THE ABC'S (AND XYZ'S) OF PEPTIDE SEQUENCING

Hanno Steen and Matthias Mann‡*

Abstract | Proteomics is an increasingly powerful and indispensable technology in molecular cell biology. It can be used to identify the components of small protein complexes and large organelles, to determine post-translational modifications and in sophisticated functional screens. The key — but little understood — technology in mass-spectrometry-based proteomics is peptide sequencing, which we describe and review here in an easily accessible format.

**Department of Systems Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA.*
‡Department of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark.
Correspondence to M.M.
e-mail: mann@bmb.sdu.dk
doi:10.1038/nrm1468

To sequence a protein ten years ago, a substantial amount had to be purified and a technique known as Edman degradation had to be used. This method, which was developed by Peer Edman, relies on the identification of amino acids that have been chemically cleaved in a stepwise fashion from the amino terminus of the protein and requires much expertise. Often no sufficiently long or unambiguous peptide sequence could be assigned and the method failed completely if the protein was acetylated at its amino terminus or was otherwise blocked to the Edman reaction, which requires a free amino terminus. During the 1990s, mass spectrometry (MS), in which biomolecules are ionized and their mass is measured by following their specific trajectories in a vacuum system, displaced Edman degradation, because it is much more sensitive and can fragment the peptides in seconds instead of hours or days[1]. Furthermore, MS does not require proteins or peptides to be purified to homogeneity and has no problem identifying blocked or otherwise modified proteins. In the last few years, further breathtaking technological advances have established MS not only as the definitive tool to study the primary structure of proteins, but also as a central technology for the field of proteomics (for recent proteomics reviews, see REFS 2–6).

Protein MS facilities have proliferated and many biologists now have access to a service to which they can submit a sample and are handed back a list of proteins that have been identified by MS. This arrangement frequently works quite well for the identification of single spots or bands, but it is our experience that biologists generally do not have the necessary background to critically interpret the results of more challenging MS experiments — in particular, the many kinds of advanced proteomic screens that are now possible. Often, the results are over interpreted. For example, a researcher might focus on the presence of an interesting signalling protein among several identified proteins in a stained gel band, even if this protein is, at best, a minor component and cannot possibly be the one that caused the gel-band staining. Similarly, long lists of proteins that are identified in proteomics experiments are published and biological conclusions are drawn when there is insufficient confidence in these identifications. These problems could be avoided if the scientists wishing to use proteomics had a solid understanding of the principal issues that are involved in peptide analysis by MS. As it is difficult and time consuming to learn about this subject from the technical literature, and most reviews have other goals, in this article we explain the principles of peptide sequencing that are important for the appreciation and interpretation of the outcome of proteomics experiments.

In the main part of this review, we describe the steps of a typical proteomic experiment (FIG. 1) and we use boxes to explain peptide ionization, peptide fragmentation, how peptides are identified by peptide-database-searching algorithms and how to judge the reliability of a peptide hit (see below). The reader is referred to other reviews for more technical detail and specialized topics (for example, see REFS 7–15).

Why are peptides, and not proteins, sequenced?

After protein purification, the first step is to convert proteins to a set of peptides using a sequence-specific

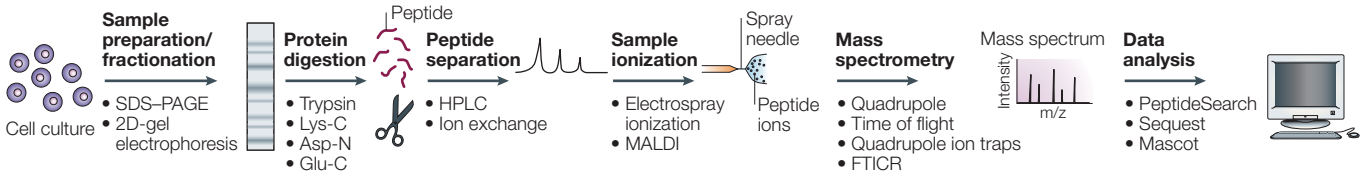


Figure 1 | **The mass-spectrometry/proteomic experiment.** A protein population is prepared from a biological source — for example, a cell culture — and the last step in protein purification is often SDS–PAGE. The gel lane that is obtained is cut into several slices, which are then in-gel digested. Numerous different enzymes and/or chemicals are available for this step. The generated peptide mixture is separated on- or off-line using single or multiple dimensions of peptide separation. Peptides are then ionized by electrospray ionization (depicted) or matrix-assisted laser desorption/ionization (MALDI) and can be analysed by various different mass spectrometers. Finally, the peptide-sequencing data that are obtained from the mass spectra are searched against protein databases using one of a number of database-searching programmes. Examples of the reagents or techniques that can be used at each step of this type of experiment are shown beneath each arrow. 2D, two-dimensional; FTICR, Fourier-transform ion cyclotron resonance; HPLC, high-performance liquid chromatography.

protease (FIG. 1). Even though mass spectrometers can measure the mass of intact proteins, there are a number of reasons why peptides, and not proteins, are analysed in proteomics. Proteins can be difficult to handle and might not all be soluble under the same conditions (it should be noted here that many detergents interfere with MS, because they ionize well and are in a huge excess relative to the proteins). In addition, the sensitivity of the mass spectrometer for proteins is much lower than for peptides, and the protein might be processed and modified such that the combinatorial effect makes determining the masses of the numerous resulting isoforms impossible. Furthermore, it is not easy to predict from the sequence what the mass of a mature, correctly modified protein will be or, conversely, which protein might have given rise to a measured protein mass. Most importantly, if the purpose is to identify the protein, sequence information is needed and the mass spectrometer is most efficient at obtaining sequence information from peptides that are up to ~20 residues long, rather than from whole proteins. Nevertheless, with very specialized equipment, it is becoming possible to derive partial sequence information from intact proteins, which can then be used for identification purposes or the analysis of protein modifications in an approach called 'top-down' protein sequencing[16–19].

With very few exceptions, trypsin is used to convert proteins to peptides. Trypsin is an aggressive and stable protease, which very specifically cleaves proteins on the carboxy-terminal side of arginine and lysine residues. This creates peptides both in the preferred mass range for sequencing and with a basic residue at the carboxyl terminus of the peptide. Such peptides result in information-rich, and easily interpretable, peptide-fragmentation spectra (see below). The endoprotease Lys-C is even more stable than trypsin and is often used before trypsin digestion under harsh, solubilizing conditions such as 8 M urea. Asp-N and Glu-C are also highly sequence-specific (but less active) proteases, which can be used to generate peptides that are complementary to the tryptic peptides. Less sequence-specific proteases are generally avoided because they divide the peptide signal into many overlapping species and generate unnecessarily complex mixtures.

Digesting the protein into a set of peptides also means that the physico-chemical properties of the protein, such as solubility and 'stickiness', become irrelevant. As long as the protein generates a set of peptides, at least some of them can be sequenced by the mass spectrometer, even if the protein itself would have been unstable or insoluble under the conditions used. It is for this reason that membrane proteins are quite amenable to MS-based proteomics. By contrast, these proteins are very difficult to work with in many other areas of protein science, because of their insolubility. However, it should be noted that the improved sequencing properties and detection efficiencies of peptide versus protein analysis by MS are achieved at the expense of sequence coverage — that is, only a low percentage of the entire sequence is analysed, which is sufficient for protein identification but not for complete protein characterization (for example, post-translational modifications, protein processing and truncations are not determined).

How do peptides get into the mass spectrometer?

The peptides that are generated by protein digestion are not introduced to the mass spectrometer all at once. Instead, they are injected onto a MICROSCALE CAPILLARY HIGH-PERFORMANCE LIQUID CHROMATOGRAPHY (HPLC) COLUMN that is directly coupled to, or is 'on-line' with, the mass spectrometer (FIG. 2). The peptides are eluted from these columns using a solvent gradient of increasing organic content, so that the peptide species elute in order of their hydrophobicity. Very hydrophilic peptides, however, might be poorly retained on the column and elute immediately, and extremely hydrophobic peptides might not elute at all when a standard gradient is used. As the mass spectrometer can distinguish the peptides by their masses, there is no need to separate them into non-overlapping chromatographic peaks and usually many peptides arrive at the end of the column at any given time. The signal intensity in the mass spectrum is directly proportional to the analyte concentration, so the peptides are eluted in as small a volume as possible. This is achieved by making the chromatographic column as small as can be packed uniformly and kept free of plugging, which is usually between 50–150 m in inner diameter. Such columns can be loaded with

MICROSCALE CAPILLARY HPLC COLUMN
High-performance liquid chromatography (HPLC) columns have inner diameters of 50–150 m and a reversed-phase stationary phase. Reversed phase means that the surface is made using long hydrophobic alkyl chains, so they retain hydrophobic compounds better than hydrophilic ones.

m/z RATIO
(mass-to-charge ratio). Mass spectrometry does not measure the mass of molecules, but instead measures their m/z value. Electrospray ionization, in particular, generates ions with multiple charges, such that the observed m/z value has to be multiplied by z and corrected for the number of attached protons (which equals z) to calculate the molecular weight of a particular peptide.

QUADRUPOLE MASS SPECTROMETER
A mass-selective 'quadrupole section' only allows the passage of ions that have a specific mass to charge (m/z) value by applying a particular sinusoidal potential. Stepping through the m/z range by applying different potentials and detecting the ions that pass through at each m/z value generates the mass spectrum.

