INNOVATION

Scoring proteomes with proteotypic peptide probes

Bernhard Kuster, Markus Schirle, Parag Mallick and Ruedi Aebersold

Abstract | Technologies for genome-wide analyses typically undergo a transition from a discovery phase to a scoring phase. In the discovery phase, the genomic universe is explored and all pertinent data are noted. In the scoring phase, relevant entities are screened to reveal groups of genes that are associated with specific biological processes or conditions. In this article, we propose that the transition from a discovery to a scoring phase is also essential, feasible and imminent for proteomics.

Proteomics attempts to analyse the entire protein complement of the genome, and it can therefore be considered a discipline of the genomic sciences. Proteomics is challenging because proteomes change quantitatively and qualitatively to reflect changes in the physiological state of cells, organs and organisms. As a consequence of post-transcriptional events (splicing) and post-translational changes (including modifications, proteolytic processing and complex formation), the number of proteins that comprise a proteome is vastly larger than the number of genes that constitute a genome. It is therefore not surprising that no complete proteome has been described to date.

At present, the most powerful proteomics techniques are based on mass spectrometry (MS). There are many variations of MS-based proteomics techniques and they have been extensively reviewed[1–3]. Although these differ in the details of their implementation, they share the process of fragmenting proteins into peptides by proteolysis in solution, followed by the use of a mass spectrometer to extract sufficient information — either from individual peptides or from the collection of peptides that is generated from a particular protein — to annotate the protein conclusively (BOX 1). Technological advances in protein chemistry, separation methods, MS and bioinformatics have contributed to the development of robust platforms that have the capacity to identify and quantify thousands of proteins in a single, albeit complex and costly, experiment.

Among the most informative proteomics applications are those that measure the effect of perturbations in the proteome or in information-rich subproteomes, such as those representing organelles or signalling networks[4–7]. Examples include the detection of characteristic patterns in body-fluid proteomes that are indicative of a particular disease[8], the detection of changes in cellular proteomes that are induced in response to pharmacological agents or physical insults[9], and the correlation of the proteomes of differentiated cells with their physiological function[10]. Such studies typically involve the repeated quantitative measurement and comparison of several proteomes or fractions of proteomes. Quantitative proteomics measurements are usually accomplished by using stable isotopes that are incorporated into the sample polypeptides by one of many strategies[1,2] (BOX 2).

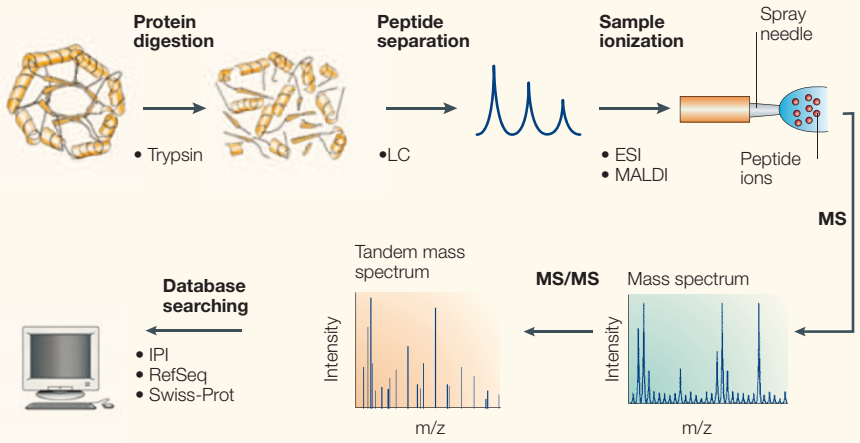
Knowledge of the number of genes in the genome of a species is one of the most striking results from completed genome-sequencing projects. For the experimental genomic sciences, the number of genes and their products define the space in which all experiments have to operate. As it made sense in geography first to chart out the globe and then learn how to navigate it using the resulting maps, it has also proved sensible first to chart out genomic space and then learn how to navigate this space. Genomic technologies have been undergoing a transition from a discovery phase, in which the genomic universe is charted out, to a scoring phase, in which all the elements are examined to identify gene constellations that are correlated with specific biological conditions. Once genomic constellations are identified, low-cost and high-throughput technologies are used to measure the constellation specifically. Gene-expression analysis and single nucleotide polymorphism (SNP) scoring are examples of this transition. These assays provide useful information, even when the whole space has not been completely charted. In this article, we propose that such a transition is also necessary, feasible and imminent for proteomics, to enable it to achieve its potential as a genomics technology that is focused at the functional level.

Limitations of current technology

In most of the common MS-based proteomics methods, purified proteins or protein mixtures are first digested with trypsin and the resulting peptides are separated by capillary chromatography. The peptides are then ionized using a process called electrospray ionization (ESI) and analysed with a tandem mass spectrometer (this process is known as liquid-chromatography–tandem MS or LC–MS/MS) (BOX 1). Tandem mass spectrometers measure the mass of the detected peptide ions and sequentially subject selected peptide ions to fragmentation in a collision cell (collision-induced dissociation or CID), which generates fragment-ion patterns that reflect the amino-acid sequence of the precursor peptide. Consequently, only those peptides that are selected by the instrument for fragmentation can yield useful information.

