JOURNAL OF MASS SPECTROMETRY
J. Mass Spectrom. 2001; **36**: 1083–1091

JMS

SPECIAL FEATURE: TUTORIAL

Proteomics: the move to mixtures

Junmin Peng and Steven P. Gygi*

Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA

Received 25 July 2001; Accepted 20 August 2001; Published online 3 October 2001

Proteomics can be defined as the systematic analysis of proteins for their identity, quantity and function. In contrast to a cell's static genome, the proteome is both complex and dynamic. Proteome analysis is most commonly accomplished by the combination of two-dimensional gel electrophoresis (2DE) and mass spectrometry (MS). However, this technique is under scrutiny because of a failure to detect low-abundance proteins from the analysis of whole cell lysates. Alternative approaches integrate a diversity of separation technologies and make use of the tremendous peptide separation and sequencing power provided by MS/MS. When liquid chromatography is combined with tandem mass spectrometry (LC/MS/MS) and applied to the direct analysis of mixtures, many of the limitations of 2DE for proteome analysis can be overcome. This tutorial addresses current approaches to identify and characterize large numbers of proteins and measure dynamic changes in protein expression directly from complex protein mixtures (total cell lysates). Copyright © 2001 John Wiley & Sons, Ltd.

KEYWORDS: mass spectrometry; tandem mass spectrometry; microcapillary liquid chromatography; proteome analysis

INTRODUCTION

In the post-genomic era, technology development for large-scale protein analysis looms large. Reductionism methods, such as studying one gene, one protein or one pathway in an organism, have contributed greatly to our understanding of many basic principles of life. However, a comprehensive picture of biology will be difficult to grasp until more integrative approaches are utilized.[1] Recently, the significant accomplishments of genomics, proteomics and bioinformatics are making the systematic analysis of all expressed cellular components a reality.[2–4]

The proteome is defined as the set of all expressed proteins in a cell, tissue or organism.[5] While it is often conceptualized that one gene produces one protein, it is known that the expressed products of a single gene in reality represent a protein population that can contain large amounts of microheterogeneity [Fig. 1(A)]. Consider the example a protein that has three potential modification states: glycosylation, phosphorylation and ubiquitination. There are eight potential protein forms that could be presented if each modification only occurs at a single site and is not mutually exclusive [Fig. 1(B)]. Each additional state (e.g. another phosphorylation, acetylation, protease cleavage, lipidation, acetylation, etc.) or modification site would add a large amount of additional diversity to the expression profile of that protein (64 potential forms for just six modifications). More than 100 modification types are recorded and additional ones are yet to be discovered.[6] All modified forms from one protein can vary in abundance, activity or location inside a cell. Some of the potential expression-diversifying events are shown in Fig. 1(A). Clearly, post-translational modification of proteins is an event with dramatic effects on the complexity of the proteome. In addition, the hallmark of a proteome is its ability to regulate dynamically protein expression in response to external and internal perturbations under developmental, physiological, pathological, pharmacological and aging conditions.

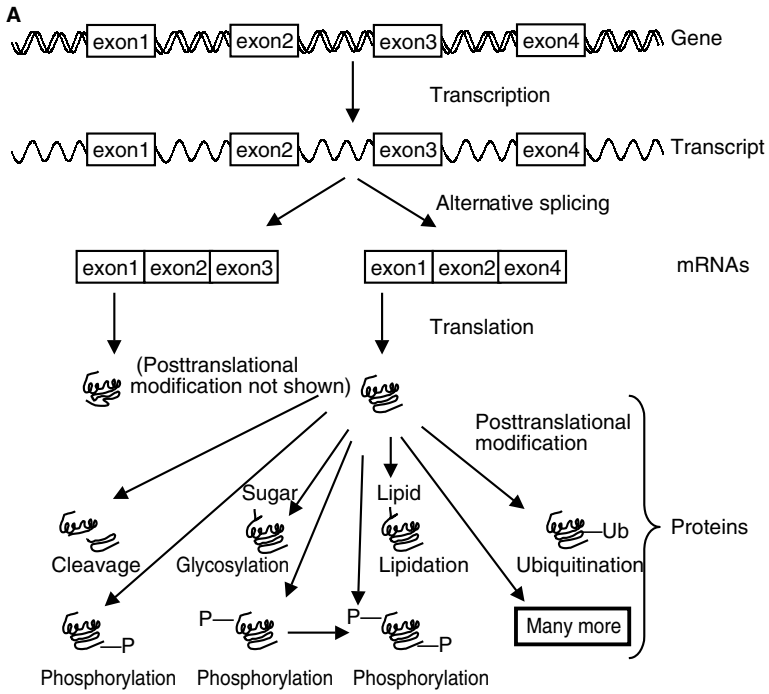
Proteome analysis presents specialized analytical problems in two major areas: (i) dynamic expression range (protein abundance)[7] and (ii) diversity of protein expression (multiple protein forms).[8] The dynamic range problem can be overcome by either increased separation power or prefractionation to enrich for lower abundance proteins.[9] Overcoming the diversity of protein expression is more involved and represents a significant challenge for proteome analysis.[10] However, it can be overcome, in some measure, by utilizing more powerful separation strategies to give a more complete picture of an individual protein's expression profile. Figure 2 presents a simplified scheme of proteome analysis. There are several separation tools that can be utilized either alone or in combination to

*Correspondence to: S. P. Gygi, Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA. E-mail: steven_gygi@hms.harvard.edu
Contract/grant sponsor: NIH; Contract/grant number: HG00041.
Contract/grant sponsor: Giovani-Armenise Harvard Foundation.
Contract/grant sponsor: Jane Coffin Childs Memorial Fund for Medical Research.

DOI: 10.1002/jms.229

JMS

A



B

# of protein forms	Glycosylation	Phosphorylation	Ubiqutination
1			
2	✓		
3		✓	
4			✓
5	✓	✓	
6	✓		✓
7		✓	✓
8	✓	✓	✓

Figure 1. Complexity of the proteome. (A) One gene can produce multiple mature mRNAs via alternative splicing of pre-mRNA transcripts. Following translation, a myriad of post-translational modifications can produce further variations in the number and types of protein forms. (B) An example of a protein that exists with three potential post-translational modifications is shown: glycosylation, phosphorylation and ubiquitination. Eight potential protein types can be formed. The presence of each modification is noted by a tick.

analyze proteins or peptides based on their physical or chemical properties: solubility, localization, charge, size, hydrophobicity and affinity to certain matrices.[11] It should be noted that tandem mass spectrometry (MS/MS) itself is a powerful separation tool of great value because a single peptide ion can be selected, isolated and sequenced in the presence of many other co-detected peptides. Collected raw data are further processed by sophisticated software to identify, quantify and characterize proteins.[12] In this tutorial, we discuss the advantages and limitations of the current method for proteome analysis based on two-dimensional polyacrylamide gel electrophoresis (2DE) and the need for alternative methods. We focus primarily on protein mixture analysis using liquid chromatography combined with tandem mass spectrometry (LC/MS/MS) as the technology base for most alternative proteome analysis strategies.

