Advances in PROTEIN CHEMISTRY

VOLUME 65

Proteome Characterization and Proteomics

ACADEMIC PRESS

Volume 58
pp. 1-474 (2001)
Novel Cofactors
Volume 57
pp. 1-405 (2001)
Prion Proteins
Volume 56
pp. 1-567 (2001)
Drug Discovery and Design
Volume 55
pp. 1-438 (2001)
Evolutionary Protein Design
Volume 54
pp. 1-481 (2000)
Analysis of Amino Acid Sequences
Volume 53
pp. 1-442 (2000)
Protein folding mechanisms

Volume 65, Pages 1-432 (2003)

Previous vol/iss Next vol/iss

Proteome Characterization and Proteomics
Edited by: Richard D. Smith and Timothy D. Veenstra
ISBN: 978-0-12-034265-5

Article List Full Abstracts
Add to my Quick Links
Display Selected Articles E-mail Articles Export Citations articles **1 - 17**

10. **Proteomics of Membrane Proteins**
Pages 271-307
Julian P Whitelegge, Stephen M G· mez and Kym F Faull
SummaryPlus | Full Text + Links | PDF (456 K) | View Related Articles

11. **Proteomics in Drug Discovery**
Pages 309-342
Rodney M Hewick, Zhijian Lu and Jack H Wang
SummaryPlus | Full Text + Links | PDF (302 K) | View Related Articles

12. **From Clone to Crystal: Maximizing the Amount of Protein Samples for Structure Determination**
Pages 343-352
Chris M Koth and Aled M Edwards
SummaryPlus | Full Text + Links | PDF (88 K) | View Related Articles

13. **Proteomics and Bioinformatics**
Pages 353-369
Carol S Giometti
SummaryPlus | Full Text + Links | PDF (180 K) | View Related Articles

14. **Author Index**
Pages 371-401
PDF (115 K) | View Related Articles

15. **Subject Index**
Pages 403-413
PDF (74 K) | View Related Articles

16. **Color Inserts**
Pages 415-418
PDF (391 K) | View Related Articles

17. **Color Inserts**
Pages 419-432
PDF (1090 K) | View Related Articles

articles **1 - 17**

PROTEOMICS IN THE POSTGENOMIC AGE

By RICHARD S. MORRISON,[*] YOSHITO KINOSHITA,[*] MARK D. JOHNSON,[*] AND THOMAS P. CONRADS[†]

[*]Department of Neurological Surgery, University of Washington School of Medicine, Seattle, Washington 98195, and [†]SAIC-Frederick, National Cancer Institute at Frederick, Frederick, Maryland 21702

I. Introduction

Advances in molecular biology and bioinformatics are making it possible to simultaneously analyze the entire complement of genes expressed in a particular cell or tissue. These advances have created unique opportunities in the field of medicine, where the results of gene expression studies are expected to help identify cellular alterations associated with disease etiology, progression, outcome, and response to therapy. These rapidly emerging technologies are also expected to result in the identification of novel therapeutic targets for a host of maladies, including infectious diseases, behavioral disorders, developmental defects, neurodegenerative diseases, aging, and cancer.

Technical advances have facilitated characterization of the three major genetic units: the genome, the transcriptome, and the proteome (Fig. 1). The *genome* describes the entire set of genes encoded by the DNA of an organism. The *transcriptome* encompasses the entire complement of messenger RNA (mRNA) transcripts transcribed from the genome of a cell. The transcriptome varies from cell to cell and fluctuates in response to numerous physiological signals, including developmental status, stress, changes in the extracellular milieu, and disease. The *proteome* describes the entire complement of proteins expressed by a cell at a point in time. Proteomic investigations also aim to determine protein localization, modifications, interactions, and, ultimately, protein function. Because the

ADVANCES IN
PROTEIN CHEMISTRY, Vol. 65

0065-3233/03 $35.00

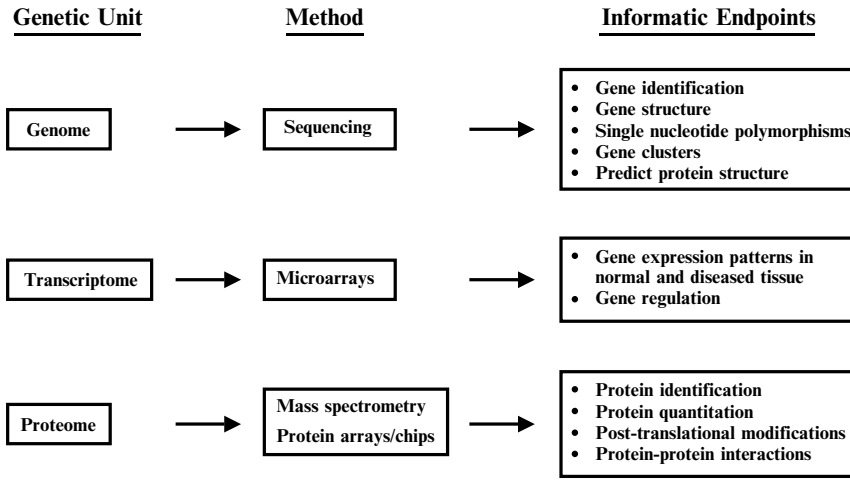


FIG. 1. Information obtained from genomic and proteomic analysis. Advances in genomic and proteomic analysis have facilitated characterization of the three major genetic units: the genome, the transcriptome, and the proteome. The data gleaned from each of these fields is in most cases unique to each genetic unit and, therefore, provide complementary information about the organization and regulation of living systems. The genome describes the entire set of genes that is encoded by the DNA of an organism and this has been obtained through a massive DNA sequencing effort. The transcriptome encompasses the entire complement of messenger RNA (mRNA) transcripts transcribed from the genome of a cell. These data have been obtained largely through the application of cDNA microarrays. The proteome describes the entire complement of proteins expressed by a cell at a point in time. Whereas two-dimensional (sodium dodecyl sulfate and isoelectric focusing) polyacrylamide gel electrophoresis has been the mainstay of proteomics analysis, the field is now embracing new techniques such as multidimensional capillary liquid chromatography coupled with tandem mass spectrometry (Link *et al.*, 1999; Washburn *et al.*, 2001), and single-dimension ultrahigh-resolution capillary liquid chromatography in combination with FTICR mass spectrometry (Jensen *et al.*, 1999). The development of high-throughput mass spectrometry methods and protein arrays will greatly accelerate the pace of proteomics research.

function of a gene is dependent on the activity of its translated protein, there has been significant impetus to develop methods that will enable high-throughput analysis of cellular proteomes. To understand the significance and impact of the rapid advances in analysis of the proteome, it is necessary to consider the utility and limitations of data obtained from analyses of the genome and transcriptome. Taken together, the study of the genome, transcriptome, and proteome provides complementary insights into a host of biological processes, and provides a greater understanding of the regulation of these processes.

II. Deciphering the Genome

Advances in nucleic acid sequencing and the software necessary to store and annotate sequence data have been instrumental in characterizing the genome of humans (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001) and other species (Fleischmann *et al.*, 1995; Blattner *et al.*, 1997; Goffeau *et al.*, 1996). The emphasis in the field of genomics has been to both generate and evaluate whole genome sequence data. The size and complexity of these genomic databases have necessitated the development of new informatic tools to organize and analyze data. The results obtained from this enormous effort will be substantial. The first major benefit of this work will be the complete identification and sequencing of the estimated 40,000 human genes that comprise the human genome (Venter *et al.*, 2001). Genes will be mapped to specific chromosomes, which will contribute to an understanding of normal development, the origin of phenotypic variability, and disease etiology and disease susceptibility in humans. Genomic data will also provide important information about intron and regulatory DNA sequences that influence such critical processes as mRNA transcription (Thieffry *et al.*, 1998) and mRNA splicing. In addition, homology alignment of new genome sequences with previously characterized genes will facilitate structural and functional predictions of expressed proteins (Gough *et al.*, 2001).

Although there is clearly much to be gleaned from genomic sequence data, interpreting the genome is complicated. It is well recognized that a single gene may encode multiple different proteins. Moreover, because coding regions are interspersed with noncoding regions of DNA and because there can be differential mRNA splicing, the genomic sequence cannot be used to reliably predict the entire spectrum of mRNA transcripts (transcriptome) or corresponding proteins expressed by a cell or tissue at any point in time.

III. Gene Expression Profiles

Most biological studies have been limited in scope to the analysis of individual mRNA transcripts or proteins. These studies employed such techniques as Northern blot analysis, RNase protection assays, reverse transcription–polymerase chain reaction, or Western blot analysis. The data derived from these techniques are specific and generally quantitative, but are limited to a small number of genes/proteins. Unfortunately, it is often difficult to appreciate how an individual gene or protein relates to an injury response, a signal transduction cascade, or a complex biological

state such as cancer. However, developments in DNA microarray technology have made it possible to simultaneously evaluate mRNA transcripts on a comprehensive level.

DNA microarrays are glass slides or nylon membranes to which cDNA sequences or oligonucleotides corresponding to select genes are affixed. Total or poly(A) RNAs are isolated from the cells or tissues being compared and reverse-transcribed to cDNAs. These are differentially labeled with fluorescent dyes or other markers. The labeled cDNAs are concomitantly hybridized to the array. The glass slides or nylon membranes are subsequently washed and scanned for intensity. A comparison of the two label intensities allows the relative expression levels of thousands of genes to be analyzed in a single experiment.

Alternatively, serial analysis of gene expression (SAGE) also provides a comprehensive and quantitative measure of gene expression (Velculescu *et al.*, 1995). SAGE is based on the generation of unique nucleotide sequence tags (10 base pairs) from a fixed position in each species of mRNA. The tags are initially prepared from mRNA that is transcribed into double-stranded cDNA and the frequency with which a tag appears in the cDNA pool reflects its relative abundance. Analyses of mRNA transcript levels, using either microarray or SAGE technology, are generally well correlated (Ishii *et al.*, 2000; Nacht *et al.*, 1999).

The most common research application of cDNA microarrays is gene expression profiling. Utilizing this approach, investigators have begun to identify subsets of genes associated with particular biological states (e.g., cancer) or that vary in response to different environmental conditions. With the completion of the Human Genome Project and through the ongoing annotation efforts, it will be possible to assess the entire subset of mRNAs (transcriptome) expressed in a tissue or cell of interest. It will also be possible to determine how the transcriptome of a cell or tissue changes with age, changing environmental conditions, or in response to injury and disease. Gene expression profiling has already proved effective in distinguishing between normal cells and tumor cells. Two subtypes of non-Hodgkin's lymphoma that could not be distinguished by traditional histological methods were distinguished by profiling 17,856 genes in patient samples (Alizadeh *et al.*, 2000). Distinct subtypes of malignant melanoma (Bittner *et al.*, 2000) and breast cancer (Perou *et al.*, 1999, 2000) have also been classified on the basis of gene expression profiling. Monitoring patterns of gene expression in malignant tissues is having a significant impact on the diagnosis and classification of many human cancers (DeRisi *et al.*, 1996; Golub *et al.*, 1999).

In a study aimed at understanding the molecular mechanisms that underlie the tumorigenesis and progression of clear cell renal cell

carcinoma (ccRCC), gene expression profiles of 29 ccRCC tumors obtained from patients with diverse clinical outcomes were analyzed with 21,632 cDNA-containing microarrays (Takahashi *et al.*, 2001). Gene expression profiles of each tumor sample were compared with cognate patient-matched normal tissue to identify gene expression alterations that occur in most ccRCCs. In addition, because all the experiments shared a ''common'' normal tissue reference, results from each experiment could be compared to identify gene expression patterns that correlated with differences in observed clinical features of the tumors. Changes in gene expression that were common to most of the ccRCCs studied and unique to clinical subsets were identified. There was a significant distinction in gene expression profiles between patients with a relatively nonaggressive form of the disease (100% survival after 5 years with 88% of the patients having no clinical evidence of metastasis) versus patients with a relatively aggressive form of the disease (average survival time of 25.4 months with a 0% 5-year survival rate). Approximately 40 genes, some of which have previously been implicated in tumorigenesis and metastasis, were identified. Moreover, the identified genes provide insight into the molecular mechanisms of aggressive ccRCC and suggest intervention strategies.

Many of the 40 genes that most effectively discriminated between patients with good outcome and those with poor outcome gave insight into the biology of the two groups of ccRCC (Takahashi *et al.*, 2001). *Sprouty*, the mammalian homolog of the *Drosophila melanogaster* angiogenesis inhibitor, was found to be exclusively upregulated in the good outcome group, which suggests that failure to properly inhibit angiogenesis may contribute to aggressive forms of ccRCC. Transforming growth factor (TGF-β), TGF-β receptor II (TGF-βRII), and its downstream effector, tissue inhibitor of metalloproteinase 3(TIMP3), were exclusively downregulated in the poor outcome group. Loss of the TGF-βII signaling pathway has previously been shown to contribute in aggressive cancer development (Engel *et al.*, 1999), and loss of TIMP3 expression by promoter methylation was shown to increase tumorigenicity (Bachman *et al.*, 1999). The identification of this pathway as downregulated in aggressive ccRCC suggests numerous targets for intervention to supplement the still low response rate of current adjuvant therapies.

There are a multitude of biological questions in addition to cancer that can be addressed by gene expression profiling. For example, this technology is being used to determine the molecular basis of apoptosis (Voehringer *et al.*, 2000) and to unlock the secrets of the aging brain (Lee *et al.*, 2000). Despite the utility that gene expression profiling provides, however, there are significant questions that cannot be answered by this

powerful technology. Genomics and gene expression profiling convey only limited information about the translated proteins that are ultimately encoded by the genome. The varied and complex properties of proteins cannot be reliably predicted by a simple linear readout of the genomic blueprint or the transcriptome.

IV. A Niche for Proteomics

Fortunately, the rapidly evolving field of proteomics (study of the proteome) is directed toward providing a comprehensive view of the characteristics and activity of every cellular protein. The proteome is clearly more complicated than the genome. The concept that one gene corresponds to one protein no longer holds true. A single gene can encode multiple different proteins. This can be attributed to (1) alternative splicing of the mRNA transcript, (2) the use of alternative translation start or stop sites, and (3) the occurrence of frame-shifting, during which a different set of triplet codons in the mRNA is translated. The net result of these activities is the generation of a proteome that contains many proteins derived from shared or overlapping genomic sequences (Fig. 2).

Another powerful impetus for moving beyond the transcriptome is the demonstration by several researchers that protein levels do not faithfully correlate with mRNA levels (e.g., Anderson and Seilhamer, 1997; Gygi *et al.*, 1999; O'Shaughnessy *et al.*, 2000). An analysis of 106 genes in the yeast *Saccharomyces cerevisiae* demonstrated that the levels of protein expression attributed to mRNA species of equal abundance could vary by as much as 30-fold. Conversely, the mRNA levels for proteins that were expressed at comparable levels varied as much as 20-fold. Experience from our own laboratory with cDNA microarray analysis yielded similar results. We identified a novel transcript in malignant mouse astrocytes, pescadillo, which was upregulated approximately 3-fold relative to nontransformed mouse astrocytes (Kinoshita *et al.*, 2001). Despite a 3-fold difference in the abundance of pescadillo transcripts, pescadillo protein levels were elevated more than 50-fold in the malignant mouse astrocytes (Y. Kinoshita, G. Foltz, J. Schuster, P. S. Nelson and R. S. Morrison, unpublished results). These results demonstrate that it is not always possible to predict changes in protein levels on the basis of changes in mRNA abundance.

One additional characteristic of proteins that is difficult to predict from genomic sequence data is the nature of their posttranslational modifications. In contrast to DNA and RNA, proteins can be modified

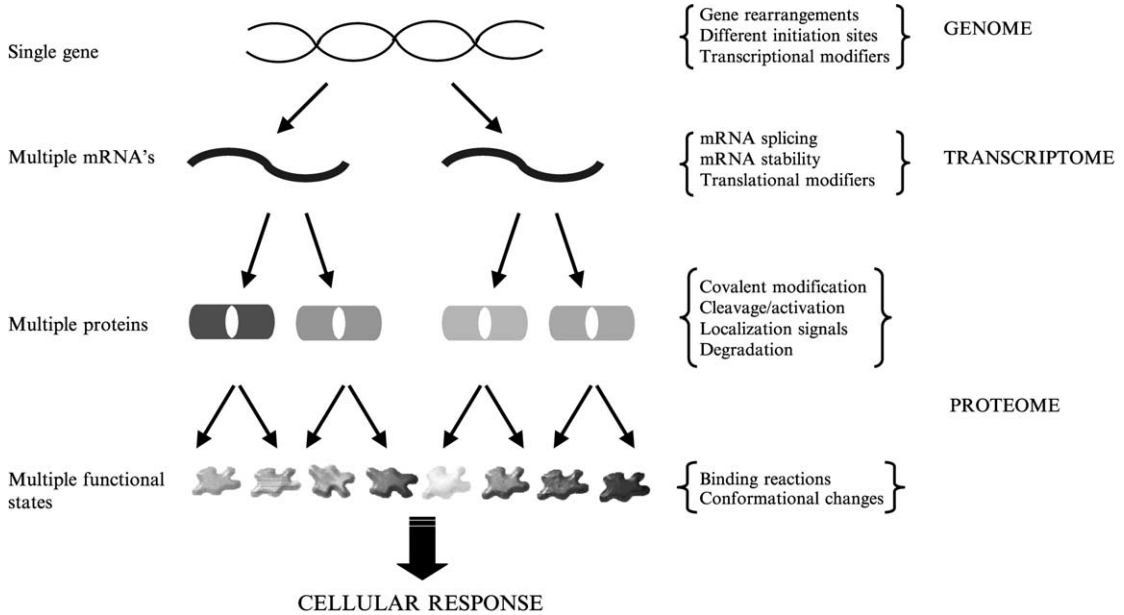


FIG. 2. The case for analyzing the proteome. At the level of the genome, a single gene may undergo rearrangement, or its transcription can be altered by several mechanisms to produce several different mRNAs. These mRNAs may then be alternately spliced, translated at varying rates, or differentially degraded to yield a potentially large and dynamically variable mRNA pool (transcriptome). Further complexity arises from a variety of posttranslational modifications and intermolecular interactions that yield additional functional protein states. Thus, dozens of unique functional protein states may be derived from a single gene. Because the process of protein production from DNA passes through so many levels of increasing complexity, it is difficult to predict the expressed complement of functional protein states (and the associated cellular response) from the genetic blueprint alone. Only a direct analysis of the proteome itself can answer these questions.

by phosphorylation, glycosylation, acetylation, nitrosation, poly(ADP-ribosyl)ation, ubiquitination, farnesylation, sulfation, linkage to glycophosphatidylinositol anchors, and SUMOylation (SUMO, small ubiquitin related modifier). In total, there are about 300 different posttranslational modifications that have been reported (Aebersold and Goodlett, 2001). These modifications can profoundly affect protein conformation, stability, localization, binding interactions, and function. Proteins are often modified at multiple sites, and it is not possible to predict from a sequence with complete certainty which sites will be modified in response to a specific set of conditions. The p53 tumor suppressor protein is a striking example of a protein that is modified at multiple sites in response to different stimuli. It is a nuclear phosphoprotein that is modified, in response to DNA damage, by the addition of phosphate to multiple seryl residues. The phosphorylation of certain seryl residues is required for p53-mediated transcription of several downstream targets associated with cell cycle arrest, including p21WAF1/Cip1 (p21) and mdm-2 (Jabbur *et al.*, 2000; Oda *et al.*, 2000). In contrast to these specific seryl residues, phosphorylation at other seryl residues regulates the transcriptional activation of apoptosis (Oda *et al.*, 2000). Moreover, the phosphorylation pattern of the p53 molecule can vary in an injury-dependent manner. Exposing cultured fibroblasts to nitric oxide induces a different pattern of p53 phosphorylation than exposure to γ irradiation, UV light, and doxorubicin (Adriamycin) (Nakaya *et al.*, 2000). p53 also requires multiple forms of posttranslational modification to manifest its activity. In addition to phosphorylation, conjugation of the ubiquitin-like molecule, SUMO-1, enhances the transcriptional activity of p53 (Gostissa *et al.*, 1999; Muller *et al.*, 2000; Rodriguez *et al.*, 1999). Although it is possible to analyze a genetic sequence for the presence of putative consensus sites for various posttranslational modifications, the mere presence of such sites does not indicate whether they are utilized, under what circumstances they are utilized, or if they are utilized simultaneously.

V. The Proteome: Greater than the Sum of its Parts

The disparity between mRNA levels and protein expression suggests that characterizing the proteome, rather than the transcriptome, under different conditions may provide a more accurate representation of the biological state of a cell. At the present time, evaluating the proteome is more difficult and labor intensive than characterizing the transcriptome. Two-dimensional (sodium dodecyl sulfate and isoelectric focusing)

polyacrylamide gel electrophoresis (2D-PAGE) is currently considered a cornerstone of proteomics techniques. 2D-PAGE produces high-resolution protein separations resulting in the display of potentially thousands of protein spots (O'Farrell, 1975). The large number of protein spots often conveys the impression that this technology provides a relatively complete view of the proteome. Unfortunately, several factors limit the resolving power of 2D-PAGE. For example, hydrophobic and membrane proteins exhibit incomplete solubility in the 2D-PAGE system, and very large or small and highly basic or acidic proteins are difficult to resolve. In addition, a single gene can represent multiple spots on a gel and, conversely, a single spot can be composed of several different gene products (Gauss *et al.*, 1999). The complexity of spot patterns arises from posttranslational modifications, alternative splicing, protein degradation, and artifacts of the 2D-PAGE system. Methods for identifying protein spots can also be problematic in that they are slow, expensive, and often insensitive. Even the combination of 2D-PAGE and mass spectrometry detects only the most abundant proteins (Gygi *et al.*, 2000), indicating that the 2D-PAGE approach does not provide sufficient coverage of the proteome.

These two methods show contrasting strategies on how to identify as much of the proteome as possible. Proteome analysis is generally driven by a two-stage process: protein fractionation and mass spectral analysis. The first strategy advocates the use of multidimensional separations so that the number of peptides analyzed by the mass spectrometer at any given time is minimized. The fewer the number of peptides within any given spectrum, the greater the increase in overall dynamic range of the measurements, thereby optimizing the detection of lower abundance peptides, even using conventional mass spectrometric instrumentation. As is discussed in Chapter 9 of this volume, there are a great number of possible combinations of separations that can be used to fractionate proteome samples. The second strategy advocates combining a single dimensional separation with advanced mass spectrometric (MS) instrumentation. Although a high-resolution separation is still necessary, the gains in dynamic range and sensitivity needed to identify low-abundance proteins are ultimately provided by the MS instrumentation. The MS developments primarily focus on methods to introduce as many peptides into the instrument as possible as well as on ways to manipulate the ion population so that lower abundance species can be exclusively measured within a complex sample. At this point in proteomic technology development it is not clear which strategy is best; however, it seems likely that the broadest proteomic coverage will come from a combination of multidimensional fractionation and advanced MS instrumentation.

Highly sensitive proteome measurements are of crucial practical importance because many important proteins, such as transcription factors, kinases, and phosphatases, are typically present in low copy number in the cell. High sensitivity also enables higher fidelity measurements of subtle changes in relative protein expression and potentially promises that much smaller cell populations or tissue sample sizes (e.g., obtained from laser-capture microdissected tissues) will be amenable to proteome analysis.

The technical limitations associated with 2D-PAGE technology have prompted the development of new proteomic methods. They are expected to enhance the identification of many additional proteins. For example, a method has been developed to analyze protein digests by multidimensional capillary liquid chromatography (LC) coupled with tandem mass spectrometry (Link *et al.*, 1999; Washburn *et al.*, 2001). Another novel method currently being evaluated involves the analysis of global proteolytic digests by a single dimension of ultrahigh-resolution capillary liquid chromatography in combination with a powerful Fourier transform ion cyclotron resonance (FTICR) mass spectrometer (Jensen *et al.*, 1999). A key aspect of using FTICR in proteomics is its ability to routinely obtain very high mass measurement accuracy (MMA) for the large numbers of proteins or proteolytic peptides delivered to the mass spectrometer by online separations. The MMA that can be routinely obtained with FTICR instrumentation is potentially ~0.1 ppm (although much more accurate measurements have been demonstrated for specialized applications), which is much higher than the 5- to 20-ppm MMA achievable with the best conventional mass spectrometer technologies and the 100- to 500-ppm accuracy of the most widely used quadrupole and ion trap MS technologies. The high MMA achievable with FTICR will allow the use of proteolytic fragments as biomarkers or accurate mass tags (AMTs) to uniquely identify a protein; an approach that is robust because of its amenability to highly posttranslationally modified proteins (as many protein fragments will not be modified and can serve as AMTs) (Conrads *et al.*, 2000). Computer simulations of this approach and initial experimental results have shown that, when the complete genome sequence information is available, an accurate mass measurement for only a single unmodified peptide allows protein identification. The online combination of high-resolution capillary LC with FTICR mass spectrometry will provide the desired high throughput, sensitivity, dynamic range, and MMA needed for wide-scale protein identification and quantitation. Although FTICR instrumentation is still not widely used, and its cost is higher than that of other types of MS instrumentation, this situation is rapidly changing because of the growing recognition of its capabilities in biological research.

Proteomics is no longer a method restricted to cataloging the presence or absence of proteins. Aebersold and colleagues have developed a novel isotope-coded affinity tag (ICAT) strategy that permits the stable isotope labeling of cysteine residues in proteins, thus facilitating a quantitative global analysis of differences in protein expression (Gygi *et al.*, 1999). In the ICAT approach, proteins are reacted with a thiol-specific reagent containing a biotin functionality group. The thiol-specific group allows for the chemical modification of reduced cysteine residues, while the biotin group is used in combination with immobilized avidin to isolate these modified peptides. Two isotopically distinct versions of ICAT reagent are available: a light isotopic version and a heavy isotopic version in which eight protons within the linker region have been substituted with eight deuterons. The ICAT approach allows for a significant reduction in the complexity of polypeptide mixtures while maintaining broad proteome coverage, and provides for proteome-wide precise quantitation of protein expression levels. Whereas metabolic labeling strategies using defined media offer the highest precision for quantitating protein abundances (because labeling occurs at the earliest possible point), the broadest proteome coverage, and minimal handling, postisolation isotopic labeling is broadly applicable to proteins extracted from any conceivable source (e.g., tissue samples). In addition, the ICAT strategy provides reduced mixture complexity because it modifies only cysteine-containing polypeptides, which can aid identification by providing an additional sequence constraint.

This technology makes it possible to evaluate a larger subset of proteins, including low-abundance proteins, in relation to any particular disease process or biological process. For example, if a subset of proteins were simply identified, but not quantified, in two biologically distinct samples (e.g., normal and diseased tissue), their involvement in the disease might be overlooked in favor of proteins that were either present or absent in one of the samples. Important, but perhaps subtle, changes in the levels of protein expression, which might be critical to a biological process, might be overlooked or ignored. With the advent of the ICAT labeling process, quantitative global measurements of proteomes can now be made in the context of a biological or pathological process.

Although the ICAT approach shows great promise for determining relative protein abundances, delineation of protein function solely on the basis of abundance changes will be limited because numerous vital activities of proteins are modulated by posttranslational modifications that may not be reflected by changes in protein abundance. One of the most important posttranslational protein modifications used to modulate protein activity and propagate signals within cellular pathways and

networks is phosphorylation (Cohen, 1982, 1992; Pawson and Scott, 1997). Studies estimate that as many as one-third of all cellular proteins derived from mammalian cells are phosphorylated (Pawson and Scott, 1997). Broad-ranging cellular processes such as protein kinase activation, cell cycle progression, cellular differentiation and transformation, development, peptide hormone response, and adaptation are all regulated by changes in the state of protein phosphorylation. Although regulation can occur at the level of protein synthesis along with concomitant proportional changes in protein phosphorylation, there are also examples of regulating protein function by phosphorylation without altering protein abundance (Huang *et al.*, 1998). Therefore the ability to broadly identify changes in the phosphorylation state of a protein may lead to discoveries related to protein activity—regardless of whether that protein is differentially expressed—and improve our understanding of cellular systems.

The predominant method used to study changes in protein phosphorylation is by labeling proteins with ^{32}P inorganic phosphate ($^{32}P_i$). To measure differences in relative abundances of phosphorylation, ^{32}P-labeled proteomes are resolved by 2D-PAGE and the relative spot intensities are compared (van der Geer and Hunter, 1994; Mason *et al.*, 1998). The use of $^{32}P_i$ to label proteins does not lend itself to high-throughput proteome-wide analysis because of the problems with handling radioactive compounds and the associated contamination of instrumentation. It would be valuable to identify other methods posing less of a risk than $^{32}P_i$, yet still able to effectively identify phosphorylated proteins and quantitate the extent of phosphorylation.

One major difficulty encountered in identification of phosphopeptides in complex mixtures concerns the challenge associated with enriching samples for these species. Although phosphospecific antibodies and metal affinity columns have been widely used, they typically result in isolation of nonphosphorylated species along with the phosphopeptides of interest. In addition, neither of these methods is capable of quantitatively determining the relative phosphorylation states of proteins isolated from different sources. Two new methods have been developed that provide for the specific enrichment and quantitation of phosphopeptides. Both methods utilize stable isotopes to differentially label samples to be compared and employ subsequent MS analysis for the identification and quantitation of the enriched phosphopeptide mixture.

The first strategy to isolate and quantitate phosphopeptides we discuss was developed concurrently, and independently, by two groups (Goshe *et al.*, 2001, 2002; Oda *et al.*, 2001). Although there are subtle differences in the specific procedures, the overall approach employed by the two

methods is inherently similar. The first step involves blocking reactive thiolates of cysteinyl residues via reductive alkylation or performic acid oxidation. In the next step, phosphate moieties are removed via hydroxide ion-mediated β elimination from the phosphoseryl (pSer) and phosphothreonyl (pThr) residues, resulting in their conversion to dehydroalanyl and β-methyl dehydroalanyl residues, respectively. The newly formed α,β-unsaturated double bond renders the β carbons in each of the newly chemically modified residues electrophilic. The next modification involves a Michael-type addition of the bifunctional reagent 1,2-ethanedithiol (EDT). The addition of EDT to either the dehydroalanyl or β-methyl dehydroalanyl residues results in the creation of a free thiolate in place of what was formerly a phosphate moiety. Iodoacetyl polyethylene oxide (PEO)-biotin can be used to covalently modify the new thiolate site, where the end result is the covalent modification of phosphoryl residues with a linker molecule that contains a terminal biotin group. The stable isotopic labeling that enables relative quantification is achieved by using commercially available sources of either a light ($HSCH_2CH_2SH$, EDT-D_0) or heavy ($HSCD_2CD_2SH$, EDT-D_4) isotopomeric version of EDT. The samples are subsequently digested with trypsin (or another proteolytic enzyme or chemical cleavage methodology) and the modified peptides are specifically extracted by immobilized avidin chromatography and analyzed by reversed-phase liquid chromatography (LC) coupled directly online with MS.

Two experimental MS strategies (MS and tandem MS) are used to identify and quantify the phosphorylation state of the phosphopeptides. In the MS mode, the masses of the intact peptide pairs are measured and the relative signal intensities provide a direct measure of the phosphorylation status of the peptide. The MS signals originating from the modified versions of the phosphopeptides are easily recognizable because they occur as pairs separated by the mass difference between the EDT-D_0 and EDT-D_4 labels (i.e., 4.0 Da).

A second labeling method to isolate and quantify phosphopeptides has been developed in the laboratory of Aebersold (Zhou *et al.*, 2001). The sequence of chemical reactions for selectively isolating phosphopeptides from a peptide mixture consists of six steps. To eliminate potential intra- and intermolecular condensation, the peptide amino groups are protected using *test*-butyl dicarbonate (tBoc) chemistry (Bodanszky, 1984). Following this the carboxylate and phosphate groups are modified via a carbodiimide-catalyzed condensation reaction to form amide and phosphoramidate bonds. The phosphoramidate bonds are then hydrolyzed via a brief acid wash to deprotect the phosphate group and cystamine is attached to the regenerated phosphate group via another

carbodiimide-catalyzed condensation reaction. A free sulfhydryl group is generated at each phosphate group by reduction of the internal disulfide of cystamine, which allows the peptides to be attached to iodoacetyl groups immobilized on glass beads. The covalent attachment of the peptides allows stringent washing conditions to be used, thereby reducing the amount of nonspecifically bound components being recovered with the phosphopeptides of interest. The phosphopeptides are recovered by cleavage of phosphoramidate bonds, using trifluoroacetic acid at a concentration that also removes the tBoc protection group, thus regenerating peptides with free amino and phosphate groups. The carboxylate groups, however, remain blocked from step 2. Although the chemistry involved is more complex, this method potentially provides greater enrichment because the phosphopeptides are covalently linked to a solid support during processing.

Zhou *et al.* (2001) noted that this method yielded mixtures highly enriched in phosphopeptides with minimal contamination from other peptides. Because this strategy does not require the removal of the phosphate group it is equally applicable to phosphoseryl, phosphothreonyl, and phosphotyrosyl residue-containing peptides. The collision-induced dissociation (CID) spectra of the modified phosphopeptides were of high enough quality to allow the peptides to be identified by sequence database searching. The CID spectra could discriminate between pSer/pThr- and pTyr-containing peptides because pSer and pThr lose an H_3PO_4 group on tandem MS (Jonscher and Yates, 1997; Qin and Chait, 1997), allowing these residues to be identified via a fragment ion corresponding to the loss of 98 Da. Phosphotyrosyl residues are more stable and do not lose their phosphate group during fragmentation. Although this strategy does not provide a direct method to quantify changes in phosphorylation state between peptides from two different samples, the blocking of the carboxylates using either normal isotopic abundance or deuterated ethanolamine (i.e., ethanolamine-d_4) does allow for incorporation of stable isotope tags that later can be differentiated and quantified by MS.

Another key attribute of these isotopic chemical modification strategies to identify phosphopeptides is that the modification remains attached to the residue during tandem MS fragmentation of the peptide. During tandem MS the intact peptide is subjected to CID, whereby the ion selected collides with an inert gas (e.g., nitrogen or helium) that causes it to fragment into smaller ions. The MS spectrum of these fragment ions typically provides partial sequence information that can be used in conjunction with commercially available computer algorithms to identify the peptide. In a typical MS experiment, however, the phosphate group

dissociates from the phosphorylated residues during the tandem MS analysis (indeed, often even in MS analyses as well), preventing site-specific assignment of the phosphate modification. The isotope labeling strategy described above, however, allows the exact phosphorylation site to be determined by tandem MS.

VI. Biological and Clinical Applications

The ability to analyze a proteome, in total or in part, should yield immense benefits for the biological and medical community. Comparing the proteomes of normal and diseased tissues may ultimately lead to the identification of new diagnostic markers. Proteomics has already proved beneficial in the study of heart disease, and has resulted in the identification of disease-specific proteins associated with dilated cardiomyopathy (Warraich *et al.*, 1999; Weekes *et al.*, 1999). The field of oncology has also derived significant benefits from the study of proteomics. This approach has resulted in the identification of proteins specifically associated with squamous cell carcinoma of the bladder, which have been used to generate antibodies capable of identifying metaplastic lesions of the bladder (Celis *et al.*, 1999). The identification of psoriacin as a protein marker of this cancer may prove useful for monitoring the status of the disease (Ostergaard *et al.*, 1999). Similar advances have been made with human breast cancer (Page *et al.*, 1999) and, in the case of prostate cancer, protein expression profiles have been used to detect androgen-regulated proteins (Nelson *et al.*, 2000).

Our own efforts have focused on characterizing the proteome of nontransformed and malignant mouse and human astrocytes. We have coupled our studies of the human astrocytoma proteome with a mouse model of astrocyte tumorigenesis (Yahanda *et al.*, 1995) because it eliminates the cellular heterogeneity observed within and between tumor samples from patients. Cultures of p53-deficient malignant mouse astrocytes represent a homogeneous group of cells that can be easily manipulated and studied under defined conditions. Because p53-deficient mouse astrocytes acquire a malignant phenotype in a reproducible temporal sequence with serial passaging in culture, it is possible to characterize the proteome of astrocytes at different stages of malignancy. By characterizing the proteome of astrocytes at different stages of malignancy, we hope to identify subsets of proteins that are causally related to the process of malignant transformation. Comparing the proteome of malignant mouse astrocytes with the human astrocytoma proteome should facilitate the identification of tumor-specific proteins.

Proteins identified in this manner will, it is hoped, improve the diagnosis of human astrocytomas and provide promising new targets for drug development.

In general, this approach could improve the identification and subclassification of tumors that display similar histological characteristics but variable clinical outcomes such as anaplastic astrocytomas of the central nervous system. Anaplastic astrocytomas are diagnosed on the basis of specific histological criteria, but patients with these tumors display dramatically different survival times. The goal of proteomics is to identify specific subsets of proteins that will help distinguish between tumors with widely divergent clinical outcomes. Moreover, in the case of human brain tumors, we are utilizing proteomics to identify subsets of proteins that will predict the likelihood that a tumor will respond to radiation and chemotherapy or recur after surgical resection.

The increasing sensitivity of mass spectrometry methods being used for proteome characterization has made it feasible to analyze the proteome of biological samples that are difficult to obtain or that contain limited amounts of protein. For example, the analysis of proteins in cerebrospinal fluid obtained from patients with brain tumors, neurodegenerative diseases, traumatic brain injury, stroke, and infections may provide new disease markers or markers that reflect the changing course of a disease. In the case of neural diseases or injuries that involve a significant degree of inflammation, proteomic analysis of cerebrospinal fluid may help identify which inflammatory mediators to target. A proteome map currently being constructed for the human hippocampus (Edgar *et al.*, 1999a) has led to the identification of 18 proteins that exhibit an abnormal pattern of expression in the brains of patients diagnosed with schizophrenia (Edgar *et al.*, 1999b). Because several of these genes are clustered, the results have implicated chromosome 6q in the etiology of schizophrenia (Edgar *et al.*, 2000).

VII. Proteomics Meets Cell Biology

Proteins are dynamic, and they are carefully regulated in response to cellular perturbations. Perturbations come in many forms, including direct cell–cell interactions, cell–substrate interactions, membrane receptor activation, and changes in extracellular ion concentrations, to name a few. The ensuing modifications to proteins are also numerous, and they govern the dynamic range of activities that occur both inside and outside of cells. Proteins respond to altered conditions by translocation, covalent modifications, cleavage, degradation, and the acquisition of new binding partners.

Proteomic strategies are allowing scientists to monitor these dynamic cellular changes. Subcellular fractionation, protein tagging, and protein purification techniques continue to improve, facilitating the separation of discrete subcellular units. The separation of cytosolic, mitochondrial, nuclear, plasma membrane, or nuclear membrane fractions followed by mass spectrometry measurements is enabling the comprehensive identification of proteins present in these different subcellular locations (Bell *et al.*, 2000; Koc *et al.*, 2000; Lopez *et al.*, 2000; Taylor *et al.*, 2000; Verma *et al.*, 2000; Wu *et al.*, 2000). For example, in one study, proteomics was used to investigate the composition of a mixture of proteins released from purified mitochondria on opening of the permeability transition pore (Patterson *et al.*, 2000). This analysis will provide insight into the soluble proteins that are released from the mitochondrial intermembrane space and matrix that are involved in the activation of cell death mediators such as caspases and nucleases. Performing these measurements at different times after a specific treatment will provide a dynamic image of how proteins change their compartmentalization and move throughout a cell.