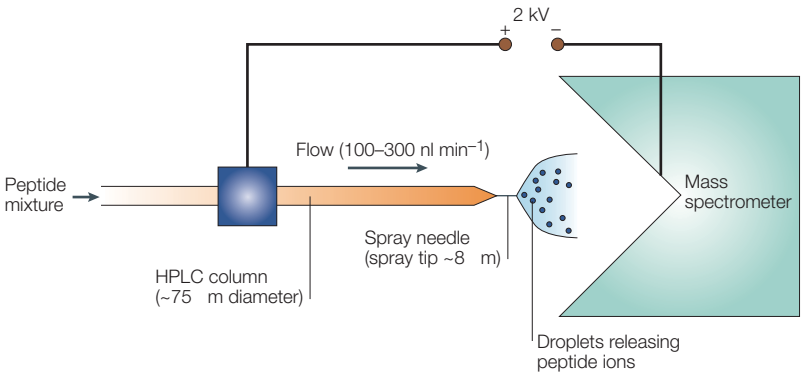


Figure 2 | **The liquid-chromatography–tandem-mass-spectrometry experiment.** A peptide mixture that has been generated by protein digestion is de-salted, concentrated and loaded onto a microscale capillary, high-performance liquid chromatography (HPLC) column using an autosampler (not shown). The peptides are ionized by electrospray ionization at the end of the capillary column. The electrospray plume is generated at atmospheric pressure in close proximity to the entrance of the mass spectrometer and, from here, the peptides are transferred into the vacuum of the instrument for further analysis.

low g amounts of total peptide and they allow flow rates on the order of 100 nl min^{-1}. The width of each peptide peak should be between 10 and 60 seconds. However, these miniaturized chromatography systems still require considerable expertise to operate. Sensitivity can be easily lost due to sub-optimal chromatography or inefficient autosampler set-up.

When a peptide species arrives at the end of the column, it flows through a needle. At the needle tip, the liquid is vaporized and the peptide is subsequently ionized by the action of a strong electric potential. This process is called 'electrospray ionization'[20] (BOX 1; see also The Nobel Prize in Chemistry 2002 in the online links box).

The single dimension of peptide separation that is provided by an HPLC column might not provide sufficient resolution if highly complex protein mixtures are analysed. In this case, the proteins can be divided into fractions and digested separately, which produces less complex peptide mixtures. Protein mixtures are often separated by SDS–PAGE and the whole lane of the gel can be excised into equally sized slices, so that the proteins in each gel slice can be analysed separately by HPLC–MS (also just called liquid-chromatography–MS or LC–MS)[21]. Advantages of this so-called 'GeLC–MS' approach include the fact that the apparent molecular weight of the proteins is known, which provides information about protein processing or modification. Furthermore, the analysis is subdivided into several independent analysis runs, which increases confidence in database identifications and the dynamic range of the measurement (the difference between the most abundant and least abundant proteins that can be identified in an experiment)[22].

As an alternative to protein fractionation, peptide mixtures can be separated in two dimensions. For example, a strong cation exchange column can be used to separate the peptides — first on the basis of their charge, and then on the basis of their hydrophobicity. This technique is known as multidimensional protein-identification technology (MudPIT)[23].

For simplicity, we focus on electrospray ionization in this article. However, biomolecules can also be ionized by matrix-assisted laser desorption/ionization (MALDI[24]; BOX 1). Although MALDI does not allow direct 'on-line coupling' to HPLC, LC fractions can be deposited in series on a metal target before automated analysis and some modern MALDI mass spectrometers are capable of peptide fragmentation as well as peptide-mass measurement.

What happens inside the mass spectrometer?

Electrosprayed peptide ions enter the mass spectrometer through a small hole or a transfer capillary. Once inside the vacuum system, they are guided and manipulated by electric fields. There are diverse types of mass spectrometer, which differ in how they determine the mass-to-charge (m/z) RATIOS of the peptides. Three main types of mass spectrometers are used in proteomics: QUADRUPOLE MASS SPECTROMETERS, TIME OF FLIGHT (TOF) MASS SPECTROMETERS and QUADRUPOLE 'ION TRAPS'. In addition, there are also mass spectrometers that combine principles, such as the popular quadrupole–TOF mass spectrometer. Each of these instruments generates a mass spectrum, which is a recording of the signal intensity of the ion at each value of the m/z scale (which has units of DALTONS (Da) per charge).

In the electrospray-ionization process, tryptic peptides usually become doubly protonated and are then designated $(M + 2H)^{2+}$, in which M is the mass of the peptide and H^+ is the mass of a proton. So, as mass spectrometers measure the m/z value, a peptide with mass of 1232.55 would be seen at (1232.55 + (2 x 1.0073))/2 = 617.28 in the mass spectrum (see the highlighted peak in FIG. 3b). Peptides can also have higher charge states if they are more than 15 amino acids long or contain further basic amino acids such as histidine, which can also be protonated. Fortunately, it is easy to determine the charge state because each peptide signal actually consists of an isotope cluster of peaks. Such peaks are separated by 1 Da, which is caused by the 1% probability of each carbon atom being the ^{13}C isotope instead of the usual ^{12}C atom (see FIG. 3b, inset). For example, if the first ^{13}C isotope peak had a difference from the ^{12}C monoisotopic peak of 1 unit on the m/z scale (618.28 without a peak at 617.78 in this example), then the charge state of the peptide ion that produced this peak cluster would be 1. However, as the difference between the first and second peaks in this example (617.28 and 617.78) is 0.5 units, the charge state of the peptide ion that produced this peak cluster must be 2.

Close inspection of the separation of any of the peptide isotope peaks reveals the resolution of the mass spectrometer. Ion traps barely resolve the isotopes of doubly charged species, whereas TOF instruments with a resolution of 10,000 (that is, the m/z value divided by the peak width at half height) show a clear baseline separation for the isotopes of even highly charged species. However, the ultimate resolution is provided by 'Fourier-transform ion cyclotron resonance' MS

TIME OF FLIGHT (TOF) MASS SPECTROMETER
This mass analyser is based on the time it takes ions to travel through an electric-field-free flight tube. In the ion source, all the ions are accelerated to the same kinetic energy. As kinetic energy is a function of mass, the lighter ions fly faster than the heavier ones and therefore reach the detector sooner.

QUADRUPOLE 'ION TRAPS'
In ion traps, the ions are first caught (trapped) in a dynamic electric field and are then sequentially — according to their mass to charge (m/z) value — ejected onto the detector with the help of another electric field. Trapped ions can also be isolated and fragmented within the trap.

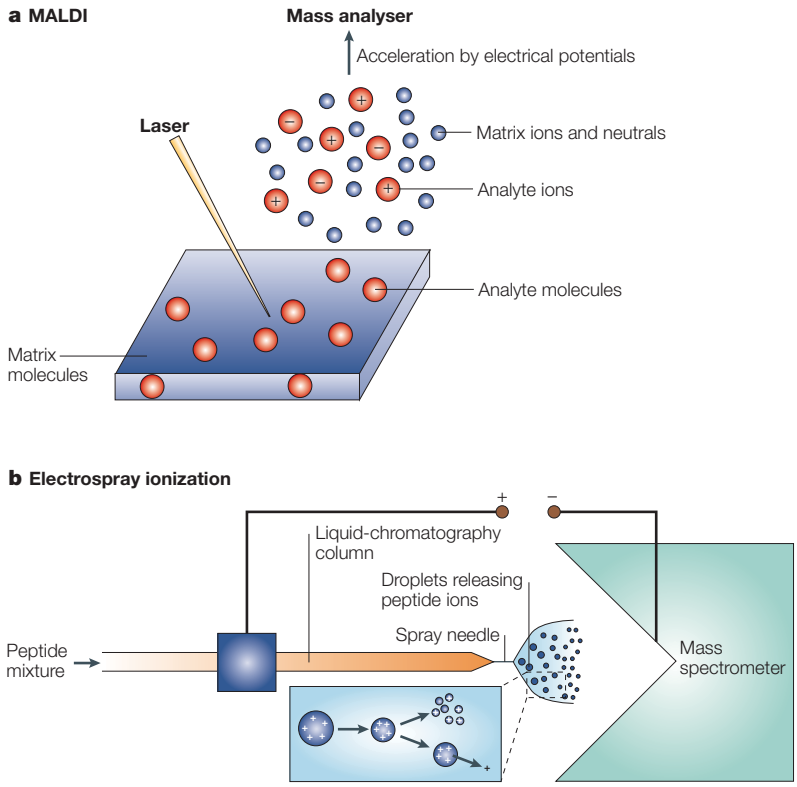
DALTON
(Da). The unit of the mass scale, which is defined as one twelfth of the mass of the mono-isotopic form of carbon, ^{12}C (1 Da = 1.6605 x 10^{-27} kg). Other commonly, but not necessarily correctly, used units of relevance to mass spectrometry are the amu (an atomic mass unit that is based on ^{16}O), the Thomson (the proposed unit for the mass to charge (m/z) scale) and the u ('unit', which is the same as Da).

(FTICR–MS or FTMS). In contrast to the quadrupole ion trap, these so-called 'Penning traps' keep the ions confined in the high magnetic field of a super-conducting magnet, so FTMS can be thought of as the MS analogue of NMR. The ions circle with frequencies that are inversely proportional to their m/z value. This circling induces an alternating current in the metal plates that make up the trap. This time-varying current constitutes a frequency spectrum of the ion motion and is converted by the mathematical operation Fourier transformation — which explains the name — into a mass spectrum. The high resolution (more than 100,000) and the mass accuracy (a few parts per million) of FTMS are due to the fact that the spectrum is acquired as a frequency measurement and frequencies can be measured exceedingly accurately.

