doi:10.2533/chimia.2016.860 Chimia 70 (2016) 860–863

Next-generation Proteomics from an Industrial Perspective

Oliver Rinner*

Abstract: Proteomics is the large-scale study of proteins. The field began with protein purification and analysis by various techniques but today is largely understood to be mass spectrometry-based highly multiplexed protein quantification. This article focusses on protein expression profiling, *i.e.* how much of each protein is in my sample or how much does the level of each protein change upon a change in the environment?

Keywords: Mass spectrometry · Proteomics

BIOGNOSYS
PROTEOMICS – NEXT GENERATION

Introduction

The term Proteomics refers to the large-scale study of proteins.[1] The proteome, a blend of the words protein and genome, usually refers to the entire set of proteins, or at least a large sub-set of all proteins produced by an organism or tissue. The proteins present in plasma for instance are referred to as the plasma proteome.

In principle proteomics as an analytical method comprises various technologies for protein purification and their analysis but it is now widely synonymous with mass spectrometry-based highly multiplexed protein quantification and this article will focus exclusively on this technology. Currently, the main application of proteomics is protein expression profiling, *i.e.* answering a simple quantitative question: How much of each protein is in my sample or how much does the level of each protein change upon a change in the environment?

The Significance of Proteomics

If genomics relates to the building plan of an organism, proteomics relates to the implementation of this plan in an actual living system. Proteins constitute most of the functions and the substance of any organism. Large-scale gene expression analysis, called transcriptomics, is in the middle between genomics and proteomics. Transcriptomics is still the most widely used proxy for protein expression and in fact due to its easy availability using microarrays or RNA-Seq, many researchers seem to equate RNA expression with protein expression.

However, in many cases it was found to be a poor proxy for protein expression.[2] The reason is that mRNA is not always translated into protein and protein levels not only depend on the transcript availability but also on protein-intrinsic factors such as degradation, or generally on post-transcriptional regulation. A comprehensive review of the relationship between RNA and protein expression can be found in the recent article of Liu *et al.*[3] In summary, RNA and protein expression are moderately correlated under steady-state conditions but can become completely disconnected after a perturbation because transcription and translation are happening on different time scales. In many cases the research question is purely related to proteins, so for instance if the density of receptor proteins on the cell surface is of interest, or if the proteome of the plasma (where virtually no RNA can be found) is addressed. Consequently, the only accurate way to analyze protein composition in an organism is to measure proteins directly.

A Short History of Proteomics

The most widely used protein separation technique is SDS-PAGE, first reported by Laemmli[4] in 1970. The step from a single-protein analysis technique to a proteomics technique was made when SDS-Page was combined with isoelectric focusing (IEF) to protein samples prior to SDS-PAGE to become two-dimensional (2D) gel electrophoresis.[5] Thanks to its visual appeal and its ability to reveal the proteome in a format familiar to researchers, 2D gel electrophoresis became very popular and until the turn of the century was the most widely used proteomics technique. However, from the beginning it was plagued with reproducibility issues and later it became apparent that it was also limited in sensitivity. Protein expression varies by orders of magnitudes. Compared to LC-MS based methods described below, 2D gel electrophoresis only scratches the surface of the proteome. For some applications, however, 2D gel electrophoresis remains the method of choice. In contrast to LC-MS it reveals intact proteins, which for instance is an advantage for the study of protein modifications.

Modern Proteomics – LC-MS/MS

Almost all modern mass spectrometry-based proteomics techniques are so-called bottom-up techniques. This term refers to the method to cut the proteins into peptides using digestion enzymes, analyze peptides and then integrate peptide level quantities computationally back to protein expression values. The reason for choosing this indirect method is that peptides are much easier to handle compared to proteins. Peptides are essentially small molecules which can be separated using a single type of established chromatographic method (usually C18 reverse phase chromatography). When injected into a mass spectrometer and subjected to MS/MS they typically fragment completely along the peptide

*Correspondence: Dr. O. Rinner
CEO Biognosys AG
Wagistrasse 25, CH-8952 Schlieren
E-mail: oliver.rinner@biognosys.ch