CURRENT STATE OF PROTEOME ANALYSIS

The emerging field of proteomics has grown out of the mature technology of 2DE for protein separation and quantification[13,14] and increasingly refined technologies for the identification of separated proteins. Today MS is overwhelmingly used as the technology base for protein identification from 2D gels[15] (Fig. 3). 2DE and protein MS now represent an integrated technology by which several thousand protein species can be separated, detected and quantified in a single operation, and hundreds of the detected proteins can be identified in a highly automated fashion by sequential analysis of the peptide mixtures generated by digestion of individual gel spots.[16] Furthermore, the additional information obtained from a 2D gel (e.g. isoelectric focusing point and molecular mass) adds validity to MS-based protein identification. However, closer examination

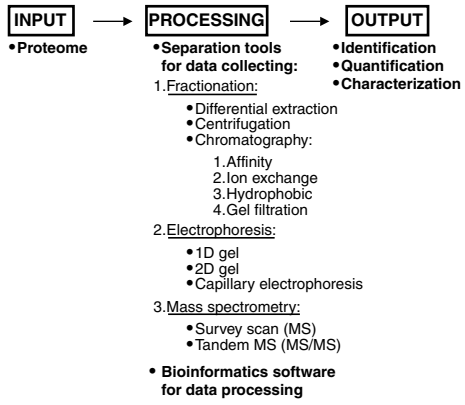


Figure 2. Schematic of proteome analysis. The input for the system is the proteome in all its complexity. Proteins or peptides are separated and analyzed by the listed tools. It is important to remember that MS/MS is also a powerful peptide separation (isolation) technique. The raw data (tandem mass spectra) are further processed by software to produce information about the identity, quantity and characteristics of the proteins detected.

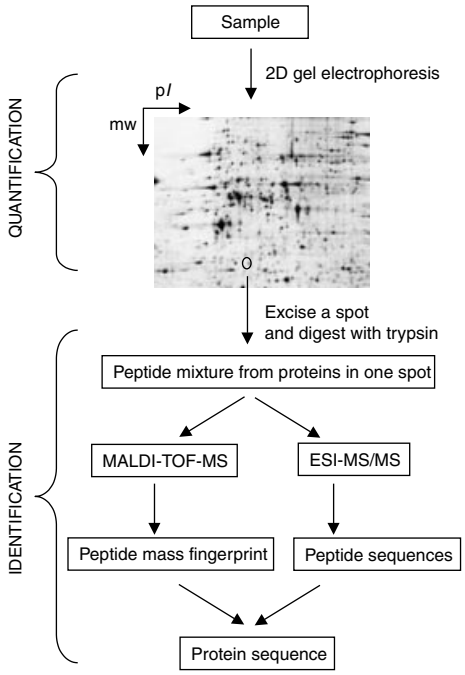


Figure 3. Proteome analysis using two-dimensional gel electrophoresis and mass spectrometry (2DE/MS). Protein from cell lysates is separated in the first dimension by isoelectric focusing (p*I*) and then in the second dimension by SDS-PAGE (molecular mass). Separated proteins are visualized by staining. Protein spots of interest are then excised and subjected to in-gel proteolytic digestion (e.g. trypsin) prior to sequence analysis by MS. In the case of matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) MS, peptide mass mapping techniques are used to create a unique mass fingerprint of the protein. For protein identification by MS/MS, a single peptide ion is isolated and fragmented to form smaller product peptide ions that contain the amino acid sequence information (see Fig. 5). The protein is then identified based on these unique identifiers (mass mapping or tandem mass spectrum) by computer sequence database matching.

of the classes of proteins routinely identified by 2DE/MS suggests that it does not represent a truly global technique. Specific classes of proteins have long been known to be excluded or under-represented in 2D gel patterns. These include very acidic or basic proteins, excessively large or small proteins and membrane proteins. In addition, by examining codon bias values of proteins identified from 2D gels, it has now been shown that the 2DE/MS approach is incapable of detecting low abundance proteins without pre-gel enrichment.[4,10]

Finally, the utilization of 2DE as the technology base for proteome analysis is severely hampered by protein co-migration. Figure 4 shows the detection of six different proteins from a single silver-stained spot from a 2D gel. This greatly confounds 2D gel databases because it is not known which protein (or a previously unidentified protein) is responsible for a measured change in expression when gels are compared.

In conclusion, 2DE/MS cannot serve as the technology base for proteome analysis when whole cell lysates are of interest (global analysis). Notwithstanding, other separation tools (shown in Fig. 2) can be used to simplify protein mixtures and to enrich proteins prior to 2DE. This greatly increases the productivity of proteome analysis utilizing 2D gels. However, pre-fractionation of the cell lysates makes quantification of proteins more complicated because one protein is often distributed into many fractions, and protein recovery rates are difficult to determine.

THE NEED FOR ALTERNATIVE STRATEGIES

To alleviate some of the limitations of 2DE, alternative separation techniques have been integrated with MS as new proteome analysis platforms. Many of these strategies rely on the ability of a tandem mass spectrometer to collect sequence information from a specific peptide, even if numerous other peptides are concurrently present in the sample. When mixtures are extremely complex, on-line reversed-phase liquid chromatography (LC) is used to concentrate and separate the peptides before sequencing by MS.[10,12,15,17,18] Sequence analysis is accomplished within the instrument by the selective isolation of the peptide ion of interest from other co-eluting peptides, gentle fragmentation of the peptide ion at peptide bonds and the recording of fragment ion masses in a tandem mass spectrum. It is these fragment ion masses that represent unique identifiers for the peptide and permit its amino acid sequence to be unambiguously determined.

A comparison between peptide sequencing by N-terminal (Edman-type) degradation and sequencing by MS/MS is instructive (Fig. 5). For N-terminal sequencing a peptide is sequenced chemically by the iterative removal and subsequent detection of the N-terminal amino acid. This process requires analysis times of ~1 h per residue with the usual outcome of the amino acid sequence of one peptide (~12 residues) being determined in an overnight analysis. In addition, N-terminal sequencing generally requires peptide amounts in the low picomole range and requires that the peptide of interest be separated from all other peptides to ensure the correct amino acid sequence is determined.

JMS

Gene	Peptide sequences identified	p*I*	MW (kDa)
GDH1	(K)VIELGGTVVSLSDSK (K)FIAEGSNMGSTPEAIAVFETAR (R)EIGYLFGAYR (K)VLPIVSVPER	5.50	49.6
RPT3	(R)ENAPSIIFIDEVDSIATK	5.32	47.9
SUB2	(R)DVQEIFR (K)LTLHGLQQYYIK (R)INLAINYDLTNEADQYLHR (K)NKDTAPHIVVATPGR (R)FLQNPLEIFVDDEAK	5.36	50.3
TIF3	(R)GSNFQGDGREDAPDLDWGAAR (R)ADLVAVLK (K)ITIPIETANANTIPLSELAHAK (R)EREEVDIDWTAAR (R)EREEPDIDWSAAR	5.17	48.5
VMA1	(K)VGHDNLVGEVIR	5.09	67.7
YFR044C	(R)YPSLSIHGVEGAFSAQGAK (K)LVYGVDPDFTR (K)FISEQLSQSGFHDIK (R)TELIHDGAYWVSDPFNAQFTAAK (K)ILIDGIDEMVAPLTEK	5.54	52.9

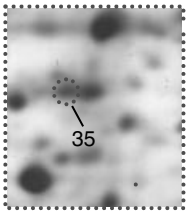


Figure 4. Co-migration of protein products is a common problem for two-dimensional gel (2D) based proteome analysis.[10] The example shows the analysis of a single spot from a silver-stained 2D gel of whole yeast lysate. Proteins in the spot were assigned by MS/MS, and the peptides identified are listed. Six different proteins were found to be present in this single 2D gel feature. On a subsequent 2D gel where the intensity of this spot was altered, it would be impossible to tell which of these proteins (or another previously unidentified protein) had changed in intensity.