Having determined the m/z values and the intensities of all the peaks in the spectrum, the mass spectrometer then proceeds to obtain primary structure (sequence) information about these peptides. This is called tandem MS, because it couples two stages of MS. In tandem MS, a particular peptide ion is isolated, energy is imparted by collisions with an inert gas (such as nitrogen molecules, or argon or helium atoms), and this energy causes the peptide to break apart. A mass spectrum of the resulting fragments — the tandem MS (also called MS/MS or MS^2) spectrum — is then generated (FIG. 3c). In MS jargon, the species that is fragmented is called the 'precursor ion' and the ions in the tandem-MS spectrum are called 'product ions' (more endearingly, but less politically correct, they used to be described as parent and daughter ions). Note that the MS^2 spectrum is the result of an ensemble of one particular precursor ion fragmenting at different amide bonds. Throughout the chromatographic run, the instrument will cycle through a sequence that consists of obtaining a mass spectrum followed by obtaining tandem mass spectra of the most abundant peaks that were found in this spectrum.

BOX 2 explains how peptides fragment and how the fragment ions are designated. The most common and informative ions are generated by fragmentation at the amide bond between amino acids. The resulting ions are called b-ions if the charge is retained by the amino-terminal part of the peptide and y-ions if the charge is retained by the carboxy-terminal part. In quadrupole or quadrupole–TOF instruments, y-ions predominate, whereas in ion-trap instruments, b- and y-ions are both observed. For an even more in-depth characterization, the peptide fragments can be further fragmented. This is known as MS^3 or, more generally, MS^n, and has recently become feasible practically in proteomics with the advent of linear ion traps — a new and improved version of the traditional three-dimensional quadrupole ion traps, in which more precursor ions can be stored for fragmentation.

Box 1 | Ionization methods



A fundamental problem in biological mass spectrometry was how to transfer highly polar, completely non-volatile molecules with a mass of tens of kDa into the gas phase without destroying them. This was solved by so-called 'soft' ionization techniques such as matrix-assisted laser desorption/ionization (MALDI)[68] and electrospray ionization[20]. The latter technique earned its inventor a share of the Nobel Prize for chemistry in 2002 (see further information in the online links box).

For MALDI (see figure, part a), the analyte is mixed with a large excess of ultraviolet-absorbing matrix, which is normally a low-molecular-weight aromatic acid. On irradiation with a focused laser beam of the appropriate wavelength, the excess matrix molecules sublime and transfer the embedded non-volatile analyte molecules into the gas phase. After numerous ion–molecule collisions in the plume of ions and molecules, singly protonated analyte ions are formed, which are accelerated by electric potentials into a mass analyser of choice.

For electrospray ionization (see figure, part b), the tapered end of a liquid-chromatography column or a metal needle is held at a high electrical potential (several kV) with respect to the entrance of the mass spectrometer. The liquid effluent containing the peptides that are eluting from the chromatography column is thereby electrostatically dispersed. This generates highly charged droplets, which are normally positively charged in proteomics experiments, due to an excess of protons. Once the droplets are airborne, the solvent evaporates, which decreases the size and increases the charge density of the droplets. Desolvated ions are generated by the desorption of analyte ions from the droplet surface due to high electrical fields and/or the formation of very small droplets due to repetitive droplet fission until each droplet contains, on average, only one analyte ion.

Peptide sequencing by mass spectrometry

As shown in FIG. 3c, each peptide fragment in a series differs from its neighbour by one amino acid. In principle, it is therefore possible to determine the amino-acid sequence by considering the mass difference between neighbouring peaks in a series, as is shown in FIG. 3. However, the difficulty lies in the fact that the information in tandem-MS spectra is often not complete and that intervening peaks, which might or might not belong to the series, can confuse the analysis. For

example, a mass difference of 114 Da might be found between two large peaks, but a very small peak might also be found at 57 Da between these two large peaks. This part of the spectrum could therefore correspond to one asparagine (residue mass = 114 Da) or two glycines (residue mass = 57 Da). In practice, experts can correctly interpret at least parts of tandem-MS spectra, whereas computer algorithms are, as yet, unreliable for determining amino-acid sequences. In either case, the success of *DE NOVO* SEQUENCING crucially depends on the quality of the data, in terms of both the mass accuracy and the resolution of the instrument, as well as the information content of the tandem-MS spectrum (for details of *de novo* sequencing software, see online supplementary information S1 (box)).

At the beginning of the 1990s, researchers realized that the peptide-sequencing problem could be converted to a database-matching problem, which would be much simpler to solve. The reason database searching is easier than *de novo* sequencing is that only an infinitesimal fraction of the possible peptide amino-acid sequences actually occur in nature. A peptide-fragmentation spectrum might therefore not contain sufficient information to unambiguously derive the complete amino-acid sequence, but it might still have sufficient information to match it uniquely to a peptide sequence in the database on the basis of the observed and expected fragment ions. There are several different algorithms that are used to search sequence databases with tandem-MS-spectra data, and they have names such as PeptideSearch, Sequest, Mascot, Sonar ms/ms and ProteinProspector (for more information, see BOX 3; see also the information on protein-identification software in online supplementary information S1 (box)). A limitation of database searching compared to *de novo* sequencing is that large-scale proteomic experiments should only be carried out using organisms that have had their genome sequenced, so that all the possible peptides are known. Organisms in which expressed-sequence-tag projects have been carried out are accessible to a lesser degree, as are the proteomes of organisms that have genes with a high homology to the proteome of a sequenced species[25–27].

Identifying proteins in small data sets

As mentioned above, proteins are often purified by one- or two-dimensional gel electrophoresis and mass spectrometry is carried out on stained bands or spots. When MALDI is used, proteins are identified by 'mass fingerprinting', which matches the tryptic peptide masses in the mass spectrum to the calculated tryptic peptide masses for each protein in a database. Although MALDI fingerprinting works well in many cases, peptide sequencing is a more specific and sensitive identification method. In this procedure, peptides are sequenced using the LC–tandem-MS experiment and each tandem mass spectrum is database searched using one of the algorithms that are described in BOX 3. Peptide identifications should be reported in terms of a probability score, as is the case for the Mascot search engine[28] and a recently modified version of the Sequest algorithm[29].

Database-searching software packages usually also indicate the scores that are considered significant (for further information on the statistical treatment of peptide-sequencing results, see online supplementary information S1 (box)). If several statistically significant

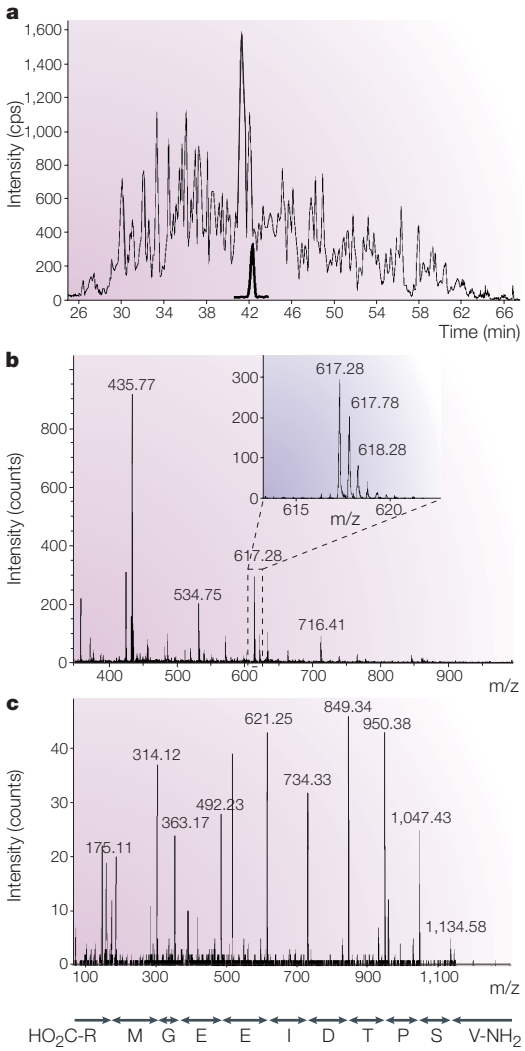


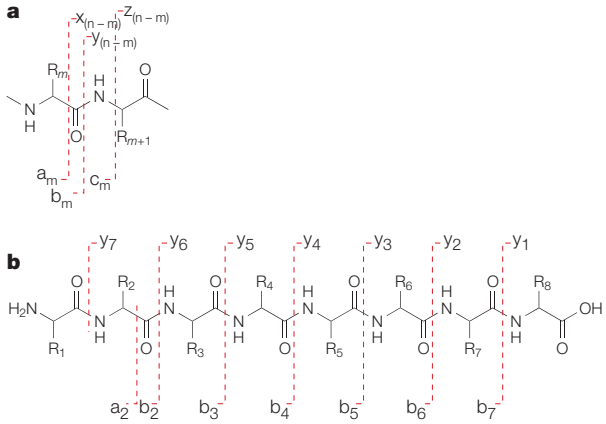
Figure 3 | **Mass-spectrometry traces. a** | The total ion intensity from all the mass spectra that were recorded during the liquid-chromatography–mass-spectrometry (MS) run is shown as a function of elution time — that is, the black trace shows the TOTAL ION CURRENT or total ion chromatogram. Shown in bold is the trace for the intensity of one particular ion, which elutes within a 40-second window approximately 42.5 minutes into the gradient — that is, the bold trace shows an EXTRACTED ION CURRENT or extracted ion chromatogram. The area under this curve represents the total signal of this peptide. **b** | The mass spectrum of the peptides that were eluted 42.4 minutes into the gradient. The insert shows the mass-to-charge values around the peptide ion of interest, which are indicative of the resolution and allow the charge state to be derived (please refer to the main text for further details). **c** | The tandem-MS (MS/MS) spectrum of the peptide ion of interest (highlighted by a dashed box in part **b**). The mass differences between this y-ion series indicate the amino-acid series, which is shown below the spectrum. As this is a y-ion series, the sequence is written in the carboxy-to-amino-terminus direction going from left to right. m/z, mass-to-charge ratio.