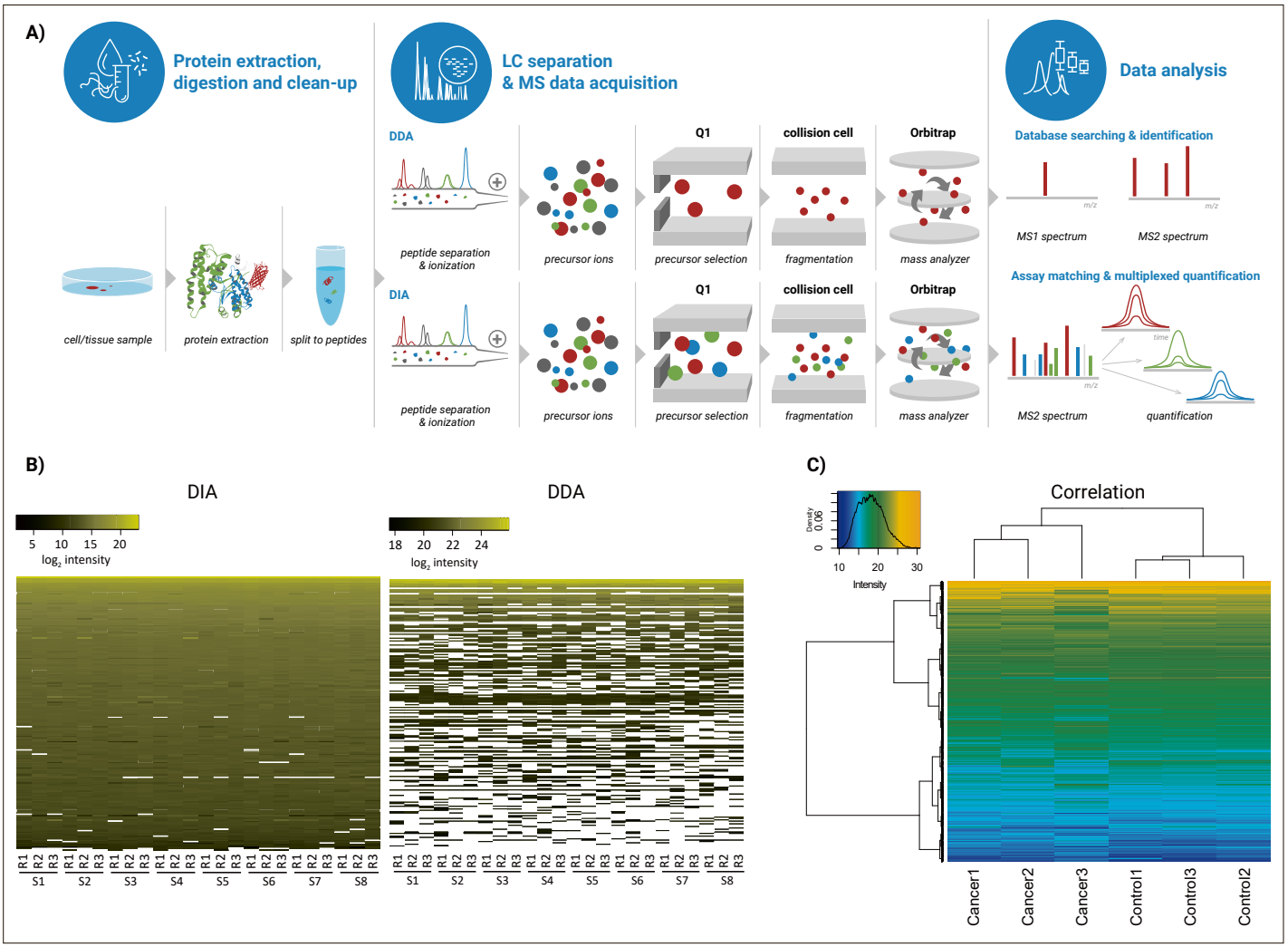

Fig. 1. A) Basic proteomics workflow for shotgun proteomics data dependent acquisition (DDA) and next-generation data independent acquisition (DIA). Proteins are extracted, digested to peptides and separated with liquid chromatography, which is connected on-line to a mass spectrometer. The main difference is the isolation of single ion species in DDA while in DIA multiple ion species are fragmented together. This results for DIA in convoluted but complete data sets, individual peptide signatures can be extracted using software algorithms.
B) Comparison of reproducibility of DIA and DDA as published by Bruderer *et al.*[16] Identical cell lysate samples were 24 times analyzed alternating between DIA and DDA. Resulting protein quantities are plotted as heat map, with more intense proteins at the top. Data derived from DDA shows many missing values (in white) as a result of the random and incomplete acquisition mode, whereas DIA data is largely complete.
C) Example of a small scale differential protein expression study (internal study). Tumor tissues from three patients are analyzed together with non-tumor tissue from the biopsy. The heat map shows that protein profiles of tumors cluster together and show large heterogeneity, while the healthy tissues from different patients show similar protein expression. The heat map is comprised of several thousand protein identifications.

bond, giving rise to information rich mass spectra.[6,7]

Liquid chromatography–mass spectrometry (LC-MS/MS) therefore combines the physical separation capabilities of liquid chromatography (or HPLC) with the mass analysis capabilities of mass spectrometry (MS).

Proteomics Using a Shotgun

The principles of mass spectrometry-based proteomics as described in the influential review of Mann and Aebersold,[8] are still valid today. Briefly, proteins are extracted using any biochemical extraction method, denatured, digested into peptides and injected on a HPLC system. As the peptides elute from the column they are ionized using electrospray ionization, a method for which John Fenn was awarded the Nobel Prize in 2002. These ionized peptides (a mixture of many peptide species) are injected into a mass spectrometer, which in the classical shotgun approach is performed in four main steps (Fig. 1a, DDA):

1. A mass spectrum of all peptide ions is recorded, which typically shows signals of hundreds of peptides at various intensity; because this happens at the first level it is called MS1.