In great contrast, peptide sequencing by MS/MS can be accomplished today by acquiring a single scan containing a tandem mass spectrum (1–2 s) at sensitivities requiring only low femtomole levels of the peptide. The fragmented peptide ions (mostly b- and y-type series ions shown in Fig. 5) allow identification of the peptide amino acid sequence by searching a protein or translated nucleotide databases.[19,20] Furthermore, mixtures of peptides usually present no problem for the tandem mass spectrometer because each peptide is isolated in the mass spectrometer prior to sequencing. The most exciting result from these types of analyses is the tremendous number of peptides that can be individually selected, isolated and sequenced during a single LC/MS/MS analysis. An example is shown in Plate 1. During an 11 s interval five different peptides were selected for fragmentation (sequenced). During the entire 30 min analysis, 800 sequencing attempts were acquired. Finally, run times can easily be extended to 2 h or more while maintaining good peptide chromatography, providing the possibility to sequence many thousands of peptides in a single analysis.

Creating the highly complex peptide mixture is straightforward. Lyophilized protein is resolubilized in a highly reducing and denaturing environment (e.g. 8 M urea, 10 mM dithiothreitol (DTT), 50 mM Tris–HCl, pH 8.3, etc.).[9,18] Cysteinyl residues are then alkylated, if desired, and the protein mixture is simply diluted 8 fold in the presence of sequencing-grade trypsin for overnight digestion. The resulting acidified mixture can be directly analyzed by LC/MS/MS techniques.

ON-LINE MICROCAPILLARY REVERSED-PHASE LC/MS/MS

Sensitivity is the driving factor in the development of HPLC columns of smaller and smaller diameter.[15,21] Working flow-rates decrease exponentially with, for example, a 1 mm i.d. column operating at ~50 μl min^{-1} and a 75 μm i.d. column operating at ~300 nl min^{-1}. Microcapillary columns are typically packed into fused-silica capillary tubing and provide peptide detection and sequencing limits routinely in the low-femtomole range.

An on-line capillary LC/MS/MS system consists of conventional HPLC pumps, transfer tubing, a pre-column flow splitter, a liquid junction, a reversed-phase microcapillary column and a tandem mass spectrometer (Fig. 6). The pumps and mass spectrometer are controlled by the same software to allow coupling between chromatography and ion detection. The flow-rate at which the pumps operate (~100 μl min^{-1}) and the optimum flow-rate for the microcapillary column (~300 nl min^{-1} for 75 μm i.d.) are vastly different. A simple flow restrictor (T-splitter) permits the gradient to form quickly and be distributed to the microcapillary column at acceptable flow-rates. A liquid junction (e.g. gold wire) with high voltage (1–2 kV) is needed to promote electrospray. The microcapillary column is typically packed into fused-silica capillary tubing with C_{18} silica beads. There are many variables here, but one proven adaptation utilizes a 75 μm i.d. capillary with 5 μm C_{18} beads and a bed length of 12 cm.[17,22,23]