DE NOVO SEQUENCING
Deriving the amino-acid sequence (primary structure) of a peptide solely from the mass-spectrometry, peptide-fragmentation data (that is, without using databases).

TOTAL ION CURRENT
The sum of all the ion signals in a mass spectrum as a function of elution time.

EXTRACTED ION CURRENT
The sum of the ion signal for a particular mass to charge (m/z) value — that is, for a particular peptide-ion species.

Box 2 | The abc's (and xyz's) of peptide sequencing



Part **a** of the figure shows the chemical structure of a peptide, together with the designation for fragment ions (the Roepstorff–Fohlmann–Biemann nomenclature) that is used when the peptide backbone is fragmented by imparting energy onto the molecule[69,70]. In the mass spectrometers that are used in proteomics, peptide fragmentation is induced by collisions with residual gas, and bond breakage mainly occurs through the lowest energy pathways — that is, cleavage of the amide bonds. This leads to b-ions when the charge is retained by the amino-terminal fragment or y-ions when it is retained by the carboxy-terminal fragment (see figure, part **b**). The fragmentation process has recently been modelled quantitatively[71].

Ions are labelled consecutively from the original amino terminus a_m, b_m and c_m, in which m represents the number of amino-acid R groups these ions contain. They are also labelled consecutively from the original carboxyl terminus $z_{(n-m)}$, $y_{(n-m)}$ and $x_{(n-m)}$, in which n – m equals the number of R groups these ions contain (n is the total number of residues, or R groups, in the peptide and m is the number of R groups that the corresponding a-, b- or c-ion would contain; see figure). Doubly charged tryptic peptides mainly yield singly charged y- and b-ions. In addition, a-ions (loss of a C=O group or a mass difference of 27.9949 Da relative to the b-ion) can occur, but this is normally only observed for the b_2-ion, which gives rise to the characteristic a_2/b_2-fragment ion pair in the lower mass range[72] (see figure, part **b**). Apart from the ion types shown, 'satellite' fragment ions due to the further loss of NH_3 or H_2O can be produced. These ions are designated, for example, $a_m - NH_3$ or $y_{n-m} - H_2O$. Fragmentation both amino-terminal to and carboxy-terminal of the same amino acid produces immonium ions, which are diagnostic of modified amino acids such as phosphotyrosine and/or hydroxyproline[73].

peptides identify the same protein, then this protein identification can be accepted without further work. However, this is not the case if the sum of many marginal peptide scores results in a seemingly significant protein score, which is a problem that frequently results in the erroneous identification of very large proteins that could produce a large number of potential peptides. Furthermore, generally only 'fully tryptic' peptides should be used in the database search — that is, peptides in which the carboxy-terminal amino acid is arginine or lysine and for which the amino acid that precedes the peptide in the protein sequence is arginine or lysine. Trypsin seems to be fully specific and only a few 'semi-tryptic' peptides are generated through protein degradation or the breakup of the peptide before tandem MS (REF. 30). Some peptides — in particular, small ones with less than seven amino acids — match more than one protein in the database, and this should be indicated by the search software and taken into account in data interpretation. Great care should be taken with proteins that are identified on the basis of a single peptide identification. If the probability score is very high, such a peptide might be sufficient to identify a protein, provided that the data are of high quality (that is, a high mass accuracy and signal-to-noise ratio). In these and all other cases in which an interesting protein will be further characterized biologically, the mass spectra should be manually inspected according to the rules in BOX 4 before proceeding (see also Supplementary material on peptide validation in the online links box). If other information, such as the apparent molecular weight of the band as determined by gel electrophoresis, does not agree with the protein identified, the raw data should be similarly checked.

Although experimenters usually intend to purify proteins to homogeneity in single electrophoretic bands, it is important to realize that such bands often contain more than one protein. Particular care should be taken if, in a single stained band, the protein of interest is identified by only a few peptides, whereas another protein is identified by many peptides and much stronger MS peaks. It is then probable that the interesting protein is a minor contaminant and did not actually cause the staining that made the band visible. An exception to this is keratin contamination, for example, through dust, hair and wool sweaters, which is caused, for instance, by working without gloves. This contamination can produce stronger peptide peaks than the protein that produced the band.

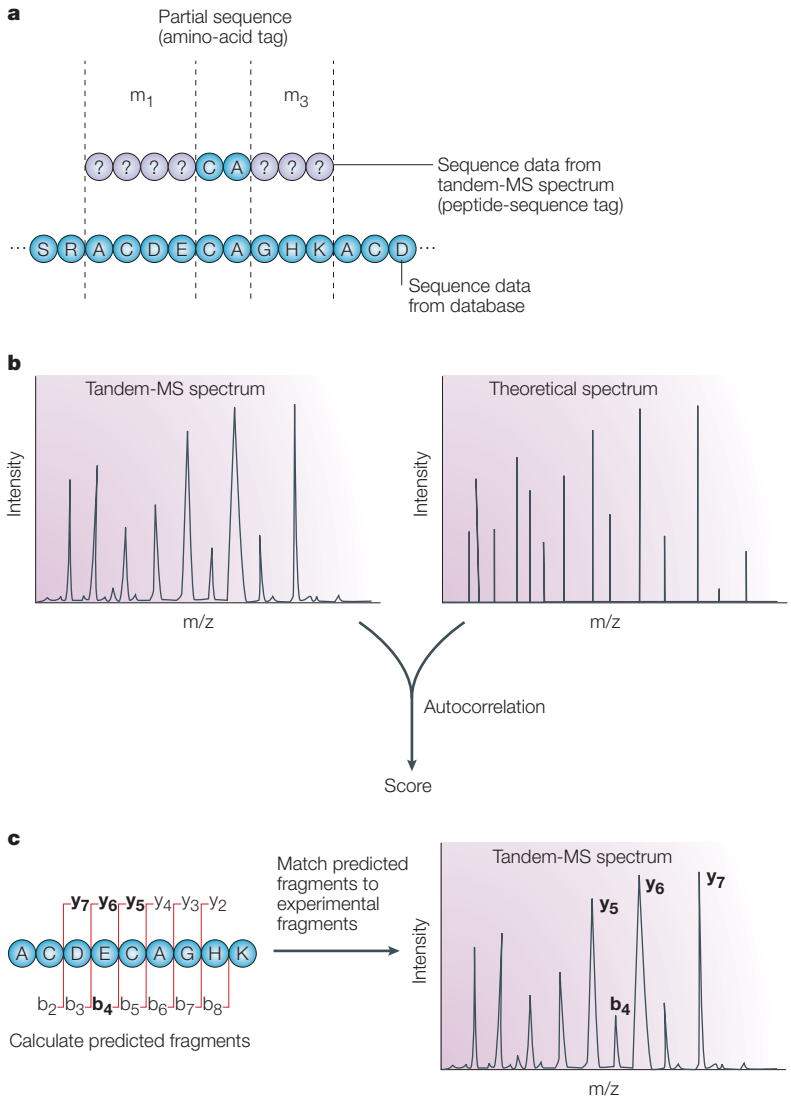
Identifying proteins in large data sets

As a result of rapidly improving technology, the identification of hundreds of proteins is not unusual, even in a single project, and determining the reliability of these protein hits is especially challenging. This is partly because even small error rates for each of the corresponding peptides can quickly add up when many thousands of peptides are being identified. Another reason is that the large-scale nature of an experiment is often used as an excuse not to carry out any critical evaluation of the outcome. As a result, many large-scale proteomic sequencing projects have an unknown, but

apparently quite high, error rate. Simply increasing the required identification scores would reduce the number of misidentifications ('false positives'). However, it would do so at the cost of an increasingly larger fraction of 'false negatives', that is, proteins that were present but which have not been confidently identified. Fortunately, all these problems have recently been addressed by a combination of experiments on defined protein mixtures, randomized databases and the application of robust statistical procedures.

Keller *et al.* used a mixture of recombinant proteins and carried out protein identification using many

Box 3 | **Database identification approaches**



The first algorithm — known as Peptide Sequence Tags, which was first implemented in the programme PeptideSearch — makes use of the fact that fragmentation spectra usually contain at least a small series of easily interpretable sequence[74]. This series constitutes an amino-acid tag. The lowest mass in the series contains information about the distance, in mass units, to one terminus of the peptide, and the highest mass contains information about the distance to the other peptide terminus. Together, the peptide-sequence tag consists of three parts — the amino-terminal mass (see m_1 in part **a** of the figure), a short amino-acid sequence (-C-A- in this example) and the carboxy-terminal mass (m_3). This construct can be matched against sequences in the database and, if desired, the peptide that is identified can be made to comply with the cleavage event of the proteolytic enzyme used (in this example, trypsin, which cleaves carboxy-terminal of arginine or lysine residues).

