Protein-interaction mapping for functional proteomics

Proteomics techniques aimed at identifying protein–protein interactions have been used successfully to characterize multiprotein complexes such as the spliceosome, the nuclear-pore complex and transient complexes in cell signaling. This has led to important biological insights. 'Interaction proteomics' is now ready for the large-scale, unbiased exploration of complexes, cellular structures and pathways.

Ole Vorm*
vorm@protana.com

Angus King*
king@protana.com

Keiryn L. Bennett*
bennett@protana.com

Thomas Leber*
leber@protana.com

Matthias Mann*†
mann@protana.com

*Protana,
Staermosegaardsvej 16, DK
5230 Odense M, Denmark.
http://www.protana.com/

†Protein Interaction
Laboratory (PIL), Department
of Biochemistry and Molecular
Biology, Odense, Denmark.

The definition of 'proteomics' has changed from being largely synonymous with differential two-dimensional (2D) gel analysis[1] to encompassing almost all of the techniques for the large-scale characterization of gene products[2]. Within the biochemical approaches of proteomics, a distinction can be made between 'expression proteomics' and 'interaction proteomics'. In the former, the total protein complement of a tissue, cell homogenate or body fluid is prepared in two physiological states (e.g. normal and diseased), separated by 2D gel electrophoresis and surveyed for differences in protein expression by staining and image analysis. Up- or downregulated proteins can subsequently be identified by mass spectrometry (MS).

Although expression proteomics holds the attraction of monitoring many cellular responses to a changed physiological state in parallel, there are several inherent disadvantages that limit its scope. The dynamic range (the variation in the abundance of proteins in total protein preparations) can exceed 10^5, making it difficult to detect, for example, low-level regulatory proteins by gel electrophoresis and/or MS. Furthermore, 2D-gel-based methods only visualize a fraction of the total protein content. Finally, expression proteomics has increasingly to compete with DNA expression arrays, which will probably relegate expression proteomics to cases that are not easily solved at the mRNA level (i.e. in which there is little correlation between mRNA- and protein-expression levels).

Interaction proteomics addresses a different question: which proteins interact with a 'bait' of interest? In this case, we are often trying to determine protein–protein interactions but the bait can also be a specific oligonucleotide sequence (to capture RNA- or DNA-binding proteins) or a small molecule (to find drug targets). A large number and a wide variety of biological questions have already been addressed using this approach. Reasons for the remarkable success of interaction proteomics include the fact that the dynamic-range problem can be circumvented: because the bait specifically retrieves the interacting proteins out of a protein mixture, the proteins are simultaneously purified and enriched, and the relative abundance of the proteins is not necessarily important. Furthermore, the manner of isolation of the protein itself immediately conveys functional information about the interacting proteins (i.e. members of a multiprotein complex with a defined function). Finally, the complexity of the protein mixture to be analysed in these affinity approaches is not overwhelming and streamlined protocols involving one-dimensional gel separation combined with automated, high-sensitivity MS identification can usually solve the analytical task.

Defining multiprotein complexes by affinity purification and MS

The central idea of this approach is to isolate a protein complex by biochemical means, to separate it into its components, to identify the constituents by MS and database searching, and, finally, to verify the actual role of the components found in the complex[3–5] (Fig. 1). Such a strategy can elucidate the function of novel genes through their interaction partners and it can also be used in an 'unbiased' way to obtain a protein-interaction map of the cell. At Protana, we combine affinity purification with high-throughput MS identification of proteins in order to elucidate the function of disease genes, to discover drug targets and to map protein interactions in pathogens.

The cell can be envisioned as a collection of multiprotein complexes, each with a defined role. To study a particular complex by proteomics, we first need a 'hook' or affinity tag for biochemical purification. There are three generic ways to obtain affinity tags. First, antibodies can be raised against one or more of the components in the complex. These antibodies can then be used to precipitate the complex from a cell extract. Antibody precipitation