More directed separations could also be performed, allowing investigators to characterize all the players in a particular functional unit. Affinity purification techniques employing specific proteins or protein domains, antibodies, metals, DNA, and other binding substrates will facilitate the enrichment of functionally defined proteins. In one striking example of this approach, anti-phosphoserine and anti-phosphotyrosine antibodies were used to extract more than 500 phosphorylated proteins from mouse fibroblasts stimulated through the platelet-derived growth factor β receptor (Soskic *et al.*, 1999). More than 100 of these phosphoproteins displayed alterations in their phosphorylation status in response to receptor activation, demonstrating that this approach provides a powerful method for identifying the protein intermediates involved in specific signal transduction cascades. In another study, matrix-assisted laser desorption/ionization time-of-flight mass spectrometry analysis (MALDI-TOF) was applied in conjunction with an anti-phosphotyrosine antibody to detect time-dependent changes in the phosphorylation status of proteins in response to tumor necrosis factor α (TNF-α) signaling (Yanagida *et al.*, 2000). Among the proteins that showed time-dependent changes in staining intensity were several proteins that had no previous known function in the TNF-α signal transduction pathway. A proteomics approach has also been used to identify novel targets associated with the mitogen-activated protein (MAP) kinase-signaling pathway (Lewis *et al.*, 2000). These results demonstrate the usefulness of proteomic methods for

comprehensive analysis of the proteins involved in signal transduction cascades.

Alternatively, proteins labeled with an epitope tag can be overexpressed, immunopurified from cells, and subjected to mass spectrometry to identify binding proteins or proteins associated with a particular protein complex. A wide variety of affinity tags are being adapted to protein arrays, enabling the implementation of high-throughput screens to quantitate specific proteins in complex solutions (Haab *et al.*, 2001; Tomlinson and Holt, 2001) and to determine protein localization, protein–protein interactions, and biochemical analysis of protein function (Zhu and Snyder, 2001).

VIII. Summary

Technical developments in the field of proteomics are poised to generate advances in our understanding of protein structure, function, and organization in complex signaling and regulatory networks. Improvements in mass spectrometry instrumentation, the implementation of protein arrays, and the development of robust informatics software are providing sensitive, high-throughput technologies for large-scale identification and quantitation of protein expression, protein modifications, subcellular localization, protein function, and protein–protein interactions. These advances have significant implications for understanding how cellular proteomes are regulated in health and disease.

References

Aebersold, R., and Goodlett, D. R. (2001). Mass spectrometry in proteomics. *Chem Rev.* **101,** 269–295.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., and Staudt, L. M. *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403,** 503–511.

Anderson, L., and Seilhamer, J. (1997). A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18,** 533–537.

Bachman, K. E., Herman, J. G., Corn, P. G., Merlo, A., Costello, J. F., Cavenee, W. K., Baylin, S. B., and Graff, J. R. (1999). Methylation-associated silencing of the tissue inhibitor of metalloproteinase-3 gene suggests a suppressor role in kidney, brain, and other human cancers. *Cancer Res.* **59,** 798–802.

Bell, A. W., Ward, M. A., Freeman, H. N., Choudhary, J. S., Blackstock, W. P., Lewis, A. P., Fazel, A., Gushue, J. N., Paiement, J., Palcy, S., Chevet, E., Lafreniere-Roula, M., Solari, R., Thomas, D. Y., Rowley, A., and Bergeron, J. M. (2000). Proteomics

characterization of abundant Golgi membrane proteins. *J. Biol. Chem.* **276,** 5152–5165.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich, K., Beaudry, C., Berens, M., Alberts, D., and Sondak, V. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406,** 536–540.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1474.

Bodanszky, A. B. M. (1984). ''The Practice of Peptide Synthesis.'' Springer-Verlag, New York.

Celis, J. E., Celis, P., Ostergaard, M., Basse, B., Lauridsen, J. B., Ratz, G., Rasmussen, H. H., Orntoft, T. F., Hein, B., Wolf, H., and Celis, A. (1999). Proteomics and immunohistochemistry define some of the steps involved in the squamous differentiation of the bladder transitional epithelium: A novel strategy for identifying metaplastic lesions. *Cancer Res.* **59,** 3003–3009.

Cohen, P. (1982). The role of protein phosphorylation in neural and hormonal control of cellular activity. *Nature* **296,** 613–620.

Cohen, P. (1992). Signal integration at the level of protein kinases, protein phosphatases and their substrates. *Trends Biochem. Sci.* **17,** 408–413.

Conrads, T. P., Anderson, G. A., Veenstra, T. D., Pasa-Tolic, L., and Smith, R. D. (2000). Utility of accurate mass tags for proteome-wide protein identification. *Anal. Chem.* **72,** 3349–3354.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14,** 457–460.

Edgar, P. F., Douglas, J. E., Knight, C., Cooper, G. J., Faull, R. L., and Kydd, R. (1999a). Proteome map of the human hippocampus. *Hippocampus* **9,** 644–650.

Edgar, P. F., Schonberger, S. J., Dean, B., Faull, R. L., Kydd, R., and Cooper, G. J. (1999b). A comparative proteome analysis of hippocampal tissue from schizophrenic and Alzheimer's disease individuals. *Mol. Psychiatry* **4,** 173–178.

Edgar, P. F., Douglas, J. E., Cooper, G. J., Dean, B., Kydd, R., and Faull, R. L. (2000). Comparative proteome analysis of the hippocampus implicates chromosome 6q in schizophrenia. *Mol. Psychiatry* **5,** 85–90.

Engel, J. D., Kundu, S. D., Yang, T., Lang, S., Goodwin, S., Janulis, L., Cho, J. S., Chang, J., Kim, S. J., and Lee, C. (1999). Transforming growth factor-β type II receptor confers tumor suppressor activity in murine renal carcinoma (Renca) cells. *Urology* **54,** 164–170.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., and Merrick, J. M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512.

Gauss, C., Kalkum, M., Lowe, M., Lehrach, H., and Klose, J. (1999). Analysis of the mouse proteome. I. Brain proteins: Separation by two-dimensional electrophoresis and identification by mass spectrometry and genetic variation. *Electrophoresis* **20,** 575–600.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 genes. *Science* **274, 546,** 563–567.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537.

Goshe, M. B., Conrads, T. P., Panisko, E. A., Angell, N. H., Veenstra, T. D., and Smith, R. D. (2001). Phosphoprotein isotope-coded affinity tag approach for isolating and quantitating phosphopeptides in proteome-wide analyses. *Anal. Chem.* **73,** 2578–2586.

Goshe, M. B., Veenstra, T. D., Panisko, E. A., Conrads, T. P., Angell, N. A., and Smith, R. D. (2002). Phosphoprotein isotope coded affinity tags: Application to the enrichment and quantitation of low-abundance phosphoproteins. *Anal. Chem.* **74,** 607–611.

Gostissa, M., Hengstermann, A., Fogal, V., Sandy, P., Schwarz, S. E., Scheffner, M., and Del Sal, G. (1999). Activation of p53 by conjugation to the ubiquitin-like protein SUMO-1. *EMBO J.* **18,** 6462–6471.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313,** 903–919.

Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19,** 1720–1730.

Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000). Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. USA* **97,** 9390–9395.

Haab, B. B., Dunham, M. J., and Brown, P. O. (2001). Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol.* **2,** 1–13.

Huang, C., Ma, W. Y., Young, M. R., Colburn, N., and Dong, Z. (1998). Shortage of mitogen-activated protein kinase is responsible for resistance to AP-1 transactivation and transformation in mouse JB6 cells. *Proc. Natl. Acad. Sci. USA* **95,** 156–161.

International Human Genome Sequencing Consortium(2001). Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.

Ishii, M., Hashimoto, S., Tsutsumi, S., Wada, Y., Matsushima, K., Kodama, T., and Aburatani, H. (2000). Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics* **68,** 136–143.

Jabbur, J. R., Huang, P., and Zhang, W. (2000). DNA damage-induced phosphorylation of p53 at serine 20 correlates with p21 and Mdm-2 induction in vivo. *Oncogene* **19,** 6203–6208.

Jensen, P. K., Pasa-Tolic, L., Anderson, G. A., Horner, J. A., Lipton, M. S., Bruce, J. E., and Smith, R. D. (1999). Probing proteomes using capillary isoelectric focusing-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **71,** 2076–2084.

Jonscher, K. R., and Yates, J. R, III. (1997). Matrix-assisted laser desorption ionization/quadrupole ion trap mass spectrometry of peptides: Application to the localization of phosphorylation sites on the P protein from Sendai virus. *J. Biol. Chem.* **272,** 1735–1741.

Kinoshita, Y., Jarell, A. D., Flaman, J. M., Foltz, G., Schuster, J., Sopher, B. L., Irvin, D. K., Kanning, K., Kornblum, H. I., Nelson, P. S., Hieter, P., and Morrison, R. S. (2001). Pescadillo, a novel cell cycle regulatory protein abnormally expressed in malignant cells. *J. Biol. Chem.* **276,** 6656–6665.

Koc, E. C., Burkhart, W., Blackburn, K., Moseley, A., Koc, H., and Spremulli, L. L. (2000). A proteomics approach to the identification of mammalian mitochondrial small subunit ribosomal proteins. *J. Biol. Chem.* **275,** 32585–32591.

Lee, C. K., Weindruch, R., and Prolla, T. A. (2000). Gene-expression profile of the ageing brain in mice. *Nat. Genet.* **25,** 294–297.

Lewis, T. S., Hunt, J. B., Aveline, L. D., Jonscher, K. R., Louie, D. F., Yeh, J. M., Nahreini, T. S., Resing, K. A., and Ahn, N. G. (2000). Identification of novel MAP kinase pathway signaling targets by functional proteomics and mass spectrometry. *Mol. Cell* **6,** 1343–1354.

Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R. (1999). Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676–682.

Lopez, M. F., Kristal, B. S., Chernokalskaya, E., Lazarev, A., Shestopalov, A. I., Bogdanova, A., and Robinson, M. (2000). High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *Electrophoresis* **21,** 3427–3440.

Mason, G. G., Murray, R. Z., Pappin, D., and Rivett, A. J. (1998). Phosphorylation of ATPase subunits of the 26S proteasome. *FEBS Lett.* **430,** 269–274.

Muller, S., Berger, M., Lehembre, F., Seeler, J. S., Haupt, Y., and Dejean, A. (2000). c-Jun and p53 activity is modulated by SUMO-1 modification. *J. Biol. Chem.* **275,** 13321–13329.

Nacht, M., Ferguson, A. T., Zhang, W., Petroziello, J. M., Cook, B. P., Gao, Y. H., Maguire, S., Riley, D., Coppola, G., Landes, G. M., Madden, S. L., and Sukumar, S. (1999). Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer. *Cancer Res.* **59,** 5464–5470.

Nakaya, N., Lowe, S. W., Taya, Y., Chenchik, A., and Enikolopov, G. (2000). Specific pattern of p53 phosphorylation during nitric oxide-induced cell cycle arrest. *Oncogene* **19,** 6369–6375.

Nelson, P. S., Han, D., Rochon, Y., Corthals, G. L., Lin, B., Monson, A., Nguyen, V., Franza, B. R., Plymate, S. R., Aebersold, R., and Hood, L. (2000). Comprehensive analyses of prostate gene expression: Convergence of expressed sequence tag databases, transcript profiling and proteomics. *Electrophoresis* **21,** 1823–1831.

Oda, K., Arakawa, H., Tanaka, T., Matsuda, K., Tanikawa, C., Mori, T., Nishimori, H., Tamai, K., Tokino, T., Nakamura, Y., and Taya, Y. (2000). p53AIP1, a potential mediator of p53-dependent apoptosis, and its regulation by Ser-46-phosphorylated p53. *Cell* **102,** 849–862.

Oda, Y., Nagasu, T., and Chait, B. T. (2001). Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nat. Biotechnol.* **19,** 379–382.

O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250,** 4007–4021.

O'Shaughnessy, R. F., Seery, J. P., Celis, J. E., Frischauf, A., and Watt, F. M. (2000). PA-FABP, a novel marker of human epidermal transit amplifying cells revealed by 2D protein gel electrophoresis and cDNA array hybridisation. *FEBS Lett.* **486,** 149–154.

Ostergaard, M., Wolf, H., Orntoft, T. F., and Celis, J. E. (1999). Psoriasin (S100A7): A putative urinary marker for the follow-up of patients with bladder squamous cell carcinomas. *Electrophoresis* **20,** 349–354.

Page, M. J., Amess, B., Townsend, R. R., Parekh, R., Herath, A., Brusten, L., Zvelebil, M. J., Stein, R. C., Waterfield, M. D., Davies, S. C., and O'Hare, M. J. (1999). Proteomic definition of normal human luminal and myoepithelial breast cells purified from reduction mammoplasties. *Proc. Natl. Acad. Sci. USA* **96,** 12589–12594.

Patterson, S. D., Spahr, C. S., Daugas, E., Susin, S. A., Irinopoulou, T., Koehler, C., and Kroemer, G. (2000). Mass spectrometric identification of proteins released from mitochondria undergoing permeability transition. *Cell Death Differ.* **7,** 137–144.

Pawson, T., and Scott, J. D. (1997). Signaling through scaffold, anchoring, and adaptor proteins. *Science* **278,** 2075–2080.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C., Lashkari, D., Shalon, D., Brown, P. O., and Botstein, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA* **96,** 9212–9217.

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature* **406,** 747–752.

Qin, J., and Chait, B. T. (1997). Identification and characterization of posttranslational modifications of proteins by MALDI ion trap mass spectrometry. *Anal. Chem.* **69,** 4002–4009.

Rodriguez, M. S., Desterro, J. M., Lain, S., Midgley, C. A., Lane, D. P., and Hay, R. T. (1999). SUMO-1 modification activates the transcriptional response of p53. *EMBO J.* **18,** 6455–6461.

Soskic, V., Gorlach, M., Poznanovic, S., Boehmer, F. D., and Godovac-Zimmermann, J. (1999). Functional proteomics analysis of signal transduction pathways of the platelet-derived growth factor β receptor. *Biochemistry* **38,** 1757–1764.

Takahashi, M., Rhodes, D. R., Furge, K. A., Kanayama, H., Kagawa, S., Haab, B. B., and Teh, B. T. (2001). Gene expression profiling of clear cell renal cell carcinoma: Gene identification and prognostic classification. *Proc. Natl. Acad. Sci. USA* **98,** 9754–9759.

Taylor, R. S., Wu, C. C., Hays, L. G., Eng, J. K., Yates, J. R., and Howell, K. E. (2000). Proteomics of rat liver Golgi complex: Minor proteins are identified through sequential fractionation. *Electrophoresis* **21,** 3441–3459.

Thieffry, D., Huerta, A. M., Perez-Rueda, E., and Collado-Vides, J. (1998). From specific gene regulation to genomic networks: A global analysis of transcriptional regulation in *Escherichia coli*. *Bioessays* **20,** 433–440.

Tomlinson, I. M., and Holt, L. J. (2001). Protein profiling comes of age. *Genome Biol.* **2**.

van der Geer, P., and Hunter, T. (1994). Phosphopeptide mapping and phosphoamino acid analysis by electrophoresis and chromatography on thin-layer cellulose plates. *Electrophoresis* **15,** 544–554.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science* **270,** 484–487.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen,

L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M. H., Wides, R., Xiao, C., and Yan, C. (2001). The sequence of the human genome. *Science* **291,** 1304–1351.

Verma, R., Chen, S., Feldman, R., Schieltz, D., Yates, J., Dohmen, J., and Deshaies, R. J. (2000). Proteasomal proteomics: Identification of nucleotide-sensitive proteasome-interacting proteins by mass spectrometric analysis of affinity-purified proteasomes. *Mol. Biol. Cell* **11,** 3425–3439.

Voehringer, D. W., Hirschberg, D. L., Xiao, J., Lu, Q., Roederer, M., Lock, C. B., Herzenberg, L. A., and Steinman, L. (2000). Gene microarray identification of redox and mitochondrial elements that control resistance or sensitivity to apoptosis. *Proc. Natl. Acad. Sci. USA* **97,** 2680–2685.

Warraich, R. S., Dunn, M. J., and Yacoub, M. H. (1999). Subclass specificity of autoantibodies against myosin in patients with idiopathic dilated cardiomyopathy: Pro-Inflammatory antibodies in DCM patients. *Biochem. Biophys. Res. Commun.* **259,** 255–261.

Washburn, M. P., Wolters, D., and Yates, J. R, III. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247.

Weekes, J., Wheeler, C. H., Yan, J. X., Weil, J., Eschenhagen, T., Scholtysik, G., and Dunn, M. J. (1999). Bovine dilated cardiomyopathy: Proteomic analysis of an animal model of human dilated cardiomyopathy. *Electrophoresis* **20,** 898–906.

Wu, C. C., Yates, J. R., Neville, M. C., and Howell, K. E. (2000). Proteomic analysis of two functional states of the Golgi complex in mammary epithelial cells. *Traffic* **1,** 769–782.

Yahanda, A. M., Bruner, J. M., Donehower, L. A., and Morrison, R. S. (1995). Astrocytes derived from p53-deficient mice provide a multistep in vitro model for development of malignant gliomas. *Mol. Cell. Biol.* **15,** 4249–4259.

Yanagida, M., Miura, Y., Yagasaki, K., Taoka, M., Isobe, T., and Takahashi, N. (2000). Matrix assisted laser desorption/ionization-time of flight-mass spectrometry analysis of proteins detected by anti-phosphotyrosine antibody on two-dimensional-gels of fibrolast cell lysates after tumor necrosis factor-α stimulation. *Electrophoresis* **21,** 1890–1898.

Zhou, H., Watts, J. D., and Aebersold, R. (2001). A systematic approach to the analysis of protein phosphorylation. *Nat. Biotechnol.* **19,** 375–378.

Zhu, H., and Snyder, M. (2001). Protein arrays and microarrays. *Curr. Opin. Chem. Biol.* **5,** 40–45.

THE TOOLS OF PROTEOMICS

By JOSEPH A. LOO

Departments of Biochemistry and Biological Chemistry, Molecular Biology Institute, University of California, Los Angeles, California 90095

I. Introduction

Proteomics is a powerful approach for biomedical research because it directly studies the key functional components of biochemical systems and the cellular targets of therapeutic agents, namely proteins. For drug discovery, understanding how drugs affect protein expression is a key goal of proteomics. Mapping proteomes, the protein complements to genomes, from tissues, cells, and organisms is being used to validate and forward new protein targets, to explore mechanisms of action or toxicology of compounds, and to discover new disease biomarkers for clinical and diagnostic applications. Once the proteins expressed within a proteome have been identified, they become a powerful means to examine global changes in protein levels and expression under changing environmental conditions. It is expected that proteomics will lead to important new insights into disease mechanisms and improved medical research strategies to produce novel therapeutics.

ADVANCES IN PROTEIN CHEMISTRY, Vol. 65
0065-3233/03 $35.00

A proteomics-based approach has been practiced for many years, in fact prior to the time the term ''proteomics'' was coined [1], to monitor changes in protein expression. The primary readout of protein expression was based on analysis of proteins separated by polyacrylamide gel electrophoresis (PAGE). However, development of novel mass spectrometry (MS) methods has made the identification of PAGE-separated proteins much more amenable and has contributed greatly to expand the range of applications to which proteomics can contribute. Provided the genome or protein sequences are known and available, MS provides a robust method for protein identification. The technology to automate the entire method of protein separation through to protein identification is available, thereby greatly increasing the throughput and overall capacity of the analysis. These virtues of MS have been a driving force in the burgeoning field of proteomics, and will continue to support these escalating efforts.

In a ''typical'' proteomics experiment, proteins are separated first by high-resolution two-dimensional (2D)-PAGE and visualized with, for example, Coomassie or silver stain [2]. To identify an individual or set of protein spots, several options are available. For protein spots that appear to be relatively abundant (i.e., >1 pmol, 10^{-12} mol), traditional protein characterization methods such as amino acid analysis and Edman sequencing can be used to provide the necessary protein identification information. The amino acid composition and/or amino-terminal sequence, combined with the approximate molecular weight and isoelectric point provided by 2D-PAGE, is often sufficient to obtain confident protein identification [3,4].

MS-based methods have become the primary technology used to identify proteins. The sensitivity of MS instrumentation allows for the identification of proteins below the 1-pmol level and in many cases in the femtomole (fmol, 10^{-15} mol) or even the attomole (amol, 10^{-18} mol) range [5]. The mass measurement accuracy afforded with current MS technology, routinely to less than 50 parts per million (ppm) and better, provides more confident protein identifications [6]. Additional sequence information provided by tandem MS (MS/MS) approaches improves the level of confidence of the identification [7]. Moreover, the speed at which the MS protein identification analysis can be obtained is unparalleled by any other biophysical technique. With these clear advantages, proteomics research has embraced MS with enthusiasm.

In this article some of the major technological advances associated with MS that have had the greatest impact on the field of proteomics are described. Most of the discussion highlights the MS-based methods related to the analysis of proteins separated by one- and two-dimensional gel

electrophoresis. Non-PAGE-based analyses are also addressed, however, as separation methods such as two-dimensional liquid chromatography have specific advantages over the PAGE approach. The developments can be divided into a few general categories, including ionization methods, analyzers, and sample processing.

II. Ionization Methods

Whereas proteins exist *in vivo* as biopolymers ranging in size from a few hundred daltons (Da) up to complexes of greater than 1×10^6 Da, proteins in proteomics studies are identified typically by analysis of their enzymatic digest products [8]. Although sample-handling issues are a major reason for digesting the proteins into smaller peptides, an even more compelling reason is the difficulty in identifying an intact protein solely on the basis of its molecular weight. The mass spectrum of a protein's peptide fragments (typically produced by digestion with an enzyme with well-defined specificity, such as trypsin) produces a ''peptide map'' or a ''peptide fingerprint'' [9]. These measured masses are then compared with theoretical peptide maps derived from database sequences (either protein or genomic data) to identify the protein that would most likely give rise to this ensemble of peptide masses (Fig. 1). For example, Fig. 2 shows the mass spectrum of a tryptic digest of an unknown protein. The measured masses are then compared with the theoretical tryptic peptide maps representing all the proteins present in the database of the particular organism from which the protein was derived. In this instance, a statistically significant number of masses agreed with the theoretical peptide map of yeast enolase. As more genomic sequence data become available, the chance for successful protein identification becomes greater. When peptide mapping does not provide sufficient information for confident identification, the most common method is to isolate peptide ions in the mass spectrometer, fragment them by collisional excitation, and measure the masses of the fragment ions to obtain partial or complete sequence information, as shown in Fig. 3. This process is more commonly referred to as tandem MS or MS/MS [10]. The measured fragment mass spectrum is then compared with theoretical MS/MS mass spectra calculated from the protein sequences in the database [11]. Today, there are a few popular choices of MS and MS/MS analysis that can be selected from this point, depending on available instrumentation and other factors.

Before the analyzer region of the spectrometer can measure peptides, they must first be ionized and propelled into the gaseous phase. Matrix-assisted

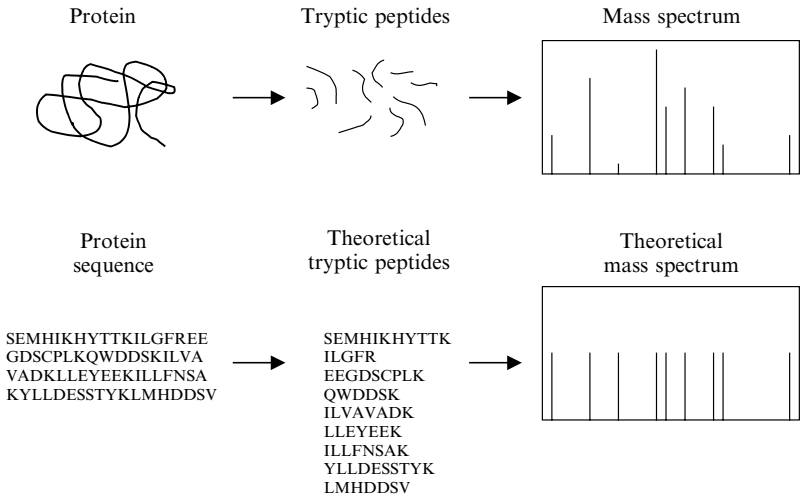


FIG. 1. Protein identification, using peptide mapping information. *Top:* In the experiment, the proteins are digested with an enzyme and the masses of the proteolytic peptides are measured by mass spectrometry. *Bottom:* In the database search, each protein sequence in the database is digested according to the specificity of the enzyme. The masses of the resulting peptides are calculated and a theoretical mass spectrum is constructed. The measured mass spectrum is compared with the theoretical mass spectrum.

laser desorption ionization (MALDI) [12] and electrospray ionization (ESI) [13] not only enabled mass spectrometers to measure molecular weights for extremely large biomolecules (i.e., >150,000), they are now the dominant methods by which peptides are ionized. Previous to the development of ESI and MALDI, ionization methods such as field desorption (FD) [14], fast atom bombardment (FAB) [15], and plasma desorption (PD) [16] could also be used to analyze insulin. The sensitivity for molecules at such large mass, however, was not terribly high. Moreover, the interface of MS with separation methods such as liquid chromatography for the analysis of biomolecules was achievable with ionization methods such as FAB and especially thermospray, an early cousin of ESI [17]. However, these MS applications were by no means ''routine.''

A. *Electrospray Ionization*

The ability to characterize proteins and peptides by MS was greatly enhanced by the development of ESI. The early practical history of ESI and MS began with the work of M. Dole, who described electrospray

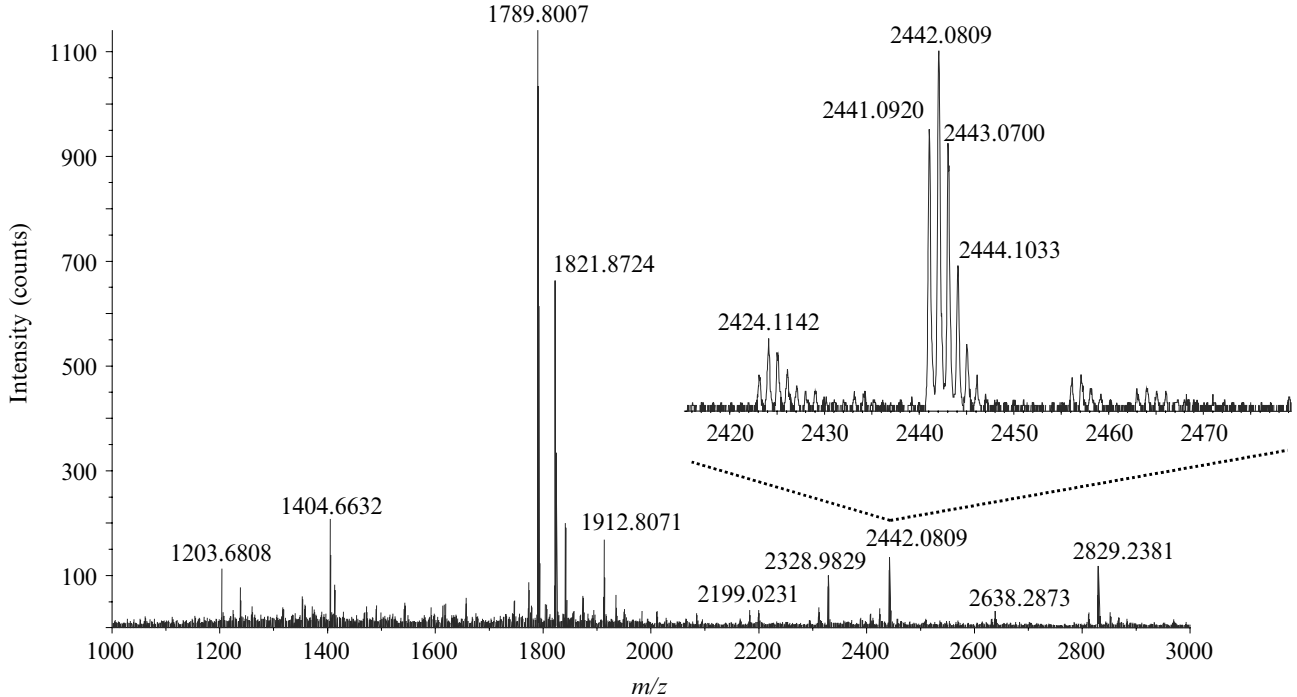


FIG. 2. MALDI-QqTOF (quadrupole time-of-flight) mass spectrum of a tryptic digest of yeast enolase (46-kDa monomer). The protein is identified by comparing the observed masses with theoretical mass spectra generated from a virtual digestion of the proteins within the yeast protein or genome database.

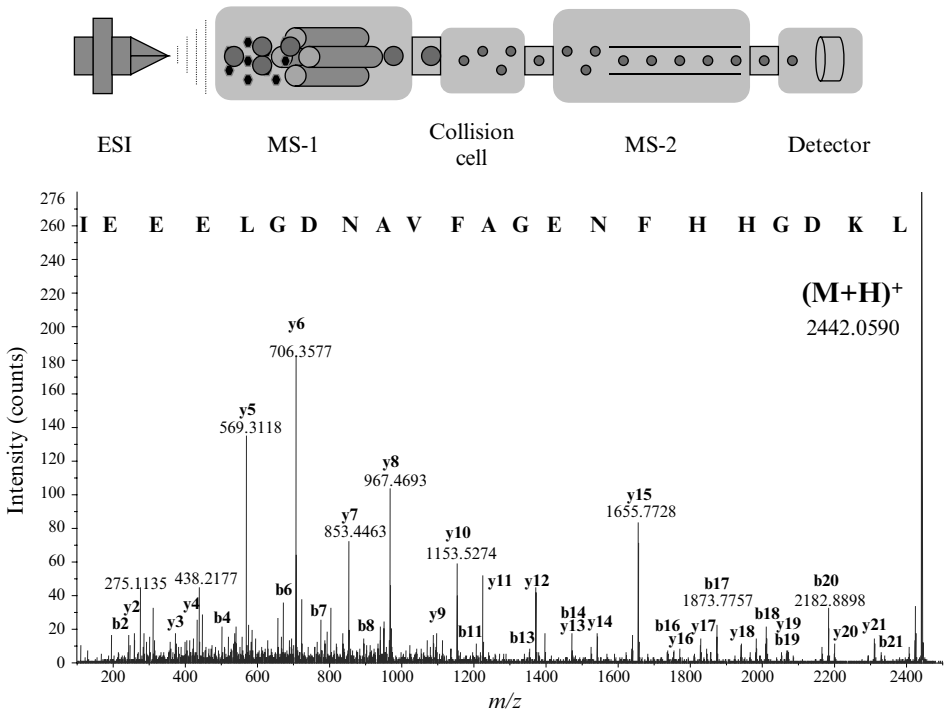


FIG. 3. Protein identification, using tandem mass spectrometry (MS/MS). *Top:* In MS/MS a parent peptide ion of interest is selected by the mass spectrometer and subjected to dissociation within a collision cell. After refocusing the fragment ions are guided to the detector. *Bottom:* MALDI-QqTOF MS/MS mass spectrum of the $(M+H)^+$ ion at *m/z* 2442 of a tryptic digest of yeast enolase (46-kDa monomer). Mass spectrum was acquired with a QSTAR Pulsar mass spectrometer (Applied Biosystems, Foster City, CA), MS/MS sequencing indicates that the peptide is the C-terminal peptide of enolase.

ionization in studies of ions from synthetic polymers of molecular weights in excess of 1,00,000 [13]. Dole conceived of using an electrospray process to produce intact high-mass polymeric ions from learning about electrospraying automobile paint while working as a consultant to a paint company. Later, Dole's experiments were extended by J. Fenn *et al.* [18] and by researchers in the former Soviet Union [19].

Although Dole's work included proteins as the analyte and Fenn's early work with electrospray demonstrated small multiple charging of small peptides, it was not until the 1988 American Society for Mass Spectrometry (ASMS) conference that the explosion of biochemical applications utilizing ESI-MS began [20]. Fenn presented initial results using electrospray ionization and a simple quadrupole mass analyzer at the June meeting in San Francisco. A relatively small audience listened and watched with amazement as Fenn showed ESI mass spectra of proteins ranging from insulin to 40-kDa alcohol dehydrogenase. The potential of ESI-MS and the ease with which online chromatographic and electrophoretic separation methodologies can be interfaced for protein analysis was obvious [21,22]. ESI mass spectra for 66-kDa bovine serum albumin and its 133-kDa dimer form were published soon afterward [23], and enhancements in sensitivity, mass range, and instrumentation proceeded quickly. The multiple charging phenomenon has been demonstrated to apply to molecules of over 200 kDa and it has permitted the measurement of relative molecular mass with a precision of better than 0.05% [24]. The number of publications in the scientific literature describing the application of ESI-MS for peptide and protein analysis rapidly increased after the 1988 ASMS Conference, as commercial mass spectrometry vendors first supplied ESI interfaces to existing instruments, and later developed dedicated ESI-MS systems.

The mechanism describing how an ESI source works, as shown in Fig. 4, is relatively simple. ESI is a solution-based ionization method; therefore it requires a sample in solution to enter the source through some type of flow stream. This stream can be generated by high-performance liquid chromatography (HPLC), capillary electrophoresis, or by direct infusion via a syringe pump. The solution passes through a stainless steel or other conductively coated needle to which a high voltage is applied. The solution, flowing in the presence of a high electric field, produces submicrometer-sized droplets on exiting the needle. Analyte ions, as well as ions originating from other solutes, are contained within these small droplets. As the droplets travel toward the mass spectrometer orifice at atmospheric pressure, they evaporate and eject charged analyte ions. This evaporation is accomplished by passing the droplets either through a heated capillary or a curtain of nitrogen gas. The desolvated ions are then sampled by the mass spectrometer for subsequent mass measurement.

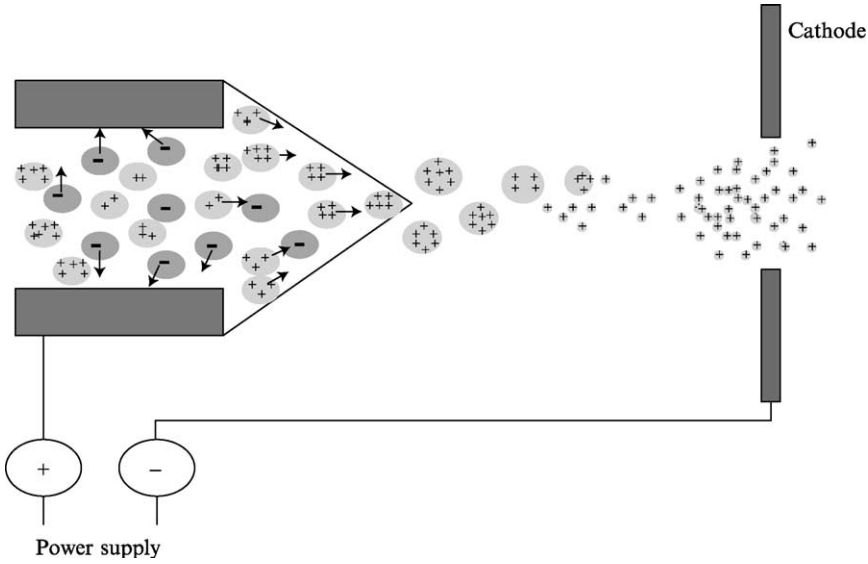


FIG. 4. Electrospray ionization (ESI) of molecules for mass spectral characterization. The sample solution is passed through a stainless steel or other conductively coated needle. A positive potential is applied to the capillary, causing positive ions to drift toward the cathode. The presence of a high electric field produces submicrometer-sized droplets as the solution exits the needle. The droplets travel toward the mass spectrometer orifice at atmospheric pressure, and evaporate and eject charged analyte ions. The desolvated ions are drawn into the mass spectrometer by the relative low pressure maintained behind the orifice.

A unique characteristic of ESI is its ability to produce multiply charged ions from large biological molecules, such as proteins, peptides, DNA, and RNA. The production of multiply charged ions makes the measurement of high molecular mass proteins amenable to instruments with limited m/z ranges, such as ion-trap and quadrupole MS analyzers. For example, a singly protonated 30,000-Da protein would yield an m/z of 30,001, making it within the detectable range only of higher m/z range instruments such as a time-of-flight (TOF) MS. A protein of this size, however, will typically accept anywhere from 15 to 30 protons depending on the solution conditions. Therefore, in the solution there will be populations of the protein that contain between 15 and 30 protons with m/z values ranging from 2001 (30,015/15) to 1001 (30,030/30). The mass spectrum of such a protein will exist as a multiply charged envelope containing signals for all of the various charge states of that protein present in solution. The same phenomenon holds for proteins measured in negative ionization mode; however, in this

case various numbers of protons are abstracted from the protein. Peptides in the range of 500 to 2500 Da typically exist as either singly, doubly, or triply charged ions, depending on their size and number of basic residues present. For a peptide of mass 1000 Da, its singly charged species (i.e., the $[M+H]^+$ ion) will have an m/z value of 1001. The doubly ($[M+2H]^{2+}$) and triply ($[M+3H]^{3+}$) charged ions will have m/z values of 501 and 334.3, respectively. For proteomics, this is the most common type of peptide analysis. Tryptic peptides are most often observed as 2+-charged species because of the basic sites on the N terminus and the C-terminal lysine or arginine residues.

B. Matrix-Assisted Laser Desorption/Ionization

At the same time as the development of ESI by Fenn, similar advances were made to laser desorption of biological molecules. Laser desorption was effective as a means to volatilize and ionize low molecular weight compounds as well as smaller peptides. However, the key development toward effective desorption/ionization was the use of organic matrices by the Hillenkamp group to produce molecular ion mass spectra of proteins with masses of greater than 10,000 Da [12]. At about the same time, Tanaka *et al.* reported the laser desorption of polymer and protein molecular ions for masses up to 22,000 Da, using glycerol polymer mixtures containing fine metallic particles as the laser absorbing matrix [25].

Matrix-assisted laser desorption ionization (MALDI) is a ''soft'' ionization process that produces (quasi)molecular ions from large nonvolatile molecules, such as proteins, oligonucleotides, polysaccharides, and synthetic polymers. MALDI generates high-mass ions by irradiating a solid mixture of an analyte dissolved in a suitable matrix compound with a pulsed laser beam. As the name implies, the laser pulse desorbs and indirectly ionizes the analyte molecules. A short-pulse (a few nanoseconds) UV laser is typically used for desorption; however, wavelengths in the infrared region have been investigated as alternatives [26].

MALDI analysis consists of two steps: sample preparation and mass spectral analysis. The sample is cocrystallized on the MALDI target plate with an appropriate matrix, which is a small, highly conjugated organic molecule that strongly absorbs energy in the ultraviolet region. Some of the most widely used matrices include α-cyano-4-hydroxycinnamic acid, 2,5-dihydroxybenzoic acid (DHB), and 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid). Before combining it with the sample, a saturated solution of the matrix is prepared in a suitable solvent such as water, acetonitrile, acetone, or tetrahydrofuran. A few microliters of this mixture

is deposited onto a substrate and dried, resulting in the integration of the peptides into a crystal lattice. The target plate containing the sample is then inserted into the high-vacuum region of the source and is irradiated with a laser pulse, as shown in Fig. 5. Many organic compounds that absorb strongly in the UV range have been evaluated as potential MALDI matrices, but only a small number function well as MALDI matrices. Although it has been more than 10 years since the birth of MALDI, the search for efficient matrices is still often carried out on a trial-and-error basis.

