to identify *bona fide* interacting partners *versus* sample contaminants. The isolation method thus plays a critical role in the ultimate success of the experiment. For example, a typical immunopurification (IP) generates significant background. While this is less important for procedures such as Western blotting, where the presence of one or more specific proteins is probed, the exquisite sensitivity of the mass spectrometer, and its ability to detect all proteins present in the sample, accentuates the contamination issue. Whereas more stringent washes may be used to reduce contaminating proteins, increasing the salt or detergent concentrations may also affect the binding of true – albeit weaker – interactors.

Antibodies typically cross-react with one or more irrelevant proteins; a typical IP will therefore contain not only cross-reactive contaminating proteins, but will also bring down their interacting partners, resulting in the identification of multiple proteins that are unrelated to the target complex. It is difficult to establish proper controls for IP protocols, in that pre-immune serum will not elicit the same type of background binding as a given specific antiserum. In addition, antibodies may interact with only a subpopulation of the protein of interest: for example, post-translational modifications may inhibit antibody binding, or, if the antibody interacts with a protein-binding domain, it may disrupt interactions with binding partners. Another common problem is that antibodies used for precipitation tend to leach from the support matrix during elution steps. High levels of immunoglobulin peptides can easily mask lower abundance peptides derived from low stoichiometry interacting partners in the MS run. This situation is typically addressed by cross-linking the antibodies to sepharose/agarose beads; however, this treatment can lead to partial or complete inactivation of the antibody (Harlow

& Lane, 1999). Regardless of these drawbacks, an advantage of the purification of endogenous proteins is that native complexes can be isolated without the possibly detrimental effects of an affinity tag, or overexpression (see below).

To overcome problems inherent to immunoprecipitation with a specific antibody directed against a protein of interest, a common strategy has been to express epitope-tagged recombinant proteins in cultured cells. Many tagging systems have been described (Jarvik & Telmer, 1998; Fritze & Anderson, 2000; Bauer & Kuster, 2003). This experimental design provides for much better controls, in that untransfected cells (or cells transfected with the tag alone) may be processed in parallel, and the purification scheme can be standardized. In addition, many of the epitope tags may be gently eluted from affinity resins via incubation with short peptides or other small molecules, which can reduce the number of non-specific background contaminants (although elution with a peptide can prevent direct analysis of the sample by MS). In general, however, single-step purification strategies can result in the identification of a significant number of contaminants (see below) making it difficult to distinguish specific from non-specific interactions.

### Tandem affinity purification strategies

The introduction of a dual purification strategy, termed TAP-tagging (tandem affinity purification; see Fig. 2), represented a major improvement in sample purification for MS (Rigaut *et al.* 1999). As with the epitope-tagging strategy, a protein of interest is fused in-frame with an N- or C-terminal tag. In this case, however, the tag is comprised of two affinity components surrounding a protease cleavage site. In the original TAP-tagging system, the distal tag is derived from the IgG binding moiety of *S. aureus* protein A (Protein A), and the proximal tag is a calmodulin-binding peptide; the tags are separated by a tobacco etch virus (TEV) protease cleavage site (Fig. 2A). A TAP-tagged protein of interest is thus expressed in a relevant cell type, and the tagged protein and its binding partners may be recovered via a tandem affinity purification protocol. The tagged protein is first immobilized on IgG-sepharose (via the Protein A moiety), and the bound proteins, along with their binding partners, are gently washed. Immobilized protein complexes are then incubated with TEV protease to specifically release the protein of interest, along with any binding partners. Proteins which bind non-specifically to the resin are left behind. In a second step, the calmodulin-binding peptide is bound to calmodulin-sepharose in the presence of calcium. Following washing, the recombinant protein and its binding partners are then specifically released via calcium chelation. Again, proteins which interact non-specifically with the support matrix are left behind (Fig. 2B).

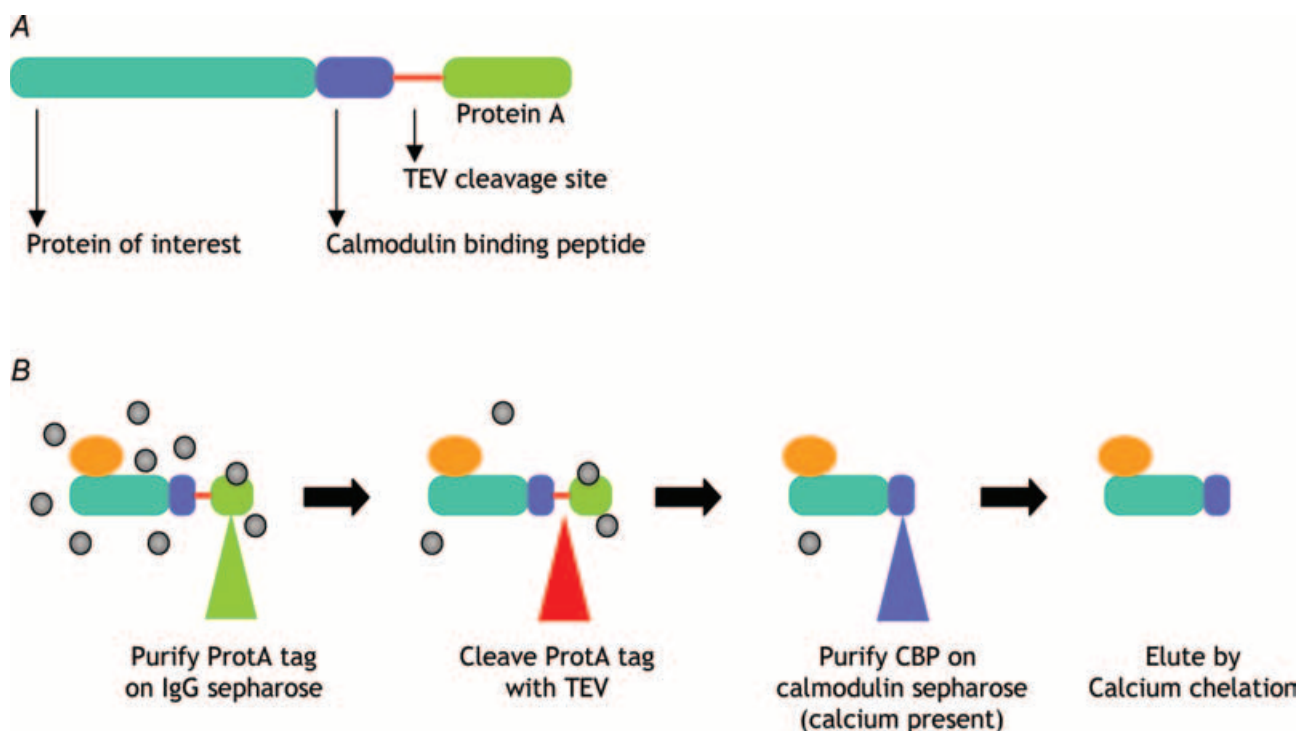
Due to the high degree of specificity conferred by the tandem purification, stringent washes (e.g. using high salt or detergent concentrations) are not necessary, thereby better preserving less stable multiprotein complexes. The number and quantity of contaminating proteins is also low. Finally, since a single purification strategy is utilized for all proteins of interest, the same population of background proteins is observed across purifications, making it relatively straightforward to generate a list of 'likely contaminants', which may easily be subtracted from the data. In rare cases, a background contaminant for one protein may be a true interacting partner for another, and more stringent statistical analyses are necessary to discriminate between these possibilities.