In a second approach, which is implemented in the Sequest algorithm[75], a signal-processing technique called autocorrelation is used to mathematically determine the overlap between a theoretical spectrum that has been derived from every sequence in the database and the experimental spectrum in question (see part **b** of the figure). The overlap is given in the form of a score, and the score to the next best matching peptide sequence is also often given. The technique has proven quite robust for low signal-to-noise spectra. It is used for low-resolution data because autocorrelation would be too computer intensive for high-resolution data.

The third main approach, which is implemented in the Mascot search engine[28], also involves calculating the theoretically predicted fragments for all the peptides in the database, and it is called probability-based matching. The predicted fragments are matched to the experimental fragments in a top-down fashion, starting with the most intense b- and y-ions (see part **c** of the figure, and note that the most intense peaks do not always result from simple b- and y-ions). The probability that the number of fragment matches is random is calculated and the negative logarithm of this number (multiplied by 10) is the identification score.

Many more bioinformatics approaches to peptide identification have been developed in recent years (for example, one involving a mathematical discipline called graph theory). For more information on protein-identification software, see online supplementary information S1 (box). MS, mass spectrometry; m/z, mass-to-charge ratio.

rounds of LC–tandem-MS (REF. 31). They could then determine the false-positive and false-negative rates of commonly used threshold criteria for peptide identifications. Researchers have also used randomized sequence databases to determine significant database identification scores[32,33]. This experiment can easily be carried out by reversing the sequences of every database entry and searching the tandem mass spectra against this 'nonsense' database. This type of exercise helps to 'tune' the significance criteria to the specific instrument and sample-preparation methods used in particular laboratories.

More generally, Keller and colleagues noticed that plotting peptide-identification scores against their frequency revealed two approximately Gaussian-shaped distributions[34]. The one centred on low score values is caused by random matches, whereas the one centred on higher scores belongs to true matches. Although the two distributions overlap, they can be mathematically fitted by two curves, and this provides a statistical way to obtain a probability for correct identification for each chosen cut-off score[35]. This, in turn, allows a desired cut-off with a known false-positive and false-negative rate to be selected depending on the requirements of the experimental question. Another advantage of this method is that it makes the data between different search engines and different laboratories comparable[36]. All that is necessary is that all peptide-identification scores — not only the ones above a particular minimum — are reported for every large-scale experiment (see also Institute for Systems Biology in the online links box).

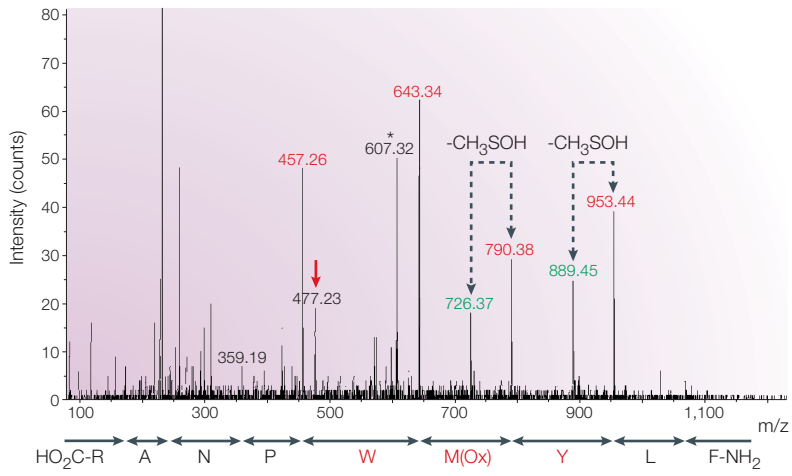
After all the peptides have been identified, they have to be grouped into protein identifications. Usually, the peptide scores are added up to yield protein scores in a straightforward manner. However, the confidence in the accuracy of a particular peptide identification increases if other peptides identify the same protein and decreases if no other peptides do so. This fact can be formulated mathematically and can be used to determine the correct probability of protein identification from the adjusted peptide-identification probability[35]. As already mentioned above, protein identifications based on single peptides should only be allowed in exceptional cases. The use of peptides that are not fully tryptic and the use of single-peptide identifications are by far the greatest causes of false-positive protein identifications.

Despite the desire for unbiased and objective criteria, these need not be applied blindly and exclusively. With high mass accuracy and high-resolution data, data interpretation by experts can add much to the interpretation process, because computers only capture certain aspects of the information in tandem-MS spectra. Furthermore, peptides with low scores are, nevertheless, often correct, so manual validation of such hits can often 'rescue' the identification of important proteins. It is improbable that human MS experience will be superseded, rather than assisted, by machine intelligence in the near future. In our opinion, the best way forward lies in the use of a combination of powerful algorithms, robust statistics and expert knowledge.

Isoforms, protein modifications and more

Proteomics experiments identify proteins on the basis of several sequenced peptides, which might or might not distinguish between all the possible isoforms of the protein. Fortunately, even a single amino-acid substitution in any observed peptide will always lead to a different

Box 4 | Validation of peptide hits



To evaluate ambiguous peptide identification, several rules of thumb can be used to assess whether a particular protein identification is reasonable or not (see also Supplementary material on peptide validation in the online links box). Some of these rules are described using the example of a tandem mass spectrum that was acquired using a quadrupole–time-of-flight instrument and the doubly protonated tryptic peptide FLYM(Ox)WPNAR (see figure; mass-to-charge ratio (m/z) = 607.32; the precursor ion is highlighted by an asterisk; M(Ox), oxidized methionine):

- It is often not possible to rationalize all the fragment ions that are observed in a tandem mass spectrum. However, in the case of doubly charged tryptic peptides, the majority of the most abundant peaks in the m/z range above and around the precursor ion should be indicative of a (short) continuous series of y-type fragment ions (see the fragment ions highlighted by red text that assign the sequence tag -Y-M(Ox)-W-). b-type fragment ions of lower intensity are expected to be present when ion traps have been used for the analysis, or if the peptide comprises an internal basic amino-acid residue.
- Peptide bonds that are amino-terminal to proline, and carboxy-terminal of aspartate, residues are particularly labile — that is, more intense fragment ions are observed for this cleavage compared to those for the cleavage of the preceding and subsequent peptide bonds. In fact, cleavage carboxy-terminal of proline, and amino-terminal to aspartate, is energetically unfavourable. Fragment ions that are derived from the labile cleavage should therefore be much more abundant than those derived from the hampered cleavage (see, for example, the signal intensities at the m/z value of 457.26 versus 359.19 in the figure).
- If a side-chain modification — such as serine/threonine phosphorylation, glycosylation and/or methionine oxidation — is present, fragment ions that comprise this modification can be accompanied by so-called 'satellite ions'. This is a result of the ready loss of modification-specific fragments — for example, phosphoric acid (98 Da) for phosphorylated species or CH_3SOH (64 Da) for oxidized methionine; see the m/z values that are labelled in green in the figure. Depending on how facile this loss is, the satellite ions can be more abundant than the related fragment ion.
- It is often assumed (by researchers and also by some of the database-searching algorithms that are used at present) that fragment ions have a lower charge state than the precursor ion from which they are derived. However, some intense fragment ions that fall in the m/z range below the precursor ion and that have not been accounted for can actually be fragment ions that have the same charge state as the precursor ion (see the fragment ion that is marked with a red arrow; this ion is the doubly protonated form of the fragment ion at the m/z value 953.44).

mass (except in the case of isoleucine/leucine), so even a low sequence coverage of a protein will generally determine which particular protein isoform it is. However, this is not necessarily the case for isoforms that are produced by alternative splicing or by protein processing if the differing protein sequences are not covered by any sequenced peptide.

The sequence databases that are used to identify proteins are still far from optimal for proteomics experiments. Ideally, we would have non-redundant databases with entries for each gene and annotations for all the isoforms, splicing variants and so on. In practice, we have to choose between large databases with much redundancy, such as the Entrez Protein database of the National Center for Biotechnology Information (NCBI; see online supplemenatary information S1 (box)) or compact and minimal databases such as the Unigene database (see online supplementary information S1 (box)). The former will list many apparently, but not actually, different proteins, whereas the latter is still in a state of flux with entries changing significantly or even disappearing as genomes become better annotated. The International Protein Index and Ensembl are other database resources that are extremely valuable for proteomics projects, because they contain not only deposited protein and translated cDNA sequences, but also information on predicted genes on the basis of genomic and expressed-sequence-tag data (see online supplementary information S1 (box)).

Post-translational modifications are generally not considered in the first round of large-scale, protein-sequencing experiments, and the identity of modified proteins can, in principle, easily be determined using any of the non-modified tryptic peptides of the protein. Slightly altered versions of the database-searching algorithms that are described in BOX 3 can deal with modifications, in effect by matching various modified versions of each peptide in the database with the spectrum. However, this is at the expense of a vast increase in the search space and leads to a corresponding decrease in the confidence of identification. On the other hand, fragmentation spectra of modified peptides have special features — for example, a prominent peak due to the loss of a phosphoryl group from phosphorylated serine- or threonine-containing peptides (BOX 4) — and these features can help to verify that a peptide is modified. However, it is difficult to obtain tandem-MS spectra of all modified peptides in the first place, as it requires much more material than is needed for identification, as well as the use of several proteases to achieve close to 100% sequence coverage[12].