2. Peptide ions of a single mass are selected from the MS1 spectrum. The mass spectrometer selects and isolates a single peptide ion starting with the most intense ion until a duty cycle is over.
3. Each single selected peptide is fragmented by subjecting it to collision with gas molecules in a fragmentation chamber. The peptide fragments along the peptide bond and gives rise to di-, tri- *etc.* peptide fragments. This second level analysis is also called MS2.
4. The combination of the mass of the intact peptide and its characteristic fragmentation pattern enables the identification of a specific peptide.

One important aspect is that this identification is not a sequencing technique but rather a matching technique where MS2 spectra are compared to a selection of all possible peptide in a sequence database. Without genome sequencing shotgun proteomics could not be done. In that sense proteomics is dependent on genomics and in fact for less common organisms, especially from the plant field, a limiting step is the availability of well annotated genomes.

Until today, shotgun proteomics is the most widely used technique and thanks to ever more sensitive mass spectrometers and nano-LC systems with extremely high peak capacity, the depth of 2D gels has long been surpassed. In a recent publication in *Nature*[9] data for more than 15'000 proteins has been provided by the group of Prof. Kuester, Technical University Munich, combining hundreds of experiments (*https://www.proteomicsdb.org*). In single injections typically 3000–4500 proteins per experiment can now be routinely identified.

Limitations of the Classical Shotgun LC-MS/MS Approach

The term shotgun proteomics originally referred to the protein digestion and re-construction of peptide to protein data but it also relates to the ion selection and isolation process in step 2, which essentially is a sequential and random sampling process. In complex samples the duty cycle of the mass spectrometer is not high enough to pick each single analyte. Ions that are not picked will not be fragmented and remain un-detected. If a sample is injected repeatedly the mass spectrometer will cumulatively 'see' all the peptides that can be detected but each single injection is inherently incomplete and irreproducible. On first glance this 'missing data' problem does not seem to be overly serious. However, in contrast to most other analytical techniques, absence of evidence is not evidence of absence. If for instance a protein in sample A is detected but not in sample B the conclusion that this protein is expressed higher in A than B is not valid.