The strength of the dual purification strategy – as compared with a single purification step with either tag alone – was demonstrated in the original publication (Rigaut *et al.* 1999). A C-terminally tagged yeast U1 snRNP subunit (Snu71) was expressed under its endogenous promoter, and the purified complexes yielded 11 protein bands by Coomassie staining. The bands included all known U1 snRNP-specific proteins, as well as some Sm proteins and a novel splicing factor, Snu30p. While the bands corresponding to these components were often visible in the single-tag purification, they frequently co-migrated with other non-specific proteins, making the identification of *bona fide* interactors much more difficult.

### Large-scale TAP tag experiments

While the TAP-tagging technique was first developed in *S. cerevisiae*, a number of tag combinations in other organisms have also been used successfully, including *S. pombe* (e.g. Gould *et al.* 2004), plants (e.g. Rohila *et al.* 2004) and mammals (see below). One of the strengths of the TAP technique is that it provides a generic purification method, enabling parallel characterization of multiple complexes, with minimal optimization required across samples. The TAP-tagging technique is thus ideal for the characterization of protein interaction networks.

Using homologous recombination, Gavin *et al.* (2002) introduced a TAP-tag cassette at the C-terminus of > 1500 yeast open reading frames (ORFs), and successfully purified 589 tagged proteins (corresponding to ~10% of all yeast ORFs). From these purifications, a map encompassing 1440 distinct gene products, or about 25% of all yeast ORFs, was generated. The 589 purifications were grouped into a reduced number (232) of biologically meaningful complexes, based on substantial overlap. Besides the sheer quantity of information generated, what is impressive about this method is the high quality of the data. Several large datasets are available for *S. cerevisiae*, allowing for a comparison of error rates associated with various high-throughput methods for identifying protein–protein interactions: two global yeast two hybrid screens (Uetz *et al.* 2000; Ito *et al.* 2001), a large-scale



**Figure 2. Tandem affinity purification (TAP) strategy**

A, structure of a recombinant C-terminally tagged fusion protein. B, isolation procedure. The orange oval represents a *bona fide* interactor, the grey circles represent contaminants.

single epitope-tag (flag) purification of overexpressed yeast proteins (Ho *et al.* 2002), and the Gavin TAP-tag study (Gavin *et al.* 2002). Whereas the error rate of the TAP-tag method was estimated at ~15%, the single epitope-tag method was ~50% and the two hybrid studies were rated at 45–80% (Dziembowski & Seraphin, 2004). It should be mentioned, however, that only the TAP-tag experiment used proteins expressed under their endogenous promoters, and that a contributing factor to the higher accuracy of this study may be related to differences in expression levels.

### Mammalian interaction networks

While homologous recombination in some species allows for rapid ORF tagging, and expression of the tagged proteins under the control of their own promoters, this approach is not yet feasible in a high-throughput mode in mammalian systems. Instead, strategies involving expression of a recombinant cDNA harbouring the protein of interest fused in-frame with a TAP-tag have been developed. Several tag combinations and vector backbones have been generated (e.g. Gavin *et al.* 2002; Knuesel *et al.* 2003; Bertwistle *et al.* 2004; Bouwmeester *et al.* 2004; Brajenovic *et al.* 2004; Jeronimo *et al.* 2004); our own vectors and protocols are described at [www.proteomecentre.org](http://www.proteomecentre.org). What is clear, regardless of the approach, is that the expression level of the tagged protein is a critical determinant in the success of the experiment. In our hands, TAP purification of various fusion proteins from transiently transfected cells (which overexpress high amounts of the recombinant protein) resulted in the isolation of large amounts of heat shock proteins and chaperones, presumably interacting with overexpressed and misfolded molecules. In addition, the stoichiometry of overexpressed proteins *versus* their binding partners was dramatically altered, making the identification of less abundant binding partners more difficult. In contrast, stable transfectants, in which recombinant proteins are expressed at much more moderate levels, generated samples from which legitimate binding partners were readily identified (Gingras *et al.* unpublished observations). Alternative approaches to stable expression have included using weaker or inducible promoters and/or virus-mediated transfer: these strategies are useful when working with difficult-to-express or toxic proteins. Another point to consider in these types of studies is that the endogenous protein is also present in the cells, competing for binding partners with the TAP-tagged fusion protein, and possibly reducing the efficiency of recovery of the tagged molecule. To circumvent this problem, Forler *et al.* (2003) established a system in *Drosophila* cells whereby RNA interference is utilized to silence the endogenous gene. At the same time, a TAP-tagged mammalian protein is expressed in the cells: because of differences in the mRNA sequence,