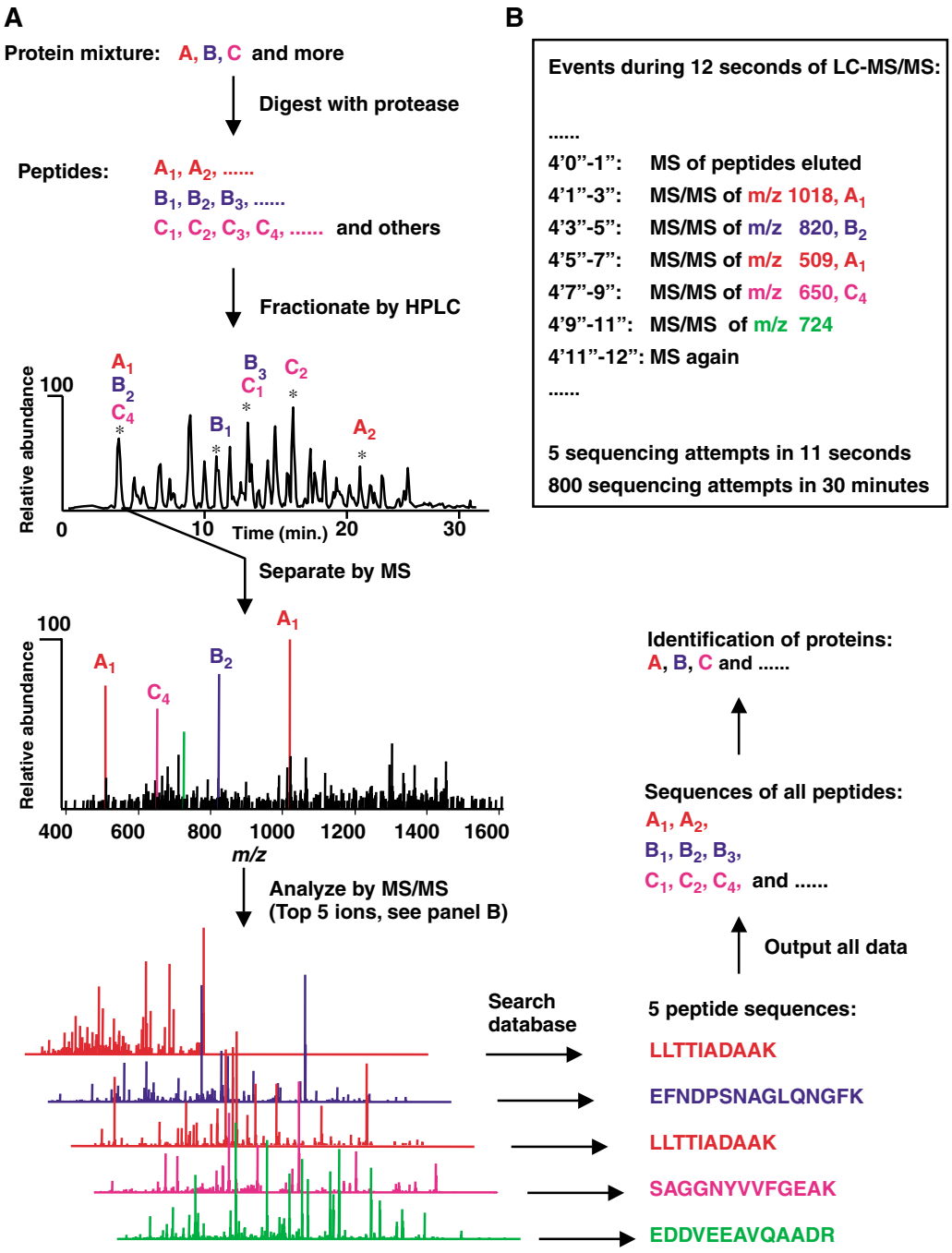


Plate 1. Direct analysis of complex peptide mixtures by on-line reversed-phase liquid chromatography tandem mass spectrometry (LC/MS/MS). (A) A protein mixture (e.g. cell lysate) is proteolyzed to a highly complex peptide mixture. The peptide mixture is separated by reversed-phase LC. Some peptides are not amenable to analysis by reversed-phase LC (e.g. C_3), but this does not affect the ability to identify unambiguously the protein because many peptides are produced by the digestion of one protein. Eluting peptides are first detected by a survey scan (single-stage MS scan). Co-eluting peptides (e.g. A_1, B_2, C_4) usually present no problem for sequence analysis by MS/MS because each peptide is first isolated from any co-present peptides prior to fragmentation. The acquired tandem mass spectrum contains the sequence information for a single peptide ion. Occasionally, several ions derived from one peptide (e.g. A_1 exists as both singly and doubly charged ions) are sequenced. Bioinformatics software can be utilized to match a theoretical (computer-produced) MS/MS of peptides from a sequence database with the acquired tandem mass spectrum. The entire process is highly automated with generally one survey (MS) scan being followed by multiple sequencing (MS/MS) scans. (B) Timeline of automated peptide sequencing by LC/MS/MS analysis. Five MS/MS scans representing the sequence analysis of five different peptide ions occur in 11 s with no user input. During a 30 min analysis, more than 800 sequencing attempts (tandem mass spectra) are acquired.

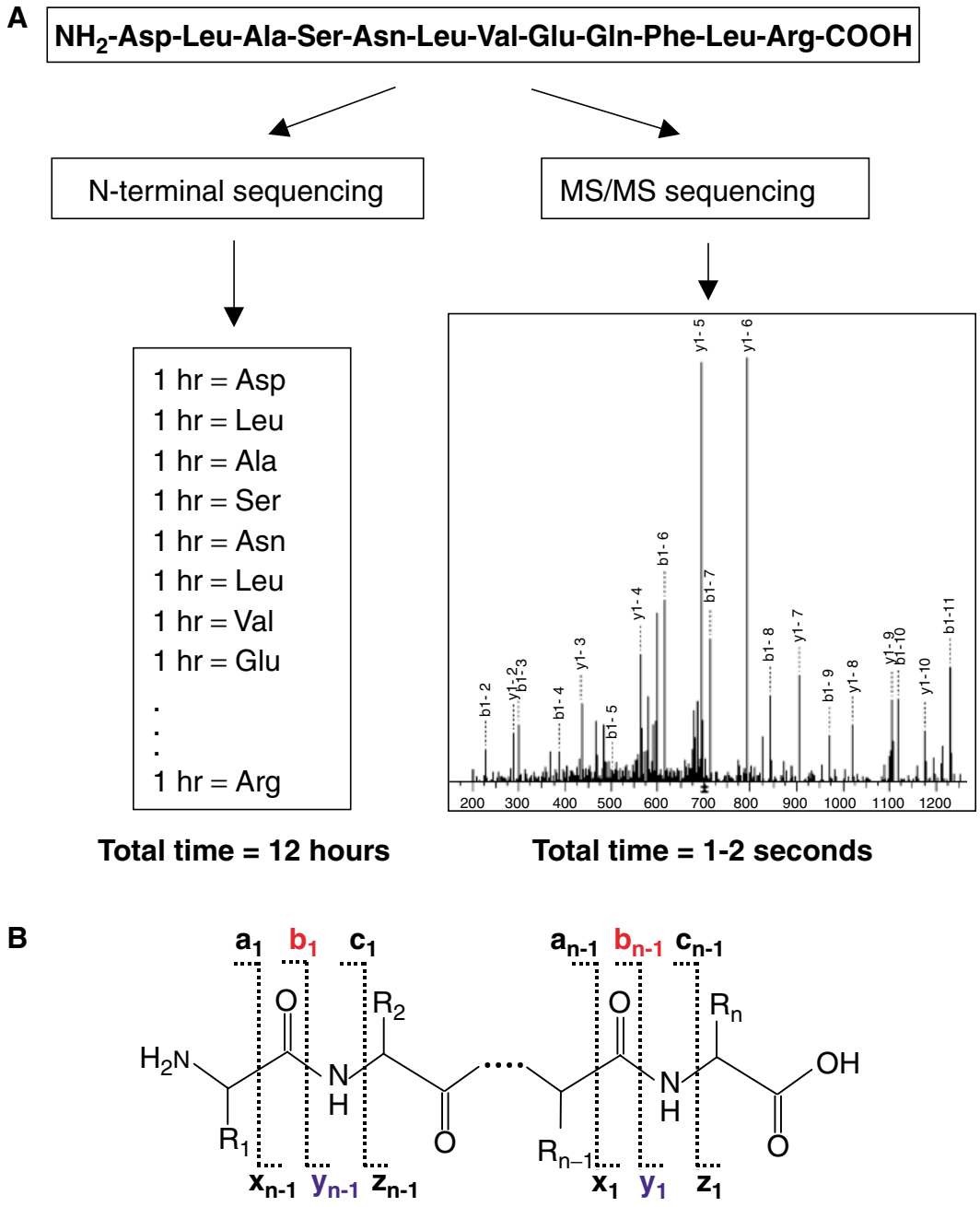


Figure 5. (A) A comparison of peptide amino acid sequence analysis by N-terminal (Edman-type) sequencing and by MS/MS. N-Terminal sequencing of a 12-mer occurs with the repetitive, chemical removal of the first (N-terminal) amino acid that is identified by HPLC with UV detection. Each cycle requires about one hour and produces the identity of one amino acid. More critically, the peptide has to be purified to homogeneity prior to sequencing. In contrast, sequence analysis by MS/MS occurs by acquiring a single MS/MS scan (1–2 s). The precursor ion (peptide ion of interest) is first isolated and then fragmented to form a product ion (MS/MS) spectrum. Because the location of peptide fragmentation along the backbone is predictable, computer software can be used to match a predicted tandem mass spectrum from a database with an acquired tandem mass spectrum. (B) Multiple ion series are present simultaneously in a tandem mass spectrum. A single fragmentation event produces two smaller peptides. Fragment ions containing the original N-terminus are termed a-, b- and c-type series ions. Fragment ions containing the original C-terminus are termed x-, y- and z-type series ions. Under low-energy conditions, the most commonly formed ions are b- and y-type series ions that represent cleavage at the peptide (amide) bond. For example, cleavage at the Val–Glu peptide bond in the peptide shown produces one singly charged b- and one singly charged y-type ion (b_{1-7} and y_{1-5}, respectively, in which the first number refers to its charge state and the second represents its residue number). Sometimes cleavage of one peptide bond can produce multiple ions (not just one b- and one y-type ion) because of the possibility of producing ions with different charge states (e.g. both y_{1-10} and y_{2-10} are present in the spectrum in Plate 2).