The source of a MALDI-MS is equipped with a laser, which fires a beam of light at the sample. In most MALDI devices, 337-nm irradiation is provided by a nitrogen (N_2) laser. Other lasers operating at different wavelengths can be used but size, cost, and ease of operation have made the nitrogen laser the most popular choice. Common nitrogen lasers

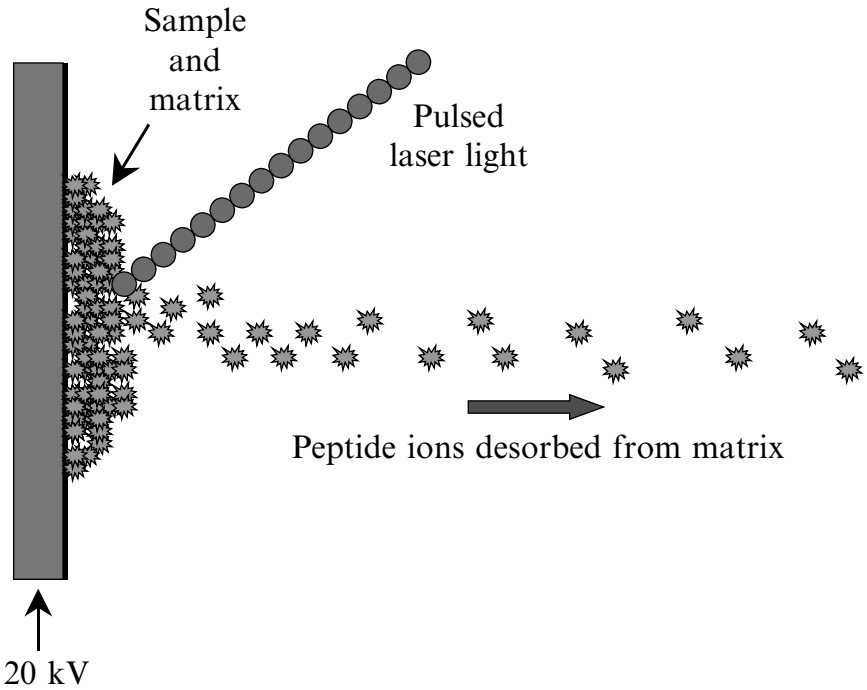


FIG. 5. Principles of matrix-assisted laser desorption ionization (MALDI). The sample is dispersed in a large excess of matrix material. Short pulses of laser light are focused onto the sample spot, causing the sample and matrix to volatilize. The matrix plays a key role in this technique by absorbing the laser light energy and causing part of the illuminated substrate to vaporize. A rapidly expanding matrix plume carries some of the analyte into the vacuum with it and aids the sample ionization process.

employed by MALDI-MS operate at 2–20 Hz. However, higher repetition rate lasers, 200–1000 Hz, have been investigated as a means to increase analysis throughput. On irradiation by the laser pulse, the matrix molecules absorb most of the laser's energy. Because of the high matrix-to-analyte concentration ratio, most of the photon energy is absorbed by the matrix and direct irradiation of the analyte is minimized. The energy absorbed by the matrix molecules is transferred into electronic excitation of the matrix. This energy is then transferred to the analytes (i.e., peptides) in the sample, which are subsequently ejected from the target surface into the gas phase.

Depending on the sample of interest, MALDI can produce both positive and negative ions. Positive ions, which are always the species of interest in peptide analysis, are formed by the acceptance of a proton as the analyte leaves the matrix. It is believed that analyte ionization occurs within the dense gas cloud that forms and expands supersonically into the vacuum region of the spectrometer. The analytes are protonated (or deprotonated) as a direct result of collisions between analyte neutrals, excited matrix ions, and protons and cations such as sodium. Because they typically pick up a single proton, most of the resulting peptide ions are singly charged. The ions formed in the source region are then directed into the analyzer region of the mass spectrometer.

More recently, MALDI sources operating at increased pressure and even atmospheric pressure have demonstrated relatively high sensitivity for peptide mass fingerprinting. By operating at atmospheric pressure, MALDI can be interfaced to analyzers typically used with ESI, such as ion traps and quadrupole time-of-flight analyzers. However, a primary advantage of atmospheric pressure MALDI is the ability to interface with analyzers that have MS/MS capabilities for peptide sequencing (see below).

III. Mass Analyzers

A. Time-of-Flight Mass Spectrometer

Time-of-flight analyzers are extremely popular in proteomics. Fast electronics and powerful computers have led to a period of rapid growth for TOF-MS. The principal factors that make TOF-MS so attractive are their high speed, sensitivity, and resolution. The TOF analyzer measures the time it takes for the ions generated in the source to fly from one end of the analyzer and strike the detector. A potential V_s (the source extraction) is applied across the source to extract and accelerate the ions from the source into the field-free ''drift'' zone, or tube, of the instrument, *d*. The

speed at which the ions fly down the analyzer tube is proportional to their m/z value. The larger the ion, the slower its velocity and thus the longer it takes to traverse the field-free drift zone [27].

The initial TOF analyzers operated in a linear mode in which the ions formed in the source were continually extracted and sent through the flight tube to the detector [28, 29]. Unfortunately, the linear mode did not provide the highest resolution because of variations in the velocities of ions of the same m/z value. The lack of resolution is a direct result of the ions having different initial energies as well as different positions as they move from the source to the analyzer region. This difficulty has been solved by the development of the reflectron [30]. The reflectron focuses ions with the same m/z values and allows them to strike the detector at the same time. The resolution of the TOF-MS was also improved by the use of pulsed-laser ionization with delayed extraction when operating in a linear mode [31, 32]. In delayed extraction, there is a slight delay between the ionization of the sample and the extraction of the ions into the flight tube. This delay allows all the ions to get an equal start time so that the ions of equal m/z will reach the detector at the same time.

Because TOF analyzers do not contain a true collision cell, they are primarily used to generate peptide maps by which proteins may be identified. Instruments equipped with a reflectron, however, can measure fragmentation products as well through a process called ''postsource decay'' (PSD) [33]. In this technique, the reflectron voltage is adjusted during the analysis so that fragment ions generated during the ionization and acceleration of the peptide are focused and detected. Although PSD analysis can be relatively slow and does not meet the high-throughput demands necessary for proteomics, it does provide useful complementary information to substantiate the identification of an intact peptide. Specifically, PSD produces immonium ions, which are useful indicators of the presence of a specific amino acid within a peptide [34].

The MALDI-TOF combination is the most popular method for generating peptide mass fingerprints. Often, a MALDI-TOF-derived peptide map is sufficient to yield a positive protein identification, especially for proteomes from previously sequenced genomes. Again, the speed and mass accuracy of the MALDI-TOF approach make it a popular choice for the initial analysis of any protein sample.

B. Triple Quadrupole Mass Spectrometer

During its early history, ESI was most commonly employed with a quadrupole mass analyzer [35]. A quadrupole mass analyzer consists of four metal rods arranged in parallel to which direct current and

radiofrequency voltages are applied. Gas-phase ions that enter the mass analyzer follow a corkscrew trajectory along the axis of the quadrupole. The voltage applied can be used to select the m/z range of the ions that are allowed to pass through the quadrupole region. All ions outside this range will fail to pass through this region and hence not be detected. Ions of increasing m/z values can be detected by sweeping the radiofrequency voltages on the quadrupoles.

Two types of quadrupole mass spectrometers are manufactured: single-stage quadrupoles and triple quadrupoles. Single quadrupole mass spectrometers have limited utility in proteomics because they lack true tandem MS abilities, although in-source CID can provide some fragmentation functionality. Triple quadrupole mass spectrometers have been used for proteomics studies [36]. The triple quadrupole instrument, as the name implies, is composed of three quadrupole regions. Two of these quadrupoles, designated Q1 and Q3, operate as described above to guide and select ions through the analyzer to the detector. These two are separated by a radiofrequency only quadrupole, q2, which acts as a collision cell. In the q2 region, collisions between ions and neutral gas ions, such as N_2 and argon, provide true MS/MS capabilities. In proteomics studies, the fragmentation resulting from these collisions is used to acquire sequence-related information for the identification of peptides.

The identification of peptides with a triple quadrupole mass spectrometer requires the instrument to alternatively switch between two different scan modes. In the first mode, a wide range of ions generated from the source region is allowed to pass through Q1. The ions that pass through Q1 also pass freely through q2 and Q3 onto the detector. This ''full-scan'' analysis provides a spectrum of all the ions generated within the source. In the second scan mode, Q1 is used as a mass filter, in which the voltage is set to allow only a specific m/z value to pass through. These ions are then subjected to fragmentation within q2. The fragment ions then pass through Q3 and are subsequently detected to provide tandem MS data on the original peptide. As with all mass spectrometers, optimal peptide fragmentation is achieved by the careful setting of the instrument parameters such as collisional gas pressure and CID energy.

The triple quadrupole mass spectrometer is capable of product ion, precursor ion, and neutral loss scanning. These instruments have been used to identify proteins extracted from 2D-PAGE gels [37], to characterize phosphopeptides [38], and to identify glycopeptides [39]. The mass measurement accuracy of triple quadrupole analyzers is at least 0.5 amu for the fragmentation ions produced by MS/MS [40]. This mass accuracy is sufficiently accurate to allow the identification of peptides by

correlating the tandem MS spectra with protein sequences obtained from biological databases.

C. *Quadrupole Time-of-Flight Mass Spectrometer*

An instrument that is having a tremendous impact in proteomics is the hybrid quadrupole-time-of-flight mass spectrometer (QqTOF) (for review see Chernushevich *et al.* [41]) The analytical community has rapidly embraced the QqTOF as a powerful and robust instrument with unique capabilities. The unique feature of this instrument is its combination of high performance of TOF analysis in both the mass spectrometry (MS) and tandem MS (MS/MS) modes, while using both ESI and MALDI methods. Although originally targeted at the analysis of peptides [42], it is now applied to problems ranging from nanospray analysis of biological samples to liquid chromatography (LC)-MS/MS of pharmaceutical preparations at high flow rates. Its rapid acceptance is due to the attractive combination of high sensitivity and high mass accuracy for both precursor and product ions, and also to the simplicity of operation for those already familiar with LC-MS analysis on quadrupole and triple quadrupole instruments.

The QqTOF mass spectrometer can be regarded either as the addition of a mass-resolving quadrupole and collision cell to an ESI-TOF, or as the replacement of the third quadrupole (Q3) in a triple quadrupole by a TOF mass spectrometer. Regardless of how it is described, the instrument combines the benefits of ion selectivity and sensitivity (quadrupole) with high mass resolution and mass accuracy (TOF) in both the MS and MS/MS modes. The QqTOF mass spectrometers typically have a resolution of 13,000–20,000 and can provide accurate molecular and product ion mass determination in the femtomole range [43]. The high mass accuracy afforded with this configuration provides the potential for real *de novo* sequencing of peptides [44].

In the usual QqTOF configuration, an additional radiofrequency quadrupole, Q0, is added to provide collisional damping, so the instrument (Fig. 6) consists of three quadrupoles, Q0, Q1, and Q2, followed by a reflecting TOF mass analyzer with orthogonal ion injection. For single MS (or TOF-MS) measurements, the mass filter Q1 is operated in the radiofrequency-only mode so that it serves merely as a transmission element, while the TOF analyzer is used to record spectra. The resulting spectra benefit from the high resolution and mass accuracy of the TOF instruments, and also from their ability to record all ions in parallel, without scanning. For MS/MS, Q1 is operated in the mass filter mode to transmit only the precursor ion of interest. The ion is then accelerated to an energy between

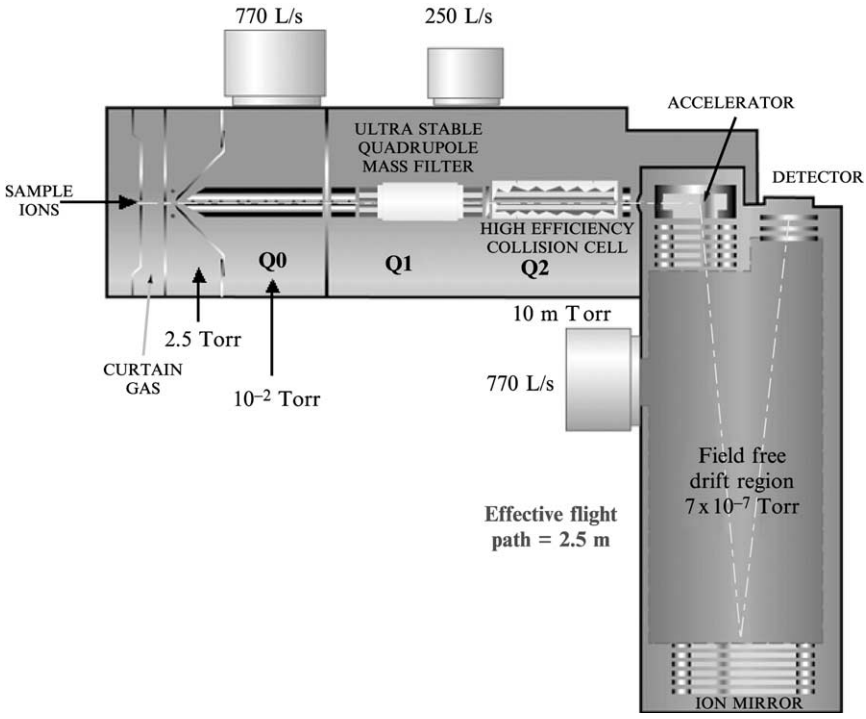


FIG. 6. Schematic of a quadrupole time-of-flight mass spectrometer. This instrument combines the ion selection and MS/MS capabilities of a triple quadrupole mass spectrometer with the high mass accuracy and resolution capabilities of a TOF.

20 and 200 eV before it enters the collision cell Q2, where it undergoes fragmentation through collisions with a neutral gas such as argon or N_2. The resulting fragment ions are collisionally cooled. Before entering the TOF analyzer, the ions are reaccelerated to the required energy, and focused. The result is a parallel beam that continuously enters the ion modulator of the TOF analyzer. A pulsed electric field is applied at a frequency of several kilohertz (kHz) to push the ions in a direction orthogonal to their original trajectory into the accelerating column. From the accelerating column, ions arrive in the field-free drift space of the TOF analyzer, where they are separated on the basis of their m/z value.

D. Ion-Trap Mass Spectrometer

The popularity of the quadrupole ion trap has its roots in the discovery and development of the mass-selective axial instability scan, for which

G. C. Stafford, Jr. was awarded the 2001 ASMS Distinguished Contribution in Mass Spectrometry Award [45]. Stafford's discovery of the mass-selective instability mode converted a simple ion storage device into the extremely versatile quadrupole ion-trap mass spectrometer that has applications across a wide variety of areas and has contributed greatly to its commercial success. Earlier quadrupole ion traps were based either on mass selective stability for mass analysis or were simply used as transmission or ion storage devices. First, they developed the mass-selective instability mode of operation. The fundamental difference between this mode of operation and previous methods is that all ions created over a given time period were trapped and then sequentially ejected from the ion trap into a conventional electron multiplier detector. Unlike the mass-selective stability mode of operation, in which only one value of m/z at a time was stored, all ions were stored while mass analysis was performed. The second breakthrough by Stafford's group was the finding that a helium gas of about 1 mtorr within the trapping volume greatly improved the mass resolution of the instrument by reducing the kinetic energy of the ions and contracting the ion trajectories to the center of the trap [45]. This phenomenon allows packets of ions of a given m/z to form, which can be ejected faster and more efficiently than a diffuse cloud of ions, thereby improving resolution and sensitivity. Both these discoveries led to the successful development of a commercial ion-trap mass spectrometer.

The ion trap works quite differently than a quadrupole mass analyzer [46]. Whereas quadrupole mass spectrometers essentially measure ions as they pass through the analyzer, an ion trap collects and stores ions, and then performs manipulations such as MS/MS (Fig. 7). The ion trap repeats the collection and storing of ions followed by scanning them out of the trap in a continuous cycle. To perform MS/MS analysis, after filling the trap with ions a particular species is selected and the trapping voltages are adjusted to eject all other ions by adjusting the trapping voltages. The applied voltages on the trap are then increased causing an increase in the energy of the remaining trapped ions. These high-energy ions then undergo collisions with He_2 in the trap, causing them to fragment. These fragments are caught in the trap and scanned out according to their m/z. Although fragmentation ions can also be retained within the trap and subjected to further rounds of MS/MS (i.e., MS/MS/MS or MS^n), such fragmentation is rarely used in proteomic studies, particularly in the area of complex mixture analysis.

The ion-trap mass spectrometer enjoys a position of prominence as the analyzer of choice for a wide variety of diverse applications in biological, pharmaceutical, environmental, and industrial laboratories. The versatility

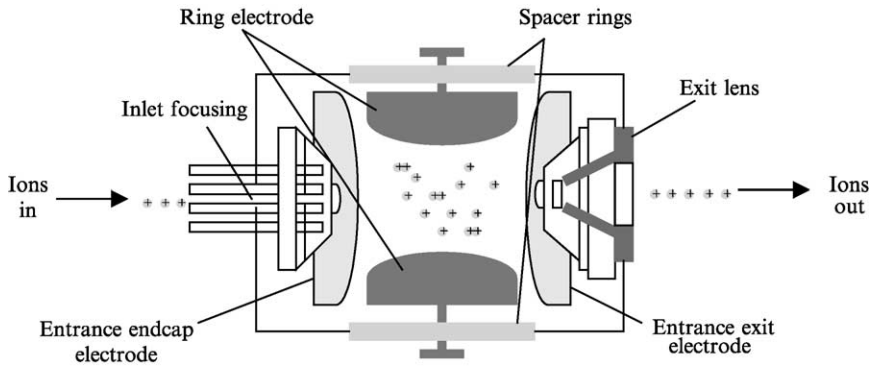


FIG. 7. The quadrupole ion-trap mass analyzer consists of the ring electrode, the entrance endcap electrode, and the exit endcap electrode. These electrodes form a cavity in which ions are trapped and analyzed. The endcap electrodes have a small hole in their centers that allow ions to either enter or exit the trap. The ring electrode is located halfway between the two endcap electrodes. Ions produced from the source enter the trap through the inlet focusing system and the entrance endcap electrode. Various voltages are applied to the electrodes to trap and eject ions according to their mass-to-charge ratios. A potential is applied to the ring electrode to produce a 3-D quadrupolar potential field within the trapping cavity, resulting in the trapping of ions in a stable oscillating trajectory within the cell. During detection, the electrode system potentials are altered to produce instabilities in the ion trajectories and thus eject the ions in the axial direction. The ions are ejected in order of increasing mass-to-charge ratio.

of the quadrupole ion trap is demonstrated by its ready interface to liquid chromatography and to ion sources such as ESI [47] and more recently MALDI [47, 48].

E. Fourier Transform Ion Cyclotron Resonance Mass Spectrometry

Fourier transform ion cyclotron resonance (FTICR) mass spectrometry was developed by Comisarow and Marshall almost 30 years ago [49]; however, it is only more recently that its many unrivaled attributes are generating great enthusiasm within the proteomics community. An FTICR-MS functions somewhat like an ion-trap analyzer; however, the trap is housed in a magnetic field [50] (Fig. 8). The presence of the magnetic field causes ions captured within the trap to resonate at their cyclotron frequency. A uniform electric field that oscillates at or near the cyclotron frequency of the trap ions is then applied to excite the ions into a larger orbit that can be

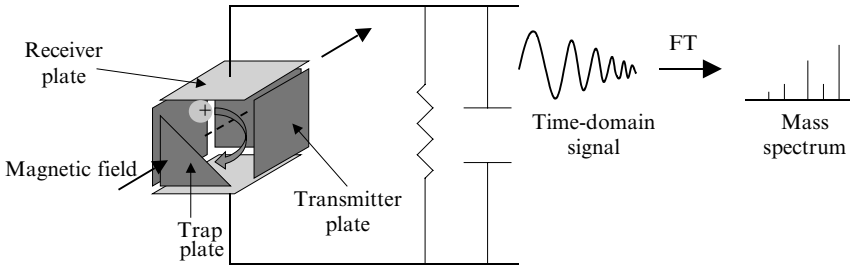


FIG. 8. Principles of Fourier transform ion cyclotron resonance mass spectrometry (FTICR). In FTICR the ion trap is placed in a strong magnetic field. The magnetic field causes ions captured within the trap to resonate at their cyclotron frequency. By applying the appropriate electric field energy the ions are excited into a larger orbit that can be measured as they pass by detector plates on opposite sides of the trap. Energy can also be applied to dissociate the ions or to eject ions from the trap by accelerating them to a cyclotron radius larger than the radius of the trap. The detector measures the cyclotron frequency of all of the ions in the trap and uses a Fourier transform to convert these frequencies into m/z values.

measured as they pass by detector plates on opposite sides of the trap. The energy applied can also be used to eject ions from the trap by accelerating them to a cyclotron radius larger than the radius of the trap or to dissociate the ions. The detector measures the cyclotron frequency of all the ions in the trap and uses a Fourier transform to convert these frequencies into m/z values (Fig. 8).

As mentioned above, the ion trap of an FTICR-MS is placed in a magnetic field. Indeed, the strength of the magnetic field is often quoted when describing a particular FTICR-MS (i.e., 9.4 T), much like the proton frequency is quoted when describing a nuclear magnetic resonance (NMR) instrument (i.e., 500 MHz). Working at higher magnetic fields benefits several parameters related to FTICR performance, the two most critical being resolution and mass accuracy. FTICR instruments have provided the highest resolution [51], mass accuracy [52], and sensitivity [53,54] in peptide and protein measurements so far achieved. These instruments can be operated with ESI [50] and MALDI [55,56] sources and are readily coupled to on-line separations such as reversed-phase LC [57] or capillary electrophoresis (CE) [58]. The true potential of FTICR is in the characterization of global proteome mixtures containing thousands of unique species. Although still an expensive technology, the advantages and increasing accessibility of FTICR will make it an important proteomics instrument in the future.

IV. Sample Fractionation

A. *Two-Dimensional Polyacrylamide Gel Electrophoresis*

Although the development of mass spectrometry technologies with higher sensitivity, dynamic range, resolution, and mass accuracy has been a driving force in the field of proteomics, there are many other advances that have contributed significantly. Two-dimensional polyacrylamide gel electrophoresis [2] was developed two decades before the coining of the term ''proteomics,'' and it still remains one of the major technologies that propels the field [59]. The combination of the resolution of 2D-PAGE and high-throughput protein identification of MS allows comparison of the expression levels of thousands of proteins in a single experiment. Regardless of opinions of the shortcomings of 2D-PAGE/MS technology, it is difficult to argue that it has been and continues to remain a cornerstone in proteomics. N. G. Anderson, one of the early developers of 2D-PAGE and a former key member of the Large Scale Biology Corporation (Vacaville, CA), remarked that ''2D gels are a pain, are central to all of proteomics, and everyone would like a dry-fingered substitute. Hence, the urgent stress on mass spectrometry, which is not a substitute for 2DE'' (*Genome Technology*, May 2002). In 2D-PAGE complex proteome samples are separated in the first dimension on the basis of their isoelectric point and in the second dimension on the basis of their molecular weight. After staining, typically with Coomassie blue or silver staining, the protein spots are excised from the gel and the protein is digested with trypsin while still within the gel matrix. After digestion the resultant peptides are extracted and analyzed by MS. Because a 2D-PAGE experiment is capable of resolving thousands of proteins the downstream sample-handling steps need to be automated for any large-scale proteomics study [60]. Many different vendors presently manufacture robotic workstations that automate the process from the staining of the gel to the preparation of the MALDI target plate for MS analysis.

B. *On-Line Separations*

Regardless of the automation done to increase the throughput of 2D-PAGE/MS, there are still many criticisms of the technique in general. The theme of many of these criticisms is based on what 2D-PAGE does not see: low-abundance proteins, low-and high-mass proteins, proteins with extreme isoelectric points, and membrane-associated proteins [59]. To circumvent these problems many groups have coupled separation

techniques directly on-line with the mass analyzer. One of these techniques is the combination of ESI-MS with reversed-phase HPLC [61]. This coupling has eased the analysis of complex mixtures because the peptides can be separated by reversed-phase HPLC with subsequent on-line mass measurement by ESI-MS. The combination of LC and MS (LC/MS) is widely used in the field of proteomics to aid in the identification of simple and complex mixtures of peptides. In a typical experiment to identify a protein, it is first digested with trypsin and the peptides are analyzed by LC/MS. The MS is able to measure the masses of the various peptides as they elute from the LC column. These recorded masses can then be compared with a sequence database to identify their protein of origin. To provide further confirmation of the identification, if a tandem mass spectrometer is available, peptide ions can be dissociated in the mass spectrometer to provide direct sequence information.

While reversed-phase LC is the separation of choice to couple directly on-line with MS, many other on-line separation techniques have also been used. Capillary electrophoresis (CE) is becoming a popular separation technique to couple on-line with MS [62]. CE offers the advantages of high efficiency, speed, and sensitivity along with its tolerance of complex sample matrices, which makes it an ideal separation for simple and complex proteome mixtures. The primary disadvantage is that much less sample can be loaded on a CE column than on a reversed-phase capillary column. With the movement toward clinical proteomics, where only small amounts of sample are available, CE-MS may play an increasingly important role. Capillary isoelectric focusing (CIEF) has been coupled directly on-line with ESI-MS to characterize simple [63] and complex [64] mixtures of intact proteins. In addition, MALDI-MS has been used to characterize fractions collected from a CIEF separation conducted off-line [65]. Other separation technologies such as capillary isotachophoresis [66], micellar electrokinetic chromatography [67], affinity capillary electrophoresis [68], and capillary electrochromatography [69] have also been coupled with mass spectrometry; however, they presently play a limited role in proteomic analyses.

Although all of the above-described technologies are simple one-dimensional (1D) separations, the complexity of the samples encountered during global proteomics samples makes it difficult to obtain sufficient resolution in a 1D separation. Investigators have begun to generate 2D LC separations in which the cumulative resolution obtained by two orthogonal separation techniques is greater than can be obtained in a single separation [70, 71]. Although a variety of separations can be coupled together, to this point the most successful has been the MudPit system developed by J. Yates [72, 73]. In the initial demonstration of this 2D on-line separation method,

complex peptide mixtures from different fractions of a *Saccharomyces cerevisiae* whole-cell lysate were loaded separately onto a biphasic microcapillary column packed with strong cation-exchange (SCX) and reversed-phase (RP) materials. The peptide mixture was loaded into the microcapillary column off-line, and the column was then inserted into the instrumental setup. Software has been developed to control the HPLC and mass spectrometer simultaneously. The peptides were first eluted from the strong cation-exchange column onto the reversed-phase column, using a salt gradient. The peptides captured on the reversed-phase column were eluted by a gradient of increasing acetonitrile concentration. The peptides were eluted directly into the mass analyzer, in this case an ion trap operating in a data-dependent tandem MS mode. Many cycles were repeated, each one using a progressively higher salt concentration to elute peptides from the strong cation-exchange column onto the reversed-phase column and then into the mass spectrometer. The tandem mass spectra generated were correlated to theoretical mass spectra generated from protein or DNA database by the SEQUEST algorithm [74]. A total of 1484 proteins were identified, using the on-line MudPit-MS/MS technology [73]. Not only was this an enormous improvement on the number of proteins identified by non-2D-PAGE-based techniques, it also identified many proteins from classes (i.e., extreme p*I*, low abundance, etc.) not easily detectable by gel-based separation methods.

C. Microfabricated Sample Preparation

With the advances in microfabrication, a natural combination to be developed is the marriage of microchip technology and mass spectrometry. Miniaturizing the scale of analytical devices provides a means to reduce sample quantities and also to incorporate novel and complex microfluidic devices for liquid flow and peptide separation. This provides also the possibility to multiplex sample introduction devices to be presented to a mass spectrometer. One can envision the future ESI source for proteomics applications: a microchip no larger than a credit card, with capabilities for ion-exchange and reversed-phase chromatography, interfaced to 384- or 1536-microwell sample plates and a single or an array of ESI emitters positioned in front of the ESI-MS opening. Given the impressive results from preliminary designs, such a device should exist in the not-too-distant future [75,76].

To reap the same advantages that microfluidic devices offer ESI-MS analysis, a microfluidic platform based on a compact disc (CD) format has been developed for parallel sample preparation before MALDI-MS [77].

Sample solutions travel through microchannels fabricated on CD discs in response to centrifugal forces as the CD spins. The samples are desalted and concentrated on a solid-phase support and are eluted onto a MALDI target area defined within the CD.

V. Software Tools

A treatise on the tools of proteomics is not complete without a description of the software tools and databases that have been developed to assist proteomics analysis. The trend in proteomics has been toward the characterization of samples of increasing complexity, which requires automated processes and results in an enormous amount of data. Automation obviously requires sophisticated software to coordinate the synchronization of the instruments, such as the HPLC and mass spectrometer, necessary to analyze proteome samples. Presently a vast majority of protein identifications made using MS data is accomplished through a comparison with known protein sequences found in a variety of different databases populated with either protein or DNA sequence information. Before the advent of the various genome projects, the number of MS spectra that would successfully identify a particular protein would have been few. Today, however, the large number of genomes that have been sequenced has contributed to the amount of sequence data that can be accessed to identify proteins through their MS spectrum.

A. *Peptide Mapping*

Because a thorough discussion of the variety of software tools developed for proteomics would require a book on its own, we focus on the most commonly used software for identifying proteins both through peptide mapping and fragmentation. Peptide mapping is the simplest method for identifying a protein. In peptide mapping a count of the number of experimentally measured peptide masses that correspond to calculated peptide masses in the theoretical mass spectrum of each protein in the database is tabulated. The protein in the database that has the highest number of matches to the experimental data is considered the most likely match. As a general rule, at least three to six experimental peptide masses must agree with the calculated masses from a single protein for a successful identification. The number of required matches, however, is highly dependent on the mass measurement accuracy of the experimental results. Several software tools are available on the Internet that use this method of ranking the proteins in the database. Some of these tools include PepSea (http://pepsea.protana.com/PA_PepSeaForm.html) [78],

PeptIdent/MultiIdent (http://us.expasy.org/tools/peptident.html) [79], and MS-Fit (http://prospector.ucsf.edu/ucsfhtml4.0/msfit.htm) [80]. While peptide mapping works well for high-quality experimental data, it usually gives higher scores to larger proteins because the probability of random matching is higher, as they usually contain a greater number of peptides. More advanced methods for identifying proteins are based on counting the number of measured peptide masses that correspond to calculated peptide masses, but they also take into account the effect of the protein size. MOWSE (http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse) considers the relative frequency of a peptide of a given molecular weight being within a protein of a given range of molecular weights [81]. The consequence of this scoring system is that matches with larger peptides (which are more likely to be unique) are more heavily weighted and that the nonrandom distribution of molecular weights in proteins of different sizes is compensated for. Although Mascot (http://www.matrixscience.com) is based on the MOWSE algorithm, it uses additional probability-based scoring [82]. The probability that a match between experimental data and a protein sequence is a random event is approximately calculated for each protein sequence in the database. The proteins are then ranked with decreasing probability of being a random match to the experimental data.

ProFound (http://prowl.rockefeller.edu/cgi-bin/ProFound) uses Bayesian theory to rank the protein sequences in the database by their probability of occurrence [83]. Unlike most other peptide-mapping algorithms, ProFound considers detailed information about each individual protein sequence in the database and allows the investigator to include additional experimental information about the protein, such as any sequence information that is already known. This gives ProFound the advantage of using different types of information to increase the sensitivity and selectivity of the algorithm. ProFound can also be used to identify proteins present in simple mixtures. A two-step approach is used in which the proteins in the database are initially ranked according to how well they match the experimental data, assuming a single protein is present. In the second step, the top ranking proteins are fused together to create entities of the top two ranking proteins, the top three ranking proteins, and so on. These fusion proteins are then ranked according to how well they match the experimental data.

B. Tandem Mass Spectrometry

The second common method to identify a protein through characterization at the peptide level is by tandem MS, which produces fragmentation spectra. Peptide fragmentation typically occurs along the polypeptide

backbone to produce products termed ''y-type'' and ''b-type'' fragment ions (in which the y ions contain the C-terminal portion of the peptide and the b ions contain the N terminus). These product ions from an MS/MS spectrum can be compared with available sequences, using powerful software tools as well. Unlike peptide mapping, which contains limited information on a large amount of the protein, tandem MS spectra contain much information about a small amount of the protein. The sequence information obtained from an MS/MS spectrum allows for the identification of a protein from a single peptide and is the method of choice for identifying proteins within a complex mixture.

One of the earliest software programs developed to identify proteins from raw tandem MS spectra is SEQUEST [74]. This program works by searching protein and nucleotide databases for peptides that match the molecular weight of the unknown peptides produced by digestion of the protein(s) of interest. In the first step of the algorithm, the precursor ion m/z value is used to select peptides of similar mass from a database. Theoretical MS/MS spectra are then generated for these peptides and a cross-correlation analysis is then performed between the experimental MS/MS spectra and each of the calculated theoretical MS/MS spectra. Highly correlated spectra result in identification of the peptide sequences, and therefore the originating protein. Although the SEQUEST search output on a single MS/MS spectrum can serve to identify a protein, having multiple MS/MS spectra match peptides from a single protein increases the confidence that the protein is present in the input sample. Similarly, for protein mixtures, if multiple peptides from multiple proteins are identified by SEQUEST then confidence increases that the multiple proteins are present in the input sample.

As described above, Mascot can use additional information about a peptide beyond its molecular weight [82]. Therefore it can be used in a sequence query, in which one or more peptide molecular masses are combined with sequence, composition, and fragment ion data to provide a powerful and sensitive identification technology. The usual source of sequence information is a partial interpretation of an MS/MS spectrum. Although it is difficult to determine a complete and unambiguous peptide sequence from an MS/MS spectrum, it is almost always possible to find a series of peaks providing three or four residues of sequence data. In the sequence query mode of Mascot the fragment ion mass data and sequence data can be supplied in any combination. Because the sequence data are specific, it takes priority over the specificity of the molecular weight, allowing for wide tolerances concerning the mass data. This can result in the identification of a peptide on the basis of sequence data that show

poor correlation between the experimental and calculated mass of the peptide. Although this may initially seem a detriment, in reality this intended behavior allows for the identification of modified peptides in which a discrepancy between the experimental and calculated masses, but not the sequence, of the peptide is expected.

In the algorithm PepSea the peptide mass can be combined with sequence information obtained by MS/MS to provide a peptide sequence tag [78]. The position of a partial amino acid sequence within the peptide is determined from peaks within the tandem MS spectrum. Once its position is determined it is then used to calculate the mass of the remaining N- and C-terminal regions of the peptide. This combination of some amino acid sequence, with masses defining the position of the sequence within the peptide, results in high search specificity. Unique hits in the database can usually be obtained with only three or four amino acids of sequence information. When a peptide has been identified, the theoretical fragmentation is predicted and compared with the MS/MS spectrum for assignment of other peaks that can validate the identification. This procedure is repeated for every fragmented peptide in the sample and leads to additional verification of the result or identification of other proteins in the sample. Although PepSea is fast, its one disadvantage is that it requires the elucidation of a peptide sequence tag before database searching.

C. Data Mining

Beyond using MS/MS spectra for simple identification, programs have been developed that use these spectra to identify posttranslationally modified peptides as well. Many software tools have been designed to examine peptide mapping and MS/MS spectra for modified peptides. In addition, individual laboratories have also designed tools to mine their own data in an attempt to maximize the amount of information retrievable from a proteomic analysis. Fortunately, many of these tools are available on the Web and a good list can be found at http://www.expasy.ch/tools/. Brief descriptions of two software programs are presented below to provide an idea of what types of data these programs utilize and the specific features they are designed to determine.

One of the latest software programs to be created is called SALSA (Scoring Algorithm for Spectral Analysis) [84]. SALSA is designed to look for specific characteristics within an MS/MS spectrum and provide a score based on how many of these characteristics and their intensities are

present. The characteristics that SALSA looks for are designed primarily to identify such things as posttranslationally modified peptides. The key characteristics SALSA detects are a product ion at a specific *m/z*, a neutral loss, a charged loss, and an ion pair. The detection of a specific product ion can be indicative of the loss of a chemical modification within the MS/MS spectrum. For example, the program could look for the presence of an ion at 98 Da, which indicates the presence of a phosphorylated peptide. The detection of a neutral loss looks for differences in the masses of a product and precursor ion of identical charge state. The presence of a charge loss is detected by SALSA as a loss of a singly charged fragment from a doubly charged precursor. Although this type of loss is most common in the formation of singly charged b and y ions from doubly charged precursor ions, such a loss can also be characteristic of some modifications. The detection of an ion pair anywhere within the MS/MS spectrum can be diagnostic of many different types of fragmentation losses. For example, in a y ion series the loss of a tryptophan residue would be represented by an ion pair separated by 186 *m/z* units. Although the losses identified by SALSA can, in principle, also be determined by visually inspecting individual MS/MS spectra, the abundance of data generated during proteomic analyses would require thousands of spectra to be examined in this manner. Therefore, the need for software analysis tools such as SALSA is critical to optimize the amount of useful data gleaned from such studies.

FindMod is a software tool designed specifically for the identification of modified peptides [85]. The program examines peptide mass fingerprinting results of known proteins for the presence of more than 20 different types of posttranslational modifications including acetylation, deamidation, farnesylation, formylation, geranyl-geranylation, O-GlcNAc, hydroxylation, methylation, phosphorylation, and sulfation. FindMol is also capable of identifying single amino acid substitutions. The program looks for mass differences between experimentally determined peptide masses and theoretical peptide masses calculated from a specified protein sequence. To run FindMod the user must input the sequence of the protein (or its protein database identifier or accession number) to be characterized as well as a list of experimentally measured peptide masses for this protein. The flexibility of the program allows the user to search for modifications not listed in FindMod by simply inputting the modifications and which amino acids types may be modified. Unlike SALSA, however, FindMod works only with peptide mapping data and not MS/MS data. This restricts FindMod use to relatively simple mixtures of proteins and cannot identify posttranslationally modified peptides in complex mixtures that would be analyzed by LC/MS/MS.

VI. Conclusions

Although technically more challenging, proteomics has and will follow many of the themes of the human genome project. Just as the human genome project seemed like an insurmountable task, proteomics faces the same skepticism. Just like genomics, however, the realization that the characterization of all the proteins expressed within a cell is possible is constantly being reinforced by the development of technologies in the areas of chromatography, mass analysis, and bioinformatics. Regardless of the ultimate success of global proteomic initiatives, such as HUPO, the Human Proteome Organisation, the analytical tools developed for these studies will have an everlasting impact on biochemical analysis. A description of some of the tools necessary for effective proteomic analysis is provided, but note that each type of tool itself requires a book to describe it in depth. Indeed, at the rate that separations, mass spectrometry technology, and bioinformatics are constantly being improved, any chapter of this type is quickly out of date.

References

1. Wilkins, M. R., Sanchez, J. C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F., and Williams, K. L. (1996). Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it. *Biotechnol. Genet. Eng. Rev.* **13,** 19–50.
2. O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250,** 4007–4021.
3. Link, A. J., Robison, K., and Church, G. M. (1997). Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. *Electrophoresis* **18,** 1259–1313.
4. Klade, C. S., Voss, T., Krystek, E., Ahorn, H., Zatloukal, K., Pummer, K., and Adolf, G. R. (2001). Identification of tumor antigens in renal cell carcinoma by serological proteome analysis. *Proteomics* **1,** 890–898.
5. Miyashita, M., Presley, J. M., Buchholz, B. A., Lam, K. S., Lee, Y. M., Vogel, J. S., and Hammock, B. D. (2001). Attomole level protein sequencing by Edman degradation coupled with accelerator mass spectrometry. *Proc. Natl. Acad. Sci. USA* **98,** 4403–4408.
6. Chaurand, P., Luetzenkirchen, F., and Spengler, B. (1999). Peptide and protein identification by matrix-assisted laser desorption ionization (MALDI) and MALDI-post-source decay time-of-flight mass spectrometry. *J. Am. Soc. Mass Spectrom.* **10,** 91–103.
7. Griffiths, W. J., Jonsson, A. P., Liu, S., Rai, D. K., and Wang, Y. (2001). Electrospray and tandem mass spectrometry in biochemistry. *Biochem. J.* **355,** 545–561.
8. Rappsilber, J., and Mann, M. (2002). What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.* **27,** 74–78.
9. Fenyo, D. (2000). Identifying the proteome: Software tools. *Curr. Opin. Biotechnol.* **11,** 391–395.