the mammalian protein is resistant to RNA interference. Although this approach reduces the problems associated with the presence of the endogenous proteins, differences between the insect and mammalian proteins may confound binding partner identification. In addition, the expression level of the TAP-tagged protein is not necessarily at endogenous levels.

While not approaching the scale of the yeast data, the TAP-tagging technique has also been successfully applied in mammalian cells to the study of several smaller protein networks. For example, a network surrounding the PAR genes, a family of proteins involved in cell polarity, was constructed in human cells (Brajenovic *et al.* 2004) and contains 60 interacting partners built around a core of 9 'bait' proteins. In a more recent and much larger example, a map of the TNF- $\alpha$ /NF- $\kappa$ B pathway, centred around 32 pathway components, was found to encompass 131 high-confidence interacting partners (Bouwmeester *et al.* 2004). This study also introduced the TAP-tagging technique as a useful tool to detect interactions modulated by a given stimulus. The authors subjected their TAP-expressing cell lines to TNF- $\alpha$  stimulation to identify a number of interactions which are modulated upon TNF- $\alpha$  treatment. This type of approach offers the exciting possibility of addressing conditional or dynamic interactions using the TAP technique.

The work on TAP-tagging in mammalian cells has been performed primarily using cell lines. However, primary cells may also be infected (or transfected in some cases) with TAP plasmids. Another exciting possibility is to knockin TAP-tags in mice: in this respect, a method was recently described for the rapid TAP-tagging of endogenous mouse genes. The speed of the technique arises from the use of recombineering and gap-repair rescue, which allow for the generation of mature transgenic mice within 3 months (Zhou *et al.* 2004). Such methods are compatible with moderately high throughput techniques for TAP-tagging many genes, and open the way to the study of tissue-specific or cell type-specific complexes.

Despite its strengths, the TAP-tagging method is not appropriate in all cases. The presence of the rather large tag (the original TAP-tag is ~20 kDa) can negatively affect protein function and/or binding partner interactions. In fact, Gavin *et al.* (2002) found that C-terminal TAP-tagging of essential yeast proteins yielded ~18% non-viable strains, suggesting that the tag at this position interferes with protein function. Several strategies may be employed to overcome this difficulty; for example, tagging the other end of the molecule, using a smaller tag, or tagging a different component of the complex of interest.

### Quantitative mass spectrometry techniques

There are clearly cases where expression of a recombinant protein is not desired, or possible. For example, one may wish to assess protein interactions in a tissue

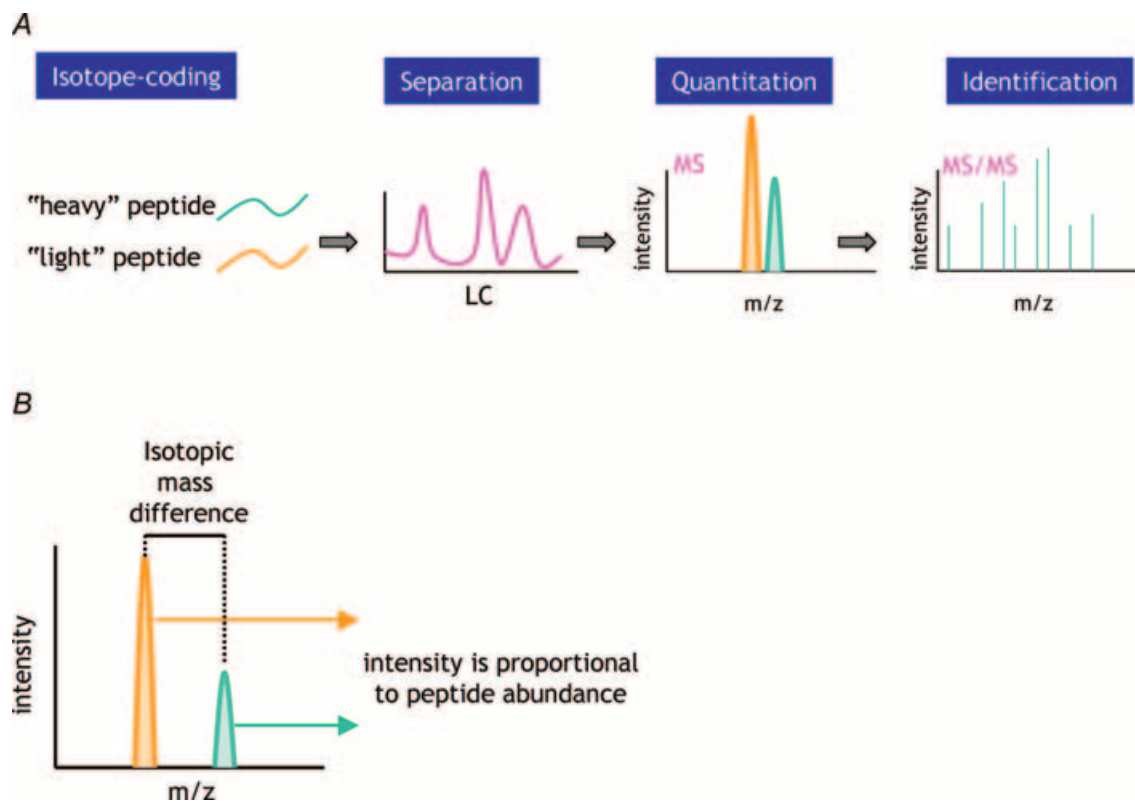
biopsy. In addition, protein complex assembly on a non-protein moiety, such as chemicals, metabolites or nucleic acids, cannot be addressed by TAP-tagging. In the following section we describe a general strategy to identify proteins associated with specific DNA sequences, based on quantitative proteomics. These techniques may also be applied to a variety of affinity purification approaches; several examples are presented.