Box 1 | **Protein identification by mass spectrometry**

Proteins can be identified by any peptide sequence that is unique to a particular protein sequence. Mass spectrometry (MS) is an analytical technique that is uniquely suited for this purpose because of its inherent speed, sensitivity and ability to analyse complex mixtures (see figure). Usually, proteins are first digested into tryptic peptides and the resulting peptide pool is partially separated using liquid chromatography (LC) methods. Two main techniques are used to generate the peptide ions that are necessary for MS detection: matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI). MALDI is a pulsed technique, in which high-energy photons are used to generate peptide ions from a sample that is embedded in a solid organic matrix, whereas ESI is a continuous ionization technique, in which charged liquid droplets are formed from which analyte ions are desorbed. Peptides that are eluting from the LC system are ionized and have their mass measured by the mass spectrometer. Some peptide ions are selected for controlled fragmentation (collision-induced dissociation) inside the instrument. The resulting fragments are further analysed by the mass spectrometer, and this process is known as tandem mass spectrometry (MS/MS). The tandem mass spectrum of a peptide is indicative of its amino-acid sequence. Mass information on the intact peptide, as well as on its fragments, can be used to query theoretical spectra for proteolytic fragments of all the sequences that are present in a protein database or a translated nucleotide sequence database (for example, International Protein Index (IPI), RefSeq and Swiss-Prot; see Online links box). For further information, see REFS 1–3. m/z, mass-to-charge ratio. Figure modified with permission from *Nature Reviews Molecular Cell Biology* REF. 3 © (2004) Macmillan Magazines Ltd.



In proteomics experiments, the peptide mixtures generated for analysis are complex, and typically contain tens to hundreds of thousands of unique peptides. This sample complexity overwhelms even the most modern tandem mass spectrometers, which can select and fragment peptide ions faster than one per second.

The limited efficiency of LC–MS/MS technology is illustrated in FIG. 1, which shows a complex peptide mixture that was analysed by LC–MS/MS. FIGURE 1a shows all the putative peptides that were detected by the mass spectrometer over the course of an LC–MS/MS experiment on a two dimensional mass/charge (m/z) versus time plot. Those features that were selected for CID (blue) can be seen in FIG. 1b, and FIG. 1c highlights the peptides that were identified with high confidence in a database search (red). Several observations can be made. First, only a fraction of the detected precursor ions were selected for CID, because too many peptide ions were concurrently presented to the mass spectrometer. Second, only a fraction of the CID attempts led to a definitive peptide identification. Third, the same peptides were repeatedly selected for CID, despite precautions to prevent this redundancy (for example, the dynamic exclusion of peptides from repeated CID selection if the same precursor had already been selected within a user-defined exclusion time window). Fourth, redundant peptides from the same protein were sequenced.

Collectively, these factors result in an under-sampling of complex peptide mixtures, limit the complexity of a proteome that can be analysed at present, increase the time required to carry out such analyses (because of the repeat experiments that are required), and unnecessarily complicate the informatics challenges posed by such experiments. It is doubtful that incremental improvements in technology will be sufficient to achieve routine, comprehensive and high-throughput proteome analysis, highlighting the need to develop a conceptually different approach. We argue that this can be accomplished by moving proteomics from a discovery to a scoring mode of operation.

From proteome discovery to scoring

LC–MS/MS and variations of this method are proteome discovery tools that perpetually rediscover substantially similar segments of the proteome, while leaving other segments undiscovered. In some fields of genomic science, high-throughput and robust analysis platforms have been achieved by moving from a perpetual discovery mode to a scoring mode of operation. Notably, this development has been successful in gene-expression array analysis and SNP scoring. For gene-expression arrays, the discovery phase establishes the transcripts that are potentially generated by a genome (the transcriptome) using experimental approaches, such as expressed sequence tag (EST) sequencing[11], and computational methods. In the scoring phase, transcript-specific probes are arranged in ordered microarrays, and the gene transcripts expressed by a cell in a specific state can be identified and quantified in a single experiment. In the discovery phase of SNP analysis, an international consortium of scientists is generating a comprehensive catalogue of the SNP space of humans, with the goal of scoring the association of specific patterns of SNPs with specific phenotypes[12].

For a comparable proteome scoring technology to be successful, the datum to be scored, as well as a platform to support the scoring has to be identified. We propose that the datum for proteome scoring should be a proteotypic peptide — that is, an experimentally observable peptide that uniquely identifies a specific protein or protein isoform. We suggest that the platform for proteome scoring should be an ordered array of proteotypic peptides that can be analysed by a (tandem) mass spectrometer. Compared with protein-based proteome scoring strategies, for example, abundance-based protein microarrays[13], proteotypic peptides avoid the main obstacles associated with protein arrays, which are the availability and specificity of reagents such as antibodies or aptamers. In the following sections, we describe experimental and computational methods for the generation of a library of proteotypic peptides and a prototypical platform for proteome scoring.