Few definitive conclusions can be drawn regarding the absence of a protein from a large-scale analysis. The protein might, in fact, still have been present, but its peptides might have co-eluted with abundant peptides from other proteins and might therefore not have been selected for sequencing by the instrument. However, it might be possible to conclude that the protein in question is not a significant component if none of its peptides has been sequenced, in particular, if all the other proteins in the protein mixture have been identified by two or more peptides.

IONIZATION EFFICIENCY
The fraction of peptides in solution that is converted to peptide ions in the gas phase.

Mass-spectrometry data for quantification

Often, we are interested not only in the identity of a peptide, but also in its quantity. Unfortunately, the intensity of the signal of a peptide ion does not directly indicate the amount of protein present. For example, when digesting a protein, the peptides that are produced should all be equimolar and might be expected to give peaks of equal height in the mass spectrum. However, accessibility to the protease, the solubility of the peptide and the IONIZATION EFFICIENCY of the peptide combine to make these signals orders of magnitude different. Fortunately, these factors are reproducible, so the peak height of the same peptides can be a good indicator of the relative amount of the related protein from one experiment to the next.

Absolute quantification. Averaging the MS response of the most abundant peptides for each protein can yield a rough measure of the absolute amount of each protein — within a factor of four if at least three peptides are taken into account (J. Rappsilber, Y. Ishihama and M.M., unpublished data). A more laborious, but precise, way of achieving absolute quantification is to include isotopically labelled 'internal standards' — that is, known amounts of peptides that mimic an expected proteolytic peptide but have a slightly different mass. The internal standard can be added to the sample before digestion and loading onto the column. This is frequently done for absolute quantification in small-molecule MS and could also be done in proteomics studies on a large scale[37–40].

Relative quantification. The most accurate way of obtaining a relative quantification of two protein populations by MS — for example, populations obtained from different cellular states or under different growth conditions — involves the use stable isotopes (no radioactivity is involved). The key idea is that two forms of a molecule that differ only as a result of stable-isotope substitution will behave identically during an MS experiment; there will just be a mass difference between them. Therefore, the ratio of the two peaks, which can be determined quite accurately, directly indicates the relative amounts of a protein that is present in the two populations. In proteomics, the stable-isotope label needs to be incorporated into the peptides and this can be done in a number of ways (see below). Hydrogen atoms (^{1}H) can be replaced by deuterium (^{2}H), ^{12}C by ^{13}C, and ^{14}N by ^{15}N, which all lead to a one-mass-unit difference per substituted atom, and ^{16}O can be replaced by ^{18}O. At least a three-mass-unit shift per peptide is desirable to separate the two isotope clusters from each other.

In metabolic labelling, which was first used in proteomics by Langen and colleagues[41] and Chait and co-workers[42], stable isotopes are included in the food source of, for example, a microorganism. This is similar to the procedure that is used to label proteins for structural

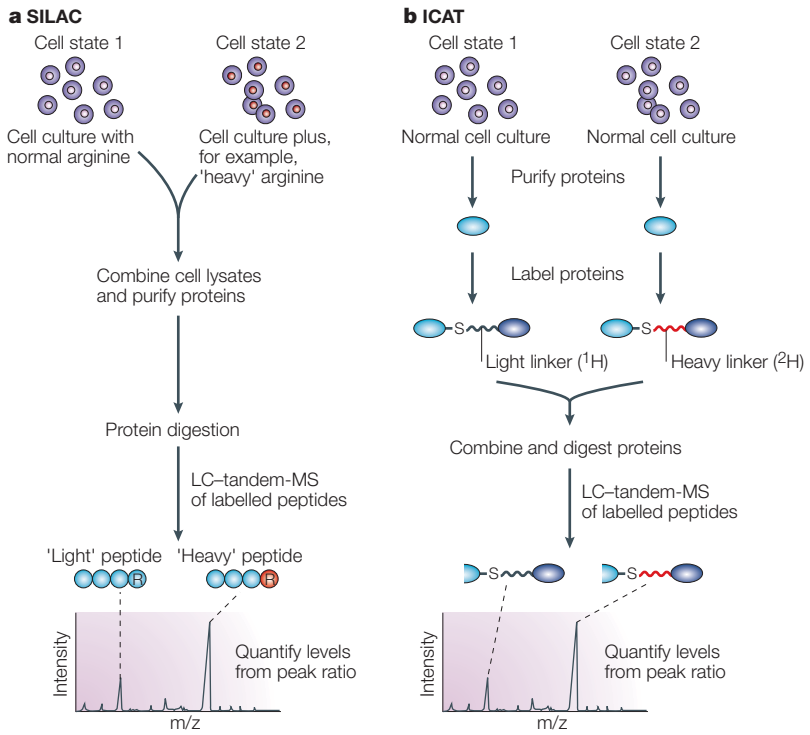


Figure 4 | **Techniques for the relative quantification of protein populations.** **a** | A schematic representation of the SILAC ('stable-isotope labelling in cell culture') method. A stably labelled amino acid in a cell-culture medium (in this case, 'heavy' arginine) is incorporated fully into the proteome of one cell population. Relative quantification experiments can easily be carried out using cells that were grown in normal media as the control state. Cell lysates from two conditions can be combined and purified through any number of steps. The proteins are then digested and if the two forms of the peptides co-elute, a peptide ratio can be obtained for each mass spectrum, which allows the protein levels in the two populations to be relatively quantified. **b** | The ICAT ('isotope-coded affinity tag') method involves the use of a label that is composed of three modular parts — an isotopically labelled linker between a biotin group and a cysteine reactant. Two different isotopically labelled linkers can be used to allow the peptides of two different protein populations from two different cell states to be relatively quantified. The biotin group allows the selective capture and analysis of only the subset of peptides that contain a cysteine residue. LC–tandem-MS, liquid-chromatography–tandem-mass-spectrometry; m/z, mass-to-charge ratio.

biology purposes. It is important that there is 100% labelling, otherwise the degree of labelling introduces another source of uncertainty. We recently described an approach called SILAC ('stable-isotope labelling in cell culture'). This approach is applicable to the cell cultures of higher organisms and uses a stably labelled amino acid in the culture media, which is incorporated fully into the proteome of a cell population[43,44] (FIG. 4a). Relative quantification experiments can easily be carried out using cells grown in normal media as the control state, and cell lysates from two conditions can be combined and purified through any number of steps. If the two forms of the peptides co-elute (which they do if the labelling is done using ^{13}C or ^{15}N), a peptide ratio can be obtained for each mass spectrum by comparing the signal intensities. Alternatively, the intensity of the chromatographic peaks for each form of the peptide (extracted ion current; FIG. 3a) can be determined separately and divided to determine the peptide ratio.

In another approach to isotopic labelling (which can be called 'post-harvest labelling'), protein samples are chemically labelled before or after proteolysis. This can be done by reacting chemical labels with particular amino-acid side chains, such as the thiol group of cysteine residues[45], or by using labels that target all newly formed peptides by reacting with primary amines — that is, the amino termini of the peptides[46]. One popular method called ICAT ('isotope-coded affinity tag') incorporates a label that is composed of three modular parts — an isotopically labelled linker between a biotin group and a thiol-specific (cysteine) reactant[47,48] (FIG. 4b). Two different isotopically labelled linkers are used to compare the peptides of two different protein populations, and the biotin group allows the selective capture and analysis of only the subset of peptides that contain the relatively rare cysteine residue. This makes the peptide mixture less complex, but proteins that lack cysteine residues cannot be quantified.

The advantages of metabolic-labelling methods include the fact that no chemical methods, which can be tedious and can reduce the sensitivity of the measurement, are involved. Furthermore, as indicated above, the cell lysates of two conditions studied can be combined and purified through any number of steps. Post-harvest labelling, on the other hand, requires keeping track of the fractions that are to be quantified against each other (FIG. 4). However, advantages of post-harvest-labelling techniques include the fact that they can be used on samples that cannot be labelled metabolically, such as human biopsies.

The accuracy of quantification is determined by the mass resolution of the instrument and the signal-to-noise ratio of the measurement, and under optimal conditions, the protein ratios can be determined to within a few per cent[44]. However, if one of the two forms of the peptide is close to the noise level, only a lower limit can be given for the protein ratio. There are many ways to assess the accuracy of quantification. If labelled peptides co-elute with their unlabelled counterparts, the peak ratios can easily be determined by comparing the signal intensities of the two peaks within each pair. As several mass spectra are acquired as the peptides elute from the LC column, several intensity ratios can be determined for each peak pair, such that the spread of these ratios is a measure of the accuracy of the quantification. If there are several peptides that quantify a protein, a standard deviation can be obtained from the separate ratios for each of these peptides. The experiment can also be carried out in reverse — that is, by swapping the label between the two states. Finally, improvements in automation and sensitivity increasingly mean that it will often be relatively easy to carry out the quantification in several independent experiments.