 1471-1931/00/$ – see front matter PII: S1471-1931(00)00010-0

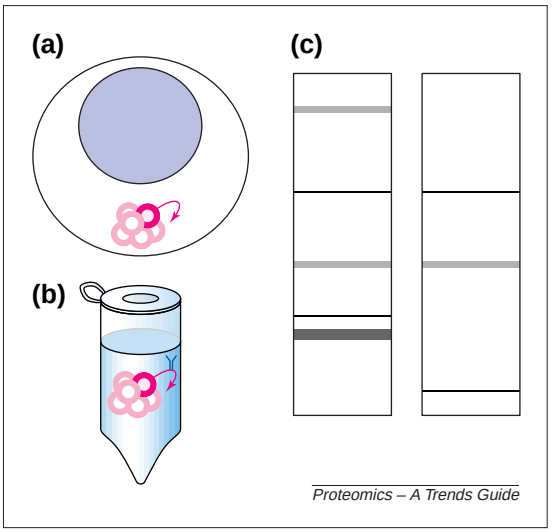

Figure 1. Defining multiprotein complexes by proteomics

(a) Cell with tagged and overexpressed protein (red) to which a short epitope has been fused (the hook). The protein is shown as a part of a multiprotein complex. **(b)** The hook is used to purify the protein complex from cell extract, in this case by antibodies that recognize the epitope. The antibodies are typically bound to beads. **(c)** The bound protein complex is specifically eluted using the properties of the hook, and separated by electrophoresis. Bands that are different between the bait (left lane) and a control (right lane) are excised and identified by mass spectrometry.

captures endogenous levels of the protein complex and is thus the method that is most likely to deliver the *in vivo* state of the complex. However, generating specific antibodies that are suitable for low-background immunoprecipitation is still time consuming. Large libraries of synthetic antibodies (e.g. phage-display antibody libraries) that would allow the selection of strongly binding antibodies against a given target would be an extremely valuable tool for interaction proteomics.

A second method is to fuse the cDNA of a complex member to an epitope tag or protein domain against which a defined antibody exists. This construct can be expressed in target cells and the affinity tag can specifically precipitate the protein together with associated binding partners. This strategy usually, but not necessarily, involves overexpression of the bait, which can be an advantage when the proteins of interest are not expressed under the growth condition being studied.

A third variation on the theme is to express an affinity-tagged version of a member of the complex and to immobilize this construct on beads, followed by incubating the beads with cell extract and retrieving the protein and its binding partners. This strategy is currently the most scalable but great care has to be taken about the design of controls and follow-up experiments to verify the interactions. In addition to these affinity-purification schemes, non-protein baits can also be used, such as oligonucleotides or small molecules that have specific but unknown protein binding partners.

The strategies outlined above can be used as one-step procedures or combined with several conventional biochemical separation methods (e.g. gradient centrifugation and various forms of chromatography that preserve the integrity of the protein complexes). In the immunoprecipitation or 'pull-down' step using, for example, antibodies coupled to magnetic beads, non-specifically binding proteins as well as physiological interactors are bound to the target and the beads. A balance must be found between minimizing the non-specific protein background by stringent washing and, at the same time, preserving weak but specific interactions. A promising advance in this area has been the development of alternative tagging systems that allow specific retrieval with little background. These involve elution by proteolytic cleavage of a specific sequence inserted between the tag and the bait, double tags or a combination of the two[6].

After elution, the components of a protein complex are separated by gel electrophoresis. Usually, the complexity of the mixture is relatively low, which allows it to be displayed on 1D rather than 2D gels. In our experience, mixtures of up to 100 proteins can often be analysed on 1D gels because MS can now easily resolve the identities of comigrating proteins. 1D SDS gels have added advantages over 2D gels because almost all proteins can be visualized, 1D gels are easy to use and 1D gel experiments can easily be scaled up to large numbers.

After staining, usually with silver, proteins are excised from gels either manually or with the help of a robotic spot picker. Subsequent analysis is also either manual or in an automated sample-handling and workflow system. The latter has the advantage of avoiding errors arising from manual tracking of spot, batch and spectrum identity. In our laboratory, protein spots are enzymatically degraded in a streamlined system using 96-well plates in order to accommodate the large numbers of proteins generated by the many pull-down experiments involved in large-scale protein-interaction mapping. The resulting peptide mixtures are screened by automated matrix-assisted laser desorption–ionization (MALDI) MS analysis, revealing the identity of a large proportion of the proteins. Protein bands that are very faint on silver-stained gels, contain complex protein mixtures or are represented only partially in a database are additionally analysed by electrospray tandem MS. Recently, we used a MALDI quadrupole time-of-flight instrument[7] that, when further developed, promises to combine these two steps into one by allowing direct sequencing of selected peptide peaks in a MALDI peptide map.

In-depth function analysis for a multiprotein complex: the spliceosome

The yeast and human spliceosome were the first protein complexes studied by the strategies outlined above. The spliceosome is a dynamic multiprotein complex that assembles on pre-mRNA and excises introns from the primary transcript to produce mature mRNA. This

process is an extremely precise reaction involving many protein and RNA components. Using gradient centrifugation and an antibody as tools to purify the U1 small nuclear RNA subcomplex of the yeast spliceosome, it was possible to visualize 20 protein bands on a 1D gel, leading to the identification of 20 gene products, some in mixtures in the same band and some occurring in multiple forms[3].