10. Martin, S. A., Rosenthal, R. S., and Biemann, K. (1987). Fast atom bombardment mass spectrometry and tandem mass spectrometry of biologically active peptidoglycan monomers from *Neisseria gonorrhoeae*. *J. Biol. Chem.* **262,** 7514–7522.
11. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994). 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.
12. Karas, M., Bachmann, D., Bahr, U., and Hillenkamp, F. (1987). Matrix-assisted ultraviolent later desorption of non-volatile compounds. *Int. J. Mass Spectrom. Ion Process.* **78,** 53–68.
13. Dole, M., Ferguson, L. D., Hines, R. L., Mobley, R. C., Ferguson, L. D., and Alice, M. B. (1968). Molecular beams of macro ions. *J. Chem. Phys.* **49,** 2240–2249.
14. Stuber, W., Hemmasi, B., and Bayer, E. (1983). Synthesis and photolytic cleavage of bovine insulin B22–30 on a nitrobenzoylglycyl-poly(ethylene glycol) support. *Int. J. Pept. Protein Res.* **22,** 277–283.
15. Dell, A., and Morris, H. R. (1982). Fast atom bombardment-high field magnet mass spectrometry of 6000 dalton polypeptides. *Biochem. Biophys. Res. Commun.* **106,** 1456–1462.
16. Sundqvist, B., Kamensky, I., Hakansson, P., Kjellberg, J., Salehpour, M., Widdiyasekera, S., Fohlman, J., Peterson, P. A., and Roepstorff, P. (1984). Californium-252 plasma desorption time of flight mass spectroscopy of proteins. *Biomed. Mass Spectrom.* **11,** 242–257.
17. Caprioli, R. M. (1988). Analysis of biochemical reactions with molecular specificity using fast atom bombardment mass spectrometry. *Biochemistry* **27,** 513–521.
18. Fenn, J. B., Mann, M., Meng, C. K., and Wong, S. F. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71.
19. Aleksandrov, M. L., Gall, L. N., Krasnov, N. V., Nikolaev, V. I., Pavlenko, V. A., and Shkurov, V. A. (1984). Ion extraction from solutions at atmospheric pressure – A method for mass spectrometric analysis of bioorganic substances. *Dokl. Akad. Nauk. SSSR* **277,** 379–383.
20. Meng, C. K., Mann, M., and Fenn, J. B. (1988). Electrospray ionization of some polypeptides and small proteins. *In:* ''36th ASMS Conference on Mass Spectrometry and Allied Topics''. San Francisco, CA, pp. 771–772.
21. von Brocke, A., Nicholson, G., and Bayer, E. (2001). Recent advances in capillary electrophoresis/electrospray-mass spectrometry. *Electrophoresis* **22,** 1251–1266.
22. Dalluge, J. J. (2000). Mass spectrometry for direct determination of proteins in cells: Applications in biotechnology and microbiology. *Fresenius J. Anal. Chem.* **66,** 701–711.
23. Guan, Z., Hofstadler, S. A., and Laude, D. A., Jr. (1993). Remeasurement of electrosprayed proteins in the trapped ion cell of a Fourier transform ion cyclotron resonance mass spectrometer. *Anal. Chem.* **65,** 1588–1593.
24. Green, B. N., Bordoli, R. S., Hanin, L. G., Lallier, F. H., Toulmond, A., and Vinogradov, S. N. (1999). Electrospray ionization mass spectrometric determination of the molecular mass of the approximately 200-kDa globin dodecamer subassemblies in hexagonal bilayer hemoglobins. *J. Biol. Chem.* **274,** 28206–28212.
25. Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., and Yoshida, T. (1988). Protein and polymer analyses up to *m/z* 100,000 by laser ionization time of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2,** 151–153.
26. Ryzhov, V., Bundy, J. L., Fenselau, C., Taranenko, N., Doroshenko, V., and Prasad, C. R. (2000). Matrix-assisted laser desorption/ionization time-of-flight analysis of

Bacillus spores using a 2.94 micron infrared laser. *Rapid Commun. Mass Spectrom.* **14,** 1701–1706.
27. Cotter, R. J. (1997). ''Time-of-Flight Mass Spectrometry: Instrumentation and Applications in Biological Research.'' *Oxford University Press, Oxford.*
28. Wang, B. H., and Biemann, K. (1994). Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry of chemically modified oligonucleotides. *Anal. Chem.* **66,** 1918–1924.
29. Huberty, M. C., Vath, J. E., Yu, W., and Martin, S. A. (1993). Site-specific carbohydrate identification in recombinant proteins using MALDI-TOF MS. *Anal. Chem.* **65,** 2791–2800.
30. Cornish, T. J., and Cotter, R. J. (1993). A curved-field reflectron for improved energy focusing of product ions in time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **7,** 1037–1040.
31. Brown, R. S., and Lennon, J. J. (1995). Sequence-specific fragmentation of matrix-assisted laser-desorbed protein/peptide ions. *Anal. Chem.* **67,** 3990–3999.
32. Juhasz, P., Roskey, M. T., Smirnov, I. P., Haff, L. A., Vestal, M. L., and Martin, S. A. (1996). Applications of delayed extraction matrix-assisted laser desorption ionization time-of-flight mass spectrometry to oligonucleotide analysis. *Anal. Chem.* **68,** 941–946.
33. Kaufmann, R. (1995). Matrix-assisted laser desorption ionization (MALDI) mass spectrometry: A novel analytical tool in molecular biology and biotechnology. *J. Biotechnol.* **41,** 155–175.
34. Kaufmann, R., Chaurand, P., Kirsch, D., and Spengler, B. (1996). Post-source decay and delayed extraction in matrix-assisted laser desorption/ionization-reflectron time-of-flight mass spectrometry: Are there trade-offs? *Rapid Commun. Mass Spectrom.* **10,** 1199–1208.
35. Yost, R. A., and Boyd, R. K. (1990). Tandem mass spectrometry: Quadrupole and hybrid instruments. *Methods Enzymol.* **193,** 154–200.
36. Li, J., Kelly, J. F., Chernushevich, I., Harrison, D. J., and Thibault, P. (2000). Separation and identification of peptides from gel-isolated membrane proteins using a microfabricated device for combined capillary electrophoresis/nanoelectrospray mass spectrometry. *Anal. Chem.* **72,** 599–609.
37. Dainese, P., Staudenmann, W., Quadroni, M., Korostensky, C., Gonnet, G., Kertesz, M., and James, P. (1997). Probing protein function using a combination of gene knockout and proteome analysis by mass spectrometry. *Electrophoresis* **18,** 432–442.
38. Steen, H., Kuster, B., and Mann, M. (2001). Quadrupole time-of-flight versus triple-quadrupole mass spectrometry for the determination of phosphopeptides by precursor ion scanning. *J. Mass Spectrom.* **36,** 782–790.
39. Carr, S. A., Huddleston, M. J., and Bean, M. F. (1993). Selective identification and differentiation of N- and O-linked oligosaccharides in glycoproteins by liquid chromatography-mass spectrometry. *Protein Sci.* **2,** 183–196.
40. Shevchenko, A., Chernushevich, I., Ens, W., Standing, K. G., Thomson, B., Wilm, M., and Mann, M. (1997). Rapid ''de novo'' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. *Rapid Commun. Mass Spectrom.* **11,** 1015–1024.
41. Chernushevich, I. V., Loboda, A. V., and Thomson, B. A. (2001). An introduction to quadrupole-time-of-flight mass spectrometry. *J. Mass Spectrom.* **36,** 849–865.
42. Morris, H. R., Paxton, T., Dell, A., Langhorne, J., Berg, M., Bordoli, R. S., Hoyes, J., and Bateman, R. H. (1996). High sensitivity collisionally-activated decomposition

tandem mass spectrometry on a novel quadrupole/orthogonal-acceleration time-of-flight mass spectrometer. *Rapid Commun. Mass Spectrom.* **10,** 889–896.
43. Shevchenko, A., Loboda, A., Shevchenko, A., Ens, W., and Standing, K. G. (2000). MALDI quadrupole time-of-flight mass spectrometry: A powerful tool for proteomic research. *Anal. Chem.* **72,** 2132–2141.
44. Loboda, A. V., Krutchinsky, A. N., Bromirski, M., Ens, W., and Standing, K. G. (2000). A tandem quadrupole/time-of-flight mass spectrometer with a matrix-assisted laser desorption/ionization source: Design and performance. *Rapid Commun. Mass Spectrom.* **14,** 1047–1057.
45. Stafford, G. C., Kelley, P. E., Syka, J. E. P., Reynolds, W. E., and Todd, J. F. J. (1984). Recent improvements in and analytical applications of advanced ion-trap technology. *Int. J. Mass Spectrom. Ion Processes* **60,** 85–98.
46. Jonscher, K. R., and Yates, J. R., III. (1997). The quadrupole ion trap mass spectrometer—a small solution to a big challenge. *Anal. Biochem.* **244,** 1–15.
47. Moyer, S. C., Cotter, R. J., and Woods, A. S. (2002). Fragmentation of phosphopeptides by atomospheric pressure MALDI and ESI/ion trap mass spectrometry. *J. Am. Soc. Mass Spectrom.* **13,** 274–283.
48. Krutchinsky, A. N., Kalkum, M., and Chait, B. T. (2001). Automatic identification of proteins with a MALDI-quadrupole ion trap mass spectrometer. *Anal. Chem.* **73,** 5066–5077.
49. Comisarow, M. B., and Marshall, A. G. (1974). Fourier transform ion cyclotron resonance spectroscopy. *Chem. Phys. Lett.* **25,** 282–283.
50. Marshall, A. G., Hendrickson, C. L., and Jackson, G. S. (1998). Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrom. Rev.* **17,** 1–35.
51. Solouki, T., Emmett, M. R., Guan, S., and Marshall, A. G. (1997). Detection, number, and sequence location of sulfur-containing amino acids and disulfide bridges in peptides by peptides by ultrahigh-resolution MALDI FTICR mass spectrometry. *Anal. Chem.* **69,** 1163–1168.
52. Bruce, J. E., Anderson, G. A., Wen, J., Harkewicz, R., and Smith, R. D. (1999). High-mass-measurement accuracy and 100% sequence coverage of enzymatically digested bovine serum albumin from an ESI-FTICR mass spectrum. *Anal. Chem.* **71,** 2595–2599.
53. Hakansson, K., Emmett, M. R., Hendrickson, C. L., and Marshall, A. G. (2001). High-sensitivity electron capture dissociation tandem FTICR mass spectrometry of microelectrosprayed peptides. *Anal. Chem.* **73,** 3605–3610.
54. Belov, M. E., Gorshkov, M. V., Udseth, H. R., Anderson, G. A., and Smith, R. D. (2000). Zeptomole-sensitivity electrospray ionization: Fourier transform ion cyclotron resonance mass spectrometry of proteins. *Anal. Chem.* **72,** 2271–2279.
55. Solouki, T., Gillig, K. J., and Russell, D. H. (1994). Mass measurement accuracy of matrix-assisted laser desorbed biomolecules: A Fourier-transform ion cyclotron resonance mass spectrometry study. *Rapid Commun. Mass Spectrom.* **8,** 26–31.
56. Mize, T. H., and Amster, I. J. (2000). Broad-band ion accumulation with an internal source MALDI-FTICR-MS. *Anal. Chem.* **72,** 5886–5891.
57. Li, L., Masselon, C. D., Anderson, G. A., Pasa-Tolic, L., Lee, S. W., Shen, Y., Zhao, R., Lipton, M. S., Conrads, T. P., Tolic, N., and Smith, R. D. (2001). High-throughput peptide identification from protein digests using data-dependent multiplexed tandem FTICR mass spectrometry coupled with capillary liquid chromatography. *Anal. Chem.* **73,** 3312–3322.
58. Hofstadler, S. A., Swanek, F. D., Gale, D. C., Ewing, A. G., and Smith, R. D. (1995). Capillary electrophoresis-electrospray ionization Fourier transform ion cyclotron

resonance mass spectrometry for direct analysis of cellular proteins. *Anal. Chem.* **67,** 1477–1480.
59. Lilley, K. S., Razzaq, A., and Dupree, P. (2002). Two-dimensional gel electrophoresis: Recent advances in sample preparation, detection and quantitation. *Curr. Opin. Chem. Biol.* **6,** 46–50.
60. Mann, M., Hendrickson, R. C., and Pandey, A. (2001). Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70,** 437–473.
61. Covey, T. R., Huang, E. C., and Henion, J. D. (1991). Structural characterization of protein tryptic peptides via liquid chromatography/mass spectrometry and collision-induced dissociation of their doubly charged molecular ions. *Anal. Chem.* **63,** 1193–2001.
62. von Brocke, A., Nicholson, G., and Bayer, E. (2001). Recent advances in capillary electrophoresis/electrospray-mass spectrometry. *Electrophoresis* **22,** 1251–1266.
63. Yang, L., Tang, Q., Harrata, A. K., and Lee, C. S. (1996). Capillary isoelectric focusing-electrospray ionization mass spectrometry for transferrin glycoforms analysis. *Anal. Biochem.* **243,** 140–149.
64. Jensen, P. K., Pasa-Tolic, L., Anderson, G. A., Horner, J. A., Lipton, M. S., Bruce, J. E., and Smith, R. D. (1999). Probing proteomes using capillary isoelectric focusing-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **71,** 2076–2084.
65. Chartogne, A., Gaspari, M., Jespersen, S., Buscher, B., Verheij, E., von Heijden, R., Tjaden, U., and van der Greef, J. (2002). On-target fraction collection for the off-line coupling of capillary isoelectric focusing with matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **16,** 201–207.
66. Smith, R. D., Fields, S. M., Loo, J. A., Barinaga, C. J., Udseth, H. R., and Edmonts, C. G. (1990). Capillary isotachophoresis with UV and tandem mass spectrometric detection for peptides and proteins. *Electrophoresis* **11,** 709–717.
67. Stubberud, K., Forsberg, A., Callmer, K., and Westerlund, D. (2002). Partial filling micellar electrokinetic chromatography optimization studies of ibuprofen, codeine and degradation products, and coupling to mass spectrometry. *Electrophoresis* **23,** 572–577.
68. Vollmerhaus, P. J., Tempels, F. W., Kettenes-Van Den Bosch, J. J., and Heck, A. J. (2002). Molecular interactions of glycopeptide antibiotics investigated by affinity capillary electrophoresis and bioaffinity electrospray ionization-mass spectrometry. *Electrophoresis* **23,** 868–879.
69. Hearn, M. T. (2001). Peptide analysis by rapid, orthogonal technologies with high separation selectivities and sensitivities. *Biologicals* **29,** 159–178.
70. Wagner, K., Miliotis, T., Marko-Varga, G., Bischoff, R., and Unger, K. K. (2002). An automated on-line multidimensional HPLC system for protein and peptide mapping with integrated sample preparation. *Anal. Chem.* **74,** 809–820.
71. Nilsson, C. L., Larsson, T., Gustafsson, E., Karlsson, K. A., and Davidsson, P. (2000). Identification of protein vaccine candidates from *Helicobacter pylori* using a preparative two-dimensional electrophoretic procedure and mass spectrometry. *Anal. Chem.* **72,** 2148–2153.
72. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R., III. (1999). Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676–682.
73. Washburn, M. P., Wolters, D., and Yates, J. R., III. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247.

74. Yates, J. R., III, Eng, J. K., McCormack, A. L., and Schieltz, D. (1995). Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **67,** 1426–1436.
75. Wachs, T., and Henion, J. (2001). Electrospray device for coupling microscale separations and other miniaturized devices with electrospray mass spectrometry. *Anal. Chem.* **73,** 632–638.
76. Li, J., LeRiche, T., Tremblay, T.-L., Wang, C., Bonneil, E., Harrison, D. J., and Thibault, P. (2002). Application of microfluidic devices to proteomics research: Identification of trace-level protein digests and affinity capture of target peptides. *Mol. Cell. Proteom.* **1,** 157–168.
77. Palm, A., Wallenborg, S. R., Gustafsson, M., Hedstrom, A., Togan-Tekin, E., and Andersson, P. (2001). Integrated sample preparation and MALDI MS on a disc: Micro Total Analysis Systems 2001. *In:* ''Proceedings of the 5th μTAS 2001 Symposium,'' October 21–25, 2001**,** Monterey, CA pp. 216–218.
78. Pappin, D. D. J., Højrup, P., and Bleasby, A. J. (1993). Rapid identification of proteins by peptide-mass finger printing. *Curr. Biol.* **3,** 327–332.
79. Wilkins, M. R., Gasteiger, E., Wheeler, C. H., Lindskog, I., Sanchez, J. C., Bairoch, A., Appel, R. D., Dunn, M. J., and Hochstrasser, D. F. (1998). Multiple parameter cross-species protein identification using MultiIdent—a world-wide web accessible tool. *Electrophoresis* **19,** 3199–3206.
80. Clauser, K. R., Baker, P., and Burlingame, A. L. (1999). Role of accurate mass measurement (10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **71,** 2871–2882.
81. Mann, M., Højrup, P., and Roepstorff, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **22,** 338–345.
82. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567.
83. Zhang, W., and Chait, B. T. (2000). ProFound—an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.* **72,** 2482–2489.
84. Hansen, B. T., Jones, J. A., Mason, D. E., and Liebler, D. C. (2001). SALSA: A pattern recognition algorithm to detect electrophile-adducted peptides by automated evaluation of CID spectra in LC-MS-MS analyses. *Anal. Chem.* **73,** 1676–1683.
85. Wilkins, M. R., Gasteiger, E., Gooley, A. A., Herbert, B. R., Molloy, M. P., Binz, P. A., Ou, K., Sanchez, J. C., Bairoch, A., Williams, K. L., and Hochstrasser, D. F. (1999). High-throughput mass spectrometric discovery of protein post-translational modifications. *J. Mol. Biol.* **289,** 645–657.

PROTEOMIC ANALYSIS BY TWO-DIMENSIONAL POLYACRYLAMIDE GEL ELECTROPHORESIS

By MING ZHOU AND LI-RONG YU

SAIC-Frederick, National Cancer Institute at Frederick, Frederick, Maryland 21702

I. Introduction

The present generation of scientists has been witness to a major revolution in how biological problems are addressed. Where possible, biological problems are no longer being pursued on the one molecular species-at-a-time approach. Many investigators are becoming more fascinated by studying megabases of DNA, thousands of mRNA transcripts, and thousands of proteins in single experiments. This new paradigm has changed much of biology from a hypothesis-driven science to a discovery-driven science. In a hypothesis-driven approach, previous experimental results are used to direct the next step in the research protocol. These approaches typically focus on a single gene, transcript, or protein per experiment. In a discovery-driven approach, little predisposed knowledge or assumptions are used to direct the research and the general goal is to use the acquired data (which is generally considerable) to provide details about the system under study. The goals are typically to sequence large numbers of genes (or the entire genome), quantitate changes in gene transcription for thousands of mRNAs simultaneously, or characterize hundreds (or thousands) of proteins in a single experiment. With these

ADVANCES IN
PROTEIN CHEMISTRY, Vol. 65
0065-3233/03 $35.00

large data sets in hand, the hope is that a global understanding of how a cell or organism functions will be revealed.

Although it has seemingly been thrust on the scientific community overnight, in reality decades of technological development have gone into fueling this revolution. The Human Genome Project could have been completed only with the many developments in cloning, amplification, and high-throughput gene sequencing [1–3] that have taken place over the past few decades. The routine analysis of thousands of gene transcripts, using mRNA arrays, was made possible by developments in such areas as mRNA isolation, sequencing, synthesis, and their covalent coupling to solid surfaces [4–6]. Although not as routine as global DNA or mRNA analysis, proteomics in its present state has also been dependent on the development of specific techniques over the past half-century. These developments have included the development of high-resolution separations of proteins, mass spectrometry (MS) technology, tandem MS of peptides, protein fractionation techniques, and bioinformatics [7–10].

Although proteomics means different things to different laboratories, the first thought that enters most people's minds when they hear the term is two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) to separate and compare protein expression, followed by MS for protein identification. There has been a concerted effort by many investigators to circumvent 2D-PAGE in their proteomic strategies; however, it is still far and away the most commonly used separation technique in this field. From the earliest days when only a few protein spots could be visualized, advances have been made that allow thousands of spots to be resolved on a single gel. Advances have also been made in the areas of protein staining to allow for more accurate and precise quantitation of proteins separated by 2D-PAGE.

The primary role of 2D-PAGE is the separation and comparison of the relative abundance of proteins from different proteome samples. This role is made possible by the high resolving power of 2D-PAGE separations of intact proteins and the development of various staining procedures to visualize these protein ''spots.'' In a routine 2D-PAGE-based experiment, the proteome from a control and treated cell type is extracted and separated on separate gels (Fig. 1). The proteins are visualized by colorimetric staining and the spots observed on both gels are aligned so that the relative staining intensity of each protein can be compared between gels. Protein spots that are more intense on one of the gels are excised from the gel and subjected to in-gel enzymatic digestion. The resultant peptides are extracted from the gel and analyzed by MS or tandem MS (MS/MS), and the MS or MS/MS data are searched against an appropriate database to identify their protein of origin. As other articles in

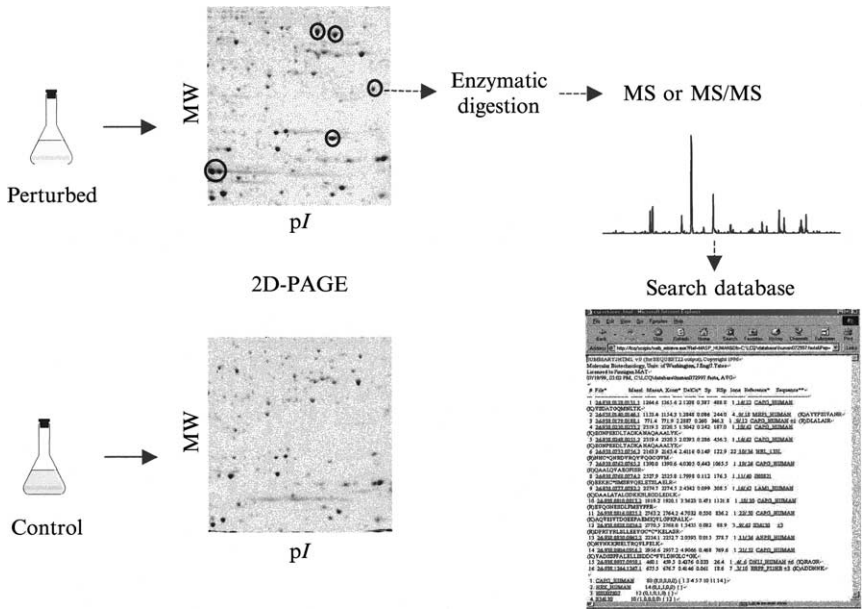


FIG. 1. Proteome analysis by two-dimensional polyacrylamide gel electrophoresis (2D-PAGE). The proteomes from distinct samples are extracted and separated by 2D-PAGE and the relative abundance of each protein in the different samples is measured by visualizing the protein spots. After excising spots with different abundances between the two gels, the spots are enzymatically digested, and the extracted peptides are analyzed by mass spectrometry (MS) or tandem MS. These data are used in conjunction with an appropriate database to identify the protein.

this volume deal with issues related to protein identification by MS, this article focuses on the technology and applications of 2D-PAGE.

II. HISTORICAL DEVELOPMENT OF 2D-PAGE

Although the initial implementation of high-resolution 2D-PAGE is attributable to O'Farrell in 1975 [11], the work that enabled his success had begun several decades before. The moving boundary method as an analytical tool for studying the electrophoresis of proteins was first demonstrated by Tiselius in 1930 and was used to resolve serum globulin into α, β, and γ components [12]. The first and original idea of separating proteins by electrophoresis in two dimensions was proposed by Smithies and Poulik in 1956 [13]. They had already recognized that using

a combination of the two electrophoretic processes on a gel at right angles would provide much higher resolution than is possible with either separately. In this example of 2D electrophoresis (2-DE), serum proteins were separated in the first dimension on the basis of their free solution mobilities, using a 5-mm-wide strip of filter paper and a buffer of pH 8.55. The strip was then inserted into a 12-cm-wide starch gel, and a second electrophoresis was carried out at right angles to the first, and the proteins were separated on the basis of their molecular size. Although the entire separation allowed them to resolve only slightly more than 15 components, it nonetheless was an improvement on existing technology and helped to usher in the development of 2-DE methods. After this report, other similar approaches to separating serum proteins were presented such as the combination of paper electrophoresis (first dimension) and cyanogum (second dimension) [14] as well as agar (first dimension) and starch (second dimension) [15]. With the advent of polyacrylamide gel [16], electrophoresis was conducted on polyacrylamide gel as a supporting medium and became a popular choice of method to separate proteins.

In 1964, the advantages of electrophoresis in flat slab gel format rather than cylindrical tubes were realized [17]. These advantages were demonstrated by Raymond, who constructed a system consisting of a vertical gel slab prepared in a specially designed cell combining a gel mold, buffer reservoirs and electrodes, and cooling plates all in one unit [17]. The advantages of using a flat slab gel format compared with a cylindrical gel were as follows: (1) the flat slab provides maximum surface area for cooling the gel; (2) the resulting patterns are easier to quantify in standard recording densitometers; (3) a large number of samples can be processed in a single gel, facilitating the direct comparison of specimens processed under identical conditions; and (4) the flat slab permits the application of two-dimensional techniques. These insightful statements are as true today as they were in 1964 and are the basis of modern 2D-PAGE. Raymond named his 2D separation technique ''orthogonal gel electrophoresis'' or ''Orthacryl.'' He applied Orthacryl to the separation of haptoglobins, isozymes, albumins, hemoglobins, and other protein systems.

At this stage a variety of different 2D gel combinations were attempted. Raymond himself experimented with the use of different acrylamide percentages and pH values in first- and second-dimensional separations. Laurell devised a procedure for the separation of serum proteins whereby electrophoresis on an agarose gel was used in the first dimension and affinity electrophoresis was used in the second dimension [18]. In this experiment, 20–50 μl of serum proteins was applied in a 2-cm-broad slit

on an agarose gel and separated by application of 10 V/cm for 3 h. A 0.5-cm-wide gel strip containing the separated components was cut from the middle of the gel and placed on another agarose gel containing antiserum. A potential gradient of 10 V/cm was applied at a 90° angle to the previous direction of electrophoretic separation for 30–90 min. The antigens as well as the antibodies started to move, and migrating precipitation zones of antigen–antibody complexes appeared within minutes. Freeman and Smith modified this procedure by not removing the protein-containing strip after the first-dimension separation [19]. After the first separation was completed, a second gel containing human antiserum was poured alongside the separated proteins, and a second electrophoresis was performed at a right angle to the first. More than 60 immunologically active proteins were resolved by this technique.

In the later 1960s, isoelectric focusing (IEF) as a principle to separate proteins was introduced into electrophoresis as the first dimension of two-dimensional polyacrylamide gel electrophoresis (IEF-PAGE) [20, 21]. Sodium dodecyl sulfate (SDS) was applied to the second dimension of IEF-PAGE in the early 1970s [22, 23]; thereby proteins migrated in polyacrylamide gel essentially on the basis of their size during electrophoresis. The previously described approaches, as well as others, laid the foundation for the development of modern 2D-PAGE. By further optimization of sample preparation and running conditions, O'Farrell introduced high-resolution 2-DE for the separation of cellular proteins under denaturing conditions, which allowed the resolution of more than 1000 proteins extracted from *Escherichia coli*. [11]. Similar methods were developed concurrently and independently by Klose [24] and Scheele [25]. These methodologies are the basis of modern high-resolution 2D-PAGE. The principle of the procedure is based on IEF, which separates the proteins on the basis of their isoelectric points (p*I*), followed by electrophoresis in the presence of SDS to separate the proteins on the basis of their molecular weights. These separation parameters allow for the resolution of proteins differing by a single charge, thereby allowing such *in vivo* modifications such as phosphorylation and certain mutations to be detected. O'Farrell's method was quickly adapted by Anderson and Anderson for the analysis of human plasma proteins [26]. They were able to resolve, on staining, approximately 300 spots, which they surmised were composed of perhaps 75–100 polypeptides. Since that time, much effort has gone into developing methods that improve the resolution of 2D-PAGE, reproducibility, and increase the ability to visualize more protein spots using more sensitive staining techniques [27–33].

III. Limitations and Advances

Ideally, every protein expressed within a cell would be resolved and detected as a single spot on a 2D-PAGE gel. Unfortunately, reality is far from this ideal. In a typical cell lysate, 10% of the different proteins comprise about 90% of the protein content. In the extreme case of serum, a single protein (albumin) makes up about 70% of the protein content. Most low-abundance proteins are below the detection limits of conventional staining techniques and hence unidentifiable except by more sensitive technologies such as MS. In addition, multiple proteins can migrate to the same position within a gel, rendering their comparative quantitation extremely difficult. For example, Gygi *et al.* identified six proteins within a single silver-stained spot in the analysis of a yeast cell extract, using a narrow-range gel [34].

Much of the developmental effort in 2D-PAGE has been to devise ways to visualize lower abundance proteins. Beyond the use of higher sensitivity staining methods described below, the use of narrow-range immobilized pH gradient (IPG) strips in the first dimension has shown much promise. A 2D-PAGE method in which proteins are applied to narrow-range IPG strips has been developed [35]. In the initial demonstration, six IPG strips covering pH ranges 3.5–5, 4.5–5.5, 5–6, 5.5–6.7, 6.2–8.2, and 7–10 were used. The strips are typically 1–3 pH units in range and overlap one another by 0.5 pH unit. After separating about 1 to 2 mg of protein from a B lymphoma cell line by IEF on each individual strip, they were then applied to individual SDS-PAGE gels and separated in the second dimension on the basis of molecular weight. For comparison, the same sample was separated on a single standard 2D-PAGE gel, using an IPG strip with a pH range of 3–10. Approximately 1500 spots were detected on this single 2D-PAGE gel; however, approximately 5000 spots were detected when using the six ultrazoom gels. Whereas only 0.8 mg of protein could be loaded onto the single 2D-PAGE gel, a total of 11 mg was loaded onto the ultrazoom gels. Wildgruber *et al.* used IPG strips with pH ranges of 4–5, 4.5–5.5, 5–6, and 5.5–6.7 to separate a complex proteome sample and compared this against gels run with IPG strips with pH ranges 3–10 and 4–7 [36]. Approximately 1.6 and 2.3 times more protein spots were detectable with the narrow-range gels than with the pH 4–7 and 3–10 IPG strips, respectively. A comparison of narrow-range IPG gels with pH 4–7 and 3–10 IPG gels is shown in Fig. 2. The use of narrow-range IPG strips provides the potential to visualize and identify low-abundance proteins and may be used as preparative gels for the isolation of species such as phosphoproteins present in low stoichiometric amounts within the cell.

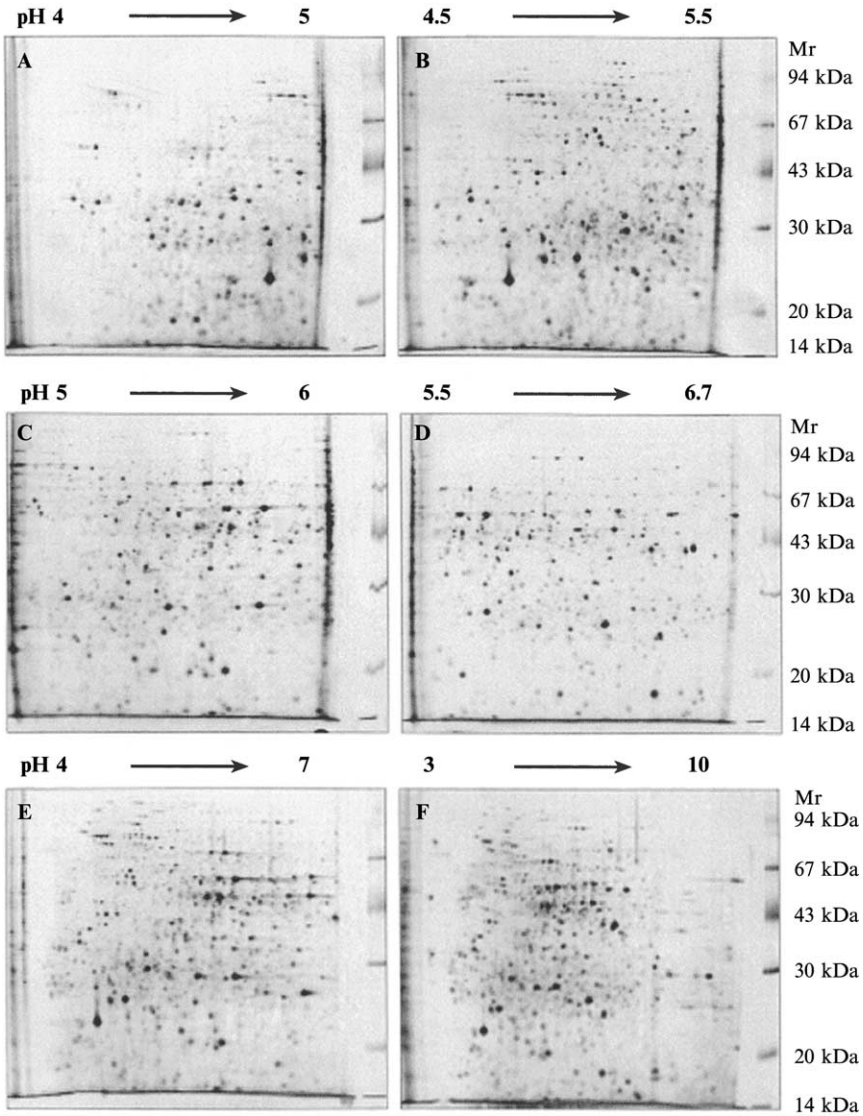


FIG. 2. Comparison of number of yeast protein spots visualized by 2D-PAGE fractionation, using immobilized pH gradient (IPG) strips with pH ranges of (A) 4–5, (B) 4.5–5.5, (C) 5–6, (D) 5.5–6.7, (E) 4–7, and (F) 3–10 in the first dimension. The second dimension was 13% SDS–PAGE and the proteins were visualized by silver staining.

Another way to detect lower abundance proteins is to enrich for these species before fractionating them by 2D-PAGE. The preenrichment method selected depends on the class of proteins of interest. For instance, Stancato and Petricoin used a panel of anti-phosphotyrosine antibodies to enrich for phosphotyrosine-containing proteins [37]. Other methods such as immobilized metal ion chromatography could also be used as a general method to enrich for phosphoproteins. Approaches for enriching for low-abundance proteins generally employ standard chromatographic techniques. These methods do not in fact select for a specific class of protein but are used to separate high-abundance from low-abundance proteins. This fractionation prevents the high-abundance proteins from obscuring the lower abundance proteins when they are visualized on a 2D-PAGE gel. Chromatographic methods such as anion-exchange and hydrophobic interaction chromatography have been used to enrich for low-abundance proteins before 2D-PAGE separation [38–40]. Other approaches have been used to prefractionate subcellular components such as membranes or mitochondria before 2D-PAGE analysis [41]. Often the role of the enrichment step is not designed to extract low-abundance proteins per se, but rather to remove the high-abundance proteins. This strategy is utilized when analyzing serum by 2D-PAGE. Serum 2D-PAGE profiles are dominated by albumin because this protein makes up about 60–70% of the protein content of serum. To be able to visualize low-abundance proteins within serum requires removing albumin from the mixture before 2D-PAGE. Although it is most often removed by immunoaffinity methods, Cibracon Blue dye chromatography has also been used to deplete serum of albumin [42].

One of the major drawbacks of 2D-PAGE is the difficulty in solubilizing membrane proteins in buffers commonly used for isoelectric focusing (IEF) [43]. To solubilize membrane proteins before IEF, solubilizers such as Triton X-114, chloroform–methanol, sodium carbonate, and detergent extractions have been used. Of the proteins identified in the analysis of rat liver Golgi complex proteins solubilized with Triton X-114 before 2D-PAGE, however, few were found to be integral membrane proteins, suggesting that many of this class still remained insoluble [44]. Work in which a combination of detergents was used to solubilize bacterial outer membrane proteins and total membrane proteins from *Haemophilus influenzae* has suggested that at present there is no way to effectively solubilize the entire membrane protein complement with a single detergent or solubilizer [45]. However, many of the membrane proteins are extremely large, causing difficulty in separating them by gel-based techniques. This kind of issue is inherent to 2D-PAGE; therefore, many large membrane proteins might not be visualized in 2D gels even if they are solubilized in lysis buffer.

IV. Protein Visualization Methods

Once proteins are separated via 2D-PAGE, they need to be visualized before MS analysis. Proteins in electrophoretic gels can be visualized by a number of staining techniques. Visualizing the protein spots provides two key parameters: the ability to locate the protein within the gel and to compare its intensity with its spot on another gel so that a sense of its relative expression between two samples can be ascertained. In addition to the conventional staining techniques described below, methods have been developed for the specific labeling or staining of proteins that are glycosylated, phosphorylated, S-nitrosylated, ADP-ribosylated, arginine methylated, and proteolytically modified. The key characteristics required of any effective staining technique are sensitivity, high linear dynamic range for quantitation measurements, and compatibility with downstream MS analysis.

A. Coomassie Blue

Coomassie blue dyes (R and G types) have been the most popular stains used in proteomics to date, because of their low cost, ease of use, and good compatibility with MS [46]. Coomassie blue R-250 dye is used to visualize proteins via a regressive staining approach in which gels are saturated with dye and then destained with an aqueous solution containing methanol and acetic acid. The proteins are visible after destaining because they have a higher affinity for the dye molecules than does the gel matrix. The gel can be further destained with the methanol–acetic acid destaining solution to get rid of Coomassie blue dyes before in-gel digestion for the subsequent MS analysis [47]. Unfortunately, staining reproducibility is difficult to control because proteins can destain to varying extents along with the gel matrix, albeit with slower kinetics. Progressive staining approaches, based on the formation of colloidal dye particles, have been introduced in which proteins are gradually stained to an end point, without significant staining of the gel matrix. Equilibrium is achieved between colloidal particles and freely dispersed dye in solution during the staining procedure. The proteins in the gel are preferentially stained by the low concentration of free dye, but the matrix remains unstained because the colloidal dye particles are excluded from the gel. The limits of detection for conventional Coomassie blue and colloidal Coomassie blue stains are 8–10 and 30–100 ng, respectively. Each dye provides a linear response with protein over a 10- to 30-fold range of concentration [48]. Therefore the primary limitations

with these stains are their poor sensitivity and limited linear dynamic range.

B. Silver Stain

In silver staining, the gel is saturated with silver ions. The less tightly bound metal ions are washed out of the gel and those bound to the proteins are reduced to form metallic silver [49]. Although silver stain will detect as little as 1 ng of protein, the linear dynamic range of the stain is restricted to a 10-fold range. The sensitivity of silver stain is at least 10-fold greater than that of Coomassie blue staining (Fig. 3). Because silver staining methods can be complex and need to be stopped at some time point to avoid overdeveloping the gel, the gel-to-gel reproducibility needs to be controlled strictly. Indeed, investigators have reported variations of 20% in spot intensity [50]. Fortunately, automated staining devices are capable of minimizing the variability; however, the price of such equipment generally makes such technology feasible only for commercial companies. The best silver stain methods use aldehyde-based fixatives; however, these fixatives interfere with the downstream analysis of gel spots by Edman sequencing or MS. While alternative silver staining methods have been developed that omit aldehydes in the fixatives, generally the detection sensitivity is poorer and background staining is less uniform [51] (Fig. 4). In addition, destaining methods are often employed to improve the compatibility of silver staining with peptide mass profiling methods [52].