Perspective for peptide-sequencing applications

As described above, peptide-sequencing technology can now rapidly generate long lists of identified proteins from virtually any source of protein material. Relative quantification between protein populations is also often achievable. Furthermore, a recent trend in proteomics

Box 5 | **Alternatives to liquid-chromatography–tandem-mass-spectrometry**

On-line liquid-chromatography–tandem-mass-spectrometry (LC–tandem-MS) is not the only technique that can be used for mass-spectrometry-based proteomics studies. Matrix-assisted laser desorption/ionization (MALDI) continues to be attractive for the identification of single protein spots or bands, because a MALDI mass fingerprint is obtained in less than one minute, whereas an LC run typically takes at least half an hour. The relatively low certainty of protein identification using MALDI fingerprinting has been addressed by the development of MALDI instruments that can also sequence peptides. In these instruments, a MALDI source is either coupled to a double time-of-flight section (MALDI–TOF–TOF), to a hybrid quadruple TOF or to an ion trap. MALDI ions are singly charged and generally give less informative mass spectra, but they are usually sufficient for identification. However, as the general trend is moving away from two-dimensional gel electrophoresis, there has been less emphasis on single-spot identification. For analysing complex protein mixtures, high-performance LC has to be coupled 'off-line' to MALDI, and this is usually done by collecting fractions onto metal 'targets' that will be introduced to the MALDI mass spectrometer. Off-line coupling is technically complex, but allows repeated analysis.

A technique that has captured the attention of many clinicians — and, notably, of Congress — is the surface-enhanced laser desorption/ionization (SELDI) method (for a review, see REF. 76). A bodily fluid, such as blood, is placed on a surface that has ion-exchange or hydrophobic properties and is analysed by MALDI to produce a pattern of peptides and small proteins. In diagnostic applications, these patterns can be linked to healthy and diseased patients by statistical techniques[77]. Despite the great clinical promise and importance of techniques for the direct diagnosis of patient samples using mass spectrometry, the SELDI technique itself has been intensely controversial because of its limited sensitivity for low-abundance components and the limited robustness of the bioinformatics analysis (for example, see REFS 78,79).

has been towards large-scale experiments and automation. In many laboratories, the HPLC column is now loaded by an autosampler, which allows the analysis of many peptide mixtures per day without too much loss of sample and sensitivity. So, where can this powerful technology be applied most usefully?

One of the most rewarding applications so far has been in the characterization of protein complexes[49,50]. There is an increasing focus on these 'molecular machines', and MS is very valuable as a first step to identify the protein members of these complexes and, possibly, their modification state. Approaches have ranged from the large-scale identification of immunoprecipitated multiprotein complexes[51,52] for the derivation of protein-interaction networks to the characterization of whole organelles[53–55]. The success of this strategy is due to the fact that the purification step enriches the protein population at the same time as it limits its complexity compared to total cell lysates. Second, and more importantly, multiprotein complexes provide a functional context, in which the proteomic results can be interpreted.

Recent developments now make it possible to determine more transient and signal-dependent interactions[56,57], by using the stable-isotope-based proteomics techniques mentioned above to encode the pull-down versus the control. For example, a phosphopeptide and the non-phosphorylated peptide can be coupled to beads and incubated with ^{13}C-Arg- and ^{12}C-Arg-encoded cell lysates, respectively. The proteins that bind to the peptide beads can then be eluted and mixed. Almost all of the peptides that are detected after digesting these proteins will produce peak pairs that have a one-to-one ratio, which indicates nonspecific binding to the phosphopeptide and the non-phosphorylated counterpart or to the beads. However, if the peptides are from proteins that specifically bind the phosphopeptide, these peptides will predominantly be detected in the ^{13}C form[58].

One of the key limitations of organelle purification has been the difficulty of distinguishing true organellar proteins from co-purifying ones[55]. This difficulty has only become worse with the increasing sensitivity and throughput of MS. However, we noticed that we could obtain a quantifiable profile of all the proteins from the centrifugation fractions in the final enrichment step of purification. True organellar proteins produced the same, characteristic profile, whereas non-organellar proteins showed quite different profiles[59]. This gives us the ability to map essentially all cellular structures, even if they can only be enriched and not purified completely.

In contrast to these targeted protein-interaction or organelle-proteomics studies, experiments that are aimed at determining protein expression in whole-cell lysates or tissues (expression proteomics) have been less successful so far. However, intense research efforts are underway at present, because such a strategy would enable the detection/identification of disease-related biomarkers. Such a measurement is essentially the equivalent of a microarray experiment, with the difference being that protein, instead of mRNA, levels are compared. MS experiments that compare protein-expression levels are much more laborious than microarray experiments, but are attractive because proteins are the active agents of the cell, whereas the mRNA population is often a poor indicator of protein levels[60]. However, it is still difficult to identify and quantify all the low-abundance proteins, especially in the presence of highly abundant proteins. Furthermore, as in microarray experiments, the results are 'noisy', because of the extremely large amounts of data, and it can be difficult to distil functional and mechanistic hypotheses from such global experiments.

That said, peptide-sequencing technology is rapidly improving. It might soon become possible to quantify most of the proteins in a cell line or tissue using high-resolution MS (REF. 61), especially as the key issues of MS-based proteomics techniques — that is, the detection

limits and dynamic range — are being pushed to new limits by ingenious hardware and software developments. As mentioned above, important research efforts are underway at present to profile highly complex protein mixtures with the aim of detecting and identifying disease-related biomarkers (see BOX 5 for developments other than those that are related to the LC–tandem-MS experiments discussed here). If proteomics could become practical for the quantification of whole-cell lysates, it would also have the advantage over microarray studies of being able to quantify proteins as a function of their modification state and their subcellular location.

Systems biology will increasingly rely on a combination of mRNA, proteome and genetic data[62]. Truly astonishing advances have been made in creating comprehensive post-genomic data sets for yeast[63–65], an organism that is now in its eighth year of post-genomics life. Integrative mRNA, proteomics and imaging techniques have also been applied to mammalian organelles, as was shown recently for mitochondria[66,67]. The unique strength that MS-based proteomics brings to this area will probably always be its ability to look at proteins in an unbiased way at their endogenous levels and in their native state.