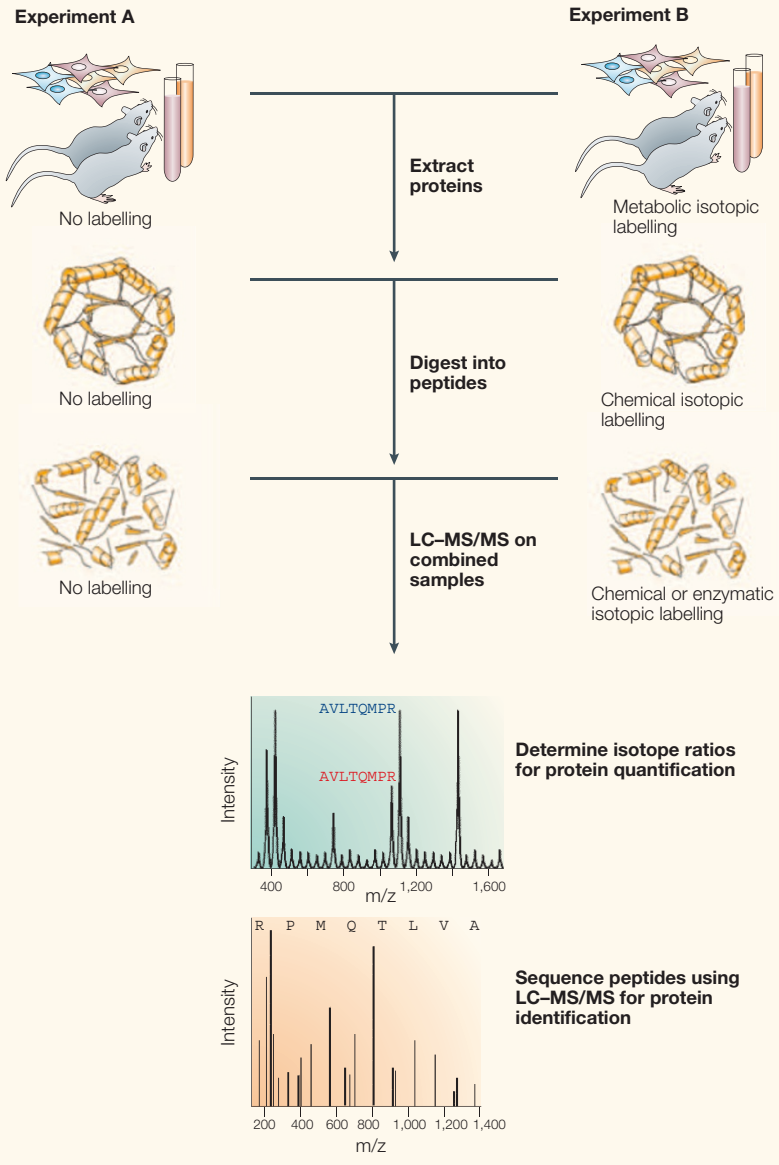
Identifying proteotypic peptides

Experimental identification. The most straightforward way to generate a list of proteotypic peptides is to collect peptide-identification data empirically and to search these data for commonly observed peptides that uniquely identify a protein. Fortunately, discovery-type proteomics projects produce enormous amounts of redundant protein identifications because they are carried out in many laboratories. Such projects cover a wide range of biological questions and experimental strategies, including the analysis of protein samples that have been extracted from different cell and tissue types, subcellular fractions, affinity- purified protein complexes, secreted proteins and proteins in body fluids. Consequently, the collective efforts of the proteomics community have discovered a sizable fraction of the human proteome, and discovery will continue undiminished until the whole proteome space that is observable using present technology is mapped out.

To transform disparate datasets into a database of validated proteotypic peptides is challenging. The data from different research groups are generated using different experimental strategies and different types of mass spectrometer. Furthermore, data analysis is inconsistent, because different database search tools are used with different scoring thresholds. In addition, it has been uncommon in the proteomics community to share raw data. To overcome these limitations, we created the first implementation of a platform that supports the import and statistical validation of proteomics data, which were generated using many experimental approaches, into a database[14]. The intent of this PeptideAtlas project (see Online links box) is to integrate data generated by the proteomics community in discovery-type projects. To assure a known and consistent level of data quality in the database, each datum (peptide) is associated with a numerical value that indicates the probability that a mass spectrum has been correctly assigned[15]. Proteotypic peptides can be easily identified from this database, and peptides can be selected that best fit the objectives of a scoring experiment, such as the identification of proteins, splice forms, SNPs, post-translational modifications and more. At present, PeptideAtlas holds peptide information that represents some 6,000 human genes. It is likely that significantly more human gene products have already been identified by public and private-sector proteomics research groups, so this represents a large body of data that could be imported into PeptideAtlas.

Box 2 | **Stable-isotope labelling**

Stable isotopes, notably ^{13}C, ^{15}N, ^{18}O and ^{2}H, can be incorporated into proteins metabolically, into extracted proteins through chemical reactions that modify reactive side chains of amino acids, or into tryptic peptides through chemical or enzymatic modification of the N or C termini (see figure). Isotope-labelled peptides have nearly the same physicochemical properties as their unlabelled counterparts, notably the same chromatographic separation behaviour and signal intensity in a mass spectrum. However, labelled peptides differ in mass by an increment that is directly related to the number of incorporated isotopes. For quantification experiments, unlabelled (see figure, experiment A) and labelled (experiment B) samples are combined and analysed in a mass spectrometer, which can precisely measure these mass increments and therefore discover pairs of unlabelled (see figure; red) and labelled (blue) peptides. It can also use the signal intensity ratios of the heavy and light forms of peptides of identical sequence to infer relative protein abundance. In this example, the ratio is approximately 1:2, which indicates a twofold increase in the abundance of the respective protein in experiment B. Furthermore, protein identification can be achieved by fragmenting the precursor ion in the collision cell of the mass spectrometer and by searching the resulting fragment ion pattern against a sequence database. LC–MS/MS, liquid-chromatography–tandem mass spectrometry; m/z, mass-to-charge ratio.