C. SYPRO Stains

SYPRO stains are relatively new compared with silver and Coomassie stains, but are becoming increasingly popular to visualize proteins separated by 2D-PAGE. SYPRO Red, SYPRO Orange, and SYPRO Tangerine dyes can detect proteins in a simple, one-step staining procedure that can be completed in less than 1 h and does not require a destaining step [53–55]. SYPRO stains can detect as little as 4–8 ng of protein, rivaling the sensitivity of rapid silver staining techniques and colloidal Coomassie blue staining methods (Fig. 5). SYPRO Ruby dye is the most popular SYPRO dye in use and has been shown to be as sensitive as the best silver staining method available [56]. For quantitative purposes, the linear dynamic range of the SYPRO Ruby stain extends over three orders of magnitude, surpassing both silver and

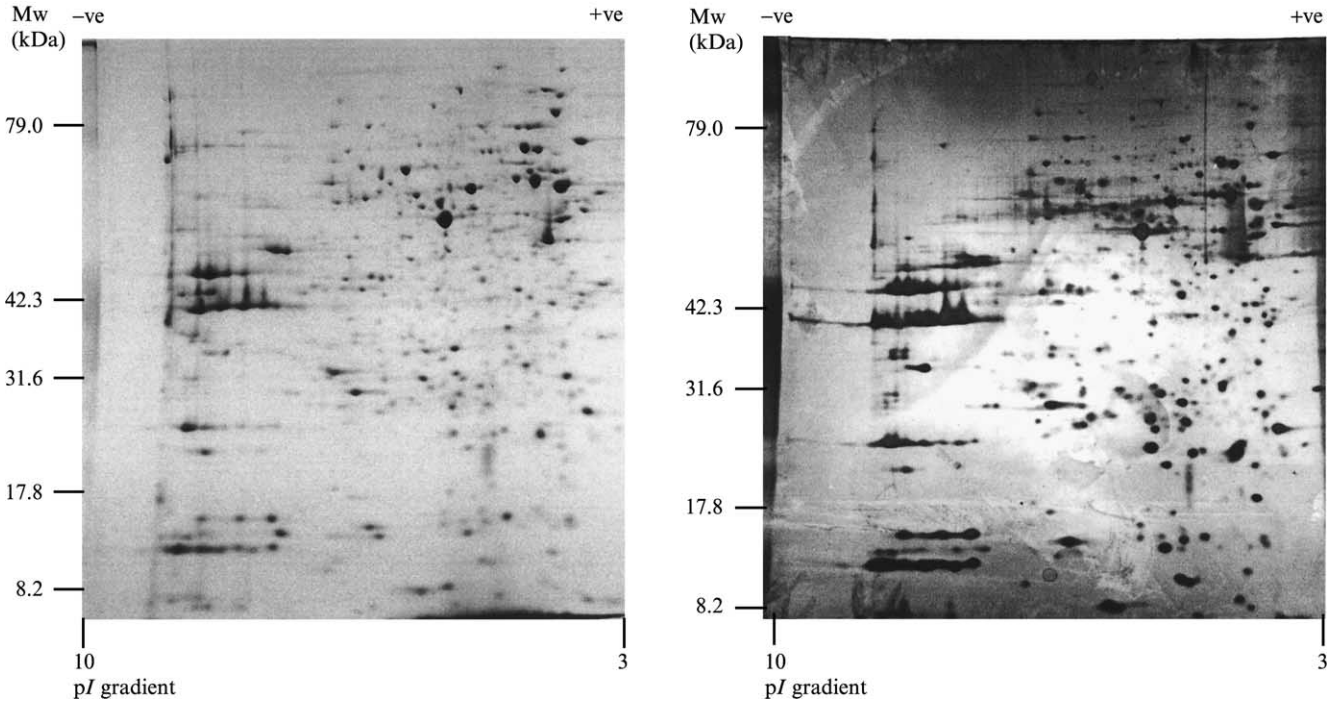


FIG. 3. Comparison of Coomassie blue and silver staining for detection of 2D-PAGE-separated proteins. Identical proteome samples extracted from rat basophil leukemia (RBL) cells were subjected to 2D-PAGE and stained with colloidal Coomassie blue (*left*) and a modified silver staining with thiosulfate enhancement (*right*). The intensity of the spots visualized by silver staining is consistently greater than that revealed by colloidal Coomassie blue staining.

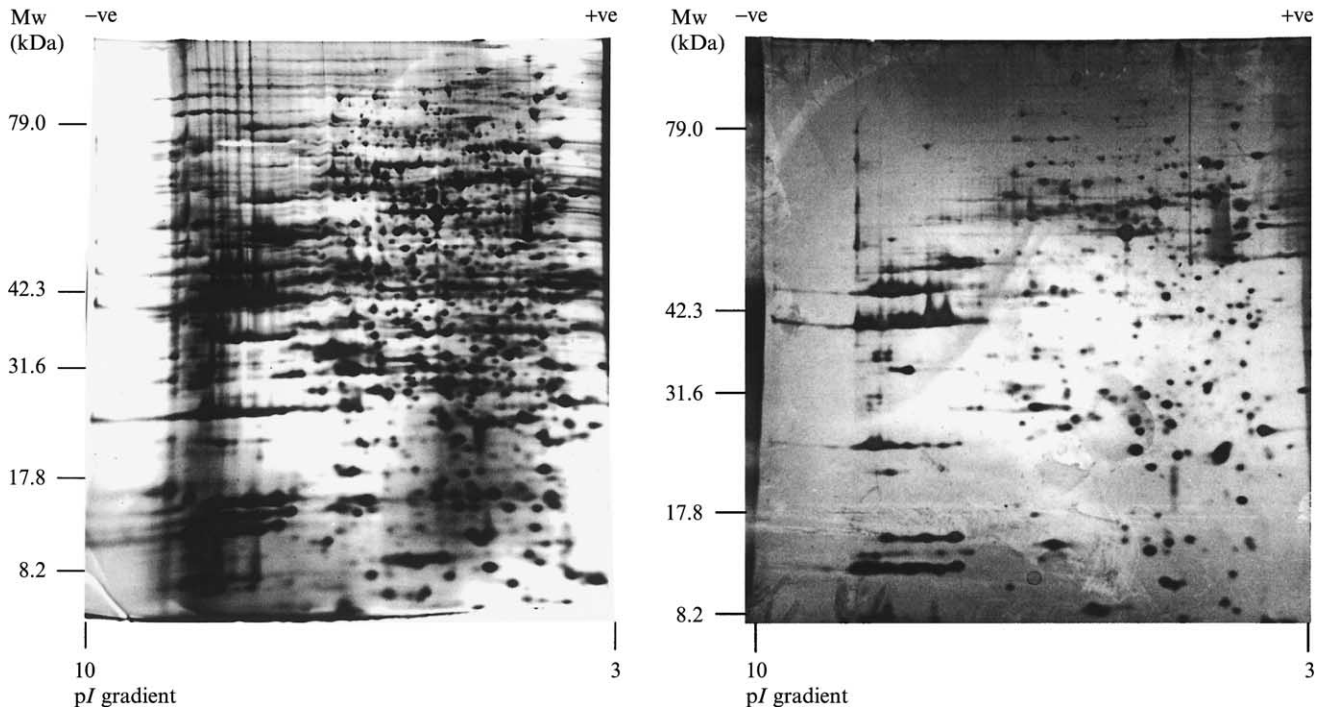


FIG. 4. Comparison of different silver-staining techniques for detection of 2D-PAGE-separated proteins. Identical proteome samples extracted from RBL cells were subjected to 2D-PAGE and stained with ammoniacal silver with glutaraldehyde fixation (*left*) and with a modified silver staining with thiosulfate enhancement (*right*). Although the glutaraldehyde-fixed gel shows many more spots, this staining technique is not compatible with downstream mass spectrometric analysis.

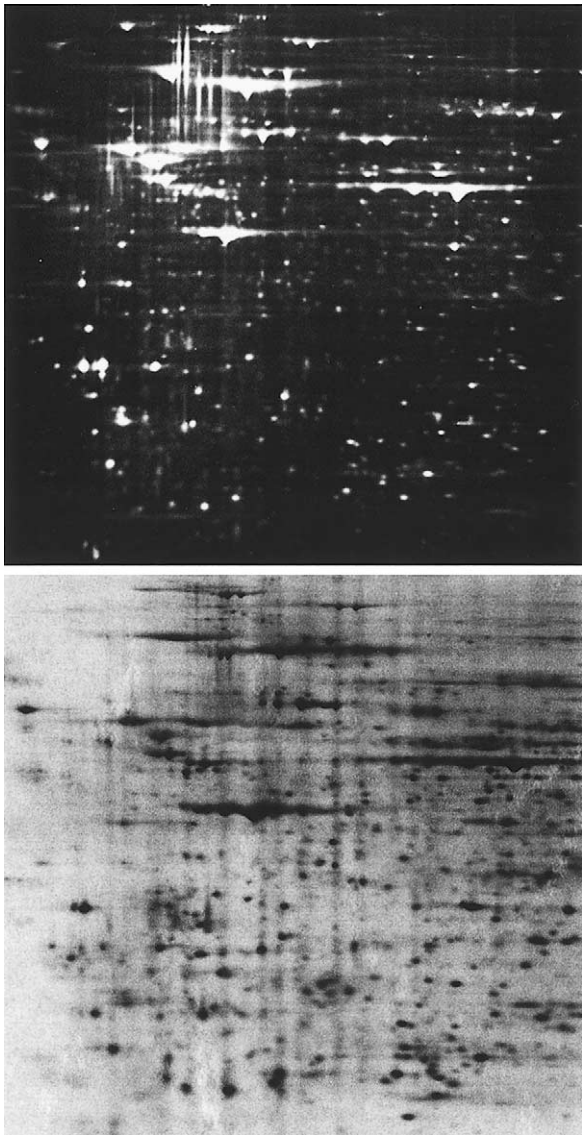


FIG. 5. Comparison of fluorescent staining with silver staining of 2D-PAGE-separated proteins. Identical proteome samples from a cultured fibroblast cell line were fractionated by 2D-PAGE and stained with SYPRO Ruby stain (*top*) or by a silver staining method that is optimized for sensitivity but is not compatible with mass spectrometric analysis (*bottom*). Similar detection sensitivity is observed between these techniques. The mass spectrometry-compatible silver staining techniques are significantly less sensitive than the mass spectrometry-compatible SYPRO Ruby protein staining.

Coomassie blue stains in performance. The dye can be excited with a standard 300-nm UV transilluminator or by imaging systems equipped with 450-, 473-, 488-, or even 532-nm lasers. Because SYPRO Ruby dye is an end-point stain, staining times are not critical and staining can be performed unattended without fear of overdevelopment. The SYPRO staining method is compatible with protein Edman sequencing and MS analysis.

D. Differential Gel Electrophoresis

An inherent problem of many 2D gel systems in comparing levels of protein expression between cells is poor reproducibility. This lack of reproducibility can make the alignment of gels difficult as well as make the relative quantitation comparisons between gels suspect. Fortunately, the succinimidyl esters of the cyanine dyes Cy2, Cy3, and Cy5 may be employed to fluorescently label as many as three different complex protein populations before mixing them together and running them simultaneously on the same 2D gel. This method is referred to as difference gel electrophoresis (DIGE) [57] (Fig. 6). Because the fluorescently labeled proteins from different samples are mixed and separated together, the same isoform of a given protein will thus migrate to the same position on the 2D gel. The relative abundance of each protein in each sample can then be obtained by scanning the gel, using excitation and emission wavelengths unique to each dye used.

In DIGE the fluorophor labels are covalently attached to lysyl residues. To maintain the solubility of the labeled proteins during electrophoresis, it is critical that only ~1–2% of the lysine residues in the proteins be fluorescently modified. If a higher degree of covalent modification is used, the solubility of the proteins is decreased, resulting in the detection of fewer protein spots. If optimized, DIGE technology is capable of detecting about half as many proteins as conventional silver staining; however, it is more sensitive than silver stain methods optimized for MS analysis [58]. While SYPRO Ruby dye staining detects 40% more protein spots compared with DIGE fluorophors, cyanine dyes detect about four times as many proteins as colloidal Coomassie blue staining [58, 59]. It is almost certain that the proteins missed by DIGE staining represent the lowest abundance species in the sample.

The true power of DIGE is in its ability to quantitate two or three proteome samples within a single gel. This eliminates the nuisance of gel-to-gel irreproducibility when comparing protein expression levels. It has been demonstrated that 20% changes in abundant proteins and 800% changes in scarce proteins can be quantified reliably at the 90%

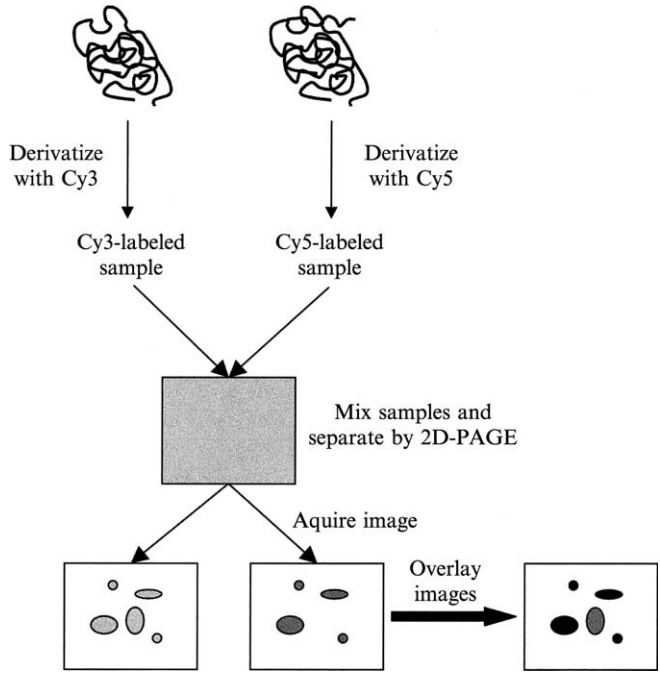


FIG. 6. Scheme of differential gel electrophoresis analysis. The proteome samples to be compared are differentially labeled with the fluorescent dye Cy2, Cy3, or Cy5. After labeling the proteins are mixed together and separated on a single 2D-PAGE gel. The gel is then visualized by fluorescence, using absorption and emission wavelengths specific to each dye. Comparison of the two images reveals proteins that are differentially expressed in the two cell types.

confidence level by DIGE technology, suggesting that quantitative changes in lower abundance proteins are likely to be overpowered by dye-dependent photophysical effects [58]. For example, protein changes in mitochondria from heart tissue of knockout mice lacking sarcomeric mitochondrial creatine kinase were studied by DIGE [60, 61]. Mitochondrial creatine kinase catalyzes the reversible phosphorylation of phosphocreatine, a critical mechanism in cellular energetics of muscle [62]. Deletion of such a vital enzyme forces the cells into metabolic crisis, causing widespread repercussions at the level of protein expression. The loss of creatine kinase, as well as 2- to 4-fold changes in a number of proteins, including mitochondrial cytochrome *c* oxidase, inorganic phosphate carrier, adenine nucleotide translocator, and voltage-dependent anion channel proteins, have been shown in the knockout mice by

conventional Western blotting [60]. Analysis of the mitochondria by DIGE technology confirmed the absence of creatine kinase in the knockout mice, as well as a minor difference in the precursor form of aconitase, but surprisingly, no other large-scale changes in heart mitochondrial proteins were detected [61]. Presumably, the quantitative changes identified in this model system by Western blotting could not be differentiated from background fluctuations associated with the DIGE labeling methodology.

E. Phosphoprotein Detection

2D-PAGE has been used primarily for the separation and quantitation of the relative abundance of proteins from different systems; methods are also being developed to characterize the phosphoproteome of cells by 2D-PAGE. The primary difference in detecting phosphorylated proteins is the staining method selected. A labeling or staining procedure must be used that is specific to phosphorylated proteins and does not interact with nonphosphorylated species. In addition, because many phosphorylation-regulated proteins are present in low amounts within the cell, the staining or labeling method must be sensitive. These criteria are met using both ^{32}P labeling (when possible) and immunostaining. The use of ^{32}P labeling is applicable only to cultured cells; however, it does represent a sensitive method of phosphoprotein detection. Immunostaining is more universally applicable and is still highly sensitive. Furthermore, antibodies can be used that can discriminate between seryl, threonyl, and tyrosyl phosphorylation. The general low abundance of phosphoproteins often requires a two-pronged approach for the detection and subsequent identification of these proteins. As shown in Fig. 7, two distinct 2D-PAGE gels of each sample, an analytical gel and a preparative gel, can be run [62]. The analytical gel is used to specifically detect the phosphoproteins. If the proteins are ^{32}P labeled then the gel is visualized by autoradiography; if immunostaining is used the proteins within this gel are transferred to a polyvinylidene fluoride (PVDF) membrane, which is then immunoblotted with the phosphospecific antibody of choice. The visualized phosphoproteins can then be traced back to their spots on the preparative gel in which more total protein has been separated, and they have been stained by a conventional staining method such as Coomassie blue or silver stain. Because gel-to-gel reproducibility is always an issue in 2D-PAGE, the immunostained membrane can also be stained to provide landmarks to pinpoint the exact spot of the phosphoprotein on the preparative gel. Some of the stains used for this include India ink and colloidal gold stain [63]. Colloidal gold stain has the advantage of being more sensitive; however, it can result in high background on nitrocellulose membranes.

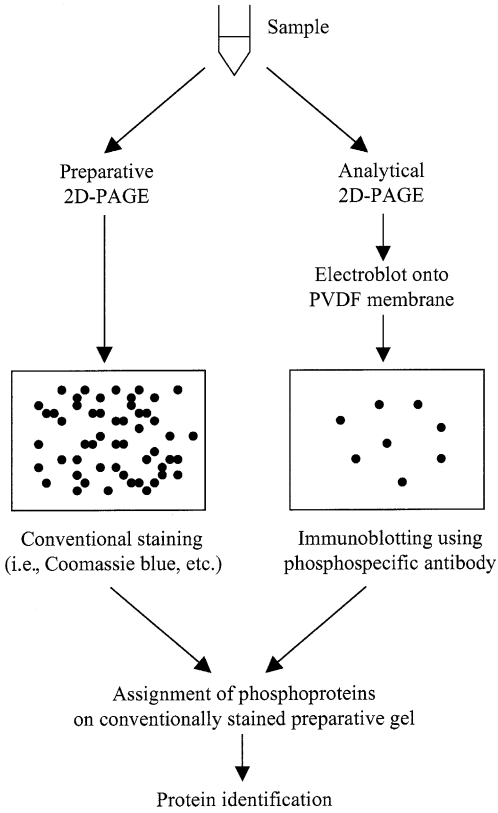


FIG. 7. The use of 2D-PAGE to identify phosphoproteins. Because of the low abundance of most phosphoproteins, preparative and analytical gels of the proteome sample can be run. Obviously, a greater amount of sample is loaded onto the preparative gel. After separation, the analytical gel is immunoblotted with a phosphospecific antibody to reveal the location of phosphoproteins within the gel. Spots visualized in this manner are aligned with their spots on the preparative gel which has been stained by a conventional method (i.e., colloidal Coomassie). The phosphoproteins can then be identified by mass spectrometry-based proteome methods.

V. PROTEOME APPLICATIONS OF 2D-PAGE

Although the basic foundation of 2D-PAGE in proteomics is the high-resolution separation and quantitation of proteins from different samples, the explosion of proteomics has resulted in many different variations on this approach. The number of applications reported that use 2D-PAGE has skyrocketed and it would be impossible to list them all within this article. For a sense of the scope of proteome projects that make use of 2D-PAGE

the reader is invited to browse the SWISS-2DPAGE database (http://us.expasy.org/ch2d/) or any of the 2D-PAGE databases listed in the WORLD-2DPAGE Web site (http://www.expasy.ch/ch2d/2d-index.html). These databases contain 2D-PAGE images from a wide variety of different cell types, organisms, and tissue types. A few applications of 2D-PAGE in proteomics are highlighted below.

A. Effects of Hypoxia on Kidney Protein Expression

One report used 2D-PAGE in combination with MS to identify proteins that were differentially regulated in rat kidney during episodic hypoxia (EH) [64]. EH is associated with systemic hypertension and is a major symptom of obstructive sleep apnea syndrome (OSAS) [65]. OSAS is a major public health problem that affects up to 5% of the adult population, often resulting in system hypertension or proteinuria and end-stage renal disease [66]. To characterize proteins differentially regulated by EH, rats were exposed to either EH or sustained hypoxia (SH) conditions for 14 or 30 days [64]. The rats were housed in special designed chambers in which a 12-h light–dark cycle was maintained. The O_2 concentration of each chamber was carefully monitored and was cycled between 10% O_2 and room air every 90 s during daylight hours for the EH rats. For the SH rats the O_2 concentration was maintained at 10% throughout the duration of the experiment.

The rats were killed after 14 and 30 days of different O_2 exposures and the kidneys were dissected and frozen in liquid N_2 and ground to powder. The individual tissue samples were resuspended in a 50 m*M* Tris, 0.3% SDS, and 200 m*M* dithiothreitol (DTT) buffer and heated to 100°C for 5 min and then put on ice. One-tenth volume of 500 m*M* Tris, 50 m*M* $MgCl_2$, DNase I (1 mg/ml), and RNase A (0.25 mg/ml) buffer was added and the sample was incubated for another 10 min. The sample was centrifuged at 12,000 rpm and the supernatant was removed. The proteins were precipitated from the supernatant by adding 10% trichloroacetic acid (TCA) and centrifuging again at 12,000 rpm. The protein pellets were then washed several times with acetone, and resuspended in sample buffer consisting of 40 m*M* Tris, 7.92 *M* urea, 0.06% SDS, 1.76% ampholytes, 120 m*M* DTT, and 3.2% Triton X-100.

The first dimension of the 2D-PAGE was run as precast carrier ampholyte tube gels, pH 3–10, with 100 m*M* NaOH as the cathode buffer and 10 m*M* H_3PO_4 as the anode buffer. A total of 100 μg of protein sample was loaded per gel. After extruding the gels from their tubes, they were incubated in Tris–acetate equilibration buffer with 0.01%

bromophenol blue and 50 m*M* DTT for 2 min. The gels were then loaded onto precast 10% homogeneous, 200 × 200 mm slab gels. After separation of the protein samples, the 2D-PAGE gels were fixed in 10% methanol and 7% acetic acid for 30 min and stained with SYPRO Ruby stain. Proteins that showed differential expression between the control, EH, and SH rats at 14 and 30 days were excised from the gel and identified by MS.

The expression profiles of 248 protein spots were evaluated on each of the gels [64]. To facilitate comparisons between the three sets of animals, an average image for each set of gels was established and used as the reference for each group. The renal proteome maps of the control animals and of the animals exposed to EH and SH for 30 days revealed significant differences. In the EH rats, all five isoforms of kallistatin were significantly downregulated whereas four of the five isoforms of this protein were downregulated in the rats exposed to SH for 30 days. A decrease in kallistatin expression is associated with reduction in the vasodilating capacity of the kidney and can cause or aggravate hypertension [67]. Interestingly, whereas all five forms of α_1-antitrypsin (A1AT) precursor were upregulated in EH rats, three of these proteins were not visible in the gel analysis of tissue from SH rats. Ferritin and β-actin were upregulated in both EH and SH rats, whereas vimentin and protein disulfide isomerase were upregulated in EH rats but downregulated in SH rats [64]. The investigators then made clever use of their proteome data to identify subsequent effects of EH and SH on kidney tissue. Because it has been demonstrated that the kallikrein/kallistatin pathway can modulate the response to bradykinin and that kallistatin can directly modulate the activity of B2-bradykinin receptor (B2R) [68], the effect of EH and SH on B2R expression was measured by immunoblotting. B2R expression was found to be increased in kidney tissue from rats exposed to both EH and SH for 14 days; however, it remained elevated only in animals exposed to SH for 30 days. The levels of kallikrein were also found to be elevated only in the SH rats exposed for 14 and 30 days. No significant change in the expression level of kallikrein was observed in the EH rats. Taken together, the proteomic and classic protein data suggest a new hypothesis that EH-induced hypertension results, in part, from decreased kallistatin levels and a lack of increasing B2R expression that would compensate for the loss of kallistatin. In addition, the data suggest that increases in kallikrein and B2R expression may act as a compensatory mechanism to inhibit hypertension during periods of sustained hypoxia.

B. *Characterization of Human Prion Protein by 2D-PAGE*

One of the major reasons for choosing 2D-PAGE for proteomics studies is its high resolving power at the protein level. Potentially thousands of proteins can be clearly resolved by this technology. A prime example of the knowledge that can be gleaned through this high-resolution fractionation is a study of the human cellular prion protein (PrP^C) [69]. PrP^C is a 209-amino acid residue glycoprotein that is highly expressed in brain cells [70]. The protein is processed through the secretory pathway in a variety of ways [71]. For example, within the endoplasmic reticulum anything from zero to two glycan groups are attached to PrP^C, whereas in the endosomal compartment the protein is cleaved to generate a C-terminal fragment. Additional fragments generated from the protein are also likely to be present within the brain. This combination of glycosylated and truncated forms of the protein suggests that many different isoforms of PrP^C are possible; however, only a few forms are thought possible, mainly on the basis of results of 1D-PAGE separations probed with a single monoclonal antibody (mAb) to PrP^C [72].

To investigate the complexity of PrP^C processing, Pan *et al.* used four different mAbs to different epitopes within PrP^C to extract protein from brain tissue [69]. Samples of the extracted protein were then separated by 1D SDS–PAGE and 2D-PAGE. In most 1D immunoblots the PrP^C proteins were separated into three distinct bands that represent the di-, mono-, and unglycosylated isoforms. This group developed 1D immunoblots of the PrP^C extracts with four different mAbs (8H4, 8B4, 3F4, and 8F9). A single band at 35 kDa was observed with mAb 8B4, corresponding to the diglycosylated, full-length PrP^C. After deglycosylation of the protein extract by PNGase F treatment, this band migrated to 29 kDa, in agreement with the estimated molecular mass of the full-length unglycosylated PrP^C (Fig. 8B, lane 2). The mAb 3F4 reacted with three PrP^C species of 35–37, 31–32, and 29–30 kDa (Fig. 8B, lane 3), which, following deglycosylation, became two bands migrating with apparent molecular masses of 29 and 21 kDa (Fig. 8B, lane 4). Both mAbs 8H4 and 8F9 demonstrated similar patterns (Fig. 8B, lanes 5 and 7), which included at least five distinct PrP^C species. Similar to patterns seen with the other mAbs, these bands shifted to significantly lower molecular weights after deglycosylation, consistent with results seen for the other mAbs.

The protein extract separated by 2D-PAGE was also subjected to immunoblot analysis with a series of mAbs. The 2D immunoblots showed a significantly more complex pattern of PrP^C proteins than visualized by 1D-PAGE. The two species migrating at 35–37 kDa recognized by mAb 8B4 (Fig. 9A.) distribute in approximately fourteen and eight spots,

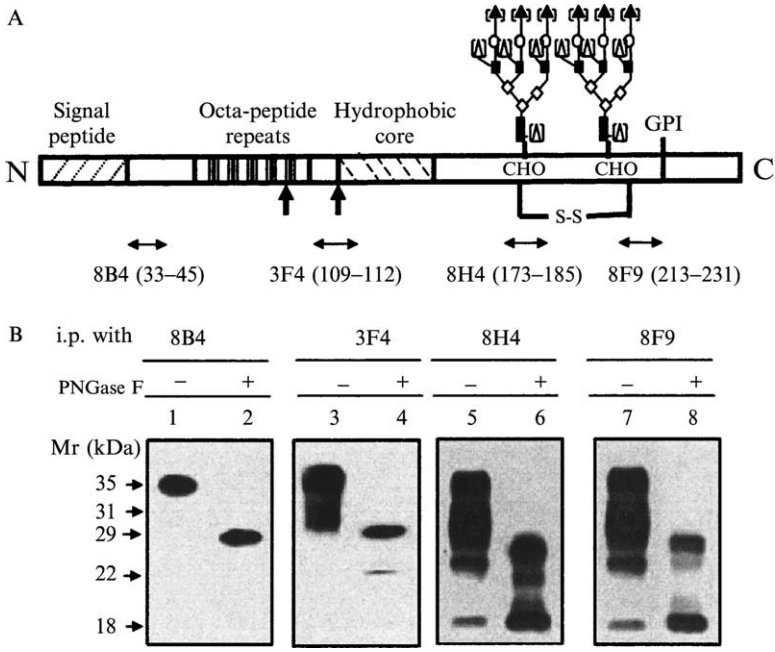


FIG. 8. Regions of the cellular prion protein (PrP^C) recognized by four monoclonal antibodies (mAbs) and one-dimensional immunoblots of PrP^C with the four mAbs. (A) The regions recognized by the four mAbs are indicated. During the processing of PrP^C, complex glycans are added to the protein in a variety of combinations. (B) PrP^C was immunoprecipitated (i.p.) from brain homogenates with the mAbs indicated, treated or not treated with PNGase F, analyzed by one-dimensional SDS–PAGE, and immunoblotted.

respectively. These isoforms span a pH range of 4.5–8.0. The mAb 3F4 identified two additional species at 36–35 and 35–33 kDa (Fig. 9B). These species were detectable in multiple spots as well. Three additional species were revealed by both mAbs 8H4 and 8F9, each resulting in about six or seven spots (Fig. 9C and D). In total, at least seven PrP^C isoforms were distinguishable by 2D-PAGE followed by immunoblotting with a series of mAbs, based on differences in their molecular weight. In addition, each of these species generated between 3 and 14 individual spots when separated by IEF. This additional dimension of separation is the primary reason that these isoforms are not detectable by 1D-PAGE, because this method does not separate proteins on the basis of their charge. Only the separation of these proteins on the basis of their p*I* allowed them to be revealed by 2D-PAGE. This example shows that more knowledge can be gained through the high resolution of 2D-PAGE and provides striking

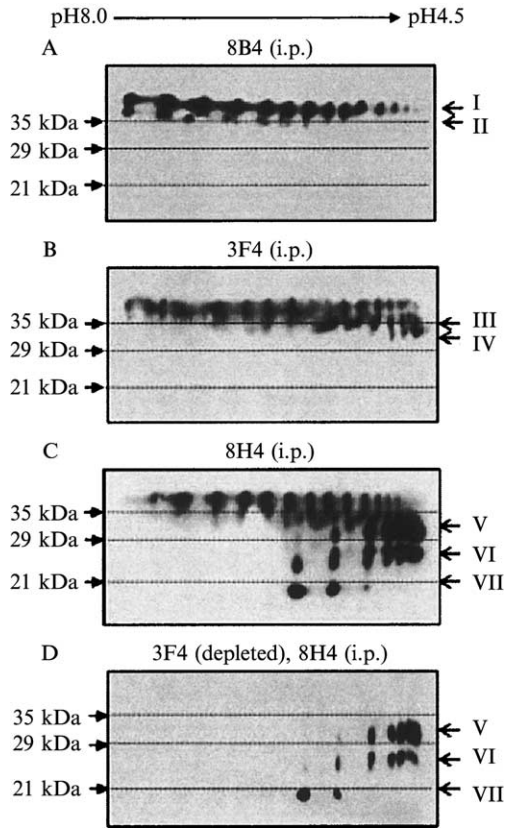


FIG. 9. Two-dimensional immunoblots of PrP^C immunoprecipitated (i.p.) with the indicated mAbs. Equal amounts of immunopurified PrP^C protein were loaded onto the first dimension of the 2D-PAGE gel. After the second-dimensional electrophoresis and electrotransferring, the PrP^C was probed with mAbs as in Fig. 8. (A–D) Seven forms of PrP^C with distinct molecular weights were observed; however, each form separates in a number of spots with different isoelectric points. Comparison with Fig. 8 reveals PrP^C to contain many more isoforms than can be resolved by simple 1D SDS–PAGE.

evidence that the processing of PrP^C is more complex than presently thought.

C. Proteome Analysis of Esophageal Cancer by DIGE

As one of the earliest applications, DIGE was used to quantify differences in protein expression between esophageal carcinoma cells and normal epithelial cells and to define cancer-specific and normal-specific protein

markers [73]. Esophageal cancer cells and normal squamous epithelium cells were procured from the same esophageal tumor sample by laser capture microdissection (LCM) [74]. Each sample, containing about 250,000 cells, was lysed and precipitated with trichloroacetic acid–acetone to remove cellular components such as lipids, nucleotides, and salts that are detrimental to 2D-PAGE resolution. The esophageal cancer cell and normal epithelial cell lysates were subsequently labeled with Cy5 dye and Cy3 dye, respectively. The labeled proteins were pooled and subjected to IEF in a Multiphor II apparatus, followed by 15% SDS–PAGE in the second dimension.

An analysis of Cy5- and Cy3-labeled gel images detected 1038 protein spots from esophageal cancer cells and 1088 protein spots from normal cells [73] (Fig. 10; see Color Insert). To compare the relative expression of proteins from the two cell types, the pixel volume of each spot was calculated on the basis of spot intensity and area. This measurement was followed by normalization with the total pixel volume of all the spots in the gel image. Of the protein spots detected, 107 were downregulated and 58 were upregulated by at least 3-fold in the esophageal cancer cells. No significant change was observed in 72.5% (916) of the protein spots.

The Cy3- and Cy5-labeled protein spots on the gel were not directly excised for identification by MS. The Cy3 and Cy5 labeling was conducted so that less than 5% of the total of each protein was fluorescently labeled. The Cy3/Cy5 visualization and measurement were used solely for quantitation. The protein was counterstained with SYPRO Ruby and the image of the up- or downregulated proteins was matched to its corresponding Cy3/Cy5 image. The Cy3/Cy5 and SYPRO Ruby images do not overlap because the Cy3/Cy5 labeling adds about 500 Da to the molecular mass of the protein. Each differentially regulated SYPRO Ruby-stained spot was excised from the gel, in-gel digested with trypsin, and analyzed by LC/MS/MS.

The group reported three identified proteins that were upregulated, downregulated, or showed no change in expression in the cancer cells. These proteins were protein tumor rejection antigen gp96 (upregulated), annexin (downregulated), and tubulin (no expression change). To confirm the results of the DIGE quantitation, the expression of each of these proteins from the two cell types was measured by Western blotting. No change was observed in tubulin expression between the two cell types. Annexin staining was observed only in the normal cells, whereas gp96 was observed only in the cancer cell lysates, confirming the results of the DIGE analysis.

D. *Proteome Analysis of Secreted Proteins*

Although 2D-PAGE is used prominently in the characterization of soluble, cytoplasmic proteins, it is limited mainly by the ability to get the proteins to migrate within the gel matrix. 2D-PAGE can be used to separate and visualize proteins from any source provided they can be solubilized and a high enough concentration of protein can be loaded. For example, Bumann *et al.* used 2D-PAGE to identify proteins secreted by *Helicobacter pylori* (*H. pylori*) [75]. The proteins secreted by *H. pylori*, a major cause of gastric and duodenal ulcers and gastric cancer [76], may potentially be involved in pathogen–host interactions [77]. To recover secreted proteins from this organism, *H. pylori* was first cultured on serum agar plates for 3 days. These bacteria were then resuspended and washed in brain heart infusion (BHI) broth. Fifteen milliliters of BHI broth was inoculated with *H. pylori* cells, and this culture was grown to an OD_{600} of 0.5 to 1. These bacteria were recovered by centrifugation and used to inoculate a second liquid culture that was grown to an OD_{600} of 0.3 to 0.5. After centrifugation, the medium from this culture was passed through a 0.45-μm pore size membrane filter to remove residual bacteria. The extracellular, secreted proteins were precipitated by a modified trichloroacetic acid method [78] and separated by 2D-PAGE. A second 2D-PAGE was performed for the entire cell lysate of *H. pylori* to compare with that of the secreted proteins.

Helicobacter pylori supernatants contained only a few protein species, and most species present in the corresponding whole cell lysates were lacking in the supernatants. There were 33 protein species reproducibly detected in the supernatants of three independent cultures. Of these, 26 were identified by using peptide mass fingerprinting and comparing the protein data with the complete genome sequence. Among the identified species, there was a weak spot that was identified as urease B, the most abundant protein in whole *H. pylori* cells. Its low abundance in supernatant, however, is consistent with the finding that this protein is not secreted by rapidly growing *H. pylori* cells [79], and that its presence in the supernatant may have resulted from a small proportion of the cells lysing during sample preparation to separate the cells from the supernatant. Sixteen of the identified species had putative signal peptide sequences for *sec*-dependent transport across the plasma membrane. Ten species lacked obvious signal sequences; however, 3 of these were homologous to flagellum-associated proteins that are transported by the type III secretion apparatus of the flagellum [78]. Four of the secreted proteins are homologous to oxidoreductases, proteins that are involved in the modification of disulfide bonds. In addition, another protein, the

function of which is undetermined, has high homology to protein-disulfide isomerases. Furthermore, γ-glutamyltranspeptidase, the primary substrate of which is glutathione, was also found in the supernatant. Indeed, a decrease in glutathione levels in the gastric mucosa is symptomatic of *H. pylori* infection [80].

VI. The Future

Although many groups ranging from small academic laboratories to large pharmaceutical departments are actively involved in proteomics, the science is still in its infancy. Indeed, gel separation of proteins has been around for decades, and has been the foundation of modern 2D-PAGE for almost 30 years; however, it is only in the more recent past that the proteomic capability of this separation tool has been realized with the development of MS approaches to identify these separated proteins (as described in other articles in this volume). There has been an active research focus on methods to obviate the need for 2D-PAGE; however, it still remains a foundation of proteomics. Will it some day become obsolete? It is difficult to say; however, the ongoing efforts to improve the resolution and sensitivity of 2D-PAGE suggest that it will be with us for many years to come.

Acknowledgments

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. NO1-CO-12400.

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

References

1. Morrow, J. F. (1979). *Methods Enzymol.* **68,** 3–24.
2. Saiki, R. K., Bugawan, T. L., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1986). *Nature* **324,** 163–166.
3. Mundy, C. (2001). *Pharmacogenomics* **2,** 37–49.
4. Wreschner, D. H., and Herzberg, M. (1984). *Nucleic Acids Res.* **12,** 1349–1359.
5. Gilham, P. T. (1970). *Annu. Rev. Biochem.* **39,** 227–250.
6. Lee, P. H., Sawan, S. P., Modrusan, Z., Arnold, L. J., Jr., and Reynolds, M. A. (2002). *Bioconjug. Chem.* **13,** 97–103.
7. Nyman, T. A. (2001). *Biomol. Eng.* **18,** 221–227.
8. Carr, S. A., Hemling, M. E., Bean, M. F., and Roberts, G. D. (1991). *Anal. Chem.* **63,** 2802–2824.

9. Issaq, H. J. (2001). *Electrophoresis* **22,** 3629–3638.
10. Maggio, E. T., and Ramnarayan, K. (2001). *Trends Biotechnol.* **19,** 266–272.
11. O'Farrell, P. H. (1975). *J. Biol. Chem.* **250,** 4007–4021.
12. Tiselius, A. (1930). Inaugural Dissertation. Almqvist & Wiksells, Uppsala, Sweden.
13. Smithies, O., and Poulik, M. D. (1956). *Nature* **177,** 1033.
14. Hermans, P. E., McGuckin, W. F., McKenzie, B. F., and Baird, E. D. (1960). *Proc. Staff Meetings Mayo Clin.* **35,** 792.
15. Ashton, G. C. (1957). *Nature* **180,** 917.
16. Raymond, S., and Weintraub, L. (1959). *Science* **130,** 711.
17. Raymond, S. (1964). *Ann. N. Y. Acad. Sci.* **121,** 350–365.
18. Laurell, C. B. (1965). *Anal. Biochem.* **10,** 358–361.
19. Freeman, T., and Smith, J. (1970). *Biochem. J.* **118,** 869–873.
20. Dale, G., and Latner, A. L. (1969). *Clin. Chim. Acta* **24,** 61–68.
21. Macko, V., and Stegemann, H. (1969). *Hoppe. Seylers Z. Physiol. Chem.* **350,** 917–919.
22. Martini, O. H. W., and Gould, H. J. (1971). *J. Mol. Biol.* **62,** 403–405.
23. Barrett, T., and Gould, H. J. (1973). *Biochim. Biophys. Acta* **294,** 165–170.
24. Klose, J. (1975). *Humangenetik* **26,** 231–243.
25. Scheele, G. A. (1975). *J. Biol. Chem.* **250,** 5375–5385.
26. Anderson, L., and Anderson, N. G. (1977). *Proc. Natl. Acad. Sci. USA* **74,** 5421–5425.
27. Bjellqvist, B., Ek, K., Righetti, P. G., Gianazza, E., Görg, A., Westermeier, R., and Postel, W. (1982). *J. Biochem. Biophys. Methods* **6,** 317–339.
28. Görg, A., Postel, W., and Günther, S. (1988). *Electrophoresis* **9,** 531–546.
29. Bjellqvist, B., Pasquali, C., Ravier, F., Sanchez, J. C., and Hochstrasser, D. (1993). *Electrophoresis* **14,** 1357–1365.
30. Bjellqvist, B., Sanchez, J. C., Pasquali, C., Ravier, F., Paquet, N., Frutiger, S., Hughes, G. J., and Hochstrasser, D. (1993). *Electrophoresis* **14,** 1375–1378.
31. Rabilloud, T., Valette, C., and Lawrence, J. J. (1994). *Electrophoresis* **15,** 1552–1558.
32. Klose, J., and Kobalz, U. (1995). *Electrophoresis* **16,** 1034–1059.
33. Görg, A., Obermaier, C., Boguth, G., and Weiss, W. (1999). *Electrophoresis* **20,** 712–717.
34. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000). *Proc. Natl. Acad. Sci. USA* **9,** 9390–9395.
35. Hoving, S., Voshol, H., and Oostrum, J. V. (2000). *Electrophoresis* **21,** 2617–2621.
36. Wildgruber, R., Harder, A., Obermaier, C., Boguth, G., Weiss, W., Fey, S. J., Larsen, P. M., and Görg, A. (2000). *Electrophoresis* **21,** 2610–2616.
37. Stancato, L. F., and Petricoin, E. F., III. (2001). *Electrophoresis* **22,** 2120–2124.
38. Butt, A., Davison, M. D., Smith, G. J., Young, J. A., Gaskell, S. J., Oliver, S. G., and Beynon, R. J. (2001). *Proteomics* **1,** 42–53.
39. Gustafsson, E., Thoren, K., Larsson, T., Davidsson, P., Karlsson, K. A., and Nilsson, C. L. (2001). *Rapid Commun. Mass Spectrom.* **15,** 428–432.
40. Bruneau, J.-M., Magnin, T., Tagat, E., Legrand, R., Bernard, M., Diaquin, M., Fudali, C., and Latgé, J.-P. (2001). *Electrophoresis* **22,** 2812–2823.
41. Fountoulakis, M., Berndt, P., Langen, H., and Suter, L. (2002). *Electrophoresis* **23,** 311–328.
42. Lollo, B. A., Harvey, S., Liao, J., Stevens, A. C., Wagenknecht, R., Sayen, R., Whaley, J., and Sajjadi, F. G. (1999). *Electrophoresis* **20,** 854–859.
43. Santoni, V., Molloy, M. P., and Rabilloud, T. (2000). *Electrophoresis* **21,** 1054–1070.
44. Taylor, R. S., Wu, C. C., Hays, L. G., Eng, J. K., Yates, J. R., III., and Howell, K. E. (2000). *Electrophoresis* **21,** 3441–3459.
45. Fountoulakis, M., and Takács, B. (2001). *Electrophoresis* **22,** 1593–1602.