In the hunt for ever larger numbers of identified proteins in the development of proteomics techniques this aspect has been largely ignored by the field. This has often led to a disconnect between scientists asking for results that can be used to understand the biology of their samples and proteomics scientists that were primarily interested in achieving large numbers of identifications. At Biognosys we sometimes use shotgun-proteomics in cases where a customer requires in fact just a list of proteins that is in a sample without quantification; in the vast majority of cases, however, the research questions of our customers are intrinsically quantitative, such as: Which proteins are different with and without treatment? How does expression of a protein change over time? Does an RNAi construct fully knock down its target protein?

Next-generation Proteomics – Parallel Sequencing Replaces Sequential Sequencing

Thanks to the pioneering work of scientists such as Prof. Ruedi Aebersold, co-founder of the Institute for Systems Biology, Seattle, and today professor at the ETH Zurich, in the recent years proteomics has become a quantitative and precise tool that finally provides the depth and the quantitative qualities that are required for quantitative biology.

The starting point for this 'new proteomics', now often called next-gen proteomics – in an obvious reference to next-gen sequencing – was not enabled by a new class of instrument but rather by a radically different approach to how peptide data is acquired and analyzed. The basic idea was developed by Ruedi Aebersold, AB Sciex – an instrument vendor, and Biognosys in 2009, while working on algorithms for improved signal processing of mass spectrometric data.[10]

The shotgun proteomics approach, despite its huge capacity for acquiring data, has essentially been an automatization of a manual approach, where an analytical scientist could in principle associate the peptide mass (MS1) with a spectrum (MS2) and by carefully combining the fragment masses derive or validate a peptide sequence.

Next-generation proteomics, also called data-independent acquisition (DIA) works in a parallel mode where peptide ions across large mass ranges (so-called SWATHs) are isolated together and also fragmented together (Fig 1A, DIA). The result is a highly convoluted MS2 spectrum that consists of fragments from many different peptides, which could not possibly be manually de-convoluted. If, however, it is known at which point in time a peptide elutes from the HPLC column and how the spectrum looks like it is possible to de-convolute it with high confidence. The collection of such template spectra is called a spectral library. The benefit is twofold: Because all randomness is gone each single experiment is highly reproducible and because all peptides that give rise to a signal at the detector are represented in the data with their MS2 fragments, many more peptides can be 'seen' in each single run (Fig. 1B).

Next-generation proteomics requires an independent spectral library for data deconvolution. We and others have published methods that describe how large libraries can be constructed in a very efficient way.[11,12] Large pre-made libraries can be used, in many cases, however, we typically generate sample specific libraries to ensure maximum coverage.

More Proteins, More Samples, Higher Precision

Next-generation proteomics has become the tool of choice for most of our contract research projects that Biognosys performs. Our focus in the past years has been on increasing the sensitivity of analysis and the throughput at the same time:

Sensitivity is important because with greater numbers of proteins more pathways are represented in the dataset. We are now able to identify and quantify in some tissue types close to 9'000 proteins in a single injection (unpublished data), which is approximately three times the number that can be achieved using classical shotgun proteomics. Compared to RNA or DNA analysis, which are amplifying techniques, the challenge for proteomics technologies is the huge dynamic range of protein expression levels. The top 99% of protein mass comes from relatively few proteins.[13] Any further increase in protein numbers requires a large increase in the sensitivity of the method. Further gains can be achieved on the software side with improved signal processing methods and on the hardware side with high-performance chromatography and mass spectrometers. The newest generation of instruments is more sensitive and provides higher scan speed and/or resolution, which not only increases the technical sensitivity but provides new possibilities for more sophisticated signal processing algorithms.

Achieving higher throughput of proteomics methods is now equally important as sensitivity, if proteomics should become as widely used as transcriptomics methods today.[14,15] Currently, a typical experiment for deep analysis of a single sample can require up to 4 hours of instrument time and even more if pre-fractionation is applied. This limits the achievable sample throughput. Consequently, in the past most of our projects with customers from pharma or biotech industry were using relatively small sample designs, typically involving less than 50 samples (see Fig. 1C for an example). New chromatographic methods and better resins have increased peak capacity and enable shorter gradients, which in turn require instruments with faster scan speed. This has enabled us to routinely process hundreds of samples and thereby provide both high-content and high-throughput at the same time.