JMS

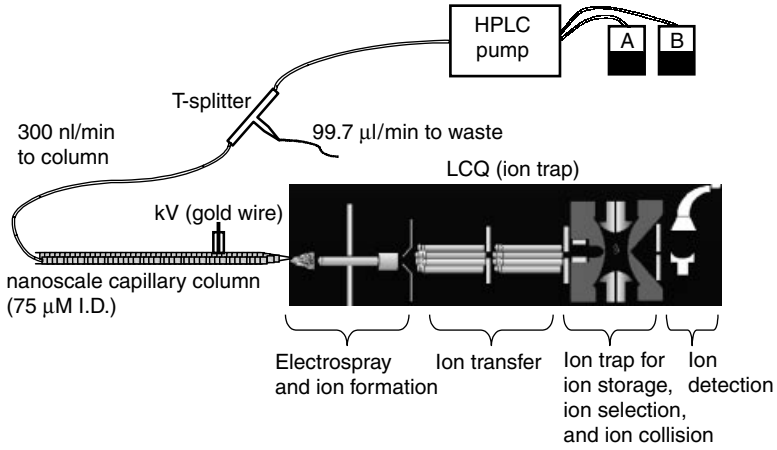


Figure 6. Schematic of on-line nanoscale microcapillary LC/MS/MS. The system consists of conventional HPLC pumps operating at 100 μl min^{-1}, a flow-restriction splitter, a fused-silica microcapillary column (75 μm i.d. 12 cm bed length, 5 μm C_{18} silica beads) and a tandem mass spectrometer. The peptide sample is loaded off-line via a pressure cell onto the nanoscale capillary column at a flow-rate of ~300 nl min^{-1}. The end of the column is then reattached to the flow-splitter. The peptides are then eluted from the column by forming a gradient from 100% solvent A (0.4% acetic acid, 5% acetonitrile) to 70% solvent B (0.4% acetic acid, 95% acetonitrile) with the pumps operating at 100 μl min^{-1}. The flow is reduced at the splitter to ~300 nl min^{-1} using a simple restriction capillary. The column is eluted at this flow-rate and peptides are ionized by electrospray ionization (~1.8 kV). Eluting peptides enter the mass spectrometer where they are focused, transferred and stored in the ion trap. For a single-stage MS scan the ions are ejected to the detector. However, for an MS/MS scan, a specific peptide ion is isolated within the trap followed by gentle fragmentation and finally fragment ion ejection and detection.

A sample can be loaded and eluted on the capillary column in different ways depending on its peptide concentration and volume. (i) It can be loaded via an injection loop on a six-port valve inserted between the T-splitter and the column. (ii) Loading off-line via a pressure-cell maximizes sensitivity but requires significant sample handling, and loading large volumes (>5 μl) is time consuming owing to the low flow-rate of 300 nl min^{-1}. (iii) To alleviate this problem, the sample can be loading on to a pre-column trap at μl min^{-1} flow-rates for concentration and fast desalting and then eluted to the reversed-phase column for separation. However, the sensitivity and robustness are often greatly decreased in this system at sub-microliter flow-rates. (iv) Recently, we have developed a strategy has that makes use of a vented microcapillary column (V-column) where the first few centimeters of the capillary column are loaded with the sample at high flow-rates exiting through the vent. After closing the vent (switching the position of a six-port valve), the bound peptides are eluted at much lower flow-rates consistent with microcapillary HPLC (300 nl min^{-1}) (L. Licklider *et al.*, in preparation), which permits completely automated analyses when combined with an autosampler.

Peptides eluting from the reversed-phase microcapillary column are ionized via electrospray ionization. The peptide ions are analyzed by a coupled tandem mass spectrometer. In the case of an ion trap mass spectrometer, peptide ions are focused, transferred and finally reach the trap where they are stored in stable orbits. During the acquisition of a full-scan mass spectrum, peptide ions in the trap are sequentially ejected based on their mass-to-charge (m/z) values to the detector where the signal strength of each ion is measured. During an MS/MS scan, a selected peptide ion is isolated within the trap by ejecting ions with other m/z values. The isolated ion is then fragmented by the addition of external energy and collisions with helium atoms. The trapped fragment ions are then ejected to the detector to produce a tandem mass spectrum containing the sequence information for that peptide. The selection of peptide ions for fragmentation can be performed in a completely automated fashion with multiple sequencing (MS/MS) scans following a single survey (MS) scan. In addition, the software controlling the process can dynamically exclude peptide ions that have been sequenced from further analysis to allow for peptides of lower abundance to be selected for sequencing.

When even greater peptide separation is desired prior to sequence analysis by tandem MS, multi-dimensional chromatography can be employed.[9,10,18,24–27] The most common implementation utilizes a separation of peptides in the first dimension by strong cation-exchange (SCX) chromatography. SCX chromatography is an excellent choice because (i) it removes intact trypsin by tight binding, (ii) peptides bind in the presence of useful molecules that carry no positive charge at pH 3.0 (e.g. SDS (or other detergents), urea, DTT, etc.) and (iii) peptides are eluted with increasing salt concentrations which are directly compatible with reversed-phase chromatography.

Peptide elution is primarily based on peptide solution charge and its hydrophobicity at pH ~3.0. A single positive charge is sufficient to adsorb the peptide to the column. With the same charge, the hydrophobic peptide is eluted later than the hydrophilic one. At pH 3.0 there are almost exclusively amine functional groups contributing to the solution charge state. The nominal charge of any peptide can be determined by adding up the number lysine, arginine and histidine residues with one additional charge contributed by the N-terminus. Tryptic peptides generally have solution charge states of 2^+ because they terminate in lysine or