In any complex sample analysed by MS/MS, the absence of evidence for a given protein does not indicate an absence of the protein from the sample. Due to limitations in peptide sampling in the mass spectrometer, the absence of a protein in the list of hits in a negative control sample is not sufficient proof that a protein identified in the experimental sample was isolated in a specific fashion. Even if one were to analyse the same complex sample via MS twice, the overlap between the peptides sequenced is never 100%, as long as the number of peptides present in the sample exceeds the number of sequencing cycles available during MS/MS analysis: the apparent absence of a peptide/protein in the negative control may be due solely to peptide undersampling.

To circumvent this problem, and to gain confidence in the specificity of the identified proteins, quantitative

proteomics techniques can be very useful. As opposed to the qualitative MS methods described above, quantitative proteomics allow for the determination of both the identity and relative quantity of particular components across different samples. Several methods for quantitative proteomics have been described and reviewed elsewhere (Goshe & Smith, 2003; Sechi & Oda, 2003). For the purposes of this discussion, we only consider stable isotope coding of proteins/peptides. Stable isotopes are ideal for use in quantitative proteomics because 'light' and 'heavy' isotopes generally exhibit identical chemical properties (i.e. they behave identically throughout any peptide purification steps and in the mass spectrometer), yet they possess a mass difference which is easily observed in the mass spectrometer (Fig. 3B). Commonly used isotopes include  $^2\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$  and  $^{18}\text{O}$ . Since, for the most part, peptides containing different isotopes of the same element co-elute in RPLC, it is possible to mix samples labelled with 'heavy' and 'light' isotope tags, and to process them together. Comparison of the relative intensities of the heavy- versus light-labelled peptides in the MS provides quantification.

Several strategies have been developed to harness stable isotopes for quantitative proteomics (reviewed in Flory



**Figure 3. Use of isotopes in quantitative proteomics**

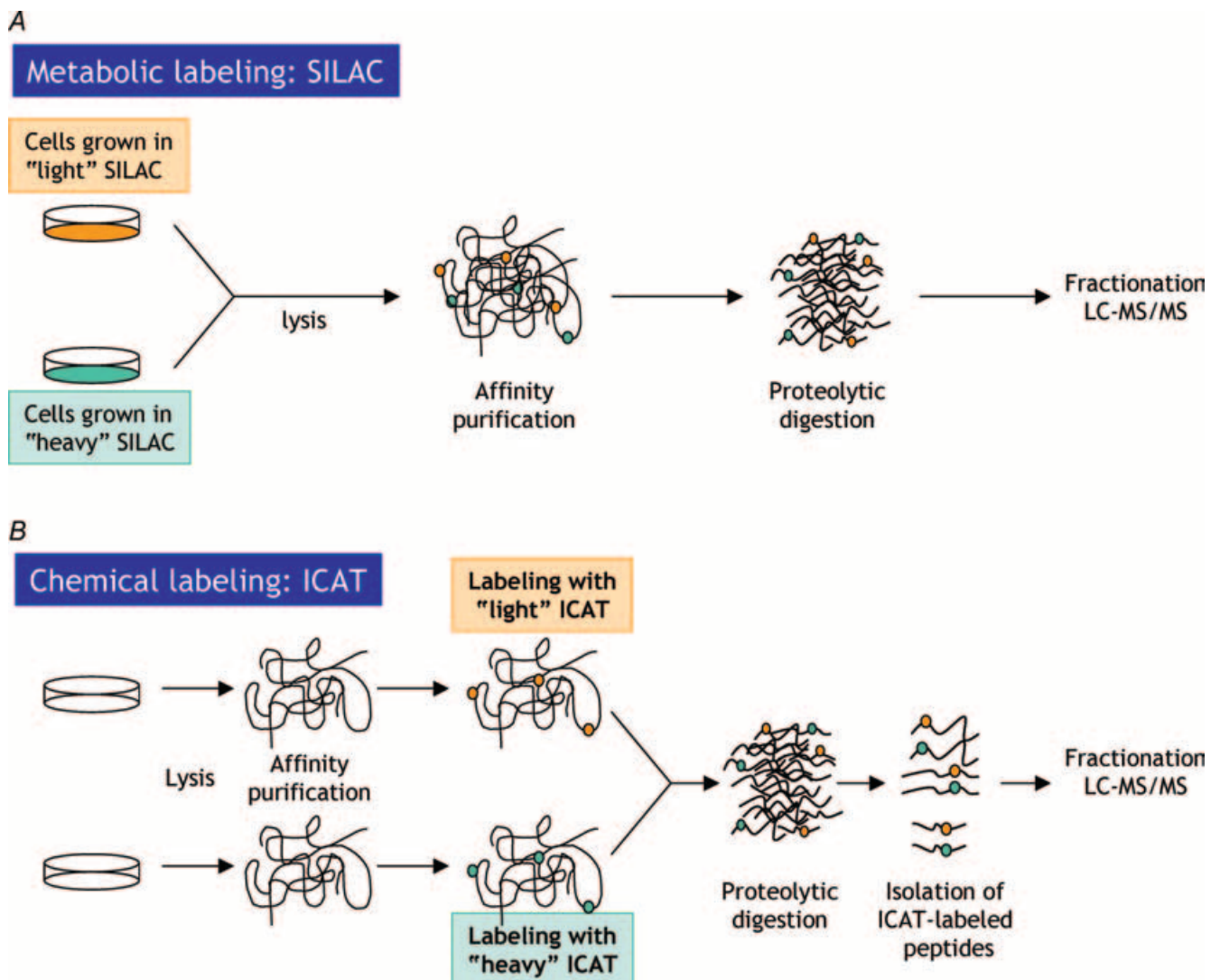
*A*, peptides containing 'heavy' or 'light' isotopes are separated by RPLC: isotopes co-elute. In the survey (MS) scan, peptides are analysed in the  $m/z$  dimension; either isotopic variant can then be selected for fragmentation and analysis by LC-MS/MS to obtain sequence identification. *B*, the isotopic mass difference allows for separate abundance measurements of the two different peptides; the relative intensity of the peaks is proportional to peptide abundance.