All the novel components found in the study were later verified as being bona fide members of the spliceosome[8]. Interestingly, a large-scale two-hybrid experiment that had been performed on the yeast spliceosome at the same time did not reveal these novel components[9]. The two-hybrid system tests for any possible pairwise interaction, and so protein complexes might not always be analysable by this method. This is especially the case if the formation of the complex also depends on other components, such as RNA components in the case of the spliceosome.

Initially, the U1 sub-complex was purified in a time-consuming, highly optimized procedure. Subsequently, improved tagging technology allowed the yeast U1 complex to be re-analysed using a generic double-tag technique with several-times-higher yield and purity[6]. More-detailed MS analysis of the yeast spliceosome also delineated components of the other subcomplexes of the yeast spliceosome[10].

The human spliceosome was purified using a biotinylated and radioactively labeled mRNA as bait. Owing to the large number of proteins, the complex was analysed both by 1D and 2D gel electrophoresis[11]. A total of 19 splicing factors were found that had not been known at the start of the investigation. Interestingly, MS sequence data (peptide sequence tags) could be associated with expressed sequence tags for these novel proteins, providing a straightforward route to obtaining full-length cDNA sequence and clones. In many cases, bioinformatic analysis of the novel proteins allowed putative functions to be assigned because homology and domain-structure information could be combined with the known role of the multiprotein complex.

Nevertheless, owing to the large number of proteins found in investigations like this, a generic way of independently verifying the presence of the protein in the complex is highly desirable. In the case of the spliceosome, cellular co-localization was chosen as a rapid follow up on the novel factors, using fusions with green fluorescent protein and immunocytochemistry. The combination of biochemical co-purification with bioinformatic analysis and an orthogonal set of data, such as co-localization, is a powerful means to obtain functional information about a large number of proteins in a multiprotein complex.

For more-detailed studies of the role of a given splicing factor, the pertinent assays (in this case splicing assays) have to be performed. However, proteomic methods can also be used at this stage. In the spliceosome study, recombinant proteins were tagged as described above and incubated with a HeLa cell extract. Sequencing of interacting proteins helped to reveal the function of the cloned proteins in more detail and led to the discovery of additional novel factors. We have found such iterative analysis of protein complexes to be very informative[12].

Other complexes and pathways successfully characterized by proteomics

A large number of protein complexes have now been analysed by an MS read out of the components of multiprotein complexes. Rout *et al.* purified the yeast nuclear-pore complex and identified several novel components; knowledge of the components of the nuclear-pore complex led to a new model of its organization[13]. The yeast spindle pole body, which organizes chromosome division in meiosis, was similarly analysed by MALDI MS from a complex protein mixture on 1D gels[14]. Proteins were verified by genomic replacement of the found genes with a tagged version that allowed the spindle pole to be seen using video microscopy. The pea chloroplast has recently been analysed by cross-species identification using the *Arabidopsis thaliana* genome[15].

Proteome analysis of the anaphase-promoting complex (APC), a complex involved in regulating the cell cycle, shows how multiprotein complexes can be characterized between model systems. APC was isolated by tagging and immunoprecipitation in yeast, leading to the identification of five novel components[16]. Homology searching with these components revealed cognate human proteins. Antibody precipitation of the human APC and MS identification revealed the vertebrate APC complex, with a previously uncharacterized component[17].

It is also sometimes possible to analyse protein complexes by digesting the unseparated protein mixture and identifying as many as possible of the proteins by MS sequencing of the peptides as they elute from a capillary column[18]. This method has been used to find a novel component in the yeast ribosome[19] and to identify components of the interchromatin-granule cluster[20].

Signal transduction is another field that involves the action of multiprotein complexes. Pandey *et al.* have treated HeLa cells with erythrocyte growth factor (EGF) to activate the EGF-receptor-associated kinase cascade[21]. Precipitation with antiphosphotyrosine antibodies will precipitate both the complex that forms at the cytosolic face of the receptor, and members of the cascade that are further downstream and have no direct contact with the