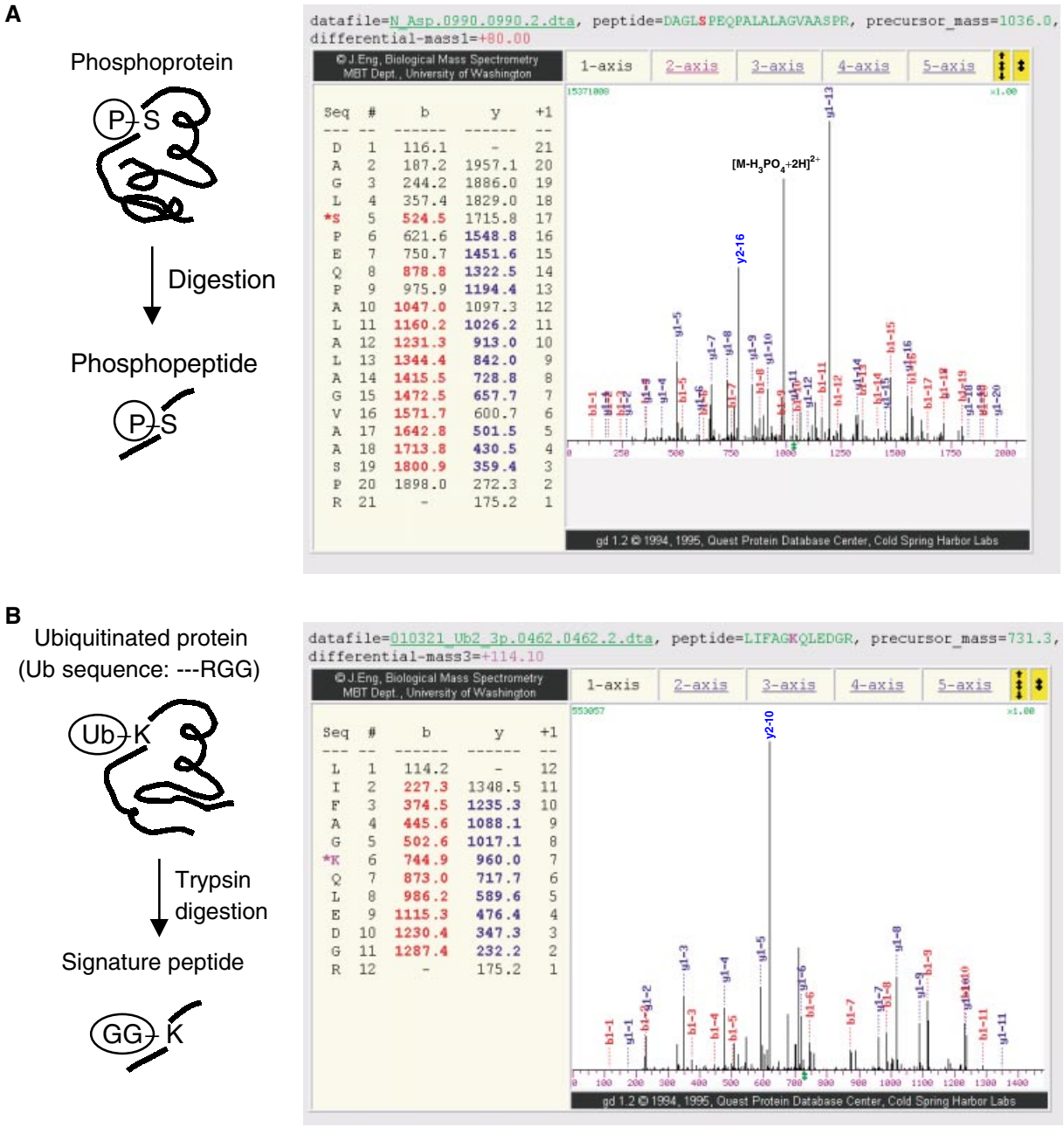


Plate 2. Identification of post-translational modifications by LC/MS/MS with database searching. (A) Phosphorylation results in a net mass change of 80 Da to an affected serine, threonine or tyrosine residue. During fragmentation, the loss of the phosphate group (as phosphoric acid, 98 Da) is common from serine and threonine residues. In the spectrum shown, the loss of phosphoric acid from the precursor ion resulted in a prominent peak at *m/z* 987. The database searching software (Sequest) was able to identify this phosphopeptide from more than 1500 tandem mass spectra generated during a 1 h LC/MS/MS analysis. In addition, the precise phosphoserine residue out of two serines present was easily determined. (B) Ubiquitination results in a polypeptide (76 amino acids) being attached to a lysine residue through an isopeptide bond. When a ubiquitinated protein is digested with trypsin, a remnant of the original ubiquitin polypeptide is left as a dipeptide (Gly–Gly) still covalently attached to the affected lysine of the original protein. By interrogating sequence database with and without the net additional mass of 114 Da, the precise site of ubiquitination site can often be defined.

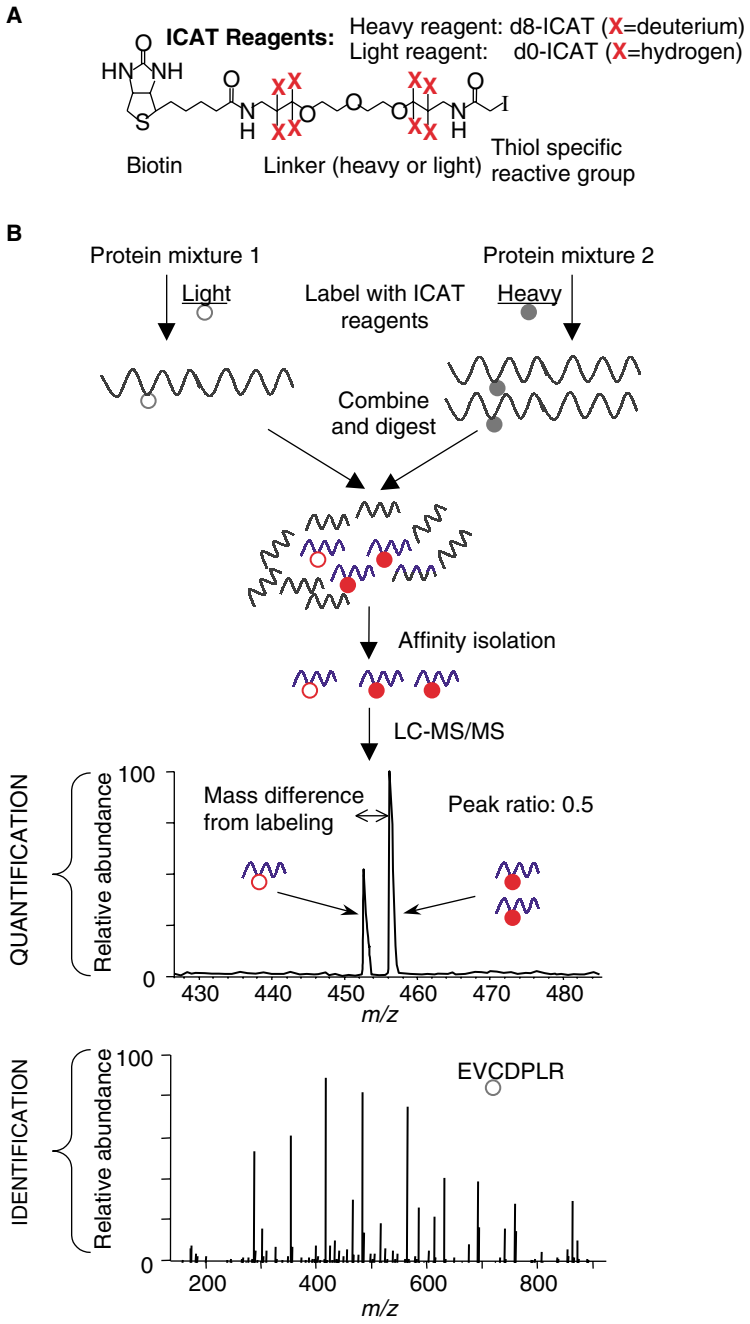


Plate 3. The isotope coded affinity (ICAT) strategy for quantifying differential protein expression. (A) Chemical structure of the ICAT reagents consisting of three features: a biotin tag, a linker for the incorporation of heavy isotopes (i.e. deuterium atoms) and a functional group (iodoacetamide). The reagents come in two forms, heavy (eight deuterium atoms in linker) and light (eight hydrogen atoms in linker). (B) Two protein mixtures representing two different cell states have been denatured, reduced and treated with the isotopically light (open circles) and heavy (filled circles) ICAT reagents, respectively; an ICAT reagent is covalently attached to each cysteinyl residue in every protein. The protein mixtures are combined, proteolyzed to peptides and ICAT-labeled peptides are isolated utilizing the biotin tag. These peptides are separated by microcapillary high-performance liquid chromatography. A pair of ICAT-labeled peptides is chemically identical and is easily visualized because they essentially co-elute and differ by 8 Da in mass measured in a scanning mass spectrometer (4 *m/z* units difference for a doubly charged ion pair). The ratios of the original amounts of proteins from the two cell states are strictly maintained in the peptide fragments. The relative quantification is determined by the ratio of the peptide pairs. Every other scan is devoted to fragmenting and then recording sequence information about an eluting peptide (tandem mass spectrum). The protein is identified by computer searching against complete protein databases. In theory, every peptide pair in the mixture is, in turn, measured and fragmented resulting in the relative quantification and identification of mixture components in a single analysis.