46. Fazekas de St. Groth, S., Webster, R., and Datyner, A. (1963). *Biochim. Biophys. Acta* **71,** 377.
47. Yu, L.-R., Zeng, R., Shao, X.-X., Wang, N., Xu, Y.-H., and Xia, Q.-C. (2000). *Electrophoresis* **21,** 3058–3068.
48. Neuhoff, V., Stamm, R., and Eibl, H. (1985). *Electrophoresis* **6,** 427–448.
49. Quadroni, M., and James, P. (1999). *Electrophoresis* **20,** 664–677.
50. Sinha, P., Poland, J., Schnolzer, M., and Rabilloud, T. (2001). *Proteomics* **7,** 835–840.
51. Yan, J. X., Wait, R., Berkelman, T., Harry, R. A., Westbrook, J. A., Wheeler, C. H., and Dunn, M. J. (2000). *Electrophoresis* **21,** 3666–3672.
52. Gharahdaghi, F., Weinberg, C. R., Meagher, D. A., Imai, B. S., and Mische, S. M. (1999). *Electrophoresis* **20,** 601–605.
53. Steinberg, T. H., Lauber, W. M., Berggren, K., Kemper, C., Yue, S., and Patton, W. F. (2000). *Electrophoresis* **21,** 497–508.
54. Steinberg, T. H., White, H. M., and Singer, V. L. (1997). *Anal. Biochem.* **248,** 168–172.
55. Steinberg, T. H., Jones, L. J., Haugland, R. P., and Singer, V. L. (1996). *Anal. Biochem.* **239,** 223–237.
56. Berggren, K., Chernokalskaya, E., Steinberg, T. H., Kemper, C., Lopez, M. F., Diwu, Z., Haugland, R. P., and Patton, W. F. (2000). *Electrophoresis* **21,** 2509–2521.
57. Unlu, M., Morgan, M. E., and Minden, J. S. (1997). *Electrophoresis* **18,** 2071–2077.
58. Tonge, R., Shaw, J., Middleton, B., Rowlinson, R., Rayner, S., Young, J., Pognan, F., Hawkins, E., Currie, I., and Davison, M. (2001). *Proteomics* **1,** 377–396.
59. Gharbi, S., Gaffney, P., Yang, A., Zvelebil, M. J., Cramer, R., Waterfield, M. D., and Timms, J. F. (2002). *Mol. Cell. Proteomics* **1,** 91–98.
60. de Groof, A. J., Oerlemans, F. T., Jost, C. R., and Wieringa, B. (2001). *Muscle Nerve* **24,** 1188–1196.
61. Kernec, F., Unlu, M., Labeikovsky, W., Minden, J. S., and Koretsky, A. P. (2001). *Physiol. Genomics* **6,** 117–128.
62. Kaufmann, H., Bailey, J. E., and Fussenegger, M. (2001). *Proteomics* **1,** 194–199.
63. Dunn, M. J. (1999). *Methods Mol. Biol.* **112,** 319–329.
64. Thongboonkerd, V., Gozal, E., Sachleben, L. R., Jr., Arthur, J. M., Pierce, W. M., Cai, J., Chao, J., Bader, M., Pesquero, J. B., Gozal, D., and Klein, J. B. (2002). *J. Biol. Chem.* **277,** 34708–34716.
65. Redline, S., and Young, T. (1993). *Ear Nose Throat J.* **72,** 20–26.
66. Partinen, M. (1995). *Curr. Opin. Pulm. Med.* **1,** 482–487.
67. Chen, L. M., Chao, L., and Chao, J. (1997). *Hum. Gene Ther.* **8,** 341–347.
68. Bader, M. (2001). *J. Cardiovasc. Pharmacol.* **38**(Suppl. 2), S7–S9.
69. Pan, T., Li, R., Wong, B.-S., Liu, T., Gambetti, P., and Sy, M.-S. (2002). *J. Neurochem.* **81,** 1092–1101.
70. Oesch, B., Westaway, D., Walchli, M., McKinley, M. P., Kent, S. B., Aebersold, R., Barry, R. A., Tempst, P., Teplow, D. B., and Hood, L. E. (1985). *Cell* **40,** 735–746.
71. Prusiner, S. B. (1999). *In* ''Prion Biology and Diseases'' (S. B. Prusiner, Ed.), p. 67. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
72. Kascsak, R. J., Rubenstein, R., Merz, P. A., Tonna-DeMasi, M., Fersko, R., Carp, R. I., Wisniewski, H. M., and Diringer, H. (1987). *J. Virol.* **61,** 3688–3693.
73. Zhou, G., Li, H., DeCamp, D., Chen, S., Shu, H., Gong, Y., Flaig, M., Gillespie, J. W., Hu, N., Taylor, P. R., Emmert-Buck, M. R., Liotta, L. A., Petricoin, E. F., III., and Zhao, Y. (2002). *Mol. Cell. Proteomics* **1,** 117–123.
74. Emmert-Buck, M. R., Bonner, R. F., Smith, P. D., Chuaqui, R. F., Zhuang, Z., Goldstein, S. R., Weiss, R. A., and Liotta, L. A. (1996). *Science* **274,** 998–1001.

75. Bumann, D., Aksu, S., Wendland, M., Janek, K., Zimny-Arndt, U., Sabarth, N., Meyer, T. F., and Jungblut, P. R. (2002). *Infect. Immun.* **70,** 3396–3403.
76. Telford, J. L., Covacci, A., Rappuoli, R., and Chiara, P. (1997). *Curr. Opin. Immunol.* **9,** 498–503.
77. Go, M. F., and Crowe, S. E. (2000). *Gastroenterol. Clin. North Am.* **29,** 649–670.
78. Komoriya, K., Shibano, N., Higano, T., Azuma, N., Yamaguchi, S., and Aizawa, S. I. (1999). *Mol. Microbiol.* **34,** 767–779.
79. Marcus, E. A., and Scott, D. R. (2001). *Helicobacter* **6,** 93–99.
80. McGovern, K. J., Blanchard, T. G., Gutierrez, J. A., Czinn, S. J., Krakowka, S., and Youngman, P. (2001). *Infect. Immun.* **69,** 4168–4173.

HIGH-PERFORMANCE SEPARATIONS AND MASS SPECTROMETRIC METHODS FOR HIGH-THROUGHPUT PROTEOMICS USING ACCURATE MASS TAGS

By RICHARD D. SMITH,* GORDON A. ANDERSON, MARY S. LIPTON, CHRISTOPHE MASSELON, LJILJANA PASA-TOLIC, HAROLD UDSETH, MIKHAIL BELOV, YUFENG SHEN, AND TIMOTHY D. VEENSTRA

Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington 99352

I. Introduction

The ability to study how the components of a biological cell or organism change and interact after a perturbation provides a foundation to understand the function(s) of its component parts, and ultimately how the system operates. Reaching this goal will require instrumental and computational methods that identify systems-level responses that recognize genes or gene products that are sensitive to change when the environment of the cell or organism is altered. Although methods to simultaneously assess the abundances of thousands of expressed genes at

*Present address: Biomedical Proteomics Program/Analytical Chemistry Laboratory, SIAC-Frederick, Inc., National Cancer Institute at Frederick, P.O. Box B, Frederick, Maryland 21702-1201.

ADVANCES IN PROTEIN CHEMISTRY, Vol. 65
0065-3233/03 $35.00

the mRNA level are now broadly applied [1, 2] with more or less success, posttranscriptional processes play a major role in determining protein abundances and modification states, and protein abundances can show poor correlation with mRNA levels [3–5]. Thus, considerable attention is now focused on the proteome, the complement of proteins expressed by a particular cell, organism, or tissue at a given time or under a specific set of environmental conditions.

The currently existing proteome analysis capability is predominantly based on protein separations using two-dimensional polyacrylamide gel electrophoresis (2D-PAGE). Although 2D-PAGE is capable of resolving thousands of proteins, proteome coverage is problematic for proteins that have high or low isoelectric points (less than ~3.5 and greater than ~9.5) and/or extremes of molecular weight, and for membrane proteins, which typically account for more than half of all the proteins expressed within a cell. It has been shown that the number of spots is poorly correlated with the number of different proteins detected, because a single gene can give rise to multiple spots [5] due to co- and posttranslational modifications, degradation intermediates, and alternative expression (e.g., alternative splicing of mRNAs, translational frame shifts). The sensitivity of 2D-PAGE is generally limited to femtomole levels [8, 9] by the need to visualize the protein spot on the gel and its subsequent processing and analysis primarily by mass spectrometry (MS) [6–8]. The largest study reported to date identified 502 proteins from *Haemophilus influenzae* [10]. Similarly, the most comprehensive yeast proteome 2D-PAGE/MS studies published to date (the broadest of which identified 279 proteins [11], and a combined total of only ~500 [7, 11–14]) provide a skewed codon bias distribution [15], indicating that only more abundant proteins were detected. Many important regulatory proteins are expressed at such low levels (e.g., <1000 copies per cell) that their detection is precluded unless 2D-PAGE is preceded by extensive fractionation of large quantities of protein and/or the processing of a large numbers of gels. Finally, the precision of protein abundance determinations by 2D-PAGE is based on comparison of protein spot intensities, limiting the capability for discerning subtle differences in protein abundances for large numbers of proteome-wide measurements.

Many of the possible alternative proteomics technologies presently being considered employ an alternative separation methodology combined with some form of MS, most typically applied after protein digestion with specific proteases (e.g., trypsin). Analysis of the peptides of size sufficiently large for protein identification, typically more than ~5- to 10-mer size based on sequence uniqueness for a specific organism, is now effectively achieved by MS [16–19]. A widely used approach involves MS

selection of a polypeptide that is dissociated to form fragments whose mass-to-charge (m/z) ratios are measured. This MS/MS analysis provides primary sequence-related information that allows the peptide (and most often its parent protein) to be identified if it is contained in an appropriate database [20]. MS/MS analysis of only one polypeptide is often sufficient for protein identification [20–24]. Washburn *et al.* demonstrated the use of this approach to identify 1484 yeast proteins by a 2D capillary liquid chromatography (LC)-MS/MS strategy, in which peptides were separated by cation-exchange LC in the first dimension into 15 fractions that were subsequently separated by reversed-phase LC [25]. This work demonstrated the potential for broad proteome coverage of highly complex polypeptide mixtures, but still leaves much to be desired in terms of speed, sensitivity, dynamic range, comprehensiveness, and the quantitative utility of the measurement method.

Here we review the technological basis and progress toward a global proteomics strategy that aims to provide large improvements in sensitivity, dynamic range, comprehensiveness, and throughput based on the use of polypeptide ''accurate mass tags'' (AMTs). The two-stage strategy exploits a single high-resolution capillary LC separation combined with Fourier transform ion cyclotron resonance mass spectrometry (FTICR) to validate polypeptide AMTs for a specific organism, tissue, or cell type. AMTs represent peptide biomarkers used to confidently identify a unique protein solely on the basis of the high mass measurement accuracy provided by FTICR. The generation of these biomarkers provides the basis for second-stage high-throughput studies using only AMTs to identify and quantify the proteins expressed within a cell system. Key attractions of the approach include the feasibility of completely automated high-confidence protein identification, extensive proteome coverage, and the capability for exploiting stable isotope-labeling methods for high-precision abundance measurements. Additional developments, including the use of multiplexed MS/MS capabilities and methods for dynamic range expansion of proteome measurements, are also described that promise to further extend the quality of measurements and their extension to much more challenging mammalian proteomes.

II. Proteome Measurement Technology and Applications

The aim of our strategy for proteome analysis is to exploit a combination of instrumental and methodological approaches to provide broad proteome coverage, high sensitivity, and the capability for greatly increased throughput compared with conventional technologies. After

initial cell lysis the recovered proteins are enzymatically digested into polypeptide fragments (e.g., using trypsin) to produce tens to hundreds of potentially detectable peptides (and modified peptides) from each protein, and perhaps 10^5 to $>10^6$ in total (depending on proteome complexity, the dynamic range of the measurements, etc.). This complex peptide mixture is then analyzed by combined high-resolution capillary LC-FTICR. Variations on sample preparation that we have explored, among the many possible, include the use of cysteine (Cys) labeling that incorporates a biotin affinity tag for isolation of the modified cysteine-containing peptides by immobilized avidin column chromatography [26–39], and the incorporation of stable isotope labeling, either by culturing in ^{15}N-labeled medium or in conjunction with the cysteine peptide labeling [26]. The capillary LC-FTICR analysis can be preceded by additional sample fractionation to both simplify the analysis and potentially provide additional information about peptide composition (e.g., surface charge in the use of ion-exchange chromatography), thus allowing more complex proteomes to be studied in greater detail. Our initial aim has been to optimize proteome coverage for the use of a single separation step so as to increase overall throughput, because every additional separation stage irrevocably leads to selective losses of some peptide and increases overall sample requirements. It is anticipated, however, that the use of 2D separations will more effectively address both the greater complexity and the dynamic range desired for measurements of mammalian proteomes.

A. Capillary LC-FTICR Measurements of Complex Global Protein Digests

The extent of proteome coverage depends substantially on the achievable dynamic range of the MS measurements, which in turn depends significantly on the resolution (or peak capacity) of the separation step(s) preceding MS analyses. The number of potentially detectable peptides from LC-FTICR measurements is the product of the capillary LC peak capacity (which can exceed 1000 in a one-dimensional separation [27]) and the maximum number of peptides detectable per spectrum. In practice the obtainable peak capacity (in terms of the number of measurable peaks) for a single FTICR mass spectrum is almost always limited by FTICR trap charge (ion) capacity rather than by the MS resolution and can potentially be as high as $\sim 10^5$ based on the charge capacity of the FTICR trap ($\sim 10^7$ charges for our 11.4-T FTICR) and the minimum number of charges that give rise to a detectable signal [$\sim$30 with a signal-to-noise ratio (S/N) > 3] when the discrete isotopic peaks

present for each species are considered. We have measured the masses of >2000 peptides in a single mass spectrum during LC-FTICR analyses. The average number of peptides that can potentially be detected at any point in a separation can be further increased by improvements in experimental dynamic range, as described below. Thus, the number of theoretically detectable species of the combined LC-FTICR approach exceeds 10^7. It is worth noting, however, that the number of distinguishable species in each spectrum is a function of resolution, mass measurement accuracy (MMA), and the MS ''spectral space,'' and exceeds 10^6 because of the high resolution obtainable by FTICR. Although difficult to quantify at the present time, it is obvious that the greater the number of distinguishable species exceeding the number of species actually detected, the better will be the level of confidence in the identifications that result. The number of distinguishable species in the combined LC-FTICR analysis is presently $\sim 10^7$, and could potentially exceed 10^8 if precise elution time information were effectively used.

To evaluate the complexity of peptide mixtures that can be addressed we analyzed a tryptic digest of the soluble proteins extracted from yeast grown to midlog phase. A 10-μg sample was separated by a gradient reversed-phase LC separation in an 85-cm-long capillary packed with 3-μm C_{18}-bonded particles. The on-line electrospray ionization (ESI)-FTICR analysis consisted of ~1200 high-resolution mass spectra. The capillary LC-FTICR total ion current (TIC) chromatogram reconstructed from the ESI-FTICR mass spectra is shown in Fig. 1. A high-efficiency separation with symmetric peaks was obtained throughout the elution under the conditions optimized for ESI-MS, and where many minor (low-abundance) species were resolved from their neighboring major (high-abundance) components. This high-efficiency separation was obtained by a simple connection of a replacable electrospray tip to the column outlet through a narrow-bore (150 μm $\times$ 1 mm channel) union, allowing convenient replacement of ESI emitters.

Examples illustrating the qualities of the separation and the MS are shown in Fig. 2. A typical single spectrum, with insets showing exploded views of several regions of the spectrum, is shown in Fig. 2A. A narrow-range (m/z 972.515 to 972.535) reconstructed ion chromatogram, where a number of both high- and low-abundance peaks eluted, with excellent peak shape, in this small m/z window during the separation, is shown in Fig. 2B. The power of the approach, however, is that this quality of information is obtained over a wide m/z range. More abundant peptides were typically observed to elute over three to five spectra (one scan takes 5.7 s), whereas minor components were observed to elute over only one or two spectra. On the basis of an average chromatographic base peak width of 25 s, the chromatographic peak capacity (for a resolution of unity)

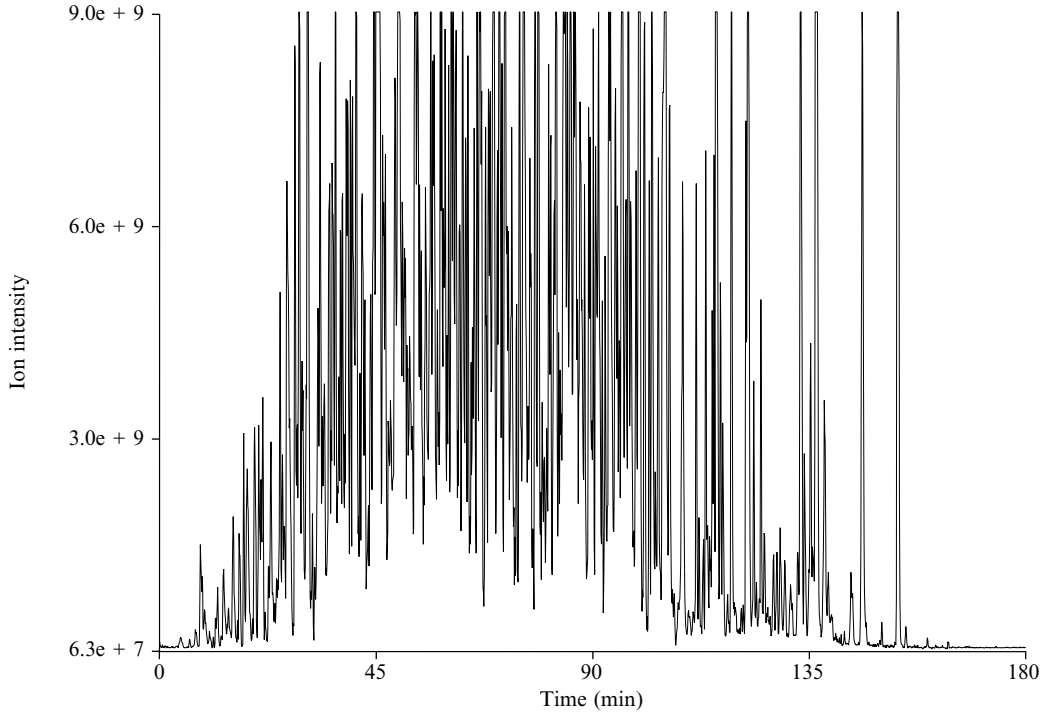


FIG. 1. Total ion current (TIC) chromatogram of capillary LC-FTICR of a global soluble yeast tryptic digest, using the multiple-capillary LC system. The separation used a pressure of 10,000 psi and a mobile phase gradient from solvent A [H_2O, 0.2% acetic acid (HOAc), 0.1% trifluoroacetic acid (TFA), v/v] to 75% solvent B [H_2O–acetonitrile (10:90), 0.2% HOAc, 0.1% TFA, v/v] over 180 min. The vertical axis is proportional to the ion signal.

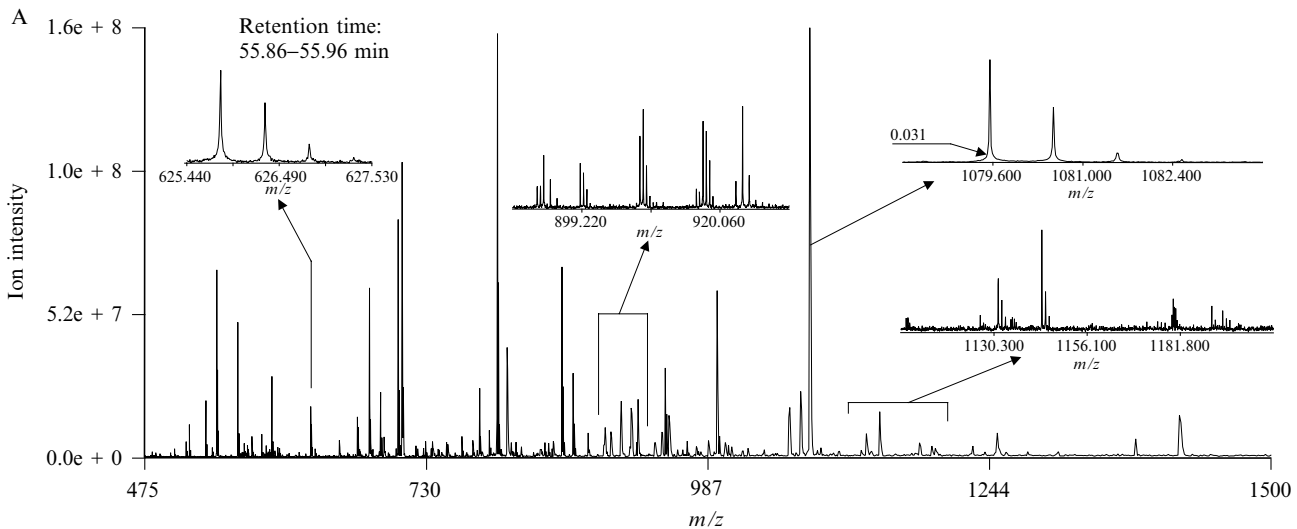


FIG. 2. (*Continued*).

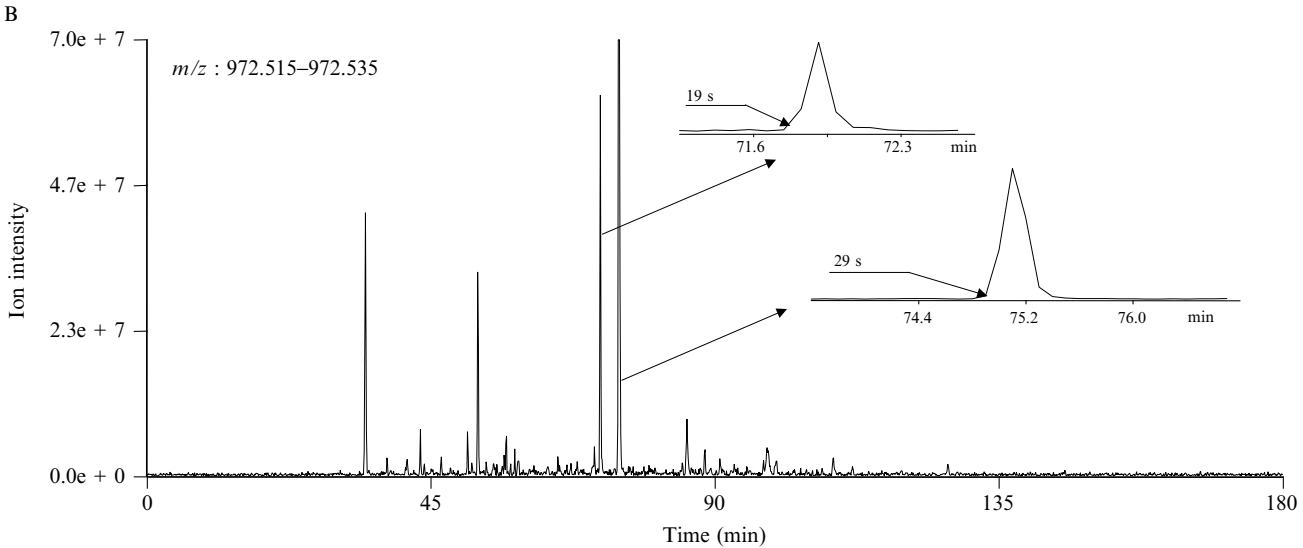


FIG. 2. Examples demonstrating the high resolution of capillary LC and 11.4-T ESI-FTICR. Conditions are as given in Fig. 1.

corresponds to $(180 \times 60/25) = 432$. Using a shorter spectrum acquisition time of 2.5 s, a chromatographic peak capacity of ~1000 has been achieved [27]. The typical data quality for a portion of the reconstructed LC chromatogram for a 3 *m/z*-unit range and the mass spectra for several peaks eluting at different time are shown in Fig. 3.

It must be noted that although FTICR can supply a resolving power of $\sim 10^5$ species, this power by itself is not sufficient to resolve the extremely complex mixtures of cellular peptides. For example, the >6000 proteins annotated within the yeast genome can potentially yield >350,000 different tryptic peptides, and ~195,000 of these have masses between 500 and 4000. Even if only ~20% of the >6000 proteins were expressed at a given time, an ideal tryptic digestion would yield ~40,000 different peptides, and a much larger number if modified and incompletely digested peptides were also considered. Many peptides will yield multiple charge states (e.g., 2+, 3+) after ESI, with each charge state comprising multiple isotopic peaks. Complexity of this type is illustrated in Fig. 2B, which shows more than 20 peaks evident in a narrow *m/z* range (0.02 Da) and having apparent LC retention factor differences of as little as 0.006. Such complexity can result in the need for even greater FTICR resolving power and/or higher efficiency separations before FTICR. Although FTICR resolution can be increased, typically at the cost of an increase in the spectrum acquisition time or magnetic field strength, the ion trap capacity imposes the greatest limits on overall dynamic range. In practice, the need to address greater levels of complexity will strongly depend on further improvements in the achievable dynamic range of proteome measurements (such as the DREAMS technology, discussed below), and will likely also involve the use of additional separation stages.

B. The Dynamic Range of Capillary LC-FTICR Analyses

As shown by the example in Fig. 4, the dynamic range obtainable in a single FTICR mass spectrum exceeds 10^3. The most highly abundant peptide eluted over 13 spectra, whereas low-abundance peptides are often detected within a single spectrum. Therefore, the effective dynamic range for detection of peptides just on this basis can be expected to approach $\sim 10^4$ if they have the same ionization or detection efficiency. Furthermore, if the aim is protein identification, then a significant (perhaps 10-fold) increase in effective dynamic range will result because of the variable ESI or detection efficiency for different polypeptide sequences. This variation in overall detection efficiency is evident in the analysis of tryptic digests of single proteins whose peptide fragments, for many possible reasons, differ

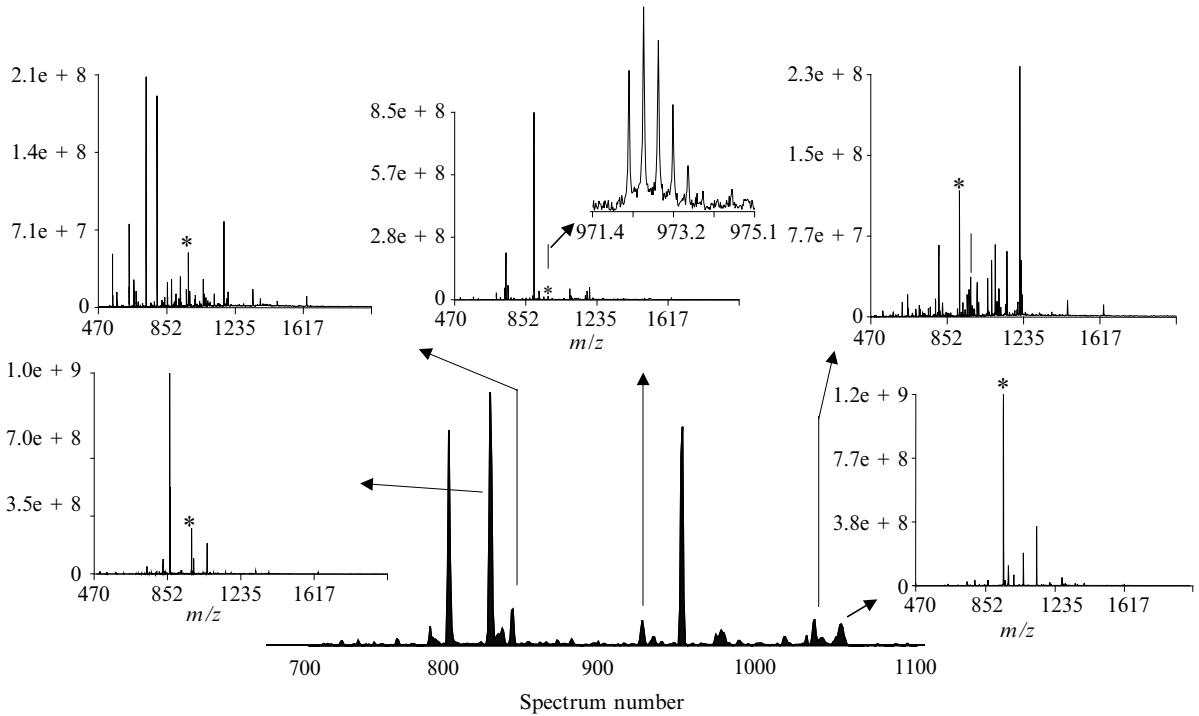


FIG. 3. Narrow-range (m/z 970.0–973.9) reconstructed partial chromatogram (*bottom center*) showing several polypeptide peaks and representative mass spectra for the indicated elution times illustrating the dynamic range and typical data quality obtained for a capillary LC-FTICR analysis of tryptic peptides from a digestion of soluble yeast proteins. The LC peak from the narrow m/z range is indicated by an asterisk in each spectrum. Note that the relatively small peak at spectrum number ~940 (*top middle*) provides high-resolution MS results for the lower level component (*inset*).

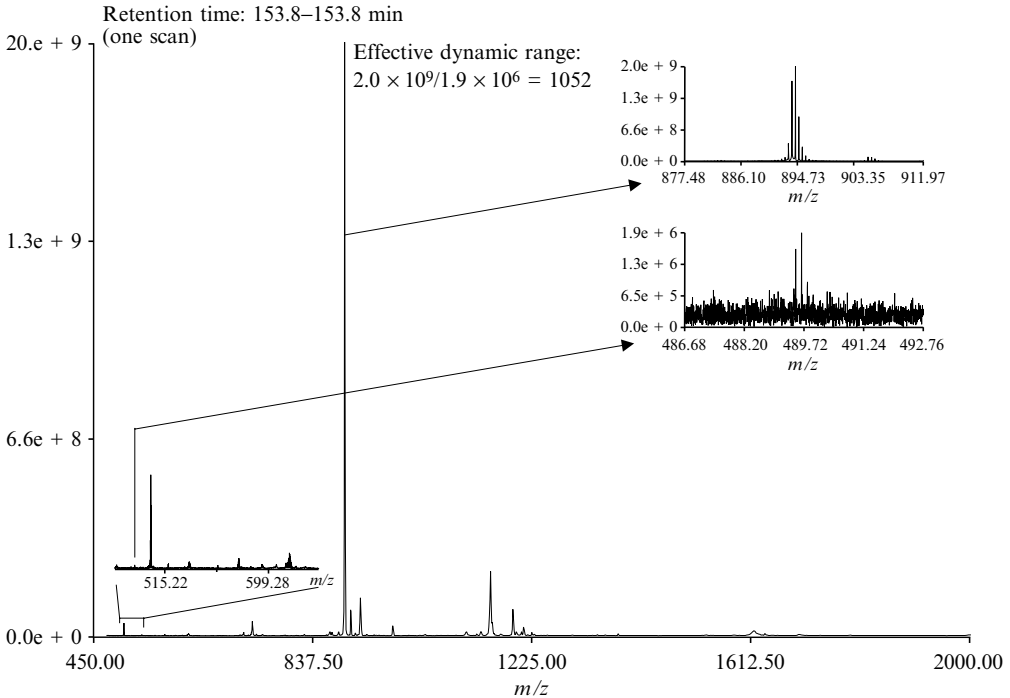


FIG. 4. Illustration of the effective dynamic range obtainable in a single FTICR mass spectrum from a capillary LC-FTICR analysis of a global yeast soluble protein tryptic digest obtained under capillary LC data acquisition conditions.

greatly in their signal intensities compared with their nominally expected equimolar abundances. More important, however, is the ion accumulation process used with the FTICR trap. Reduced space charge results in more efficient ion introduction from the ESI source into the ''external'' ion accumulation trap region. Although we have not yet quantified all these effects, it is clear that the best achievable sensitivity is greatly improved when low-abundance peaks are chromatographically separated from high-abundance peaks. This contribution can account for at least an order of magnitude increase in dynamic range, and can potentially be much greater if ion bias effects that result from ''overfilling'' of the ion accumulation region can be mitigated [40, 41] or avoided by the automated control of accumulation times. Considering these factors, we estimate that the dynamic range currently achieved is approximately 10^4 to 10^5 [42], and believe that an order of magnitude further gain is achievable on the basis of automated selection of ion accumulation times. Of course, many issues (e.g., contaminants, background signals from low levels of ion dissociation) can prevent this from being realized.

We have directed considerable effort to achieving an extended dynamic range for proteome measurements. As can be seen from the above discussion, the dynamic range of a single FTICR mass spectrum is limited by the charge capacities of both the ''external'' ion accumulation region and the FTICR analyzer trap. As noted earlier, the useful charge capacity of the external accumulation quadrupole trap is on the order of 10^7 charges if undesirable effects due to overfilling are to be avoided [43]. These combined effects include bias due to charge stratification in the accumulation quadrupole and ''coalescence'' of closely spaced m/z ion packets in the FTICR trap. Improvements in the ESI source design and use of an electrodynamic ion funnel now allow currents of several nanoamperes of useful ions to be transmitted to the ion accumulation region. This corresponds to $>10^{10}$ charges/s, a factor of $\sim 10^3$ in excess of the ion population than can currently be analyzed in a single spectrum even if only a 20% transfer efficiency is assumed with the present 11.4-T FTICR instrument. Significant efforts have thus been directed toward establishing a routinely useful ''active'' dynamic range enhancement capability in which the information from a proceeding spectrum is used to remove the high-abundance species in a radio frequency (RF)-only quadrupole just before the ion accumulation quadrupole region. In this fashion every other spectrum would ''dig deeper'' into the proteome and provide more information about lower abundance species, and the overall dynamic range should significantly exceed 10^6. However, its practical utility remains to be fully demonstrated. We discuss the initial demonstration of this capability later in this review.

C. *Protein Identification Using AMTs*

The power of MS for protein identification derives from the specificity of mass measurements for either the intact peptides or their fragments after dissociation in MS/MS measurements, and is implicitly based on the relatively small number of possible polypeptide sequences for a specific organism compared with the total number of possible sequences (see Table I). MS/MS measurements typically provide partial sequence information, and in relatively rare cases complete sequence information, and make a large fraction of the $\sim 10^{26}$ possible 20-mer polypeptide sequences distinguishable and potentially identifiable. Thus, highly confident polypeptide identifications using MS/MS methods can often be achieved from only limited sequence data because of the enormously smaller numbers of peptides predicted for an organism; for example, only 3463 different 20-mer peptides are predicted from an ''ideal'' tryptic digestion of all yeast proteins. The distinctiveness of polypeptide sequences increases with size, but in practice the utility of increased size for identification is mitigated by the increased likelihood that a peptide will be unpredictably modified. Indeed, we have found that exact molecular weight measurements alone have limited utility for the identification of intact proteins [33].

TABLE I
Number of Possible Peptides and Number Predicted for Three Organisms after Digestion with Trypsin

	Possible		Predicted number of peptides[a]		
Length	Sequences[b]	Masses[c]	*Deinococcus radiodurans*	*Saccharomyces cerevisiae*	*Caenorhabditis elegans*
10-mers	10^{13}	2×10^{7}	3471	11,275	30,623
20-mers	10^{26}	7×10^{10}	1292	3463	9475
30-mers	10^{39}	2×10^{13}	494	1278	3602
40-mers	10^{52}	1×10^{15}	195	405	1295

[a]Predicted from the identified open reading frames and applying the cleavage specificity of trypsin.

[b]Assumes 20 possible distinguishable amino acid residues.

[c]The number of peptides of length r potentially distinguishable by mass based on the number of possible combinations of n different amino acids: $(n + r - 1)!/r!\ (n - 1)!$. The actual number of possible masses is somewhat smaller due to some mass degeneracy. The number of distinguishable peptides in actual measurements depends on the MS resolution.

Although much smaller than the number of possible sequences, the number of potentially distinguishable peptide masses, given sufficient resolution and accuracy, also dwarfs the number of predicted peptides from any organism. For example, the number of potentially mass-distinguishable 30-mer peptides, estimated from the number of possible combinations, is $>10^{13}$, compared with the 494, 1278, and 3602 predicted for tryptic digestion of all predicted proteins for *Deinococcus radiodurans*, *Saccharomyces cere visiae*, and *Caenorhabditis elegans*, respectively. As shown in Table II, an ideal tryptic digestion of all yeast proteins would produce 194,239 peptides having masses between 500 and 4000 Da, the range typically studied by MS. Of these, 34% are unique at 0.5 ppm MMA. (A larger fraction is unique if constrained by additional information resulting from any prior sample fractionation steps or the use of LC elution times.) These distinctive peptide masses would cover 98 and 96.6% of all predicted *S. cerevisiae* and *C. elegans* proteins, respectively.

Thus, given sufficient MMA, a polypeptide mass measurement can often be confidently attributed to a single protein within the constraints provided by a single genome sequence and its predicted proteome (i.e., serve as an accurate mass tag; an AMT). The limited MMA achievable by conventional MS technologies generally requires extensive separations (e.g., using 2D-PAGE) or the use of MS/MS methods for protein identification. The AMT strategy obviates the routine need for MS/MS, and thus reduces sample requirements. Because the masses of many

TABLE II
Predicted Number of Peptides[a] for Ideal Global Tryptic Digestions

	Peptides[a]			Cysteine peptides[a]		
Organism	Number	Unique[b] (%)	ORF coverage[c] (%)	Number	Unique[b] (%)	ORF coverage[c] (%)
Deinococcus radiodurans	60,068	51.4	99.4	4906	87.2	66
Escherichia coli	84,162	48.6	99.1	11,487	83.6	80
Saccharomyces cerevisiae	194,239	33.9	98	27,483	72.7	84
Caenorhabditis elegans	527,863	20.9	96.6	108,848	52.5	92

[a] Peptides or cysteine peptides in mass range of 500 to 4000 Da, assuming ideal trypsin cleavage specificity.
[b] Percent unique to 0.5 ppm (by mass, not using elution time).
[c] Percentage of ORFs (or predicted proteins) covered by unique peptides.

peptides will generally be obtained in each mass spectrum, requiring equivalent or less time than one MS/MS measurement, the increase in throughput is, at the least, equal to the average number of peptides in each spectrum. In practice the increase in either throughput or proteome coverage is even greater because the lower abundance peptides are often not analyzed by conventional MS/MS approaches, or require the need for additional time for extended ion accumulation or spectrum averaging to yield spectra of sufficient quality. Thus, the AMT approach provides increased sensitivity, coverage, and throughput, and facilitates quantitative studies involving many analyses of different perturbations or time points.