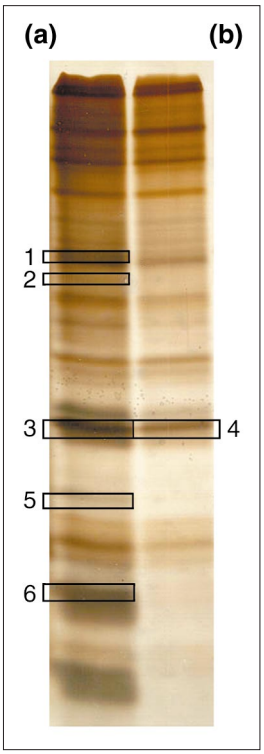

Figure 2. Unbiased exploration of protein complexes

Hypothetical genes in the *Mycobacterium tuberculosis* genome were amplified with PCR, expressed as fusion proteins with a double-tag construct and incubated with lysates of tuberculosis-infected cells. Interacting proteins were separated by one-dimensional gel electrophoresis and differences between the **(a)** experimental and the **(b)** control lanes were identified by a combination of automated matrix-assisted laser desorption–ionization mass spectrometry peptide mapping and nanoelectrospray mass spectrometry. The tagged hypothetical protein interacts with two other hypothetical proteins and two tuberculosis proteins with a known function or a function ascribed by homology [bands 1, 2, 5, 6 (bands 3 and 4 are nonspecifically binding proteins)].

receptor. In the EGF experiment, nine proteins were found to be different between EGF-treated and untreated cells. MS sequencing of these bands revealed seven proteins that were already known to be part of the pathway, one known protein that was placed in the EGF pathway for the first time by this experiment and one novel protein whose role in EGF-receptor signaling is currently under investigation. It appears that these and similar strategies will be generally applicable to the identification of novel members of a number of signaling pathways.

Large-scale affinity approaches with MS read outs

Most of the examples cited above had a known starting point (i.e. a protein of interest or a protein complex with at least one known component). However, it is also possible to apply the methods described here in an 'unbiased' way. In this case, a subset, or even all of the genes of a genome are expressed as tagged constructs and the methods for multiprotein characterization are performed on all the expressed constructs (Fig. 2). The technology is now at hand to perform such genome-wide interaction experiments in microorganisms such as pathogens and yeast. It is also already possible to screen for interaction partners of a subset of human genes such as disease-related genes. In this case, the analysis of multiprotein complexes can be extended to elucidate the differences between diseased and normal states, and to reveal potential targets for future drug development.

Conclusion

Combining affinity purification of proteins with MS read out of the interacting proteins is a powerful strategy for analysing the functions of genes. Analysis can be targeted highly specifically to a protein complex of interest or it can be performed in a genome-wide fashion by tagging a large collection of genes. The identified co-purifying proteins need to be verified as *in vivo* members of the complex by independent, generic experiments such as co-localization.

As the characterization of protein complexes does not involve the same dynamic-range challenges as whole-proteome analysis, crude peptide mixture analysis and stable-isotope methods[18,22] might be particularly applicable to interaction proteomics. The concept of interaction proteomics can be refined into a multitude of alternative strategies by, for example, using combinations of small molecules or interaction mapping in different cell compartments. As all the component technologies are scaled up, interaction proteomics will contribute even more to our understanding of fundamental cellular processes as well as to the definition of drug targets.

Acknowledgements

We thank our colleagues at Protana for discussions and support. G. Neubauer and J. Rappsilber performed the proteome analysis of the spliceosome in M.M.'s group. PIL's laboratory at the University of Southern Denmark is supported by a generous grant from the Danish National Research Foundation.

References

1 Wilkins, M.R. *et al.* (1997) *Proteome Research: New Frontiers in Functional Genomics*, Springer
2 Pandey, A. and Mann, M. (2000) Proteomics to study genes and genomes. *Nature* 405, 837–846
3 Neubauer, G. *et al.* (1997) Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* 94, 385–390
4 Lamond, A.I. and Mann, M. (1997) Cell biology and the genome projects – a concerted strategy for characterizing multi-protein complexes using mass spectrometry. *Trends Cell Biol.* 7, 139–142
5 Blackstock, W.P. and Weir, M.P. (1999) Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* 17, 121–127
6 Rigaut, G. *et al.* (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* 17, 1030–1032
7 Shevchenko, A. *et al.* (2000) MALDI quadrupole time-of-flight mass spectrometry: a powerful tool for proteomic research. *Anal. Chem.* 72, 2132–2141
8 Gottschalk, A. *et al.* (1998) A comprehensive biochemical and genetic analysis of the yeast U1 snRNP reveals five novel proteins. *RNA* 4, 374–393
9 Fromont-Racine, M. *et al.* (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.* 16, 277–282
10 Gottschalk, A. *et al.* (1999) Identification by mass spectrometry and functional analysis of novel proteins of the yeast [U4/U6.U5] tri-snRNP. *EMBO J.* 18, 4535–4548
11 Neubauer, G. *et al.* (1998) Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat. Genet.* 20, 46–50
12 Shevchenko, A. and Mann, M. (1999) Deciphering functionally important multiprotein complexes by mass spectrometry. In *Mass Spectrometry in Biology and Medicine* (Burlingame, A. *et al.*, eds), pp. 237–269, Humana Press
13 Rout, M.P. *et al.* (2000) The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* 148, 635–651
14 Wigge, P.A. *et al.* (1998) Analysis of the *Saccharomyces* spindle pole by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry. *J. Cell Biol.* 141, 967–977