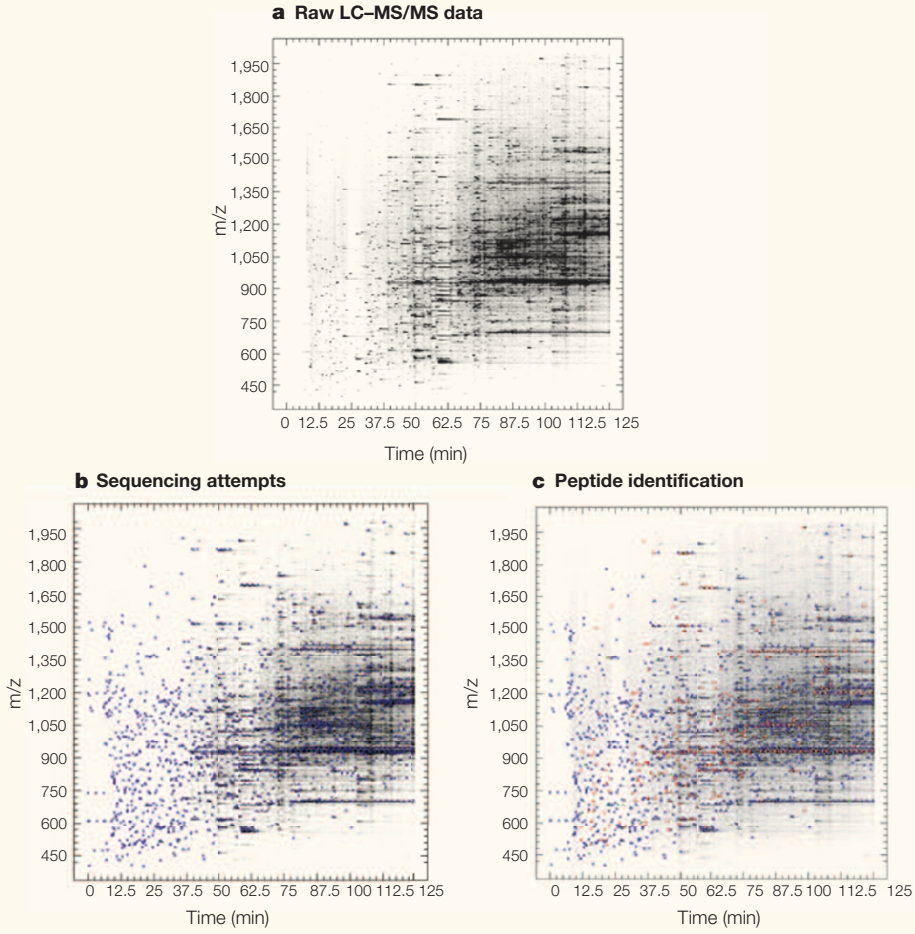


Figure 1 | **Current approaches for scoring proteomes.** The current liquid-chromatography–tandem mass spectrometry (LC–MS/MS) approaches are inefficient for the rapid scoring of proteomes. **a** | A representation of the data that were obtained from a typical LC–MS/MS experiment, in which a complex tryptic protein digest was analysed. Every spot on the two-dimensional graph represents the detection of a tryptic peptide of a particular mass-to-charge ratio (m/z) at a particular chromatographic elution time and with a particular intensity (the darker the spot, the greater its intensity). The graph contains 2,720 such spots. **b** | This panel highlights the 1,633 spots (blue) for which the mass spectrometer attempted to generate a tandem mass spectrum (for peptide sequencing). **c** | This panel highlights the 363 spots (red) that resulted in unambiguous peptide identification. Therefore, of the 2,720 spots that were detected in the sample, only 13% could be identified and, of these, many redundantly identified the same peptide or protein. Repeated analysis of the same sample might increase this value, but is prohibitive because of the analysis time and volume of data generated. This example therefore highlights the need for a conceptually different approach for the rapid scoring of proteomes. Reproduced with permission from REF. 26 © (2004) American Chemical Society.

Multiple repeat identifications of peptides are a fundamental weakness of discovery proteomics projects. However, for the identification of proteotypic peptides, redundant identifications are essential, as they highlight the peptides that are consistently observable under specific or diverse experimental conditions. We have observed that most proteins are identified on the basis of the detection of one to a few peptides observed at a high frequency (FIG. 2). Although the exact number and frequency of such peptides might vary between proteins and with the MS method used, it is almost always possible to find peptides that are common to all methods and that enable 'basic' proteome scoring (that is, the detection of a gene product). More sophisticated proteome scoring tasks, such as the identification of splice forms, SNPs and post-translational modifications, might require more specialized sets of peptides and, potentially, more refined analytical approaches.

Computational identification. The identification of proteotypic peptides from experimental data is restricted to proteins that have been identified repeatedly, and therefore allow peptide frequencies to be calculated. So, until PeptideAtlas or similar databases[16] reach saturation, proteotypic peptides are unavailable for the fraction of proteins that have not yet been observed in discovery projects. Furthermore, for an increasing number of species with completed genome sequences, little proteomics data have been collected. It would therefore be advantageous if proteotypic peptides could be computationally identified whether or not they have been empirically detected in discovery experiments.

On the basis of previous observations regarding the existence of proteotypic peptides, we have searched for physicochemical properties that can distinguish proteotypic peptides from peptides that are not observed ('unobserved' peptides). We have carried out a pilot study to discover some of these properties and developed a tool to predict the proteotypic peptides of a protein. First, a set of proteotypic peptides and a set of unobserved peptides were defined for four of the common MS proteomics experimental designs — isotope-coded affinity tag (ICAT)–ESI[17], PAGE–ESI[18], multidimensional protein-identification technology (MudPit)–ESI[19], and PAGE–matrix-assisted laser desorption/ionization (MALDI)[20]. Only proteins that were observed with a high confidence in several experiments were included in the study. A peptide was declared proteotypic if it was observed in more than 50% of the experiments in which the corresponding protein was observed. A set of approximately 500 numeric, physicochemical property scales (such as secondary-structure propensity, hydrophobicity and charge) for amino-acid residues has been identified[21], so that any peptide can be described in terms of an array of its properties (FIG. 2e). Having defined a set of properties for our proteotypic and unobserved peptides, we used statistical methods to identify a subset of these properties that best distinguished the set of proteotypic peptides from the set of unobserved peptides. For each of our four sets of proteotypic peptides that were generated by the four common types of proteomics experiment, and for a composite set, four to six properties were discovered to be the most discriminating and could be used to score the likelihood of a peptide being proteotypic or not. On a set of peptides that were excluded from this training exercise, the performance of the classifiers were typically 85% accurate in their ability to predict a proteotypic peptide (P.M. *et al.*, unpublished observations).