The practical utility of high MMA in defining AMTs for global proteomic measurements is obviously dependent on the complexity of the system being studied. As indicated above, if the ''universe'' of all possible peptides were to be considered, essentially nothing could be identified by MS (as conventionally practiced) with high confidence. When the use of AMTs is confined to a specific system, in which a much more constrained set of possible peptide masses exists, the situation becomes much more tractable. If the set of protein sequences could be predicted with confidence, were not modified during or after translation, and were digested by proteolytic enzymes such as trypsin in an ideal fashion to cleave only certain sites and do so with 100% efficiency, the use of AMTs would be relatively straightforward. In the real world, however, the situation is intermediate between these extremes. First, if the organism is fully sequenced, then a set of possible polypeptide masses from a tryptic digestion can be predicted with a facility that is relatively straightforward (but not flawless) for microorganisms, and is substantially more difficult for mammalian systems. This defines, for better or worse, what can be considered the best case scenario. At the opposite extreme, it is possible to consider the range of possible masses that can result from all possible polypeptide cleavage sites, all possible modifications of these peptides, sequence variants (e.g., all possible single amino acid residue substitutions for each predicted peptide), contaminants, and so on. This, situation similarly becomes intractable unless constrained in some fashion.

The uniqueness for peptides within the predicted proteome of *D. radiodurans* as a function of peptide molecular weight and for three different levels of mass measurement accuracy (1, 10, and 100 ppm) is compared in Fig. 5. The calculations used the 3187 proteins predicted from the DNA sequence by White *et al.* [44] and assumes an ''ideal'' tryptic digestion involving protein cleavages only after lysyl and arginyl residues. The number of peptides having molecular masses between 500 and 4000 varies is 60,068. As shown in Fig. 5, an MMA of 1 ppm provides unique mass tags for a substantial fraction of the peptides generated for the ideal

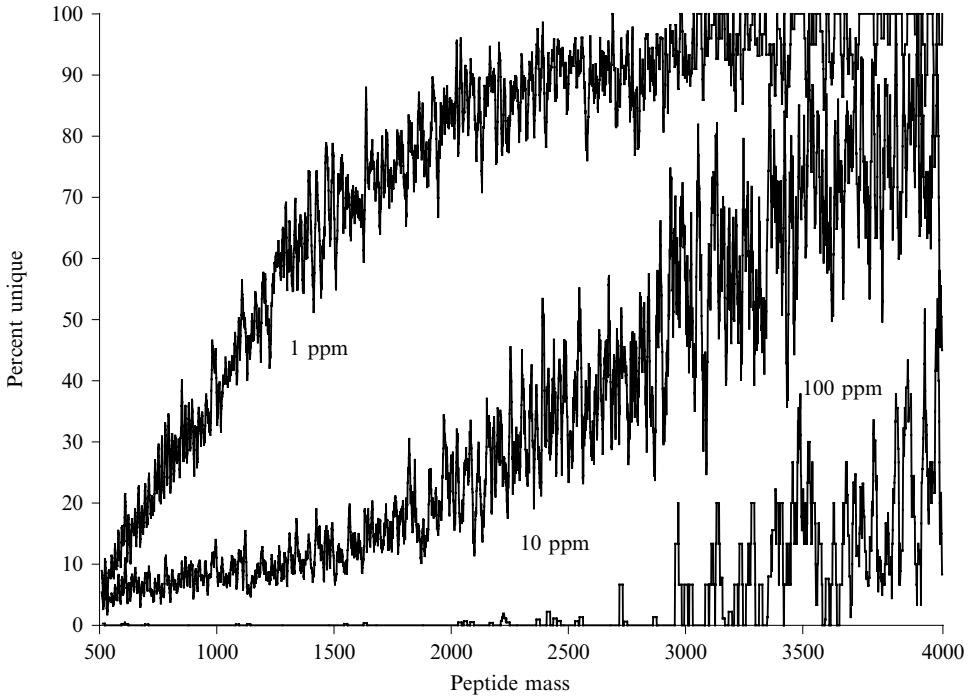
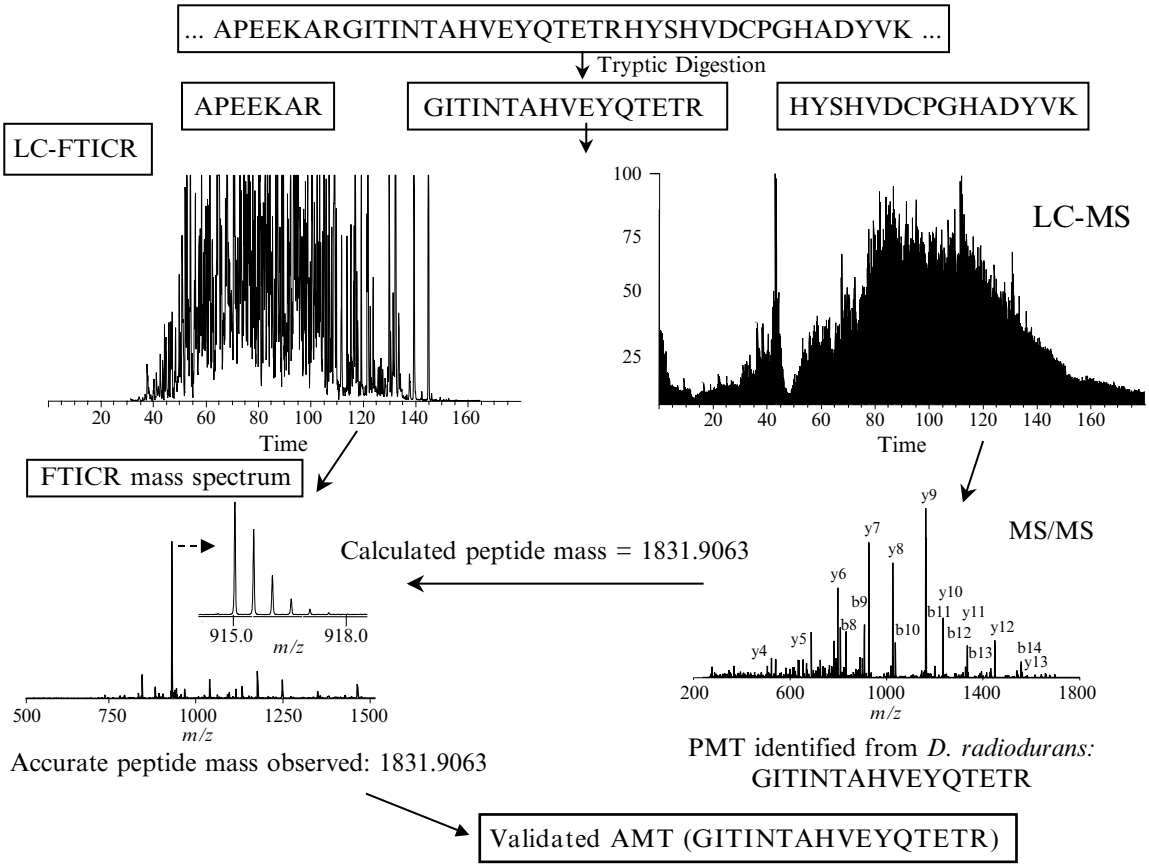


FIG. 5. The number of unique peptides as a function of molecular weight for three different levels of mass measurement accuracy in the case of a whole proteome ideal tryptic digestion of the microorganism *D. radiodurans*. The ''real world'' complexity of proteomes benefits from an approach whereby the use of AMTs is further augmented by the use of elution time data, as well as the initial validation of AMTs by tandem mass spectrometry.

case. Measurements at 10 ppm MMA retains some utility, but an MMA of >100 ppm (a level more typical of conventional mass spectrometers) is of limited value except for the largest peptides. In a more realistic situation, however, ''missed cleavage'' sites and various posttranslation modifications will make the MMA requirements somewhat more demanding than shown in Fig. 5. The unknown extent of complexity beyond that suggested by Fig. 5 needs to be addressed in some fashion. We do so by two additional aspects of our approach: the use of elution time information to increase the specificity of the measurements and the initial use of MS/MS measurements to initially validate the use of accurate mass and elution time data for subsequent studies.

D. The Validation of AMTs from PMTs Generated by Tandem MS

The above-described analysis neglects the use of elution time information for the identification of peptides, and includes consideration of many peptides that will not ever be observed. The development of a capability for the prediction of peptide separation times would be an extremely valuable adjunct to the mass analysis, but a useful capability for this does not yet exist. Thus, our AMT strategy uses tandem MS/MS for the initial screening for AMTs as well as for the (initially limited at this point) identification of modified or otherwise ''unexpected peptides'' [that potentially arise because of sequence errors, frame shifts, open reading frames (ORFs) missed by gene calling software, etc.]. FTICR MS/MS measurements are also used to establish the initial set of highly confident polypeptide ''lock masses'' for FTICR spectrum calibration and for identification of peptides when higher resolution, mass accuracy, or sensitivity is needed. The generation of most AMTs by our approach presently uses a two-stage process (Fig. 6). The proteome sample is digested (e.g., with trypsin) and analyzed by high-efficiency capillary LC-MS/MS, using either a conventional [LCQ ion trap or quadrupole time-of-flight (Q-TOF)] mass spectrometer operating in a data-dependent mode or FTICR (using multiplexed MS/MS, as described later). In the data-dependent mode a single MS scan is followed by three consecutive MS/MS analyses in which three parent ions from a mass spectrum are sequentially selected for MS/MS analysis based on predefined criteria designed to minimize repeated analysis of the same species. While the conventional ion trap MS/MS measurements yield ''potential mass tags'' (PMTs) that are subsequently validated as AMTs if the accurate mass of the predicted peptide is observed by FTICR in a corresponding sample and at an equivalent elution time, the peptides identified by FTICR MS/MS are immediately assigned as AMTs, as a result of the high MMA of FTICR.


... APEEKARGITINTAHVEYQTETRHYSHVDCPGHADYVK ...
Tryptic Digestion
APEEKAR
GITINTAHVEYQTETR
HYSHVDCPGHADYVK
LC-FTICR
Time
LC-MS
Time
FTICR mass spectrum
915.0
918.0
m/z
Calculated peptide mass = 1831.9063
MS/MS
m/z
Accurate peptide mass observed: 1831.9063
PMT identified from D. radiodurans:
GITINTAHVEYQTETR
Validated AMT (GITINTAHVEYQTETR)

Ion trap MS/MS-generated PMTs are presently identified using ''scores'' produced by the SEQUEST search program based on the similarity of the spectrum with a set of peaks predicted on the basis of the known, most common peptide fragmentation processes. Because of the nature of the analysis, the results will invariably span the range from low scores, where identifications are highly doubtful, to high scores, where identifications are quite reliable, with no clear line of demarcation. If only the highest scores for identification are used, fewer proteins will be identified; however, uncritical use of lower scores will result in many false identifications. Conventionally, many MS/MS spectra as well as the program search results need to be manually examined for the less confident identifications to evaluate both spectrum quality and the ranking of peptide scores so as to establish acceptable confidence for identifications. This process generally results in discarding a substantial fraction of the peptides identified with lower scores, and serves to increase the confidence to an extent that is difficult to quantify. In our approach the use of highly accurate mass measurements provides an additional, high-quality ''test'' for tentative peptide identifications that can be applied in the data analysis using software developed at our laboratory. A consequence of the automated validation of AMTs from PMTs is the increased confidence in the peptide identifications that results.

As mentioned above, promoting PMTs to validated AMTs also exploits the distinctive elution times of the peptides. Although run-to-run variations in the gradient capillary LC separations limited the use of elution times to 10 to 20% in our initial work, better calibration procedures and flow control will likely reduce this to less than a few percent and significantly increase the utility of elution times, and the ability to exploit AMTs having

FIG. 6. Experimental steps involved in establishing an accurate mass tag (AMT), illustrated by the identification of an AMT for elongation factor Tu (EF-Tu). Peptides are automatically selected for collision-induced dissociation (CID) and tentatively identified as a potential mass tag (PMT) by an automated search program (SEQUEST). In this example a tryptic peptide from EF-Tu (in boldface) was identified by tandem MS (MS/MS), using an ion trap mass spectrometer. The accurate mass of this PMT was calculated on the basis of its sequence and its elution time was recorded. In the second stage, the same proteome sample is analyzed under the same LC-MS conditions, using a high-field FTICR mass spectrometer. An AMT is established when a peptide eluting at the same time and corresponding to the calculated mass (e.g., within 1 ppm) of the PMT identified in the first stage is observed. This peptide is then considered an AMT for EF-Tu for *D. radiodurans* and functions as a biomarker to identify this particular protein in all subsequent experiments. In the LC-FTICR analysis of the same sample, a doubly charged peptide was observed at this same elution time, having a mass consistent with the calculated mass of this peptide within a specified MMA.

otherwise indistinguishable masses. The use of the same ''lock masses'' that correspond to peptide AMTs would serve as LC elution time calibrants, allowing correction for any differences in mobile phase gradient and flow rate that may occur between separations.

In our initial studies with *D. radiodurans* described below, a peptide is validated as an AMT if its observed mass, as measured by FTICR, agrees with the theoretical calculated mass of the PMT within 1-ppm mass accuracy. In addition, the LC elution times in the experiment wherein the PMT was first identified must agree with that observed in the FTICR experiment to within some predetermined accuracy, which can be $<5\%$. Only when these criteria were met was a peptide designated as an AMT, and subsequently used to confidently identify a specific protein in subsequent proteome studies from the same strain. Without the need to reestablish the identity of a peptide by time-consuming MS/MS analyses, multiple high-throughput studies to measure changes in relative protein abundances between two (or more) different proteomes can be completed in rapid fashion solely on the basis of the highly accurate mass measurements provided by FTICR. Once a protein has been identified with AMTs, its subsequent identification (and quantitation) in other studies is based on FTICR measurements (and its elution time), which provide much greater sensitivity than the conventional MS instrumentation.

An example of the generation of an AMT for the protein elongation factor Tu (EF-Tu) is shown in Fig. 6. A tryptic digest of *D. radiodurans* proteins was analyzed by capillary LC-MS/MS, using a conventional ion-trap mass spectrometer. In this example the peptide selected for dissociation was identified as GITINTAHVEYQTETR from the protein EF-Tu and is considered a PMT. The theoretical mass of this peptide is then calculated on the basis of its amino acid sequence and its LC elution time recorded. Next, the same or similarly derived digested proteome sample is analyzed by LC-FTICR under identical LC separation conditions. In the LC-FTICR analysis, the test involves whether a peptide with an observed mass that closely agrees with the calculated theoretical mass of this peptide (e.g., within 1 ppm) and has a corresponding elution time is actually detected. If a peptide of predicted accurate mass and elution time is detected, it is now considered to confidently function as an AMT for EF-Tu. Although we have focused our initial efforts toward obtaining highly confident protein identifications of unmodified tryptic fragments, the AMT validation approach can be used to identify any class of peptide or modified peptide, provided that the modified peptide can be characterized by MS/MS.

E. Increased Confidence in Protein Identifications Using AMTs

The generation of AMTs for a specific proteome has two key advantages over conventional protein identification practices. First, correlating PMT-calculated theoretical masses to observed accurate masses by high mass accuracy FTICR measurements provides a much higher level of confidence in those peptides identified. Using the search/identification program SEQUEST and a minimum cross-correlation score of 2, a large number of polypeptide PMTs have been identified for *D. radiodurans* (see below); however, only ~70% were then validated as AMTs. An analysis of the conversion of PMTs to AMTs as a function of SEQUEST quality score is shown in Fig. 7 for all peptides identified from ion trap MS/MS measurements and also for the charge states of the parent peptides. This analysis shows that greater than 95% of the PMTs identified with a SEQUEST cross-correlation (X_{corr}) value greater than 4.0 are converted into AMTs. A rapid decrease in the conversion of PMTs to AMTs, however, is observed for PMTs identified with an X_{corr} value less than 3.0. This result illustrates the effectiveness of the conventional approach to peptide identification with high X_{corr} values, but also reveals a rapid decrease in peptide identification confidence at X_{corr} values <3, the latter accounting for the majority of peptides analyzed. The factors leading to low SEQUEST scores include poor peptide ion intensities, simultaneous selection of multiple species for MS/MS, the presence of peptide modifications, as well as dissociation pathways for some peptide sequences that do not yield the types of products used by the search program for identification. The results shown in Fig. 7 support the need for validating peptide identifications, using programs such as SEQUEST (we have observed similar results with the program MASCOT), and demonstrate that a significant increase in confidence in peptide identifications is obtained through applying the AMT validation criteria.

The key downstream advantage in generating AMTs is their applicability for high-throughput studies designed to compare changes in the relative abundances of proteins between two separate proteome samples (i.e., control versus treated) solely on the basis of accurate mass measurements provided by FTICR. Once an AMT has been established, it can be used to confidently identify a specific protein in subsequent proteome studies. Without the need to reestablish the identify of a peptide by MS/MS analyses, multiple high-throughput studies focused on measuring changes in relative protein abundances between two (or more) different proteomes are facilitated. In such comparative studies, stable isotope labeling methods can be used to provide a means to measure protein

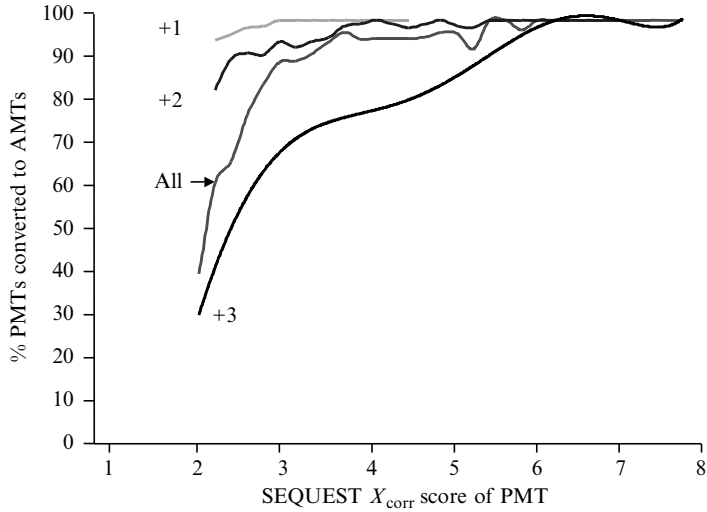


FIG. 7. An analysis of the conversion of *D. radiodurans* PMTs identified from ion trap tandem MS measurements to AMTs as a function of SEQUEST cross-correlation score for all peptides and for peptides of charge states +1, +2, and +3. Almost all of the PMTs identified with a SEQUEST cross-correlation (X_{corr}) score greater than 3.0 were converted into AMTs. This conversion rate dramatically drops off for PMTs with X_{corr} scores less than 3.0, resulting in elimination of about one-third of all PMTs, and particularly those with lower cross-correlation scores. The AMT validation step results in a significant improvement in the confidence of the protein identifications.

relative abundances, a process that also benefits from the resolution and sensitivity of the FTICR measurements.

F. Increasing Proteome Coverage Using AMTs

Several strategies have been applied in our initial work with *D. radiodurans* to increase the number of AMTs so as to subsequently routinely allow lower level proteins to be analyzed by this approach. First, samples were analyzed several times by the same capillary LC-MS/MS strategy, but with different *m/z* ranges and with the ''exclusion'' of parent ions that were previously selected for MS/MS, resulting in the selection of different peptides and the generation of additional PMTs. Beyond variations in instrumental approaches, proteome samples extracted from cells harvested at different growth phases (i.e., midlog phase, stationary phase) or cultured under a variety of different conditions (i.e., nutrients, perturbations) were also analyzed. By varying growth conditions and harvesting stages, the potential pool of PMTs increases significantly because

the absolute number of proteins collectively present in the different samples is significantly greater than the number expressed by the organism under a single growth condition. Finally, because any additional sample fractionation will increase the overall dynamic range achievable, we also analyzed peptide fractions first separated off-line by ion-exchange chromatography, and that again results in the generation of large numbers of additional PMTs for peptides that would otherwise have too-low abundance for conventional MS/MS analyses. It should be noted that any number of alternative sample fractionation and analysis strategies can be performed to increase the number of PMTs and AMTs generated, and that the extra efforts at this stage are more than offset by the resulting ability to make subsequent comprehensive proteome measurements with much greater sensitivity and speed. We continued PMT generation efforts for *D. radiodurans* for more than 200 different ion trap MS/MS runs, and until the rate of generation of novel PMTs decreased significantly. Because the analysis procedure can be totally automated, this corresponds to a one-time effort requiring approximately 3 weeks using a single ion trap instrument, and additional experience should significantly reduce the number of runs required for PMT generation.

Thus, although a significant number of samples is analyzed to generate the MS/MS spectra used for generating the set of AMTs for a specific organism, this initial investment of effort obviates the need for routine use of MS/MS in future analyses. The dividends for such an investment are realized in proteome studies designed to quantify changes in the relative abundance of proteins as a function of time or environment. In addition, a significant fractionation of samples before LC-MS/MS analysis is necessary to generate a large number of useful spectra by conventional instruments, but these lower level peptides can then be routinely detected by FTICR without the need for sample fractionation. Without the need to reidentify the expressed proteins through time-consuming MS/MS analyses and extensive sample fractionation, studies designed to quantify the relative abundances of proteins between two distinct proteome samples can be completed in a high-throughput manner by the exclusive use of the high mass accuracy measurements afforded by FTICR and with low attomole level sensitivity.

G. *Identification of* Deinococcus radiodurans *Proteins*

The use of AMT approach has been initially demonstrated for the prokaryotic organism *D. radiodurans*, a gram-positive, nonmotile, red-pigmented bacterium whose most distinguishing feature is its

effectiveness in coping with insults that cause DNA damage (i.e., desiccation, ultraviolet, or ionizing radiation) [45]. *Deinococcus radiodurans* has an ~3.1-Mb genome that initial annotation efforts predict to encode 3116 unique proteins [44]. An ideal tryptic digest of all proteins would yield 60,068 peptides between 500 and 4000 Da, of which ~51% would be unique at 1-ppm MMA, affording coverage of >99% of all predicted proteins. Analysis of proteins extracted from the organism cultured under a number of different growth conditions typically resulted in the detection of 20,000 to >50,000 peptides by capillary LC-FTICR (our unpublished data). Using capillary LC with MS/MS measurements (including ion trap MS/MS measurements that generated >9000 PMTs), 6997 peptides were validated as bona fide AMTs. These AMTs provide confident identification of 1910 predicted proteins (with an average of >3 AMTs per protein), covering ~61% of the predicted proteome and spanning every category of predicted protein function from the annotated genome. This comprehensiveness of proteome coverage exceeds that achieved for any other organism to date, and allowed (depending on culture conditions) 15 to 25% of the predicted proteome for *D. radiodurans* to be identified from AMTs detected in single FTICR runs. In a typical single FTICR analysis we detect ~1500 AMTs, corresponding to ~500 to >800 unique ORFs, representing ~15 to 25% of the predicted proteome of *D. radiodurans*. Greater proteome coverage would likely be obtained by separately processing and analyzing the insoluble, membrane protein fraction (see below).

The peptides detected by capillary LC-FTICR for *D. radiodurans* grown in a defined minimal medium and harvested at midlog phase are illustrated in a two-dimensional display of molecular mass versus spectrum number in Fig. 8. In this single LC-FTICR analysis of a *D. radiodurans* proteome sample, >22,000 putative peptides were observed. The inset within Fig. 8 shows detail for a typical region where peptides identified as AMTs are annotated with the ORFs that encode the parent proteins as well as the tryptic peptides corresponding to the AMT. For example, the spots labeled as DR0309.t25 (upper left) corresponds to tryptic peptide 25 from the N terminus of EF-Tu. The proteins identified in this small segment of the 2D display are listed in Table III. Although a significant fraction of the species detected was not validated as AMTs, many also do not correspond to the masses of ''possible'' peptides predicted from the sequences genome.

The *D. radiodurans* proteins observed included a majority of the predicted proteins in most of the functional categories defined by the Institute for Genomic Research, including 88 and 78% of the proteins associated with protein synthesis and transcription, respectively (Table IV). Predicted proteins involved in various *D. radiodurans* metabolic pathways were commonly identified with multiple different AMTs, suggesting their high

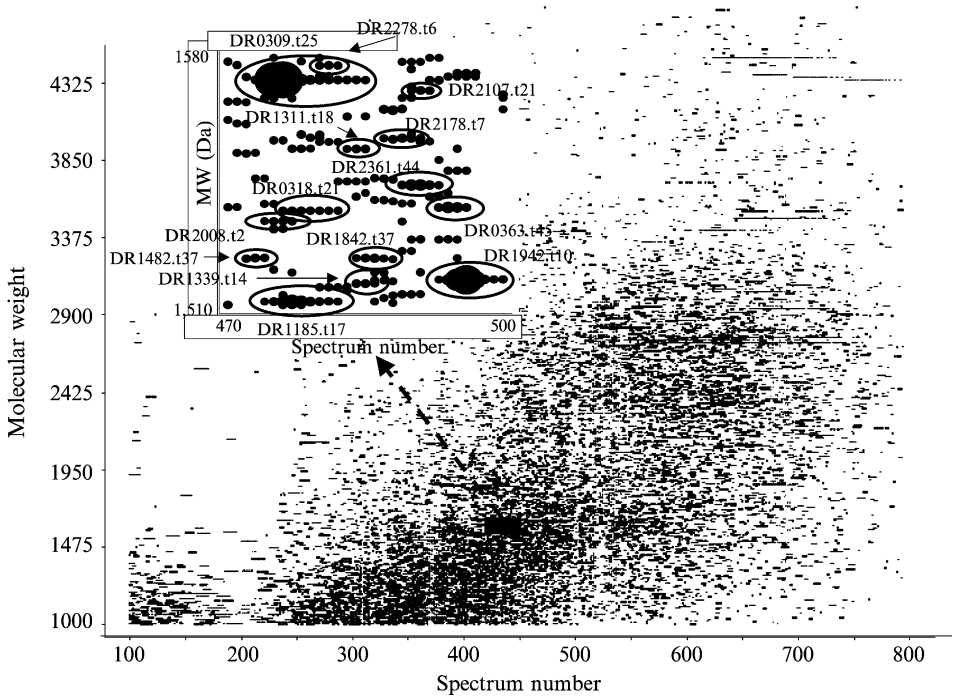


FIG. 8. Two-dimensional display of peptides based on their molecular weight (MW) and elution order (i.e., FTICR spectrum number) and identified AMTs from *Deinococcus radiodurans*. *Inset*: Circled spots identified as AMTs in the enlarged region. The spots are labeled on the basis of their annotation within the organism's genome sequence (i.e., DR0309; elongation factor Tu) and the tryptic peptide of the protein that was identified (i.e., t25; tryptic peptide 25, counting from the amino terminus, based on complete digestion). A list of the AMTs identified in the inset region, along with their calculated accurate mass and protein of origin, is given in Table III.

TABLE III
Accurate Mass Tags and Their Corresponding Proteins Identified[a]

AMT	Peptide sequence	Calculated AMT M_r[b]	Protein of origin
DR0309.t25	VQDEVEIVGLTDTR	1572.7994	Elongation factor TU
DR0318.t21	VMFEVAGVTEEQAK	1536.7493	Ribosomal protein L16
DR0363.t45	TMALPDSFPGYDPK	1537.7122	Putative peptide ABC transporter
DR1185.t17	APGFADYTTTITVR	1511.7619	S-layer-like array-related protein
DR1311.t18	TGDIGHAIQSLAESR	1553.7797	Methionine aminopeptidase
DR1339.t14	TATADDAEELAAAIR	1516.7368	Triosephosphate isomerase
DR1482.t37	ETYEIMNAELVGR	1523.7289	2-Isopropylamalate synthase
DR1942.t10	FGVTIPDEAAETIR	1517.7725	Acyl carrier protein
DR2008.t2	VAIVGATGAVGHELLK	1533.8878	Asparate-semialdehyde dehydrogenase
DR2107.t21	GENLGGLIITHYQR	1569.8263	ABC transporter
DR2178.t7	HGEVPAEAHAALVQK	1555.8106	Adenylosuccinate lyase
DR2278.t11	GSDGQVQGFDIDIAR	1576.7481	Amino acid ABC transporter
DR2361.t44	AAQQLGSITMVIGQK	1543.8391	Putative acyl-CoA dehydrogenase

[a] Peptide assignment corresponds to the open reading frame (ORF) reference number from the inset in Fig. 8.
[b] Monoisotopic molecular weight.

abundance. We verified the expression of all the predicted proteins corresponding to the vacuolar type (V-type) proton ATP synthase, as well as the predicted components of the organism's TCA cycle enzymes. In addition, 80% of the predicted proteins involved in glycolysis and the pentose phosphate shunt were detected. A slightly smaller fraction of electron transport proteins was identified, including several integral membrane proteins.

The proteins identified from *D. radiodurans* by our approach include many predicted to be present in low abundance based on their predicted codon adaptation index (CAI) [46]. The CAI has been shown to be a crude but useful predictor of protein abundance; proteins with high CAI values tend to be highly expressed, and those with low CAI values tend to be expressed at low levels. The distribution of CAI values for the identified *D. radiodurans* proteins compared with the distribution for all proteins predicted from the genome of this organism is shown in Fig. 9. In general, CAI values for the proteins identified by AMTs group in a Gaussian-like distribution similar to that for all of the predicted proteins, with more than 90% of predicted proteins having CAI values >0.8 being detected. In addition, we detected 18 of the 54 predicted *D. radiodurans* proteins that have a CAI value <0.2, further indicating that a good representation for expressed proteins is obtained.

TABLE IV
Deinococcus radiodurans Proteome Coverage by Functional Category[a]

Category	Total	Observed	Percent coverage
Amino acid biosynthesis	80	70	88
Biosynthesis of cofactors, prosthetic groups, and carriers	61	41	67
Cell envelope	77	62	81
Cellular processes	89	64	72
Central intermediary metabolism	154	111	72
DNA metabolism	81	55	68
Energy metabolism	199	152	76
Fatty acid and phospholipid metabolism	53	40	75
Conserved hypothetical	499	276	55
Phage-related and transposon proteins	47	9	19
Protein folding, modification, and secretion	86	65	76
Protein synthesis	114	100	88
Purines, pyrimidines, nucleosides, and nucleotides	53	42	79
Regulatory functions	126	76	60
Transcription	28	22	79
Transport and binding proteins	191	138	72
Unknown function	176	107	61
Hypothetical	1002	479	48

[a]Based on functional assignments in reference 44.

In this initial work we applied methods solely on the basis of a ''global'' tryptic digestion that would be expected to be much less effective for membrane proteins. Interestingly, we find that a significant fraction of the most hydrophobic proteins is still detected. The percentage of proteins detected as a function of the portion of each protein predicted to reside in a membrane decreases by ~50% for the most hydrophobic proteins, as shown in Fig. 10. Further refinements of sample solubilization and digestion methods, such as those utilized by Washburn *et al.*, should further increase membrane protein coverage.

H. Quantitative High-Throughput Proteome Measurements

Useful proteome measurements often require comparing protein abundances between two cellular populations resulting from, for example, some insult or perturbation. The predominant method for measuring

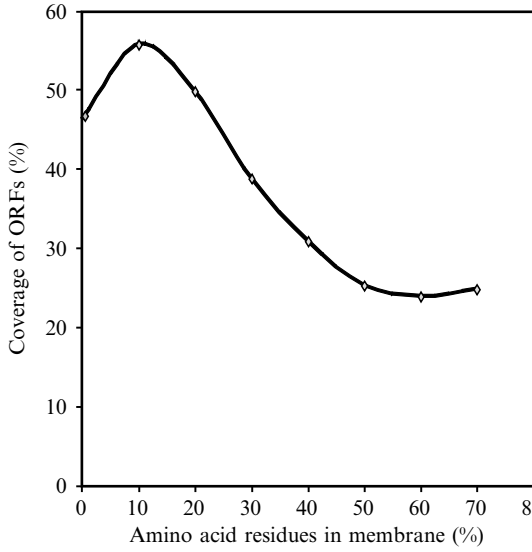


FIG. 9. The percentage of detected *D. radiodurans* proteins as a function of the portion of each protein predicted to reside in a membrane is shown, indicating there is a bias for detection against the most hydrophobic proteins, attributed to the sample processing and digestion conditions used in this initial work.

changes in protein expression levels by current proteomic technology is to compare the intensities of the corresponding 2D-PAGE spots. Attempts to infer absolute peptide abundances based on MS signal intensities can be problematic for reasons that include variations in ionization efficiencies and losses during sample preparation and separations, and although useful for large differences in abundances, are unsuited to study more subtle variations. The generation and use of AMTs enable high-throughput and high-precision expression studies based on stable isotope labeling by directly comparing two proteomes in the same analysis (e.g., utilizing a ''reference proteome'' to which perturbed systems are compared). A stable isotope-labeled reference proteome, for example, provides an effective internal standard for each protein, and hence their tryptic peptides, allowing changes in protein abundances based on the relative abundances of AMTs to be assessed, potentially to precisions better than 10% [38, 48, 49]. Although such measurements require both versions of the protein or peptide to be present, it should be feasible to combine this information with absolute peak intensity data to provide less

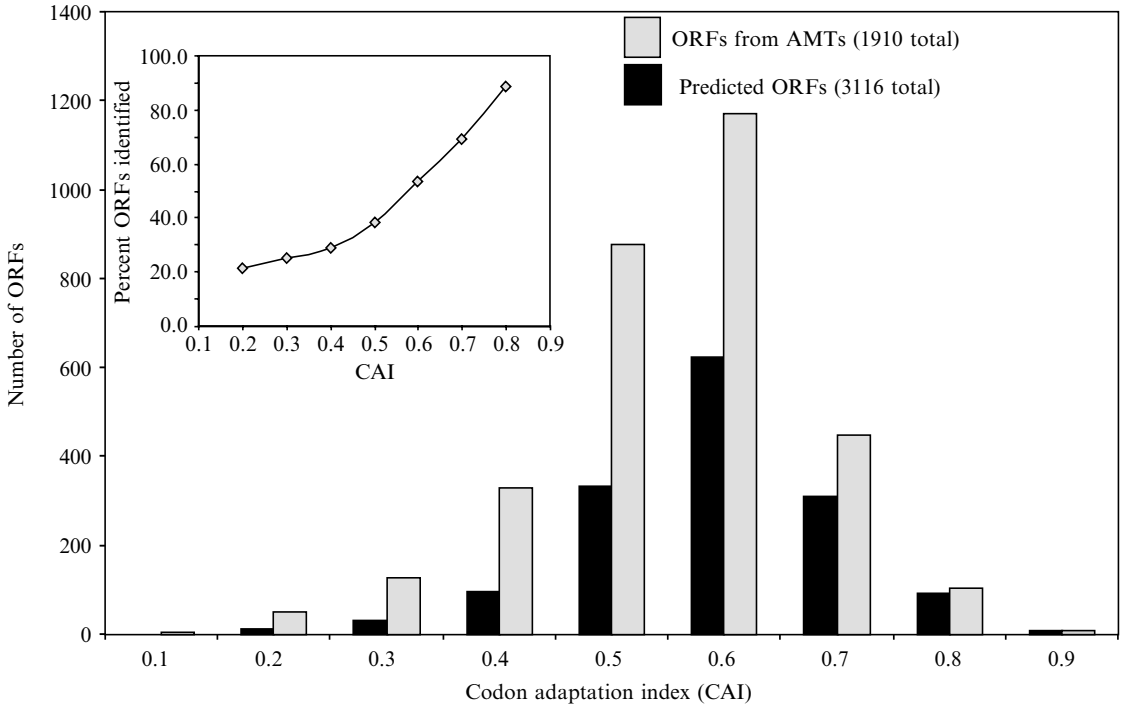


FIG. 10. Distribution of codon adaptation index (CAI) values for detected *D. radiodurans* proteins compared with the distribution for all predicted proteins [46]. This comparison indicates that, although there is some bias for high-CAI proteins, many proteins with low CAI values, and thus predicted to be expressed at only low levels, are observed.

precise abundances when only one peptide is detected, and also to establish approximate absolute abundances (albeit with less precision).

A 2D display is shown in Fig. 11 (see Color Insert), where the relative peptide abundances for *D. radiodurans* control cells are compared with those of cells treated with H_2O_2. The colored spots represent the measured relative abundance levels of peptides (green represents a decrease in abundance, black unchanged, and red an increase). In this preliminary study, *D. radiodurans* cells were cultured in both normal and ^{15}N-enriched growth media; the inset in Fig. 11 shows the corresponding mass spectral region for two spots corresponding to AMTs for catalase and S-layer protein. As previously observed by conventional methods [50, 51], significant differences in a large fraction of peptide abundances after H_2O_2 exposure are observed. Similarly, Fig. 12 (see Color Insert) shows a 2D display for *D. radiodurans* undergoing recovery after exposure to 17.5 kGy, where each spot corresponds to a set of isotopically distinctive peptides from which an expression ratio can be determined (Fig. 12, inset A).

Although significant methodological refinement is required to fully exploit this approach and also to define its limitations, this general approach should provide a global view of the response of a proteome to a perturbation. The ability to conduct many such experiments will significantly contribute to an increased understanding of the functions of proteins and their interactions. Work has also highlighted the promise of applying such methods to selected subproteomes, for example, by isotopic labeling and affinity isolation of phosphopeptides [52–54].

III. Technology Advances for Expanding Proteome Coverage

To meet future demands and expectations of global proteomics will continue to require the development of technologies that increase the sensitivity and throughput of protein measurements. Although increases in sensitivity can be advanced by different means, dynamic range expansion is a key factor when dealing with complex mixtures in which several species may be detected in a single spectrum. In mixtures with the complexity described above, high- and low-abundance species will invariably coelute into the MS. With the use of conventional MS technologies that have limited dynamic range the lower abundance species may go undetected. The biology of the cell drives the need for increased dynamic range. Many key functional proteins such as kinases, phosphatases, and transcription factors, as well as proteins that may play crucial roles in cell function in disease states, are expressed at low levels.

The bottleneck to increasing the throughput of proteomic measurements remains the amount of time required to identify the large number of proteins present within these complex mixtures. Although the AMT approach provides the potential to obviate the need for MS/MS in subsequent studies of a species-specific proteome, it also requires a significant identification effort. Tandem MS will continue to play an important role in the unambiguous identification of proteins for the foreseeable future. Whereas the MS spectra of several peptides can be acquired simultaneously, MS/MS analysis is limited to a single peptide per experiment. We have developed instrumental methods that are aimed at allowing the MS/MS spectra of multiple peptides to be recorded concurrently. As described below, the high MMA of FTICR is a key factor allowing the correct assignment of the resulting complex spectrum.