15 Peltier, J.B. *et al.* (2000) Proteomics of the chloroplast: systematic identification and targeting analysis of lumenal and peripheral thylakoid proteins. *Plant Cell* 12, 319–342

16 Zachariae, W. *et al.* (1996) Identification of subunits of the anaphase-promoting complex of *Saccharomyces cerevisiae*. *Science* 274, 1199–1204

17 Grossberger, R. *et al.* (1999) Characterization of the DOC1/APC10 subunit of the yeast and the human anaphase-promoting complex. *J. Biol. Chem.* 274, 14500–14507

18 Washburn, M.P. and Yates, J.R., III (2000) New methods of proteome analysis: multidimensional chromatography and mass spectrometry. In *Proteomics: A Trends Guide* (Blackstock, W.P. and Mann, M., eds), pp. 27–30, Elsevier

19 Link, A.J. *et al.* (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 17, 676–682

20 Mintz, P.J. *et al.* (1999) Purification and biochemical characterization of interchromatin granule clusters. *EMBO J.* 18, 4308–4320

21 Pandey, A. *et al.* (2000) Analysis of receptor signaling pathways by mass spectrometry: identification of Vav-2 as a substrate of the epidermal and platelet-derived growth factor receptors. *Proc. Natl. Acad. Sci. U. S. A.* 97, 179–184

22 Gygi, S.P. and Aebersold, R. (2000) Using mass spectrometry for quantitative analysis. In *Proteomics: A Trends Guide* (Blackstock, W.P. and Mann, M., eds), pp. 31–36, Elsevier

Protein arrays: a high-throughput solution for proteomics research?

High-density DNA and protein arrays are small flat surfaces that allow the simultaneous analysis of thousands of molecular parameters within a single experiment. DNA array technologies have resulted in smaller sample volumes, more efficient analyses and higher throughput. As proteins are more complex and more diverse compared with nucleic acids, development of similar platforms for proteomics has proved difficult. This review outlines current techniques used in the generation and applications of high-density protein arrays, with emphasis on recent developments and applications in proteomics.

Dolores J. Cahill
cahill@molgen.mpg.de.

Max-Planck-Institute of Molecular Genetics, Ihnestrasse 73, D-14195 Berlin, Germany.

Microarray production is a highly automated process, using either pin-based or microdispensing liquid handling robots to arrange biological samples on a flat surface for multiple analysis. For the generation of DNA arrays, PCR techniques are well established and now play a central role in large-scale genome analysis[1]. The surface used for arraying varies with the application, for example, if large numbers of samples are to be analysed for their interaction with the same ligand, DNA is arrayed onto either filter membranes (e.g. nitrocellulose, nylon, polyvinylidene difluoride) or glass slides coated with various reagents (e.g. poly-L-lysine or polyacrylamide).

Generation of arrays

Similar technology has been used to generate high-density protein arrays[2,3] and micro-arrays[4] (Fig. 1). This method involves using gridding robots that transfer either DNA or the corresponding protein expressed, from microtitre plates onto nylon (Hybond N+, Amersham, for DNA analysis) or polyvinylidene difluoride (Hybond–PVDF, Amersham, for protein analysis) membranes in high-density grids[5]. The spotting robot carries a 384-pin head on a servo-controlled three axis linear drive system which can be positioned with an accuracy of 5 μm and produces densities of approximately 300 spots/cm^2. The tip diameter depends on the spotting application but can vary between 150 and 450 μm (Ref. 4). *In situ* expression of recombinant fusion proteins and expression products is then detected using an antibody against a His-tag-containing epitope. Antibodies binding non-specifically are then washed away and the filters are incubated with the appropriate labelled secondary antibody and substrate. An image is taken of the filter using a charge coupled (CCD) camera, on exposure of the filter to UV. Custom image analysis software is then used to score positive clones[2]. DNA filters, where clones or PCR products are gridded onto nylon membranes, can be re-used at least 20 times, without significant loss of signal intensity. By contrast, protein arrayed on polyvinylidene

 1471-1931/00/$ – see front matter PII: S1471-1931(00)00006-9