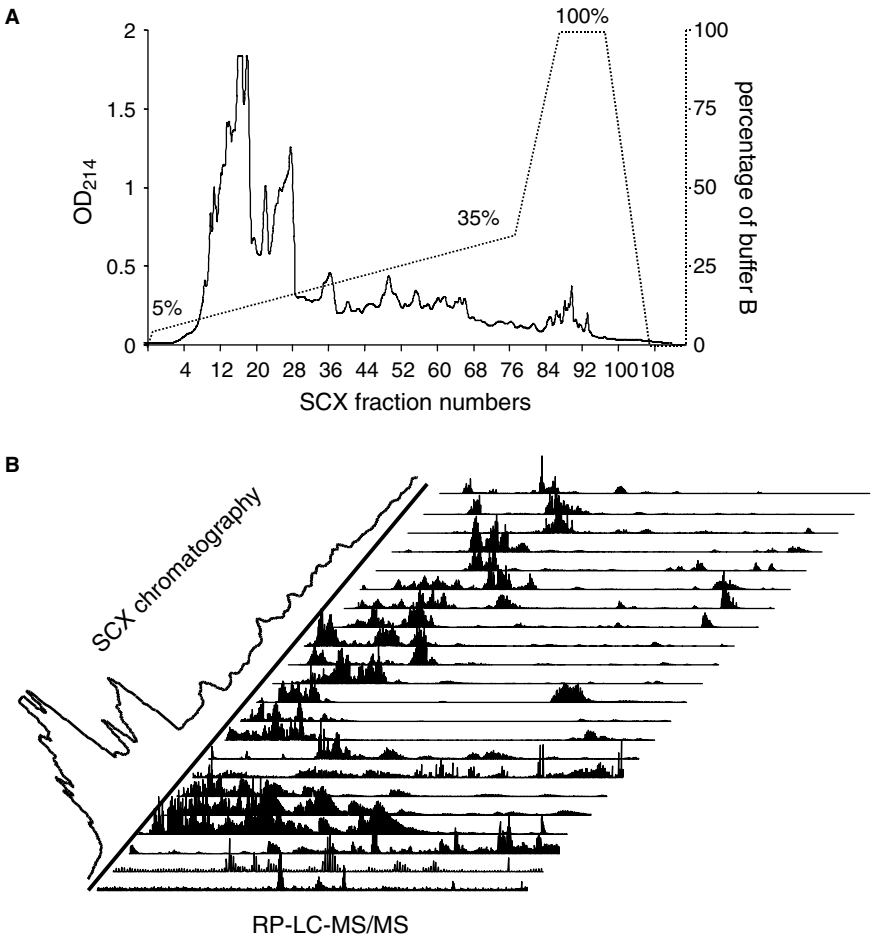


Figure 7. Multi-dimensional peptide chromatography permits the analysis of thousands of proteins from a single sample. In the example shown, 1 mg of whole cell yeast extract was proteolyzed with trypsin under reducing and denaturing conditions. (A) The highly complex peptide solution was separated in the first dimension by strong cation-exchange (SCX) chromatography with UV detection and fraction collection every minute (solvent A, 5 mM phosphate buffer, 25% acetonitrile, pH 3.0; solvent B, the same as A with 350 mM KCl). (B) The collected fractions were then analyzed individually by the nanoscale microcapillary LC/MS/MS techniques described in Fig. 6. As a result, more than 12 000 unique peptides were sequenced and more than 1600 unique proteins were identified from 80 fractions using previously established criteria for database matching.[9]

arginine and have a free N-terminus. A solution charge state of 3^+ is seen for tryptic peptides containing one histidine residue. Tryptic peptides carrying a single charge in solution at pH 3.0 are highly specialized representing either the C-terminal peptide from the protein, the N-terminal peptide that is blocked (acetylated) or phosphorylated peptides. The elution of peptides with solution charge states of 4^+ or more are also specialized (e.g. disulfide-linked tryptic peptides, missed cleavages, etc.).

Figure 7 shows the separation of 1 mg of tryptic peptides from whole yeast lysate by SCX chromatography. Fractions were collected each minute and analyzed in the second dimension by the same reversed-phase LC/MS/MS techniques as already described. Because each fraction represented a unique subset of peptides for analysis, the combined effect is the ability to identify automatically thousands of proteins (from many thousands of peptides) from a single sample. In the example shown in Fig. 7, more than 12 000 unique peptides were assigned and ~1600 unique proteins were identified (J. Peng *et al.*, in preparation) using criteria described previously.[9]

In a typical experiment, about 10–20% of tandem mass spectra lead to the identification of unique peptides with reasonable confidence because of several factors. (i) The majority of peptides are eluted in a relatively narrow window (20–30% acetonitrile), resulting in a significant amount of sequencing attempts being acquired when no peptides are present (background ions). (ii) A peptide can be selected for sequencing more than once because it is able to generate several ions with different charge states (e.g. 1^+, 2^+ and 3^+) and for very abundant ions, the isotope peaks may also be selected. (iii) In order to identify a peptide without ambiguity, only tandem mass spectra meeting stringent criteria from database searching are considered a match.

PEPTIDE SEQUENCING VIA DATABASE SEARCHING OF TANDEM MASS SPECTRA

Because the sequence information for a single peptide is contained in one tandem mass spectrum, it is theoretically possible to use a mass ruler to determine based on peak mass differences the precise amino acid sequence of the

JMS

peptide. Determining the sequence in this way (*de novo*) is simply not possible for the majority of spectra obtained because (i) more than one ion series is usually present (i.e. sequencing the peptide from the N- and C-termini simultaneously), (ii) ion series are rarely complete, (iii) fragment ions are present in varying abundances, often with associated losses of water and/or ammonia, and (iv) some amino acids have very similar or identical molecular masses (e.g. isoleucine and leucine, glutamine and lysine, etc.). For these reasons, database-searching algorithms have become an indispensable tool for the interpretation of tandem mass spectra.

Several approaches are available for database searching based on MS/MS data.[20,28–31] The most useful software algorithms allow for the identification of peptide sequences directly from an LC/MS/MS analysis. For example, the software algorithm Sequest matches a peptide sequence with a tandem mass spectrum using the following steps:[20] (1) peptides with molecular masses matching that of the peptide ion sequenced in the tandem mass spectra are extracted from a protein database; (2) each peptide is given a preliminary score by examining the number of predicted fragment ions from the database peptide that match the acquired fragment ions in the tandem mass spectrum; and (3) the top 500 best-matching peptides undergo a more rigorous ion-matching algorithm that generates a cross-correlation score. A list of peptides with good correlation is returned to the user with the top-scoring peptide being considered the best candidate.

These database-searching algorithms can also support post-translational modifications of specific amino acids (Plate 2). If, for example, the precise amino acid site of phosphorylation within a peptide is the desired outcome, the database search can differentially include the change in mass associated with a phosphate molecule (80 Da) to each serine, threonine and tyrosine residue encountered in the database. Likewise, a ubiquitination event can be determined because after trypsin digestion the modified lysine residue of a ubiquitinated protein now only contains a dipeptide remnant glycine–glycine of the original ubiquitin which can be identified by a change in mass to an affected lysine residue of 114 Da (J. Peng *et al*., in preparation).