*et al.* 2002; Goshe & Smith, 2003; Ong *et al.* 2003). One such strategy, SILAC (stable isotope-labelled amino acids in cell culture), involves metabolic incorporation of isotopically heavy amino acids into proteins (Fig. 4A; Ong *et al.* 2002). Essentially, two populations of cells are grown in the same type of culture medium, except that in one set, one or more essential amino acids are replaced by a version containing heavy atoms, such as  $^{13}\text{C}$ . With each cell doubling, a proportion of all proteins is labelled with the heavy amino acid. After several doublings, almost all proteins are saturated with heavy amino acids. Cells may then be subjected to a desired stimulus or treatment, after which the 'heavy' and 'light' samples are combined, and all subsequent steps are performed on a pooled lysate. Purified proteins of interest are submitted to proteolytic

digestion and LC-MS/MS. The first MS measurement (the survey or parent ion scan) is used to quantify the relative abundances of the heavy *versus* light versions of each peptide (Fig. 3A). If heavy arginine and/or lysine are used in the labelling procedure, only a single amino acid (on average) is labelled in each peptide (since trypsin cleaves C-terminally to these residues). This makes the analysis relatively straightforward, as the 'heavy' peptides will display a mass difference caused by a single incorporated heavy amino acid. The relative intensities of the heavy *versus* light peptides serve to establish the quantification. Peptides are then subjected to MS/MS for identification.

As opposed to the *in vivo* labelling technique described above, post-lysis *in vitro* labelling may also be performed. One such *in vitro* quantitative method involves the



**Figure 4. Alternative methods for isotopic labelling of peptides**

*A*, metabolic labelling via SILAC. Coloured circles represent heavy or light amino acids incorporated into proteins. *B*, chemical labelling via ICAT. Coloured circles represent heavy or light isotope-coded affinity tags covalently bound to cysteine (or other reactive) groups in proteins.

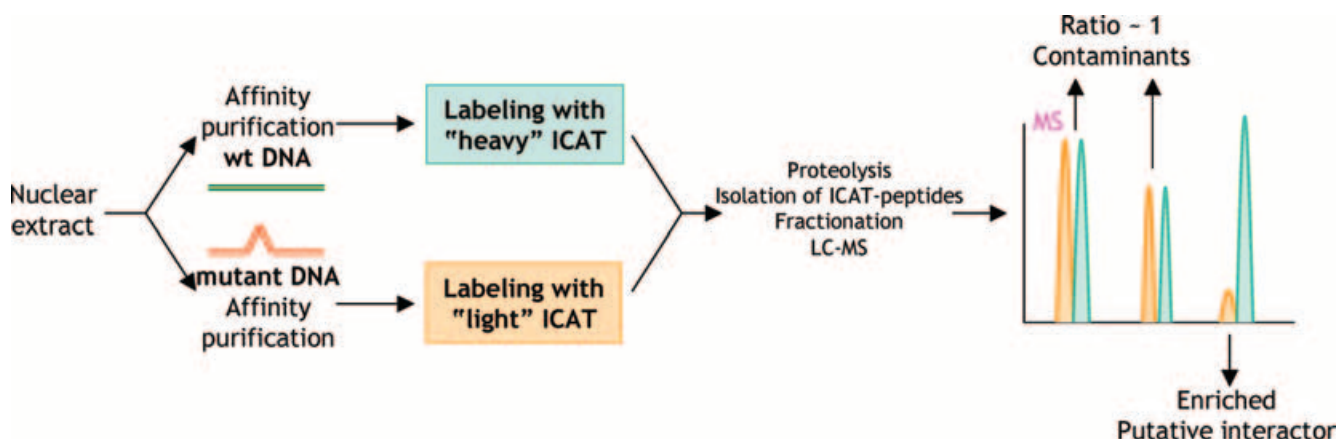
chemical attachment of isotopic tags to proteins or peptides in solution, a strategy referred to as isotope-coded affinity tagging (ICAT; Gygi *et al.* 1999). While the nature of the ICAT tag may vary, these reagents are generally composed of three moieties: a reactive group (used to covalently attach the tag to peptides possessing a specific chemical feature), a linker group (containing 'heavy' or 'light' isotopes), and an affinity handle (such as biotin, used for the purification of the tagged peptides). The general labelling procedure is described in Fig. 4B. An advantage of ICAT over other isotope-labelling systems is that it may also be used to dramatically simplify complex peptide mixtures. For example, the original ICAT is cysteine-reactive; thus, only cysteine-containing peptides react with the reagent. ICAT-labelled peptides are separated from unlabelled peptides via a biotin affinity handle, resulting in a > 10-fold simplification in the mixture. The immediate result of such a simplification is that peptides of relatively lower abundance may be detected, making the ICAT approach ideally suited for the study of complexes of medium to high complexity. As with the SILAC approach, isolated peptides are subjected to LC-MS/MS, such that relative quantification as well as peptide identification are obtained.