The possibility of predicting proteotypic peptides for proteins that have not yet been observed in discovery projects significantly expands the scope of proteome scoring, and

makes the approach applicable to all species for which nucleotide sequence data are available.

A platform for proteome scoring

Proteome scoring technology requires a type of datum to be scored — in this case, proteotypic peptide data — and a platform that carries out the scoring operation. The objective of such a platform is to identify and quantify, robustly and reliably, the proteins that are contained in a protein sample. This platform should also be able to score other features of the proteome, such as the products of differential splicing and post-translational modification.

We have described such a platform, which is based on the generation of ordered arrays of proteotypic peptides and the use of MALDI–MS/MS[22] (FIG. 3). To produce ordered peptide arrays, protein samples are subjected to tryptic digestion and are combined with a mixture of defined amounts of isotopically labelled proteotypic peptides. These reference peptides are generated by chemical synthesis and are labelled with heavy stable isotopes, either through the incorporation of heavy amino acids during synthesis or through post-synthesis chemical modification. The combined peptide mixture is separated by capillary reverse-phase liquid chromatography and the eluting peptides are deposited on a MALDI sample plate to form an ordered peptide array.

In this array, each spot contains peptides that are derived from the digested sample proteins and/or from the mixture of reference peptides. For the detection and quantification of target polypeptides (that is, those proteins for which a reference peptide was added to the sample), the sample is analysed using a MALDI tandem mass spectrometer. The instrument acquires a mass spectrum for each array element, which generates two types of signal: single peaks, which represent peptides that lack a reference peptide or a reference peptide without a natural counterpart; and paired signals, which represent those peptides for which a proteotypic reference peptide was added. These paired signals have a mass difference that precisely corresponds to the mass difference encoded by the stable-isotope. The relative signal intensity of the differentially labelled peptides can be used to calculate the abundance of the target peptide and, if required, CID can be used to confirm the sequence of the target peptide.

In this proteome scoring method, the mass spectrometer is focused on the targeted analysis of the information-rich proteotypic peptides. Therefore, the redundancy in data collection that is inherent in discovery projects is eliminated, the instrument is used more productively, and the analysis of the collected data simplified. Current MALDI mass spectrometers are equipped with 200 Hz lasers, which allow a proteome that is arranged in an array of 200 elements to be scored in less than 30 minutes. With the imminent introduction of lasers of 1 kHz or faster, it is anticipated that proteomes will be scored in just a few minutes.

a MS-compatible peptides

```
MSRSKRDNNF YSVEIGDSTF TVLKRYQNLK PIGSGAQGIV CAAYDAILER NVAIKKLSRP FQNQTHAKRA YRELVLMKCV
NHKNIIGLLN VFTPQKSLEE FQDVYIVMEL MDANLCQVIQ MELDHERMSY LLYQMLCGIK HLHSAGIIHR DLKPSNIVVK
SDCTLKILDF GLARTAGTSF MMTPYVVTRY YRAPEVILGM GYKENVDLWS VGCIMGEMVC HKILFPGRDY IDQWNKVIEQ
LGTPCPEFMK KLQPTVRTYV ENRPKYAGYS FEKLFPDVLF PADSEHNKLK ASQARDLLSK MLVIDASKRI SVDEALQHPY
INVWYDPSEA EAPPPKIPDK QLDEREHTIE EWKELIYKEV MDLEERTKNG VIRGQPSPLG AAVINGSQHP SSSSSVNDVS
SMSTDPTLAS DTDSSLEAAA GPLGCCR
```

b Observed peptides

```
MSRSKRDNNF YSVEIGDSTF TVLKRYQNLK PIGSGAQGIV CAAYDAILER NVAIKKLSRP FQNQTHAKRA YRELVLMKCV
NHKNIIGLLN VFTPQKSLEE FQDVYIVMEL MDANLCQVIQ MELDHERMSY LLYQMLCGIK HLHSAGIIHR DLKPSNIVVK
SDCTLKILDF GLARTAGTSF MMTPYVVTRY YRAPEVILGM GYKENVDLWS VGCIMGEMVC HKILFPGRDY IDQWNKVIEQ
LGTPCPEFMK KLQPTVRTYV ENRPKYAGYS FEKLFPDVLF PADSEHNKLK ASQARDLLSK MLVIDASKRI SVDEALQHPY
INVWYDPSEA EAPPPKIPDK QLDEREHTIE EWKELIYKEV MDLEERTKNG VIRGQPSPLG AAVINGSQHP SSSSSVNDVS
SMSTDPTLAS DTDSSLEAAA GPLGCCR
```

c Proteotypic peptides

```
MSRSKRDNNF YSVEIGDSTF TVLKRYQNLK PIGSGAQGIV CAAYDAILER NVAIKKLSRP FQNQTHAKRA YRELVLMKCV
NHKNIIGLLN VFTPQKSLEE FQDVYIVMEL MDANLCQVIQ MELDHERMSY LLYQMLCGIK HLHSAGIIHR DLKPSNIVVK
SDCTLKILDF GLARTAGTSF MMTPYVVTRY YRAPEVILGM GYKENVDLWS VGCIMGEMVC HKILFPGRDY IDQWNKVIEQ
LGTPCPEFMK KLQPTVRTYV ENRPKYAGYS FEKLFPDVLF PADSEHNKLK ASQARDLLSK MLVIDASKRI SVDEALQHPY
INVWYDPSEA EAPPPKIPDK QLDEREHTIE EWKELIYKEV MDLEERTKNG VIRGQPSPLG AAVINGSQHP SSSSSVNDVS
SMSTDPTLAS DTDSSLEAAA GPLGCCR
```