QUANTITATIVE PROTEOME ANALYSIS

2DE-based proteome analysis provides information about protein abundance at the gel level by comparing staining intensities. However, when peptide mixtures are analyzed directly by LC/MS/MS techniques, the original quantitative information is lost. Recently, quantitative proteome analysis techniques based on direct mixture analysis have been introduced incorporating stable isotope labeling. Stable isotope dilution involves the addition to the sample of a chemically identical form of the analyte(s) containing stable heavy isotopes (e.g. ^{2}H, ^{13}C, ^{15}N, etc.) as internal standards. Because ionization efficiencies are highly variable for peptides, the best-suited internal standard for a candidate peptide is that same peptide labeled with stable isotopes. Therefore, protein profiling is accomplished if two protein mixtures are compared where one serves as the reference sample, containing the same proteins as the other sample but at different abundances and labeled with stable isotopes. In theory, all peptides from the combined, proteolyzed samples then exist as analyte pairs of identical sequence but differing masses. The peptide pairs have the same physico-chemical properties and behave similarly under any conceivable isolation or separation step. Thus, the ratio between the intensities of the lower and upper mass components of these pairs of peaks provides an accurate measure of the relative abundance of the peptides (and hence the protein) in the original cell lysates. Many groups have independently reported measuring protein profiles based on stable isotopes.[23,32–35] The techniques differ in the method of incorporation of stable isotopes and in the analytical procedures used.

An example of one such technique is the isotope coded affinity tag (ICAT) strategy shown in Plate 3. Proteins from two mixtures are labeled with either a heavy or a light reagent and then mixed together. At this point, any following fractionation can be employed to reduce the complexity of the mixture or enrich for proteins of interest. The relative quantities between two mixtures remain constant. In the experiment shown, the mixed proteins are digested with trypsin and subjected to biotin-affinity chromatography for selective isolation of only labeled peptides, which greatly reduces the complexity of the peptide mixture. The peptides are further separated and analyzed by LC/MS/MS. The relative ratio of labeled peptide pairs is quantified by comparing their peak areas in the LC elution profile. Recently, this technique allowed the profiling of ~300 proteins from yeast in a single experiment.[2] Using the same approach, ~600 membrane proteins from mammalian cells were compared under normal and apoptotic conditions (D Han., personal communication).

CONCLUSIONS

MS-based sequencing of peptides is fast becoming a mature technology capable of providing the platform for proteome analysis. It is now possible to sequence thousands of peptides representing hundreds to thousands of proteins from a single sample taken directly from a cell lysate. In addition, protein expression profiles between two cell states can be measured directly from the mixtures by the incorporation of stable isotopes.

Acknowledgements

We thank Carson Thoreen for expert computer programming, Dr Larry Licklider for encouraging discussions and Dongmei Cheng for her work on phosphorylated proteins. We also thank Tingting Yao and Dr Robert Cohen for providing the ubiquitinated samples. This work was supported in part by NIH grant HG00041 (S.P.), the Giovani-Armenise Harvard Foundation (S.P.) and the Jane Coffin Childs Memorial Fund for Medical Research (J.P.).

REFERENCES

1. Vidal M. *Cell* 2001; **104**: 333.
2. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L. *Science* 2001; **292**: 929.
3. Pandey A, Mann M. *Nature (London)* 2000; **405**: 837.

4. Gygi SP, Aebersold R. *Proteomics: A Trends Guide.* 2000; 32.
5. Pennington SR, Wilkins MR, Hochstrasser DF, Dunn MJ. *Trends Cell Biol.* 1997; **7**: 168.
6. O'Donovan C, Apweiler R, Bairoch A. *Trends Biotechnol.* 2001; **19**: 178.
7. Corthals GL, Wasinger VC, Hochstrasser DF, Sanchez JC. *Electrophoresis* 2000; **21**: 1104.
8. Harry JL, Wilkins MR, Herbert BR, Packer NH, Gooley AA, Williams KL. *Electrophoresis* 2000; **21**: 1071.
9. Washburn MP, Wolters D, Yates JR. *Nature Biotechnol.* 2001; **19**: 242.
10. Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R. *Proc. Natl. Acad. Sci. USA* 2000; **97**: 9390.
11. Cordwell SJ, Nouwens AS, Verrills NM, Basseal DJ, Walsh BJ. *Electrophoresis* 2000; **21**: 1094.
12. Yates JR. *J. Mass Spectrom.* 1998; **33**: 1.
13. O'Farrell PH. *J. Biol. Chem.* 1975; **250**: 4007.
14. Gorg A, Obermaier C, Boguth G, Harder A, Scheibe B, Wildgruber R, Weiss W. *Electrophoresis* 2000; **21**: 1037.
15. Mann M, Hendrickson RC, Pandey A. *Annu. Rev. Biochem.* 2001; **70**: 437.
16. Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm M, Vorm O, Mortensen P, Shevchenko A, Boucherie H, Mann M. *Proc. Natl. Acad. Sci. USA* 1996; **93**: 14440.
17. Yates JR, Carmack E, Hays L, Link AJ, Eng JK. *Methods Mol. Biol.* 1999; **112**: 553.
18. Gygi MP, Licklider LJ, Peng J, Gygi SP. In *Peptides Separation and Mass Spectrometry*, Simpson R (ed.). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2001.
19. Yates JR, Eng JK, McCormack AL. *Anal. Chem.* 1995; **67**: 3202.
20. Eng J, McCormack AL, Yates JR. *J. Am. Soc. Mass Spectrom.* 1994; **5**: 976.
21. Yates JR. *Trends Genet.* 2000; **16**: 5.
22. Gygi SP, Rochon Y, Franza BR, Aebersold R. *Mol. Cell. Biol.* 1999; **19**: 1720.
23. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. *Nature Biotechnol.* 1999; **17**: 994.
24. Washburn MP, Yates JR. *Proteomics: A Current Trends Guide Supplement.* 2000; 28.
25. Davis MT, Beierle J, Bures ET, McGinley MD, Mort J, Robinson JH, Spahr CS, Yu W, Luethy R, Patterson SD. *J. Chromatogr. B* 2001; **752**: 281.
26. Opiteck GJ, Lewis KC, Jorgenson JW, Anderegg RJ. *Anal. Chem.* 1997; **69**: 1518.
27. Opiteck GJ, Jorgenson JW. *Anal. Chem.* 1997; **69**: 2283.
28. Mann M, Wilm M. *Anal. Chem.* 1994; **66**: 4390.
29. Clauser KR, Baker P, Burlingame AL. *Anal. Chem.* 1999; **71**: 2871.
30. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. *Electrophoresis* 1999; **20**: 3551.
31. Zhang W, Chait BT. *Anal. Chem.* 2000; **72**: 2482.
32. Oda Y, Huang K, Cross FR, Cowburn D, Chait BT. *Proc. Natl. Acad. Sci. USA* 1999; **96**: 6591.
33. Pasa-Tolic L, Jensen PK, Anderson GA, Lipton MS, Peden KK, Martinovic S, Tolic N, Bruce JE, Smith RD. *J. Am. Chem. Soc.* 1999; **121**: 7949.
34. Ji J, Chakraborty A, Geng M, Zhang X, Amini A, Bina M, Regnier F. *J. Chromatogr. B* 2000; **745**: 197.
35. Munchbach M, Quadroni M, Miotto G, James P. *Anal. Chem.* 2000; **72**: 4047.