#### Quantitative approaches to DNA-protein interactions

Identification of specific binding complexes on a given DNA sequence is a particularly difficult task. In addition to the challenging background issues inherent in characterizing all protein-protein interactions, the binding of sequence non-specific DNA-binding proteins and other positively charged proteins to a DNA sequence of interest largely prohibits the use of a single-step DNA-affinity isolation protocol in

protein identification strategies. However, since sequence non-specific DNA-binding proteins presumably have similar affinity for wild type (WT) and mutant DNA sequences, a simple strategy was devised to discriminate between sequence-specific interactions and contaminants (Fig. 5).

In one such application, Himeda *et al.* (2004) used a DNA-affinity-based approach to identify a binding factor for the transcriptional regulatory element (Trex) in the muscle creatine kinase enhancer. Double-stranded DNA oligonucleotides harbouring either a WT or mutant Trex sequence were coupled to magnetic beads. DNA binding proteins from a HeLa nuclear extract were then purified with either WT or mutant Trex oligos, and recovered proteins were labelled with the heavy (WT) or light (mutant) ICAT reagents. Samples were combined, proteolysed, separated into multiple fractions by strong cation exchange (SCX) and avidin affinity chromatography, and analysed by LC-MS/MS. The relative abundances of the heavy- versus light-labelled peptides were measured in the MS survey scan (MS mode, no peptide fragmentation). Whereas non-sequence-specific DNA-binding proteins are expected to be in roughly the same abundance in the WT and mutant samples, the proteins that associate specifically with the Trex element are expected to be increased in the WT DNA sequence sample. Of 893 proteins (1904 peptides) identified, only 3 displayed abundance ratios > 2-fold in the WT versus mutant samples. One of these proteins, Six4 (1 cysteine-containing peptide detected; 2.4-fold enriched), is a homeodomain transcription factor, and was demonstrated to be a *bona fide* Trex-binding factor by subsequent gel shift and transactivation assays. This was the first example – but certainly not the last – of the use of quantitative proteomics techniques for the identification



**Figure 5. Using quantitative proteomics to identify site-specific DNA-binding proteins**

Parallel purifications using WT or mutant DNA sequences are performed; one of the samples is labelled with 'heavy' ICAT, the other with 'light' ICAT. Samples are processed as above, and analysed via MS. The heavy/light ratios are utilized to identify sequence-specific DNA-binding proteins (or *bona fide* interacting partners; ratio > 1) versus sequence non-specific DNA-binding contaminants (ratio ~1).



of transcription factor complexes interacting with a DNA sequence of interest from mammalian cells.

Using a related strategy, Ranish *et al.* (2003) purified the RNA polymerase II (Pol II) pre-initiation complex, which assembles on Pol II promoters in a TATA-binding protein (TBP)-dependent manner. Yeast nuclear extract lacking functional TBP (the negative control) was produced from a strain harbouring a temperature-sensitive version of TBP. Half of the TBP-deficient nuclear extract was supplemented with recombinant TBP (+ rTBP). These two extract preparations were then subjected to parallel purifications with an immobilized promoter DNA molecule. Proteins isolated from the rTBP-supplemented pool were labelled with the 'heavy' ICAT reagent, while proteins isolated from the negative control sample were labelled with the 'light' ICAT reagent. The two samples were then combined and processed together, as described above. MS analysis indicated that these samples were highly complex: 326 proteins were identified, and 206 proteins were successfully quantified. A majority of the identified proteins appeared to interact with DNA in a sequence non-specific fashion. However, many previously identified core Pol II factors could be distinguished from the background: 45 of the 49 proteins whose abundance ratios were > 1.9 represented previously identified Pol II core proteins. Interestingly, one of the three unknown proteins whose abundance ratio was > 1.9 was found to be a novel, tenth subunit of the TFIID complex, termed TFB5 (Ranish *et al.* 2004). In addition to its role in RNA Pol I and Pol II transcription, TFIID has been implicated in DNA damage repair in yeast and humans. Mutations in human TFIID subunits are associated with DNA repair-deficient trichothiodystrophy (TTD), a rare photo-hypersensitive syndrome. In a subset of TTD patients (TTD-A), none of the nine previously known subunits of TFIID was mutated, yet a highly purified WT TFIID complex could correct *in vitro* the DNA repair defect associated with the syndrome. Expression of TFB5 in mutant TTD-A lines also corrected the defect *in vivo*, and inactivating mutations in TFB5 were discovered in three unrelated families with TTD-A (Giglia-Mari *et al.* 2004). This study thus indicated that large DNA binding complexes can also be successfully analysed via quantitative proteomics methods.

### Quantitative approaches for dynamic interactions

Several novel and interesting quantitative proteomics approaches have been utilized to characterize the dynamics of complex assembly. Blagoev *et al.* (2003) applied the SILAC strategy to the analysis of proteins that differentially associate with the SH2 domain of the Grb2 adapter protein in an epidermal growth factor (EGF)-dependent manner. Unstimulated cells were maintained in  $^{12}\text{C}$ -Arg, while another population was grown in  $^{13}\text{C}$ -Arg; the latter population was stimulated with EGF for 10 min

immediately prior to harvest and lysis. Lysates were combined and purified on an affinity matrix harbouring the Grb2 SH2 domain. Eluted proteins were digested with trypsin, and analysed by LC-MS/MS: 228 proteins were identified, 28 of which were enriched following EGF stimulation. Enriched proteins included known signalling molecules, as well as proteins involved in signal attenuation and cytoskeletal functions. Importantly, several novel proteins were identified which had not previously been reported to be involved in EGF signalling, highlighting the power of this technique for studying signalling pathways.