d Frequency analysis

Proteotypic peptide	Frequency of presence (%)	Cumulative presence (%)
APEVILGMGYK	81.6	81.6
NIIGLLNVFTPQK	71.1	94.7
ILDFGLAR	52.6	97.4

e Prediction

Property	Peptide sequence										
	K	I	L	D	F	G	L	A	R	Total	Average
Frequency in turn	0.1	0.06	0.07	0.08	0.07	0.15	0.07	0.06	0.09	0.75	0.08
Hydrophobic moment	5.7	1.2	1.0	1.9	1.1	0.0	1.0	0.0	10.0	21.9	2.4
Negative charge	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.1

Figure 2 | **Experimental and computational approaches for the identification of proteotypic peptides.** The concept of proteotypic peptides is illustrated using human mitogen-activated protein kinase-8 (MAPK8) as an example. **a** | The fraction of the sequence (64%) that is expected to be compatible with mass spectrometry (MS) analysis is shown in blue (the tryptic peptide mass range is 800–3,000 Da). **b** | All of the peptides that were detected in a set of approximately 100 experiments that identified MAPK8 (53% of the sequence) are highlighted in green. **c** | Peptides that were detected in at least 50% of all experiments are highlighted in red (proteotypic peptides). **d** | This panel shows a tabulation of the frequencies with which the three proteotypic peptides were detected, as well as their cumulative presence — that is, in 97.4% of the cases in which MAPK8 was identified, at least one of the three proteotypic peptides was detected. **e** | A computational approach to identify proteotypic peptides. Peptides are described by a numerical matrix of physicochemical properties that are associated with each amino acid (3 of ~500 properties are shown). By comparing the values for proteotypic peptides with those for peptides that have not been observed, the most discriminating properties can be identified and used to build a predictor. This facilitates the selection of proteotypic peptides for proteins that have not yet been observed by MS.

Realization of concept and projections

At this time, the concept described here has not been fully translated into practice, despite some successful studies that have used isotopically labelled reference peptides to quantify specific proteins[23,24]. Clearly, much of the technical details still need to be refined, but given the current technological status we believe that the transition from the discovery to the scoring phase of proteomics is both timely and feasible, and that there is a clear path towards the robust and economical implementation of a proteome-scoring technology. The basic elements of the technology are in place and will be further refined. It can be expected that the content of PeptideAtlas, or similar data repositories, will grow substantially, and the added content will enhance the empirical and computational identification of an optimal

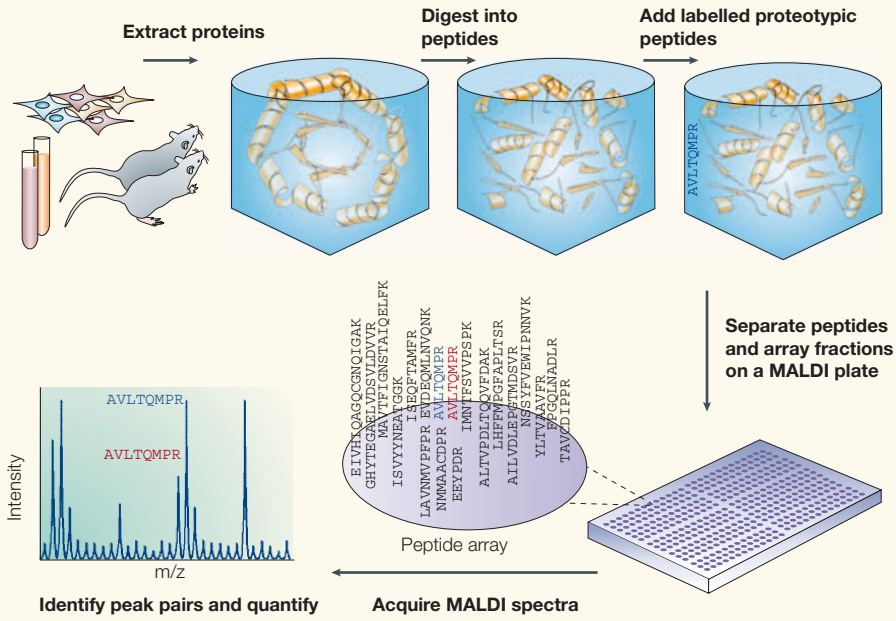


Figure 3 | **Analysis platform for scoring proteomes.** Proteotypic peptides that are labelled with stable isotopes are added to the tryptic digest of a proteome that is to be scored for the presence of specific proteins. Labelled proteotypic peptides are generated by chemical synthesis, or through the post-synthesis chemical modification of reactive groups with heavy-isotope-containing labelling reagents. The complex digest is separated using liquid chromatography, and a matrix-assisted laser desorption/ionization (MALDI) sample plate functions as a fraction collector. As the unlabelled, endogenous peptides (red sequence) and labelled, exogenous proteotypic peptides (blue sequence) behave identically during liquid chromatography, they are deposited on the same MALDI plate spot, and therefore form an ordered array of peptides on the MALDI plate. MALDI spectra are then acquired for each spot and the data are analysed for the presence of peaks that correspond to pairs of endogenous, 'light' peptides (red) and added proteotypic, 'heavy' peptides (blue). The relative protein abundance can be inferred from the intensity ratio of the peak pair. If required, the peptide sequence can be subsequently verified in a collision-induced-dissociation experiment. m/z, mass-to-charge ratio.

set of proteotypic peptides. It can also be expected that the performance of mass spectrometers will continue to improve, so that proteomes can be scored faster, more accurately and at higher sensitivity.