The Future of Proteomics

From our commercial work with customers across a wide range of research questions we know that next-gen proteomics provides valuable functional protein level data. But there are several chal-

lenges connected with the technical complexity of nano-LC systems and programming of mass spectrometers that have to be overcome before proteomics becomes as accessible to non-experts as for instance next-gen sequencing is today.

The vendors of high-performance chromatography systems and mass spectrometers are aware of these challenges and are currently working towards the goal of having easy to use desktop systems, and as the demand for proteomics data further increases it can be expected that such systems become available in the near future.

The arguably largest challenge when dealing with proteomics is the deluge of data that is produced at ever greater quantities. The digital peptide ion maps generated with next-gen proteomics technology are in the range of 10 GB/sample. To biologists without strong bioinformatics background signal processing, data analysis and finally data interpretation of dataset with 100'000 analytes or even more can be a daunting task. To promote development of the field we provide our Spectronaut software as free academic version, which is able to efficiently process and analyze such large datasets.[16]

Connecting the Proteome with the Genome

Transcriptomics and proteomics are sometimes seen as competing technologies. From a biological point of view, however, they work on different levels of regulation in biological systems and as such are complementary. In a recent publication of Williams *et al*.[17] proteomics, transcriptomics, metabolomics, and genomics data were combined to establish causal links between genotype and phenotype in mouse liver function, which could not be deduced from each data source alone. The study also shows that the data-structures of RNA-Seq and proteomics data are very compatible and it can be expected that in the future proteogenomics becomes the method of choice for comprehensive system level analysis.

Functional Proteomics

So far we have discussed how proteomics technology can provide precise quantities of peptides and proteins. Functional proteomics goes beyond expression profiling and enables the researcher to gain insight into structure, signaling, protein-protein interaction and even protein complex stoichiometry. In most cases this is achieved by specific workflows on the sample preparation level. Examples for such functional proteomics workflows are:

- Protein crosslinking.[18] Here a bivalent cross-linker is applied to native protein lysates before digestion. After digest peptides that were formerly close in tertiary protein structure are still connected and can be identified. This provides direct distance constraints, which can be used for modeling of protein complex structures from low resolution cryo-electron microscopy structures.[19]
- In phospho-proteomics experiments phosphorylated peptides are separated from non-modified peptides using special affinity resins. Thousands of phospho-sites can be identified in a single experiment and provide information about activity of signaling pathways.[20]
- With limited proteolysis binding events (*e.g.* to a drug) can be observed. Subtle structural changes upon binding affect the digestion kinetics. The presence or absence of a peptide correlated with binding can provide direct evidence of the protein that interact with a molecule.[21]
- Large-scale protein–protein interaction studies have been performed where binding partners of proteins are identified by immune-precipitation with thousands of baits followed by proteomics analysis of pull-downs. For a review see for example Gringras *et al*.[22]

Outlook

This article is focused on the development of proteomics technology in the research field. Currently there is a strong interest of instrument vendors and clinical researchers to apply proteomics methods in clinical practice, *e.g.* as high-dimensional diagnostic tests in the context of personalized or precision medicine. It can be expected, however, that it will take several years before proteomics will be broadly used in a regulated environment. The reasons are not primarily technological problems but rather the need to address issues of instrument certification, the relationship to existing regulations of authorities such as the FDA, and with questions of reimbursement. There is no doubt that proteomics and other -omics technologies will find their place in clinical practice. But until this can be achieved we should focus on the possibilities that come with the new quality and depth of data that next-gen proteomics provides, to develop better drugs and diagnostics, and ultimately to better understand life.