ICAT approaches may also be used for the study of dynamics of complex formation. For example, Brand *et al.* (2004) characterized changes in the protein complexes associating with the MafK transcription factor (involved in  $\beta$ -globin transcription) upon erythroid differentiation. Two populations of proerythroblast MEL cells were prepared: undifferentiated *versus* differentiated (by exposure to DMSO for 4 days). Immunoprecipitation on immobilized anti-MafK resin was performed in parallel. Eluates were labelled with the heavy (differentiated) or light (undifferentiated) cysteine-specific ICAT reagents, and samples were digested and analysed by LC-MS/MS, as above. Interestingly, the population of MafK-binding partners differed significantly before and after differentiation, consistent with a novel role for MafK as a dual function molecule, which has the ability to switch from a repressed to an activated state.

Recent and exciting developments in quantitative proteomics allow for multiplex experiments (i.e. more than two samples quantified simultaneously) to be conducted. The Mann group, using three isotopic forms of a single essential amino acid (with mass differences of +6 and +10 Da), investigated the early events of EGFR signalling in HeLa cells (Blagoev *et al.* 2004). Cells were maintained in three different pools of culture medium (containing  $^{12}\text{C}_6^{14}\text{N}_4$ -Arg,  $^{13}\text{C}_6^{14}\text{N}_4$ -Arg, or  $^{13}\text{C}_6^{15}\text{N}_4$ -Arg), and exposed to EGF prior to lysis. Proteins phosphorylated on tyrosine residues were precipitated with an anti-phosphotyrosine antibody, and samples were digested and analysed by LC-MS. A time course of EGF stimulation consisting of five time points was obtained by performing the stimulation/labelling twice (once with 0, 1 and 10 min of stimulation; and once with 0, 5 and 20 min), and using the 0 time point as a common reference. Most proteins previously known to be involved in EGF signalling were detected and quantified. In addition, several proteins not previously reported to be involved in EGFR signalling were identified.

Multiplexing experiments using chemical labelling is now also possible: a set of four amino group-reactive reagents was recently introduced (Ross *et al.* 2004). These reagents are isobaric, meaning that a mass difference is not

observed in the mass spectrometer survey scan. Instead, the quantification occurs at the MS/MS level. The first report using these reagents described the expression profiling of three yeast strains (one WT and two different mutants), but the technique should be easily adapted to the study of protein–protein interactions.

## Conclusions

As described above, multiple strategies have been developed to more efficiently and accurately characterize protein complexes using the mass spectrometer. By dramatically reducing background levels during sample purification, the TAP-tagging technique has allowed for standardized processing and identification of high confidence protein–protein interactions, and assembly of protein interaction networks. By discriminating between true interactors and noise, quantitative proteomics techniques have been successfully applied to the study of protein–protein and DNA–protein interactions. In addition, quantitative proteomic approaches are extremely versatile, and can allow the study of RNA–protein, chemical–protein (e.g. Oda *et al.* 2003), or metabolite–protein interactions. In addition to defining static interaction networks, both types of approaches are compatible with the analysis of the dynamics of protein complex formation.

The abundance of interaction data is also driving bioinformatics advances, as it is necessary to integrate, display and interpret information from different sources. A standard representation format for protein interaction data has been proposed (Hermjakob *et al.* 2004), and has been largely adopted by interaction database providers (a list of current repositories can be found at [www.hgmp.mrc.ac.uk/GenomeWeb/prot-interaction.html](http://www.hgmp.mrc.ac.uk/GenomeWeb/prot-interaction.html)). Visualization software, such as Cytoscape ([www.cytoscape.org](http://www.cytoscape.org)) or Osprey ([biodata.mshri.on.ca/osprey/servlet/Index](http://biodata.mshri.on.ca/osprey/servlet/Index)) can be used to display and explore interaction networks. Finally, methods are being developed to define the organization of interaction networks, through decomposition into functional modules (e.g. Gagneur *et al.* 2004). Software must also be adapted to record, analyse and visualize dynamic interactions, and integrate information concerning direct/indirect interactions.

It is likely that in the years to come we will see more large-scale efforts to chart interactomes from different species and tissues. Superimposition of other information (such as structural analysis and binary interactions) on the interaction networks will allow for a better understanding of how complexes are assembled. In addition, analysis of post-translational modifications will need to be combined with interaction data, paving the way for a molecular understanding of the dynamics of complex assembly.