The main challenge associated with the strategy is the chemical synthesis and distribution to the community of the peptide reagents that are required for proteome scoring. At present, the cost associated with synthesizing a proteotypic peptide for each human gene will require an investment of several million US dollars. Such peptide collections will also require a storage and distribution system that is similar to those available for EST clones, cell lines or tissue samples. Examples from other areas of genomics — gene chips, PCR primers and, more recently, small interfering RNAs — show that production, storage and distribution problems are manageable and present important commercial opportunities. The cost of synthesizing a single unlabelled peptide is typically in the order of US$250 per µmol, and isotopically labelled peptides are available at approximately $50–100 per nmol (see Sigma–Aldrich and Thermo Electron Corporation in the Online links box for commercial sources of isotope-labelled peptides). However, given that a single data point in a proteome-scoring experiment will require less than 100 fmol of a peptide, the cost for each data point becomes almost negligible.

However, before committing to investments of this order, it would be prudent to carry out proof-of-concept studies of limited scope. We suggest that the human serum proteome and the human kinome could function as valuable pilot studies to show the potential power of the approach in diagnostics and drug discovery. Both proteomes are sufficiently well characterized to warrant the reliable selection of approximately 500 proteotypic peptides, and we anticipate that such pilot studies could be completed in 12–18 months. A recent serum proteome study from our laboratory shows that the approach is feasible[22]. The selective scoring of information-rich subproteomes is biologically and clinically important and is expected to be the main application of the method.

The strategy is not limited to the human proteome, to the proteomes of species for which there are extensive amounts of MS data, or to the exclusive use of MALDI–MS. With the increasing volume of discovery-type proteomics data, the predictors for proteotypic peptides will become sensitive enough to allow the prediction of peptides from any protein, regardless of the species of origin, as long as the nucleotide sequence is available. Computational approaches will also provide collections of proteotypic peptides that not only address protein-identification issues, but that can also be used to identify splice variants, isoforms and SNPs. However, it should be noted that these will probably require special attention, because simple trypsinolysis might not immediately yield peptides that are compatible with MS for these purposes. Finally, it can be anticipated that ESI–MS will also be used for proteome scoring using proteotypic peptides, possibly through the extension of the accurate mass-and-time-tag concept that was pioneered by R. D. Smith's group[25].

Conclusions

For the quantitative profiling of proteomes to reach the robustness and throughput levels that are required to study systematically cellular perturbations in response to environmental changes or panels of clinical samples, we propose that discovery-driven proteomics will pave the way to a scoring phase in proteomics. Such a transition can be facilitated by arrays of proteotypic peptides. The realization of a project of this scale depends on an appropriate infrastructure for providing reagents and the required data analysis tools, and on the cooperation of interested scientists. Although it is clear that such a joint effort has considerable political and economic implications, similar efforts in the genomics sciences show that such a scenario is indeed realistic and of potentially large impact.

Bernhard Kuster and Markus Schirle are at the Department of Analytical Sciences and Informatics, Cellzome AG, Meyerhofstrasse 1, 69117 Heidelberg, Germany.

Parag Mallick is at the Louis Warschaw Prostate Cancer Center, Cedars-Sinai Medical Center, 8631 West Third Street, Suite 1001E, Los Angeles, California 90048, USA.

Ruedi Aebersold is at the Institute for Molecular Systems Biology, ETH Hönggerberg HPT E 78, Wolfgang Pauli-Strasse 16, CH-8093 Zürich, Switzerland.

Correspondence to R.A. and B.K.
e-mails: aebersold@imsb.biol.ethz.ch; bernhard.kuster@cellzome.com