1. Wilm, M. *et al*. Femtomole sequencing of proteins from polyacrylamide gels by nano electrospray mass spectrometry. *Nature* **379**, 466–469 (1996).
Showed that MS could identify gel-separated proteins using a much smaller quantity of the sample than was required by chemical techniques such as Edman degradation.
2. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422**, 193–197 (2003).
3. Zhu, H., Bilgin, M. & Snyder, M. Proteomics. *Annu. Rev. Biochem.* **72**, 783–812 (2003).
4. Phizicky, E., Bastiaens, P. I., Zhu, H., Snyder, M. & Fields, S. Protein analysis on a proteomic scale. *Nature* **422**, 208–215 (2003).
5. Sali, A., Glaeser, R., Earnest, T. & Baumeister, W. From words to literature in structural proteomics. *Nature* **422**, 216–225 (2003).
6. Hanash, S. Disease proteomics. *Nature* **422**, 226–232 (2003).
7. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
8. Figeys, D. Proteomics in 2002: a year of technical development and wide-ranging applications. *Anal. Chem.* **75**, 2891–2905 (2003).
9. Romijn, E. P., Krijgsveld, J. & Heck, A. J. Recent liquid chromatographic–(tandem) mass spectrometric applications in proteomics. *J. Chromatogr. A* **1000**, 589–608 (2003).
10. Lin, D., Tabb, D. L. & Yates, J. R. 3rd. Large-scale protein identification using mass spectrometry. *Biochim. Biophys. Acta* **1646**, 1–10 (2003).
11. Wu, C. C. & Yates, J. R. 3rd. The application of mass spectrometry to membrane proteomics. *Nature Biotechnol.* **21**, 262–267 (2003).
12. Mann, M. & Jensen, O. N. Proteomic analysis of post-translational modifications. *Nature Biotechnol.* **21**, 255–261 (2003).
13. Patterson, S. D. & Aebersold, R. H. Proteomics: the first decade and beyond. *Nature Genet.* **33** (Suppl.), 311–323 (2003).
14. Ferguson, P. L. & Smith, R. D. Proteome analysis by mass spectrometry. *Annu. Rev. Biophys. Biomol. Struct.* **32**, 399–424 (2003).
15. Mo, W. & Karger, B. L. Analytical aspects of mass spectrometry and proteomics. *Curr. Opin. Chem. Biol.* **6**, 666–675 (2002).
16. Mørtz, E. *et al*. Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. *Proc. Natl Acad. Sci. USA* **93**, 8264–8267 (1996).
17. Horn, D. M., Zubarev, R. A. & McLafferty, F. W. Automated *de novo* sequencing of proteins by tandem high-resolution mass spectrometry. *Proc. Natl Acad. Sci. USA* **97**, 10313–10317 (2000).
18. Sze, S. K., Ge, Y., Oh, H. & McLafferty, F. W. Top-down mass spectrometry of a 29-kDa protein for characterization of any posttranslational modification to within one residue. *Proc. Natl Acad. Sci. USA* **99**, 1774–1779 (2002).
19. Taylor, G. K. *et al*. Web and database software for identification of intact proteins using 'top down' mass spectrometry. *Anal. Chem.* **75**, 4081–4086 (2003).
20. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64–71 (1989).
21. Lasonder, E. *et al*. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419**, 537–542 (2002).
22. Schirle, M., Heurtier, M. A. & Kuster, B. Profiling core proteomes of human cell lines by 1D PAGE and LC–MS/MS. *Mol. Cell. Proteomics* **2**, 1297–1305 (2003).
23. Washburn, M. P., Wolters, D. & Yates, J. R. 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.* **19**, 242–247 (2001).
Established the 'shotgun' technology by showing that many proteins in a yeast-cell lysate could be identified in a single experiment.
24. Hillenkamp, F., Karas, M., Beavis, R. C. & Chait, B. T. Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal. Chem.* **63**, 1193A–1202A (1991).
25. Mann, M. A shortcut to interesting human genes: peptide sequence tags, ESTs and computers. *Trends Biochem. Sci.* **21**, 494–495 (1996).
26. Taylor, J. A. & Johnson, R. S. Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **11**, 1067–1075 (1997).
27. Liska, A. J. & Shevchenko, A. Expanding the organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. *Proteomics* **3**, 19–28 (2003).
This and other papers from this group address the important issue of using cross-species identification for proteins if the genome of the organism of interest has not been sequenced (see also reference 32).
28. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
29. MacCoss, M. J., Wu, C. C. & Yates, J. R. 3rd. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* **74**, 5593–5599 (2002).
30. Olsen, J. V., Ong, S. E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3**, 608–614 (2004).
Shows that trypsin is an exceedingly specific protease (non-tryptic peptides are produced by protein degradation or by the decomposition of peptides at labile bonds before tandem MS).
31. Keller, A. *et al*. Experimental protein mixture for validating tandem mass spectral analysis. *Omics* **6**, 207–212 (2002).
32. Shevchenko, A. *et al*. Charting the proteomes of organisms with unsequenced genomes by MALDI–quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* **73**, 1917–1926 (2001).
33. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC–MS/MS) for large-scale protein analysis: the yeast proteome. *J. Proteome Res.* **2**, 43–50 (2003).
Reports the large-scale identification of yeast proteins and, using searches in sequence-reversed databases, it establishes a statistical description for false-positive identification. Finally, by re-analysing the data with the cut-off values that have been used in some studies, they show that error rates can be very high.
34. 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).
35. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003).
36. Nesvizhskii, A. I. & Aebersold, R. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discov. Today* **9**, 173–181 (2004).
References 34–36 establish an objective and powerful statistical framework to assess the probability of correct protein identification in proteomics experiments. The procedures can be used on any data set independent of the type of mass spectrometer used and could be the basis of a common identification standard in proteomics.
37. Barr, J. R. *et al*. Isotope dilution — mass spectrometric quantification of specific proteins: model application with apolipoprotein A-I. *Clin. Chem.* **42**, 1676–1682 (1996).
38. Stemmann, O., Zou, H., Gerber, S. A., Gygi, S. P. & Kirschner, M. W. Dual inhibition of sister chromatid separation at metaphase. *Cell* **107**, 715–726 (2001).
39. 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).
References 37–39 introduce the so-called 'AQUA' (absolute quantification) technology for absolute peptide quantification, which involves mixing stable-isotope-labelled peptide analogues into the peptide mixture.
40. Aebersold, R. Constellations in a cellular universe. *Nature* **422**, 115–116 (2003).
41. Lahm, H. W. & Langen, H. Mass spectrometry: a tool for the identification of proteins separated by gels. *Electrophoresis* **21**, 2105–2114 (2000).
42. Oda, Y., Huang, K., Cross, F. R., Cowburn, D. & Chait, B. T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl Acad. Sci. USA* **96**, 6591–6596 (1999).
43. Ong, S. E. *et al*. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).
44. Ong, S. E., Kratchmarova, I. & Mann, M. Properties of 13C-substituted arginine in stable isotope labeling by amino acids in cell culture (SILAC). *J. Proteome Res.* **2**, 173–181 (2003).
45. Sechi, S. & Chait, B. T. Modification of cysteine residues by alkylation. A tool in peptide mapping and protein identification. *Anal. Chem.* **70**, 5150–5158 (1998).
46. Munchbach, M., Quadroni, M., Miotto, G. & James, P. Quantitation and facilitated *de novo* sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety. *Anal. Chem.* **72**, 4047–4057 (2000).
47. Gygi, S. P. *et al*. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.* **17**, 994–999 (1999).
Introduces the ICAT technology — the first demonstration of a global, quantifiable MS technique that is applicable to mammalian samples.
48. Tao, W. A. & Aebersold, R. Advances in quantitative proteomics via stable isotope tagging and mass spectrometry. *Curr. Opin. Biotechnol.* **14**, 110–118 (2003).
49. Lamond, A. I. & Mann, M. Cell biology and the genome projects — a concerted strategy for characterizing multi-protein complexes using mass spectrometry. *Trends Cell Biol.* **7**, 139–142 (1997).
50. Neubauer, G. *et al*. Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl Acad. Sci. USA* **94**, 385–390 (1997).
51. Ho, Y. *et al*. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).

52. Gavin, A. C. *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
Large-scale immunoprecipitations in references 51 and 52 show that protein–protein interaction maps can be obtained by MS and that much of the yeast cell is organized into protein complexes.
53. Dreger, M. Subcellular proteomics. *Mass Spectrom. Rev.* **22**, 27–56 (2003).
54. Taylor, S. W., Fahy, E. & Ghosh, S. S. Global organellar proteomics. *Trends Biotechnol.* **21**, 82–88 (2003).
55. Brunet, S. *et al*. Organelle proteomics: looking at less to see more. *Trends Cell Biol.* **13**, 629–638 (2003).
56. Blagoev, B. *et al*. A proteomics strategy to elucidate functional protein–protein interactions applied to EGF signaling. *Nature Biotechnol.* **21**, 315–318 (2003).
57. Ranish, J. A. *et al*. The study of macromolecular complexes by quantitative proteomics. *Nature Genet.* **33**, 349–355 (2003).
58. Schulze, W. X. & Mann, M. A novel proteomic screen for peptide–protein interactions. *J. Biol. Chem.* **279**, 10756–10764 (2004).
References 56–58 show that quantitative methods can identify functionally important protein interactions in the presence of a large excess of background proteins.
59. Andersen, J. S. *et al*. Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426**, 570–574 (2003).
Protein-correlation profiling is introduced as a technology to distinguish true members of complexes and organelles from co-purifying background proteins on the basis of their fractionation profiles.
60. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
61. 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).
62. Ideker, T. *et al*. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
63. Bader, G. D. *et al*. Functional genomics and proteomics: charting a multidimensional map of the yeast cell. *Trends Cell Biol.* **13**, 344–356 (2003).
64. Ghaemmaghami, S. *et al*. Global analysis of protein expression in yeast. *Nature* **425**, 737–741 (2003).
65. Huh, W. K. *et al*. Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691 (2003).
66. Mootha, V. K. *et al*. Identification of a gene causing human cytochrome *c* oxidase deficiency by integrative genomics. *Proc. Natl Acad. Sci. USA* **100**, 605–610 (2003).
67. Mootha, V. K. *et al*. Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **115**, 629–640 (2003).
References 66 and 67 illustrate the power of combined organelle proteomics and mRNA co-regulation data.
68. Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular mass exceeding 10,000 daltons. *Anal. Chem.* **60**, 2299–2301 (1988).
69. Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.* **11**, 601 (1984).
70. Biemann, K. Mass spectrometry of peptides and proteins. *Annu. Rev. Biochem.* **61**, 977–1010 (1992).
71. Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **76**, 3908–3922 (2004).
72. Schlosser, A. & Lehmann, W. D. Five-membered ring formation in unimolecular reactions of peptides: a key structural element controlling low-energy collision-induced dissociation of peptides. *J. Mass Spectrom.* **35**, 1382–1390 (2000).
73. Steen, H., Kuster, B., Fernandez, M., Pandey, A. & Mann, M. Detection of tyrosine phosphorylated peptides by precursor ion scanning quadrupole TOF mass spectrometry in positive ion mode. *Anal. Chem.* **73**, 1440–1448 (2001).
74. Mann, M. & Wilm, M. S. Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390–4399 (1994).
75. Eng, J. K., McCormack, A. l. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
76. Tang, N., Tornatore, P. & Weinberger, S. R. Current developments in SELDI affinity technology. *Mass Spectrom. Rev.* **23**, 34–44 (2004).
77. Wulfkuhle, J. D., Liotta, L. A. & Petricoin, E. F. Proteomic applications for the early detection of cancer. *Nature Rev. Cancer* **3**, 267–275 (2003).
78. Sorace, J. M. & Zhan, M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics* **4**, 24 (2003).
79. Baggerly, K. A., Morris, J. S. & Coombes, K. R. Reproducibility of SELDI–TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* **20**, 777–785 (2004).

Acknowledgements

We thank our colleagues at the Center for Experimental BioInformatics (CEBI) and Harvard Medical School for fruitful discussions and for critically reading the manuscript. Work at the CEBI is supported by generous grants from the Danish National Research Foundation (Grundforskningsfond) and the European Union sixth framework programme.

Competing interests statement

The authors declare no competing financial interests.

Online links

FURTHER INFORMATION

American Society for Mass Spectrometry (ASMS): http://www.asms.org
Center for Experimental BioInformatics (CEBI): http://www.cebi.sdu.dk
Human Proteome Organisation (HUPO): http://www.hupo.org
Institute for Systems Biology: http://www.systemsbiology.org
SpectroscopyNOW.com, Proteomics: http://www.spectroscopynow.com/Spy/basehtml/SpyH/1,1181,10-0-0-0-0-home-0-0,00.html
Supplementary material on peptide validation: http://www.cebi.sdu.dk/Steen_Mann_NRM_Suppl_PeptValid.pdf
The Nobel Prize in Chemistry 2002 (for mass spectrometry): http://www.nobel.se/chemistry/laureates/2002/index.html

SUPPLEMENTARY INFORMATION

See online article: S1 (box)
Access to this links box is available online.