## References

- Bauer A & Kuster B (2003). Affinity purification-mass spectrometry. Powerful tools for the characterization of protein complexes. *Eur J Biochem* **270**, 570–578.
- Bertwistle D, Sugimoto M & Sherr CJ (2004). Physical and functional interactions of the Arf tumor suppressor protein with nucleophosmin/B23. *Mol Cell Biol* **24**, 985–996.
- Blagoev B, Kratchmarova I, Ong SE, Nielsen M, Foster LJ & Mann M (2003). A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling. *Nat Biotechnol* **21**, 315–318.
- Blagoev B, Ong SE, Kratchmarova I & Mann M (2004). Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat Biotechnol* **22**, 1139–1145.
- Bouwmeester T, Bauch A, Ruffner H, Angrand PO, Bergamini G, Coughton K *et al.* (2004). A physical and functional map of the human TNF- $\alpha$ /NF- $\kappa$ B signal transduction pathway. *Nat Cell Biol* **6**, 97–105.
- Brajenovic M, Joberty G, Kuster B, Bouwmeester T & Drewes G (2004). Comprehensive proteomic analysis of human Par protein complexes reveals an interconnected protein network. *J Biol Chem* **279**, 12804–12811.
- Brand M, Ranish JA, Kummer NT, Hamilton J, Igarashi K, Francastel C, Chi TH, Crabtree GR, Aebersold R & Groudine M (2004). Dynamic changes in transcription factor complexes during erythroid differentiation revealed by quantitative proteomics. *Nat Struct Mol Biol* **11**, 73–80.
- Dziembowski A & Seraphin B (2004). Recent developments in the analysis of protein complexes. *FEBS Lett* **556**, 1–6.
- Eng J, Martin D & Aebersold R (2005). Tandem mass spectrometry database searching. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*, ed. Dunn M, Jorde L, Little P, Subramaniam S, (in press). John Wiley & Sons Ltd, Chichester, UK.
- Flory MR, Griffin TJ, Martin D & Aebersold R (2002). Advances in quantitative proteomics using stable isotope tags. *Trends Biotechnol* **20**, S23–29.
- Forler D, Kocher T, Rode M, Gentzel M, Izaurralde E & Wilm M (2003). An efficient protein complex purification method for functional proteomics in higher eukaryotes. *Nat Biotechnol* **21**, 89–92.
- Fritze CE & Anderson TR (2000). Epitope tagging: general method for tracking recombinant proteins. *Meth Enzymol* **327**, 3–16.
- Gagneur J, Krause R, Bouwmeester T & Casari G (2004). Modular decomposition of protein–protein interaction networks. *Genome Biol* **5**, R57.
- Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A *et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147.
- Giglia-Mari G, Coin F, Ranish JA, Hoogstraten D, Theil A, Wijgers N *et al.* (2004). A new, tenth subunit of TFIIH is responsible for the DNA repair syndrome trichothiodystrophy group A. *Nat Genet* **36**, 714–719.
- Goshe MB & Smith RD (2003). Stable isotope-coded proteomic mass spectrometry. *Curr Opin Biotechnol* **14**, 101–109.

- Gould KL, Ren L, Feoktistova AS, Jennings JL & Link AJ (2004). Tandem affinity purification and identification of protein complex components. *Methods* **33**, 239–244.
- Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH & Aebersold R (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**, 994–999.
- Harlow E & Lane D (1999). *Using Antibodies: a Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A *et al.* (2004). The HUPPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. *Nat Biotechnol* **22**, 177–183.
- Himeda CL, Ranish JA, Angello JC, Maire P, Aebersold R & Hauschka SD (2004). Quantitative proteomic identification of six4 as the trex-binding factor in the muscle creatine kinase enhancer. *Mol Cell Biol* **24**, 2132–2143.
- Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL *et al.* (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183.
- Itou T, Chiba T, Ozawa R, Yoshida M, Hattori M & Sakaki Y (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569–4574.
- Jarvik JW & Telmer CA (1998). Epitope tagging. *Annu Rev Genet* **32**, 601–618.
- Jeronimo C, Langelier MF, Zeghouf M, Cojocaru M, Bergeron D, Baali D *et al.* (2004). RPAP1, a novel human RNA polymerase II-associated protein affinity purified with recombinant wild-type and mutated polymerase subunits. *Mol Cell Biol* **24**, 7043–7058.
- Knuesel M, Wan Y, Xiao Z, Holinger E, Lowe N, Wang W & Liu X (2003). Identification of novel protein–protein interactions using a versatile mammalian tandem affinity purification expression system. *Mol Cell Proteomics* **2**, 1225–1233.
- Nesvizhskii AI & Aebersold R (2004). Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov Today* **9**, 173–181.
- Oda Y, Owa T, Sato T, Boucher B, Daniels S, Yamanaka H *et al.* (2003). Quantitative chemical proteomics for identifying candidate drug targets. *Anal Chem* **75**, 2159–2165.
- Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A & Mann M (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**, 376–386.
- Ong SE, Foster LJ & Mann M (2003). Mass spectrometric-based approaches in quantitative proteomics. *Methods* **29**, 124–130.
- Ranish JA, Hahn S, Lu Y, Yi EC, Li XJ, Eng J & Aebersold R (2004). Identification of TFB5, a new component of general transcription and DNA repair factor IIIH. *Nat Genet* **36**, 707–713.
- Ranish JA, Yi EC, Leslie DM, Purvine SO, Goodlett DR, Eng J & Aebersold R (2003). The study of macromolecular complexes by quantitative proteomics. *Nat Genet* **33**, 349–355.
- 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.
- Rohila JS, Chen M, Cerny R & Fromm ME (2004). Improved tandem affinity purification tag and methods for isolation of protein heterocomplexes from plants. *Plant J* **38**, 172–181.
- Ross PL, Huang YN, Marchese J, Williamson B, Parker K, Hattan S *et al.* (2004). Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3**, 1154–1169.
- Sechi S & Oda Y (2003). Quantitative proteomics using mass spectrometry. *Curr Opin Chem Biol* **7**, 70–77.
- Steen H & Mann M (2004). The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* **5**, 699–711.
- Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR *et al.* (2000). A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627.
- Zhou D, Ren JX, Ryan TM, Higgins NP & Townes TM (2004). Rapid tagging of endogenous mouse genes by recombineering and ES cell complementation of tetraploid blastocysts. *Nucleic Acids Res* **32**, e128.

## Acknowledgements

This work is supported in whole or in part by Federal funds from the National Heart, Lung and Blood Institute, National Institutes of Health, under contract No. N01-HV-28179. A.-C.G. is supported by a fellowship from the Canadian Institutes of Health Research (CIHR). We are grateful to Jeff Ranish and Bernd Wollscheid for critical reading of the manuscript.

## Author's present address

R. Aebersold: Institute of Biotechnology, Swiss Federal Institute of Technology, ETH Hönggerberg HPT E 78, CH-8093, Zurich, Switzerland.