Received: September 13, 2016

[1] N. L. Anderson, N. G. Anderson, *Electrophoresis* **1998**, *19*, 1853.
[2] S. P. Gygi, Y. Rochon, B. R. Franza, R. Aebersold, *Mol. Cellul. Biol.* **1999**, *19*, 1720.
[3] Y. Liu, A. Beyer, R. Aebersold, *Cell* **2016**, *165*, 535.
[4] U. K. Laemmli, *Nature* **1970**, *227*, 680.
[5] P. H. O'Farrell, *J. Biol. Chem*. **1975**, *250*, 4007.
[6] W. H. McDonald, J. R. Yates, *Dis. Markers* **2002**, *18*, 99.
[7] J. K. Eng, A. L. McCormack, J. R. Yates, *J. Amer. Soc. Mass Spectrom.* **1994**, *5*, 976.
[8] R. Aebersold, M. Mann, *Nature* **2003**, *422*, 198.
[9] M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J. H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, B. Kuster, *Nature* **2014**, *509*, 582.
[10] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, R. Aebersold, *Mol. Cellul. Proteom.* **2012**, *11*, O111.016717.
[11] G. Rosenberger, C. C. Koh, T. Guo, H. L. Röst, P. Kouvonen, B. C. Collins, M. Heusel, Y. Liu, E. Caron, A. Vichalkovski, M. Faini, O. T. Schubert, P. Faridi, H. A. Ebhardt, M. Matondo, H. Lam, S. L. Bader, D. S. Campbell, E. W. Deutsch, R. L. Moritz, S. Tate, R. Aebersold, *Sci. Data* **2014**, *1*, 140031.
[12] R. Bruderer, O. M. Bernhardt, R. Gandhi, L. Reiter, *Proteomics* **2016**, *16*, 2246.
[13] M. Beck, A. Schmidt, J. Malmstroem, M. Claassen, A. Ori, A. Szymborska, F. Herzog, O. Rinner, J. Ellenberg, R. Aebersold, *Mol. Syst. Biol.* **2011**, *7*, 549.
[14] T. Guo, P. Kouvonen, C. C. Koh, W. E. Wolski, H. L. Röst, G. Rosenberger, B. C. Collins, L. C. Blum, S. Gillessen, M. Joerger, W. Jochum, R. Aebersold, *Nat. Med.* **2015**, *21*, 407.
[15] L. C. Gillet, A. Leitner, R. Aebersold, *Ann. Rev. Anal. Chem.* **2016**, *9*, 9.1.
[16] R. Bruderer, O. M. Bernardt, T. Gandhi, S. M. Miladinovic, L. Y. Cheng, S. Messner, T. Ehrenberger, V. Zanotelli, Y. Butscheid, C. Escher, O. Vitek, O. Rinner, L. Reiter, *Mol. Cell. Proteom*. **2015**, *14*, 1400.
[17] E. G. Williams, Y. Wu, P. Jha, S. Dubuis, P. Blattmann, C. A. Argmann, S. M. Houten, T. Amriuta, W. Wolski, N. Zamboni, R. Aebersold, J. Auwerx, *Science* **2016**, *352*, doi:10.1126/science.aad0189.
[18] O. Rinner, J. Seebacher, T. Walzthoeni, L. Mueller, M. Beck, A. Schmidt, M. Mueller, R. Aebersold, *Nat. Meth.* **2008**, *5*, 315.
[19] K. Lasker, F. Förster, S. Bohn, T. Walzthoeni, E. Villa, P. Unverdorben, F. Beck, R. Aebersold, A. Sali, W. Baumeister, *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 1380.
[20] K. Sharma, R. C. J. D'Souza, S. Tyanova, C. Schaab, J. R. Wisniewski, J. Cox, M. Mann, *Cell Rep.* **2014**, *8*, 1583.
[21] Y. Feng, G. De Franceschi, A. Kahraman, M. Soste, A. Melnik, P. J. Boersema, P. P. de Laureto, Y. Nikolaev, A. P. Oliveira, P. Picotti, *Nat. Biotechnol.* **2014**, *32*, 1036.
[22] A. C. Gingras, M. Gstaiger, B. Raught, R. Aebersold, *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 645.