doi:10.1038/nrm1683
Published online 15 June 2005

1. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
2. Patterson, S. D. & Aebersold, R. H. Proteomics: the first decade and beyond. *Nature Genet.* **33**, 311–323 (2003).
3. Steen, H. & Mann, M. The abc's (and xyz's) of peptide sequencing. *Nature Rev. Mol. Cell Biol.* **5**, 699–711 (2004).
4. Andersen, J. S. *et al.* Directed proteomic analysis of the human nucleolus. *Curr. Biol.* **12**, 1–11 (2002).
5. Blagoev, B. *et al.* A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling. *Nature Biotechnol.* **21**, 315–318 (2003).
6. Bouwmeester, T. *et al.* A physical and functional map of the human TNF-α/NF-κB signal transduction pathway. *Nature Cell Biol.* **6**, 97–105 (2004).
7. Brand, M. *et al.* Dynamic changes in transcription factor complexes during erythroid differentiation revealed by quantitative proteomics. *Nature Struct. Mol. Biol.* **11**, 73–80 (2004).
8. Rosenblatt, K. P. *et al.* Serum proteomics in cancer diagnosis and management. *Annu. Rev. Med.* **55**, 97–112 (2004).
9. Liotta, L. A., Ferrari, M. & Petricoin, E. Clinical proteomics: written in blood. *Nature* **425**, 905 (2003).
10. Shiio, Y. *et al.* Quantitative proteomic analysis of Myc oncoprotein function. *EMBO J.* **21**, 5088–5096 (2002).
11. Adams, M. D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651–1656 (1991).
12. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
13. LaBaer, J. & Ramachandran, N. Protein microarrays as tools for functional proteomics. *Curr. Opin. Chem. Biol.* **9**, 14–19 (2005).
14. Desiere, F. *et al.* Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **6**, R9 (2005).
15. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383–5392 (2002).
16. Craig, R., Cortens, J. P. & Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **3**, 1234–1242 (2004).
17. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.* **17**, 994–999 (1999).
18. Schirle, M., Heurtier, M. A. & Kuster, B. Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography–tandem mass spectrometry. *Mol. Cell. Proteomics* **2**, 1297–1305 (2003).
19. Washburn, M. P., Wolters, D. & Yates, J. R. III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.* **19**, 242–247 (2001).
20. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
21. Kawashima, S., Ogata, H. & Kanehisa, M. AAindex: amino acid index database. *Nucleic Acids Res.* **27**, 368–369 (1999).
22. Pan, S. *et al.* High-throughput proteome-screening for biomarker detection. *Mol. Cell. Proteomics* **4**, 182–190 (2005).
23. Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W. & Gygi, S. P. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl Acad. Sci. USA* **100**, 6940–6945 (2003).
24. Lu, Y., Bottari, P., Turecek, F., Aebersold, R. & Gelb, M. H. Absolute quantification of specific proteins in complex mixtures using visible isotope-coded affinity tags. *Anal. Chem.* **76**, 4104–4111 (2004).
25. Lipton, M. S. *et al.* Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc. Natl Acad. Sci. USA* **99**, 11049–11054 (2002).
26. Li, X. J. *et al.* A tool to visualize and evaluate data obtained by liquid chromatography–electrospray ionization-mass spectrometry. *Anal. Chem.* **76**, 3856–3860 (2004).

Competing interests statement
The authors declare no competing financial interests.

Online links

FURTHER INFORMATION
International Protein Index: http://www.ebi.ac.uk/IPI
PeptideAtlas: http://www.peptideatlas.org
PeptideProphet: http://sourceforge.net/projects/peptideprophet
RefSeq: http://www.ncbi.nlm.nih.gov/RefSeq
Seattle Proteome Center: Proteomics Tools: http://tools.proteomecenter.org/software.php
Sigma–Aldrich: http://www.sigmaaldrich.com
Swiss-Prot: http://www.expasy.org/sprot
Thermo Electron Corporation: http://www.thermo.com
Access to this interactive links box is free online.

OPINION

The actin cytoskeleton: a key regulator of apoptosis and ageing?

Campbell W. Gourlay and Kathryn R. Ayscough

Abstract | Evidence from many organisms has shown that the accumulation of reactive oxygen species (ROS) has a detrimental effect on cell well-being. High levels of ROS have been linked to programmed cell death pathways and to ageing. Recent reports have implicated changes to the dynamics of the actin cytoskeleton in the release of ROS from mitochondria and subsequent cell death.

The study of pathways that trigger apoptosis and ageing is crucial for increasing our understanding of these fundamental processes. Whereas apoptosis is probably triggered by a range of both extracellular and intracellular factors, its morphological characteristics are notably consistent. From yeast to human cells, the hallmarks of this programmed cell death pathway include chromatin condensation, degeneration of mitochondrial membrane potential and exposure of phosphatidylserine at the plasma membrane[1,2] (BOX 1).

Crucial to our current understanding of apoptosis is the role of the mitochondrion. Both the release of cytochrome *c* and the production of reactive oxygen species (ROS) from mitochondria are involved in apoptosis in yeast and mammalian systems. Despite evidence for a role of the accumulation of ROS in both apoptosis and ageing, relatively few studies have explored the relationship between the pathways that culminate in cell death in these two processes.

Data from yeast indicate that cell death as a result of ageing probably occurs through an apoptotic pathway[3–5], but this has yet to be definitively shown in higher eukaryotes. Given the level of evolutionary conservation between yeast and mammalian cells in many other fundamental pathways — from membrane trafficking to cell-cycle control — it might be expected that cell death pathways will be conserved. However, despite a substantial amount of data, the idea that single-celled organisms can be programmed to trigger their own demise seems to be an anathema to many researchers. Mounting evidence is, however, driving the acceptance of the idea that apoptosis and ageing occur in yeast, and even suggesting that studies in yeast might reveal previously unknown pathways that are involved in these processes.

The actin cytoskeleton functions in the generation and maintenance of cell morphology and polarity, in endocytosis and intracellular trafficking, and in contractility, motility and cell division. The assembly and disassembly of actin filaments, as well as their organization into functional higher-order networks, are regulated by several actin-binding proteins, many of which are conserved from yeast to humans. A key feature of actin is its ability to bind and hydrolyse ATP. The transition from ATP–actin to ADP-bound actin is accompanied by a conformational change that is crucial to the dynamic turnover of actin filaments. Recent studies from us and from other groups have indicated that changes in the dynamic state of actin — caused by direct mutation of actin, addition of actin-binding drugs, or by mutations in actin-binding proteins — can change cell fate. Increased turnover of filamentous (F)-actin can promote cell longevity, whereas decreased actin turnover seems to trigger cell death through an apoptosis-like pathway.

Here, we explore the evidence from several eukaryotic cells that implicates the actin cytoskeleton as a physiological regulator of ROS release from mitochondria and as a key element in the upstream activation of cell death pathways.

Respiration and ROS production

Superoxide (O_2^-) is formed when an oxygen molecule accepts an additional electron. This occurs regularly, through a non-enzymatic mechanism, as a by-product of the mitochondrial electron transport chain. Complexes I and III of the chain are the main sites of O_2^- production as a result of donation from