A. DREAMS FTICR for Expanded Dynamic Range Proteome Measurements

The large variation of protein relative abundances having potential biological significance in mammalian systems (more than six orders of magnitude) presents a major challenge for proteomics. Although FTICR mass spectrometry has demonstrated a capability for ultrasensitive characterization of biopolymers (e.g., achieving subattomole detection limits) [30], as noted earlier the maximum dynamic range for a single mass spectrum (i.e., without the use of spectrum averaging or summation) is typically constrained to about 10^3. An important factor conventionally limiting achievable FTICR sensitivity and dynamic range is the maximum charge capacity of either the external ion accumulation device or the FTICR trap mass analyzer region itself. Prolonged ion accumulation would be helpful during the LC elution of low-abundance components (i.e., during the ''valleys'' in chromatograms), potentially allowing measurable signals to be obtained for otherwise undetectable species, and increasing the effective overall dynamic range of proteome measurements. Unfortunately, ''overfilling'' of external multipole ion traps by other high-abundance species often results in a biased accumulation process in which parts of the m/z range are selectively retained or lost [55] and/or extensive ion activation and dissociation occur well before sufficient populations of low-level species can be accumulated. When analyzed in conjunction with capillary LC separations, both the total ion production rate for peptides from ESI and the complexity of the mixture at any point can vary by more than two orders of magnitude. This temporal variation in ion production rate and spectral complexity constitutes a major challenge for proteome analyses. For example, the elution of highly abundant peptides can restrict the detection of lower level coeluting peptides. If the ion accumulation

time is optimized for the most abundant peaks, the accumulation trap will not be filled to capacity during the elution of lower abundance components, and the overall experimental dynamic range will be significantly constrained. If, however, longer accumulation times are used, the conditions conventionally used result in an ''overfilling'' of the external accumulation trap in many cases, which will be manifested by biased accumulation or extensive activation and dissociation. Thus, we have attempted to develop methods that avoid or minimize the undesired artifacts associated with overfilling the external accumulation trap [42] or that actively select ion accumulation times (i.e., implement automated gain control), and thus effectively expand the dynamic range of measurements.

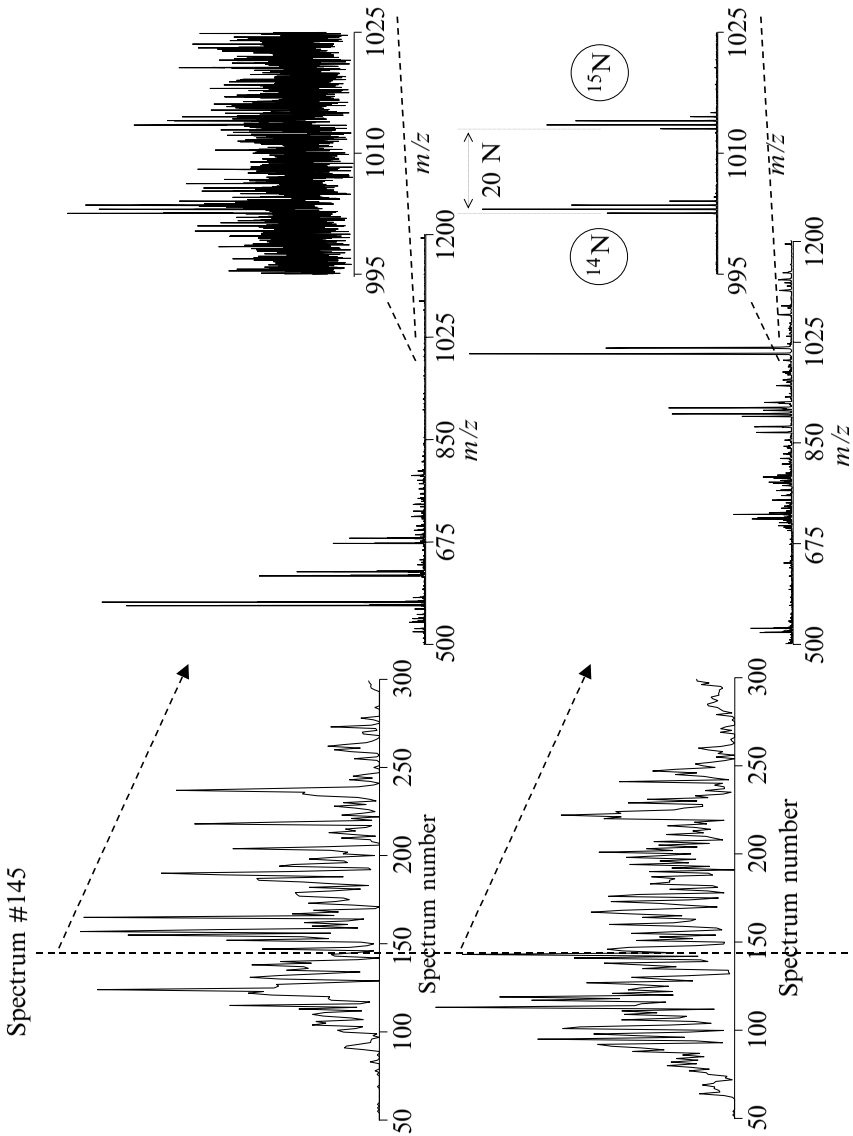
A generally useful approach to dynamic range expansion involves the use of ion ejection from a linear quadrupole device external to the FTICR that is accomplished by resonant RF-only dipolar excitation [56]. The effective removal of the major species from a spectrum allows the lower abundance species to be accumulated for extended periods, resulting in an increase in the dynamic range. This dynamic range enhancement applied to mass spectrometry (DREAMS) approach thus provides the basis for a significant gain in the coverage of proteomic measurements.

The DREAMS methodology involves acquisition of sets of mass spectra during the nonselective accumulation, in which each spectrum is followed by software-controlled selection of the most abundant ion peaks based on their quadrupole secular frequencies and then selective RF-only ejection of the most abundant species before external accumulation (for the next spectrum immediately following the nonselective ''normal'' spectrum). We initially evaluated the data-dependent selective external ion ejection with a mixture of peptides, and then evaluated the DREAMS approach for the characterization of a global yeast proteome tryptic digest using a low-field (3.5-T) FTICR mass spectrometer [56]. Examination of the mass spectra acquired by automated DREAMS data-dependent selective external ion accumulation showed that the experimental mass resolution during actual LC separation for RF-only ion ejection from the selection quadrupole was in the range of 30 to 50, depending on m/z. This initial demonstration of the DREAMS FTICR method generated two data sets comprising spectra for the detected peptide isotopic distributions from the nonselective and selective DREAMS accumulations. To evaluate the approach these data sets were processed and compared with a data set acquired in a separate capillary LC-FTICR analysis using only the standard nonselective external ion accumulation method. Throughout the LC separations the intensities of the most abundant ion species were found to vary by approximately two orders of magnitude (consistent with variation in a chromatogram obtained by ultraviolet detection). The

results obtained for the two data sets acquired by alternating sequences in one LC-FTICR run (i.e., the nonselective and DREAMS selective external ion trapping) were compared with the number of putative peptides identified in a separate LC run using nonselective external ion accumulation. It was found that the number of peptides detected with the alternating sequences (30,771 after subtraction of species detected in both) was greater by about 35% than that acquired by nonselective ion accumulation (22,664). The same methodology was subsequently applied with data-dependent selective ion ejection of the two and three most abundant ion species. A 40% increase in the number of peptides was achieved when combining the nonselective ion accumulation with data-dependent selective ion ejection of the three most abundant ion species [56].

We have also implemented the DREAMS approach with high-performance capillary reversed-phase liquid chromatography (RPLC) separations and high magnetic field electrospray ionization FTICR mass spectrometry for obtaining improved mammalian proteome measurements [57]. We prepared a global tryptic digest of ^{14}N- and ^{15}N-labeled peptides from soluble proteins extracted from mouse B16 cells and mixed at a ratio of approximately 1:1. In the analysis every second mass spectrum resulted from the accumulation of ions that remained after the dipolar ejection of the most abundant ions (i.e., the DREAMS spectrum). Thus, it was possible to reconstruct two different ion chromatograms for this experiment, corresponding to the ''normal'' and the DREAMS spectra sets.

The dynamic range enhancement capability using DREAMS by comparing the two mass spectra for one point in a capillary LC separation is illustrated in Fig. 13. The data allow two chromatograms to be reconstructed from the data corresponding to normal and DREAMS spectra (Fig. 13, left top and bottom). As shown by the comparison of the full mass spectra, the normal spectrum (Fig. 13, middle top) is dominated by a number of major peptide ions, most prominently three pairs of ^{14}N/^{15}N-labeled peptides in the $500 < m/z < 700$ range. The information from this spectrum was used ''on the fly'' to apply dipolar RF excitation in the 2D quadrupole at the secular frequencies corresponding to the m/z of these major ions during ion accumulation for the next spectrum. As a result, the species in this m/z region were effectively ejected before accumulation and the FTICR spectrum is now dominated by a much different set of species (Fig. 13, middle bottom). As a single example selected from many similar cases, the inset to the right shows that the signal-to-noise ratio (S/N) for a peptide pair at m/z ~1000 is greatly improved from a level at which no effective identification could be


Spectrum #145
Spectrum number
Spectrum number
m/z
m/z
m/z
m/z
20 N
^{14}N
^{15}N

obtained, to a level at which a precise relative abundance ratio (AR) can be determined for the peptide pair, with the gain in the S/N in this case being ~50.

In this initial implementation the overall speed of the DREAMS process was limited by both software (for data processing, data transfer times, and the generation of the DREAMS RF waveforms) and the ion cooling times used to improve FTICR spectrum quality. Thus, the total time required for acquisition of both the normal and DREAMS spectra pair is ~20 s. To provide a basis for a more quantitative evaluation of the additional information acquired by DREAMS analysis, we used longer high-performance reversed–phase gradient separations in which the minimum peak width is about 1 min. Chromatograms reconstructed from the FTICR spectra acquired during a single high-pressure capillary LC separation of 1:1 mixture of ^{14}N- and ^{15}N-labeled mouse B16 tryptic peptides are shown in Fig. 14. The average LC peak widths in this case spanned from 3 to more than 10 spectra for both the normal and DREAMS chromatograms. It is evident from the quite different chromatographic profiles that many additional species were detected with much greater signal intensities in the DREAMS spectra, particularly at longer retention times. However, the chromatograms provide only a qualitative view of the added information from the DREAMS approach, and one that is obviously limited to the more abundant species.

To further evaluate the utility of the DREAMS approach for providing additional information about proteomes we use as a figure of merit the number of peptide pairs that could be confidently assigned to both the normal and DREAMS spectra sets for the separation shown in Fig. 14. Any given mass that appeared at two distinctive elution times was counted only once, thus leading to a small potential underestimation of the number of peptides actually detected. This list was then searched for pairs whose mass difference corresponded to a multiple of the ^{14}N–^{15}N mass difference that

FIG. 13. Capillary LC-FTICR results showing total ion chromatogram (TIC) and portions of the mass spectra acquired during the normal spectrum acquisition process and the alternating DREAMS spectrum acquisition process. *Top*: TIC reconstructed from the FTICR spectra acquired during RPLC separation of a mixture of identical aliquots of a natural isotopic abundance and ^{15}N-labeled version of mouse B16 melanoma cells and representative spectrum obtained using broadband mode acquisition and a 100-ms accumulation time. *Bottom*: Corresponding TIC and representative spectrum obtained using RF-only selective acquisition and 300-ms accumulation time. Resonant frequencies for RF-only dipolar excitation were identified during a broadband ion acquisition (*Top*) and up to five species having relative abundance >10% were then data-dependently ejected during a selective acquisition that immediately followed (*Bottom*).

could correspond to possible peptides (i.e., the maximum and minimum nitrogen content for a peptide of a given mass), and for which there was also an overlap in the elution time. From this smaller set we also discarded any possible peptide pairs that had a peptide assigned to more than one peptide pair (i.e., an ambiguous pair assignment). Finally, we also discarded a small set of possible peptide pairs that had relative abundances that were outside the range of 0.4 to 1.6 (because an approximately 1:1 ratio was expected). These criteria resulted in a conservative determination of the number of peptide pairs that were detected in the normal and DREAMS spectrum sets. This number represented a lower bound for the number of cases in which quantitative comparisons of the relative intensities of such pairs could be derived, and thus the cases in which stable isotope-labeling methodology could potentially be applied to obtain quantitative measurements for relative protein abundances.

A 2D display that includes spots for only the peptide pairs measured from the capillary LC-FTICR analysis of Fig. 14 is shown in Fig. 15. Figure 15 is a result of substantial postprocessing of the capillary LC-FTICR analysis to convert the mass spectral information into molecular masses and to identify the subset of $^{14}N/^{15}N$-labeled peptide pairs as described above. The analysis of the normal spectra resulted in detection of a total of 9896 $^{14}N/^{15}N$-labeled peptide pairs (Fig. 15, left). The average AR for the peptide pairs was 1.05, only a slight deviation for the nominal value of 1.0 expected. The standard deviation for the peptide pairs was 0.28. The second set of DREAMS mass spectra revealed 8856 $^{14}N/^{15}N$-labeled peptide pairs (Fig. 15, right), of which 7917 were ''new'' peptide pairs not detected in the normal spectra, and 939 that were lower level peptide pairs also observed in the normal spectra. The average AR for the peptides detected in the DREAMS portion of the analysis was 1.015, and the standard deviation was 0.31. It should be noted that the assigned peptide pairs accounted for less than 25% of the total number of species detected in the analysis and that presumably contains contaminants, other mixture components, and pairs of peptides excluded by the criteria set above. Finally, the combined total of 17,813 unambiguous and unique peptide pairs gave an average AR of 1.035 and a standard deviation of 0.29.

Although different approaches for the analysis of these data, primarily the use of different peptide pair selection criteria, will lead to slightly different results, the key conclusion is that the number of detected peptides is greatly increased by the use of the DREAMS methodology. Indeed, the number of peptides for which quantitative information could be obtained from the relative abundances of peptide pairs was increased by 80%, using the criteria described above.

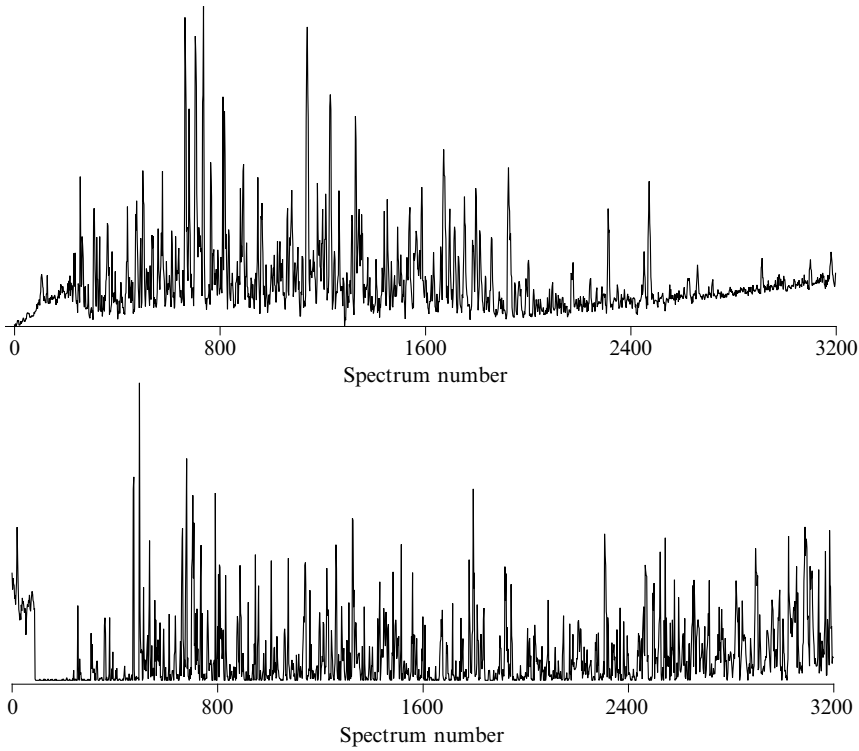


FIG. 14. Total ion chromatograms (TICs) reconstructed from FTICR mass spectra acquired using broadband mode acquisition with 100-ms accumulation time (*top*) and with RF-only selective (i.e., DREAMS) acquisition and 1-s accumulation time (*bottom*) during a capillary LC separation of soluble proteins that were digested with trypsin from a 1:1 mixture of ^{14}N- and ^{15}N-labeled mouse B16 cells.

Additional efforts are planned toward improvements of the resolution of the ion ejection step (currently 30 to 50) and also to implement the DREAMS analysis approach as a routine part of our AMT strategy for quantitative proteome measurements. In particular, the effective integration of this step with the use of a variable ion accumulation time for automated gain control is in progress. This step will be important to enable the use of absolute peak intensities, and also to eliminate small errors associated with AR measurements caused by possible variations in sensitivity during peak elution and the small elution time offsets associated with the different stable isotope-labeled version of each peptide. Because this capability would more effectively exploit the optimum trap capacity for each spectrum, we anticipate that this combination of capabilities

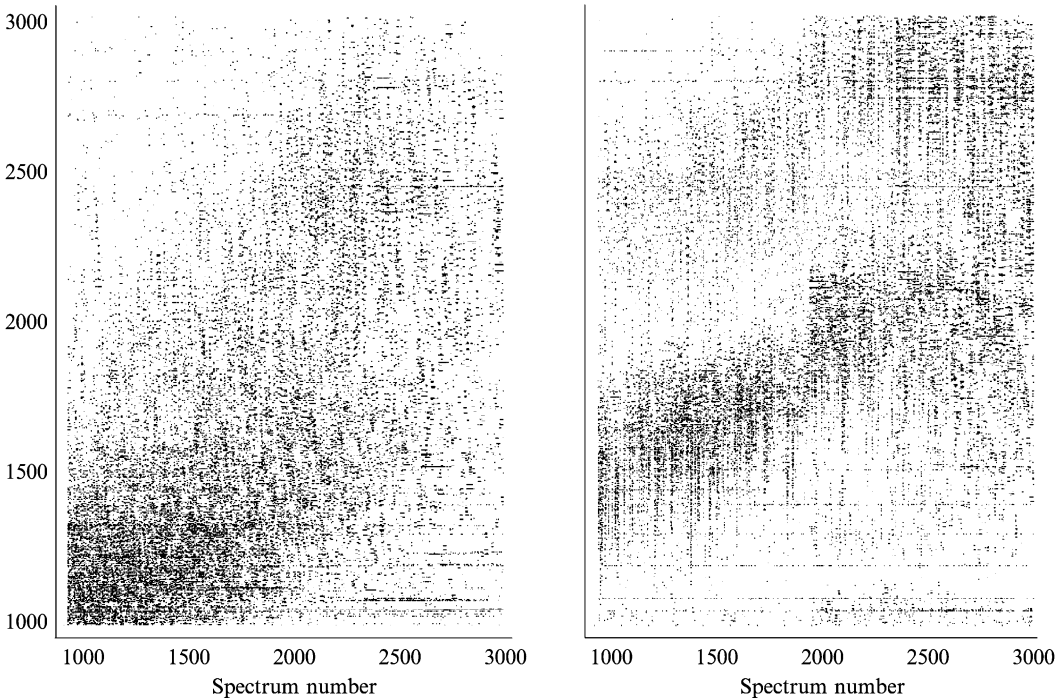


FIG. 15. Two-dimensional display showing the molecular weights (MW) and separation number for peptide pairs detected from a 1:1 mixture of ^{14}N- and ^{15}N-labeled peptides from mouse B16 cells (see Fig. 14). *Left*: Peptide pairs detected in the normal spectrum set. *Right*: Peptide pairs detected in the DREAMS spectrum set. A total of 17,813 unique peptide pairs was detected using the conservative criteria described in text, an 80% increase over the number observed from the normal spectrum set alone.

would further increase the dynamic range in proteome measurements. Finally, the speed of the methodology needs to be increased. At present we utilize a 5-s ion cooling step to optimize the quality of the resulting data. When this is done in conjunction with DREAMS analysis (or multiplexed MS/MS; see below), however, the result is either the need for an extended analysis time or effectively decreased chromatographic resolution (and significantly decreased proteome coverage). We have shown the potential to decrease the ion cooling time by up to an order of magnitude, using an adiabatic cooling scheme that involves the ramping of the FTICR ion trapping well voltages [58]. The combination of this step with decreased data transfer and processing times promises to result in substantially reduced overall analysis times.

B. Multiplexed MS/MS for High-Throughput and Targeted Peptide Identification

Because of the complexity of the proteomic samples multiple peptides generally coelute and often hundreds can be observed in a single FTICR spectrum, even for the highest resolution LC separations. On the other hand, conventional MS/MS analysis is sequential (i.e., can address only one peptide at a time), and the data acquisition rate typically fails to allow selection and fragmentation of all detected peptides in the time available for analysis (i.e., the elution time of a peak). Consequently, the dynamic range is effectively reduced because typically the low-abundance ions are not selected for MS/MS analysis even when ''dynamic exclusion'' methods are used to prevent repetitive selection of the same peak. To help alleviate the problem, various ''peak-parking'' schemes have been developed in which the chromatographic or electrophoretic peak elution time is extended to allow additional MS/MS experiments to be conducted [59–61]. For example, Martin *et al.* described a variable-flow HPLC apparatus for on-line tandem mass spectrometric analysis of tryptic peptides [62]. Although such an approach alleviates the problem to some extent, comprehensive MS/MS analysis remains impractical for complex proteome digest samples. The reduction of LC flow rates not only significantly increases the overall separation time but may also decrease sensitivity for MS detection (particularly for the use of small inner diameter capillaries, where the ESI efficiency is already close to its maximum [63]) and the resolution obtained in the separation, and is less useful for high-efficiency LC separations that are based on the use of high pressures [42].

One approach for addressing this issue involves the effective collection of LC effluent fractions for subsequent analysis using matrix-assisted laser desorption ionization (MALDI) MS, an approach that allows as much time as needed for MS/MS analyses, or the use of approaches that presume to identify the ''interesting'' peptides that will then by identified. However, many issues related to the throughput, dynamic range, sensitivity, sample preparation, storage stability, matrix/bias effects, and so on, remain to be addresses to reasonable access the viability of such an approach.

We have developed an approach that utilizes the multiplexing capability derived from the high resolution and MMA of FTICR for the simultaneous MS/MS analysis of multiple peptides [64] during on-line capillary LC separations. A unique attribute of FTICR is the ability to select and simultaneously dissociate multiple precursor peptides. Although repeating MS/MS experiments with different subsets of parent ions allows the assignment of all fragments to the corresponding parent species, these methods require a large amount of sample, and are too slow for use with on-line separations. Our multiplexed MS/MS strategy simultaneously obtains sequence information for multiple peptides, providing both enhanced sensitivity and a gain in throughput.

As an initial demonstration of the use of the multiplexed MS/MS approach for protein identification from complex proteome samples, a 2D display for the capillary LC-FTICR analysis of tryptic peptides from *D. radiodurans* whole cell lysate, in which the >13,000 peptide ''spots'' detected are shown on the basis of their molecular mass and LC elution time, is shown in Fig. 16A. Figure 16B shows an example of a mass spectrum obtained in a single MS acquisition corresponding to the dotted line in the 2D display, whereas Fig. 16C shows the corresponding (i.e., immediately following) multiplexed MS/MS spectrum with the dissociation products from the four most abundant parent ions selected from the spectrum shown in Fig. 16B. A significant number of sequence-specific fragments was attributed to each selected peptide. The extensive fragmentation allowed the identification of the four selected tryptic peptides, with each identifying a unique *D. radiodurans* protein. For example, the peptide with M_r = 1628.86 Da was identified as LLDSGMAGDNVGVLLR from elongation factor Tu (DR2050 and DR0309, which happen to be duplicated open reading frames) and the peptide with M_r = 1696.94 Da was identified as a tryptic fragment from glyceraldehyde-3-phosphate dehydrogenase (DR1343). We have found that peptides having significant differences in abundances (visualized by variation of the spot size in 2D display and the ion intensity in

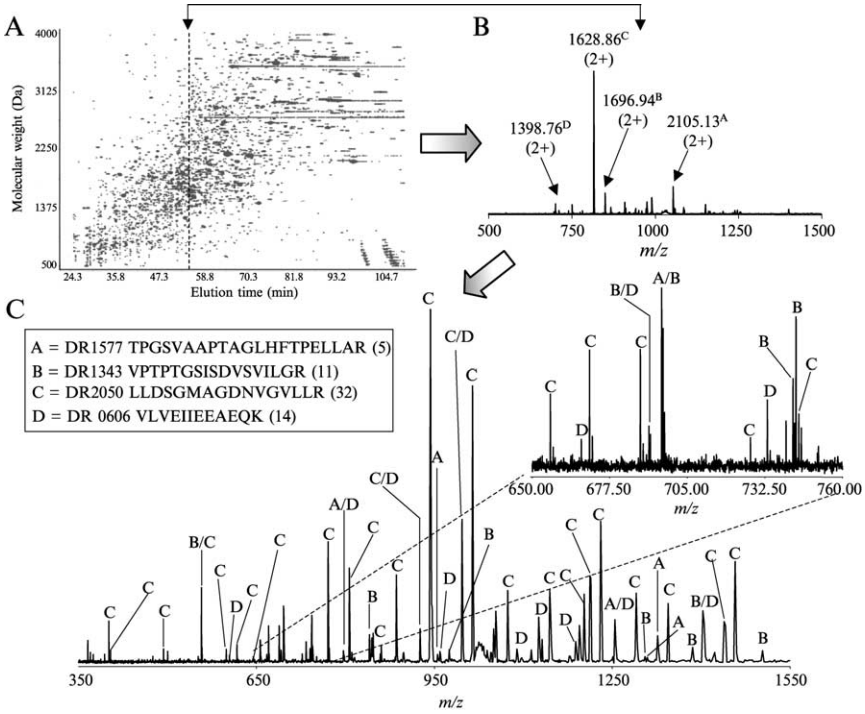


FIG. 16. (A) Two-dimensional (2D) display reconstructed from FTICR spectra obtained during a capillary LC-FTICR analysis of a global tryptic digest of a *D. radiodurans* cell lysate. More than 13,000 peptide ''spots'' (i.e., isotopic distributions) were detected. (B) A mass spectrum [indicated by the dotted line in the 2D plot shown in (A)], with the four most abundant ions selected for the subsequent MS/MS acquisition (C) Multiplexed MS/MS spectrum of the four peptide species selected in (B) with fragment ions attributed to each individual peptide after searching against the *D. radiodurans* protein database. The four proteins uniquely identified from this spectrum are listed in the *inset box* with their open reading frame reference number (e.g., DR1577). The numbers listed in parentheses after each peptide indicate the number of fragment ions detected for each tryptic peptide. An expanded view of *m/z* 650–760 with sequence-specific fragment ions labeled is shown in the *in set* on the right.

MS spectrum) could be readily fragmented to yield useful sequence information.

Obviously the initial stages of proteomic research with any cell or tissue type by this approach will require a greater effort to establish and validate AMTs. These efforts, together with the need to unambiguously identify modified peptides or unexpected (e.g., due to frame shifts) peptides, can be significantly facilitated by the multiplexed MS/MS approach. The ability to selectively eject the most abundant species before external ion

accumulation and the transfer of ions to the ion cyclotron resonance (ICR) cell should allow us to obtain broader proteome coverage by measuring low-abundance species that are undetectable by other methodologies. The coupling of DREAMS active dynamic range enhancement methodology with multiplexed MS/MS provides a basis for protein identification to be performed with much greater sensitivity and speed.

C. More DREAMS for the Future

The possibilities for application of the new quantitative proteome measurement technology and methods described in this work are broad. Although major challenges remain to realize the full potential of the technology, particularly for obtaining useful quantitative measurements for low-abundance proteins, and for extending the approach to modified proteins, it is clear that many useful applications to microbial systems are already tractable. The combination of sensitivity, dynamic range, and throughput should enable new types of studies to be contemplated. In particular, studies that would otherwise demand excessive quantities of protein or present too much complexity should now be tractable.

Most of our work to date has focused on relatively simple microbial systems. Mammalian proteomes pose a much greater challenge relative to microorganisms because of their greater complexity and range of relative protein abundances. Mouse B16 cells cultured in both ^{14}N- and ^{15}N-enriched media and affinity tagged with iodoacetyl-PEO-biotin have been used to isolate two cysteine-containing versions of each peptide having masses that differ by the number of nitrogen atoms [26]. Both this and the Isotope-coded affinity tag (ICAT) strategy significantly reduce the complexity by isolating cysteine peptides without significantly decreasing the proteome coverage. As shown in Table II, approximately 90% of *C. elegans* (and presumably human) proteins are potentially identifiable using cysteine peptides. Initial results for mammalian proteomes have detected fewer peptides than were observed for yeast, likely because of both the reduced complexity due to affinity selection of the cysteine peptides and the greater dynamic range for proteins, including the presence of a number of highly abundant proteins. The DREAMS capability presently under development for use with LC separations and the implementation of protein fractionation steps before LC-FTICR analysis should substantially increase the number of detectable species (exceeding the 100,000 detectable components we have experimentally demonstrated for soluble yeast proteins [27]). The human proteome can

be estimated to yield a peptide mixture having a complexity on the order of several million components for an ideal tryptic digest. Although the actual complexity is certainly higher, only a fraction of the possible proteins will be expressed in any single cell type. A fractionation method that yields 10 distinct fractions combined with the capillary LC-FTICR approach we have demonstrated would provide a combined theoretical peak capacity $>10^8$, potentially allowing mammalian proteomes to be studied without the need to apply cysteine peptide selection methods.

We believe, for example, that the DREAMS FTICR technology is an important component of an approach that provides the basis for a significant gain in the coverage of proteomic measurements. It is clear that such technology development efforts can have a significant impact on the practice of proteomics, particularly for applications where measurements of the highest quality and broadest scope are beneficial. In future the initial application of this new technology to the study of microbial systems should clarify its role.

Acknowledgments

We acknowledge the contributions of Kim Hixson, Ron Moore, Rui Zhao, Richard Harkewicz, David Anderson, Nikola Tolic, Lingjun Li, Ken Auberry, Keqi Tang, Deanna Auberry, Nicolas Angell, and Brian Thrall (all of PNNL) and of Thomas Conrads of the Biomedical Proteomics Program at NCI/Frederick to the work reviewed here. We thank the U.S. Department of Energy Office of Biological and Environmental Research for long-time support of *D. radiodurans* research and FTICR technology development, as well as the National Institutes of Health, through the NCI (CA81654), NINDS (NS39617), and NCRR (RR12365), for support of portions of this work. The Pacific Northwest National Laboratory is operated by the Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC06-76RLO 1830.

References

1. Adams, M. D. (1996). Serial analysis of gene expression: ESTs get smaller. *Bioessays* **18,** 261–262.
2. Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E. J., Hieter, P., Vogelstein, B., and Kinzler, K. W. (1997). Characterization of the yeast transcriptome. *Cell* **88,** 243–251.
3. Anderson, L., and Seilhammer, J. (1997). A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18,** 533–537.
4. Haynes, P. A., Gygi, S. P., Figeys, D., and Aebersold, R. (1998). Proteome analysis: Biological assay or data archive? *Electrophoresis* **19,** 1862–1871.
5. Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000). Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. USA* **97,** 9390–9395.

6. Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996). Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379,** 466–469.
7. Shevchenko, A., Wilm, M., Vorm, O., and Mann, M. (1996). Mass spectrometric sequencing of proteins from silver stained polyacrylamide gels. *Anal. Chem.* **68,** 850–858.
8. Yates, J. R. I., Speicher, S., Griffin, P. R., and Hunkapiller, T. (1993). Peptide mass maps: A highly informative approach to protein identification. *Anal. Biochem.* **214,** 397–408.
9. Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Boucherie, H., and Mann, M. (1996). Linking genome and proteome by mass spectrometry: Large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA* **93,** 14440–14445.
10. Langen, H. *et al.* (2000). Two-dimensional map of the proteome of *Haemophilus influenzae. Electrophoresis* **21,** 411–429.
11. Perrot, M. (1999). Two-dimensional gel protein database of *Saccharomyces cerevisiae. Electrophoresis* **20,** 2280–2298.
12. Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S., and Garrels, J. I. (1999). A sampling of the yeast proteome. *Mol. Cell Biol.* **19,** 7357–7368.
13. Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.* **19,** 1720–1730.
14. Garrels, J. I., McLaughlin, C. S., Warner, J. R., Futcher, B., Latter, G. I., Kobayashi, R., Schwender, B., Volpe, T., Anderson, D. S., Mesquitafuentes, R., and Payne, W. E. (1997). Proteome studies of *Saccharomyces cerevisiae*: Identification and characterization of abundant proteins. *Electrophoresis* **18,** 1347–1360.
15. Kitayama, S., and Matsuyama, A. (1971). Mechanism for radiation lethality in *M. radiodurans. Int. J. Radiat. Biol. Relat. Stud. Phys. Chem. Med.* **19,** 13–19.
16. Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., and Watanabe, C. (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA* **90,** 5011–5015.
17. Pappin, D. J., Hojrup, P., and Bleasby, A. J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3,** 327–332.
18. Mann, M., Hojrup, P., and Roepstorff, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **22,** 338–345.
19. James, P., Quadroni, M., Carafoli, E., and Gonnet, G. (1993). Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* **195,** 58–64.
20. Yates, J. R., McCormack, A. L., and Eng, J. (1996). Mining genomes with MS. *Anal. Chem.* **68,** A534–A540.
21. McCormack, A. L., Schieltz, D. M., Goode, B., Yang, S., Barnes, G., Drubin, D., and Yates, J. R. (1997). Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal. Chem.* **69,** 767–776.
22. Ducret, A., Vanoostveen, I., Eng, J. K., Yates, J. R., and Aebersold, R. (1998). High throughput protein characterization by automated reverse-phase chromatography electrospray tandem mass spectrometry. *Protein Sci.* **7,** 706–719.
23. Link, A. J., Hays, L. G., Carmack, E. B., and Yates, J. R. (1997). Identifying the major proteome components of *Haemophilus influenzae* type-strain NCTC 8143. *Electrophoresis* **18,** 1314–1334.

24. Yates, J. R. (1998). Mass spectrometry and the age of the proteome. *J. Mass Spectrom.* **33,** 1–19.
25. Washburn, M. P., Wolters, D., and Yates, J. R. I. (2001). Large-scale analysis of the yeast proteome by multidimensional protein indentification technology. *Nat. Biotechnol.* **19,** 242–247.
26. Conrads, T. P., Alving, K., Veenstra, T. D., Belov, M. E., Anderson, G. A., Anderson, D. J., Lipton, M. S., Pasa-Tolic, L., Udseth, H. R., Chrisler, W. B., Thrall, B. D., and Smith, R. D. (2000). Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and ^{15}N-metabolic labeling. *Anal. Chem.* **73,** 2132–2139.
27. Shen, Y., Zhao, R., Belov, M. E., Conrads, T. P., Anderson, G. A., Tang, K., Pasa-Tolic, L., Veenstra, T. D., Lipton, M. S., Udseth, H. R., and Smith, R. D. (2001). Packed capillary reversed-phase liquid chromatography with high-performance electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for proteomics. *Anal. Chem.* **73,** 1766–1775.
28. Kim, T., Tolmachev, V., Harkewicz, R., Prior, D. C., Anderson, G. A., Udseth, H. R., Smith, R. D., Bailey, T. H., Rakov, S., and Futrell, J. H. (2000). Design and implementation of a new electrodynamic ion funnel. *Anal. Chem.* **72,** 2247–2255.
29. Belov, M. E., Gorshkov, M. V., Udseth, H. R., Anderson, G. A., Tolmachev, A. V., Prior, D. C., Harkewicz, R., and Smith, R. D. (2000). Initial implementation of an electrodynamic ion funnel with FTICR mass spectrometry. *J. Am. Soc. Mass Spectrom.* **11,** 19–23.
30. Belov, M. E., Gorshkov, M. V., Udseth, H. R., Anderson, G. A., and Smith, R. D. (2000). Zeptomole-sensitivity electrospray ionization-Fourier transform ion cyclotron resonance. *Anal. Chem.* **72,** 2271–2279.
31. Belov, M. E., Nikolaev, E. N., Anderson, G. A., Auberry, K. J., Harkewicz, R., and Smith, R. D. (2001). Electrospray ionization-Fourier transform ion cyclotron mass spectrometry using ion pre-selection and external accumulation for ultra-high sensitivity. *J. Am. Soc. Mass Spectrom.* **12,** 38–48.
32. Marshall, A. G., Hendrickson, C. L., and Jackson, G. S. (1998). Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spectrom. Rev.* **17,** 1–35.
33. Jensen, P. K., Pasa Tolic, L., Anderson, G. A., Horner, J. A., Lipton, M. S., Bruce, J. E., and Smith, R. D. (1999). Probing proteomes using capillary isoelectric focusing-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **71,** 2076–2084.
34. Bruce, J. E., Anderson, G. A., Brands, M. D., Pasa-Tolic, L., and Smith, R. D. (2000). Obtaining more accurate FTICR mass measurements without internal standards using multiply charged ions. *J. Am. Soc. Mass Spectrom.* **11,** 416–421.
35. Li, L., Masselon, C., Anderson, G. A., Pasa-Tolic, L., Lee, S.-W., Shen, Y., Zhao, R., Lipton, M. S., Conrads, T. P., Tolic, N., and Smith, R. D. (2001). High-throughput peptide identification from protein digests using data-dependent multiplexed tandem FTICR mass spectrometry coupled with capillary liquid chromatography. *Anal. Chem.* **73,** 3312–3322.
36. Yates, J. R., Eng, J. K., McCormack, A. L., and Schieltz, D. (1995). Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **67,** 1426–1436.
37. Spahr, C. S., Susin, S. A., Bures, E. J., Robinson, J. H., Davis, M. T., McGinley, M. D., Kroemer, G., and Patterson, S. D. (2000). Simplification of complex peptide mixtures for proteomic analysis: Reversible biotinylation of cysteinyl peptides. *Electrophoresis* **21,** 1635–1650.

38. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999.
39. Gygi, S. P., Rist, B., and Aebersold, R. (2000). Measuring gene expression by quantitative proteome analysis. *Curr. Opin Biotechnol.* **11,** 396–401.
40. Belov, M. E., Nikolaev, E. N., Harkewicz, R., Masselon, C., Alving, K., and Smith, R. D. (2001). Ion discrimination during ion accumulation in a quadrupole interface external to a Fourier transform ion cyclotron resonance mass spectrometer. *Int. J. Mass Spectrom.* **208,** 205–225.
41. Belov, M. E., Gorshkov, M. V., Alving, K., and Smith, R. D. (2001). Optimal pressure conditions for unbiased external ion accumulation in a 2D RF-quadrupole for FTICR mass spectrometry. *Rapid Commun. Mass Spectrom.* **15,** 1988–1996.
42. Shen, Y., Tolic, N., Zhao, R., Pasa-Tolic, L., Li, L., Berger, S. J., Harkewicz, R., Anderson, G. A., Belov, M. E., and Smith, R. D. (2001). High-throughput proteomics using high efficiency multiple-capillary liquid chromatography with on-line high performance ESI FTICR mass spectrometry. *Anal. Chem.* **73,** 3011–3021.
43. Tolmachev, A. V., Udseth, H. R., and Smith, R. D. (2000). The charge capacity limitations of radio frequency ion guides in their use for improved ion accumulation and trapping in mass spectrometry. *Anal. Chem.* **72,** 970–978.
44. White, O., Eisen, J. A., Heidelberg, J. F., Hickey, E. K., Peterson, J. D., Dodson, R. J., Haft, D. H., Gwinn, M. L., Nelson, W. C., Richardson, D. L., Moffat, K. S., Qin, H., Jiang, L., Pamphile, W., Crosby, M., Shen, M., Vamathevan, J. J., Lam, P., McDonald, L., Utterback, T., Zalewski, C., Makarova, K. S., Aravind, L., Daly, M. J., Minton, K. W., Fleischmann, R. D., Ketchum, K. A., Nelson, K. E., Salzberg, S., Smith, H. O., Venter, J. C., and Fraser, C. M. (1999). Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286,** 1571–1577.
45. Makarova, K. S., Aravind, L., Wolf, Y. I., Tatusov, R. L., Minton, K. W., Koonin, E. V., and Daly, M. J. (2001). Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomes. *Microbiology* **65,** 44–79.
46. Sharp, P., and Li, W. (1987). The codon adaptation index: A measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15,** 1281–1295.
47. Emmert-Buck, M. R., Strausberg, R. L., Krizman, D. B., Bonaldo, M. F., Bonner, R. F., Bostwick, D. G., Brown, M. R., Buetow, K. H., Chuaqui, R. F., Cole, K. A., Duray, P. H., Englert, C. R., Gillespie, J. W., Greenhut, S., Grouse, L., Hillier, L. W., Katz, K. S., Klausner, R. D., Kuznetzov, V., Lash, A. E., Lennon, G., Linehan, W. M., Liotta, L. A., Marra, M. A., Munson, P. J., Omstein, D. K., Prabhu, V. V., Prange, C., Schuler, G. D., Soares, M. B., Tolstoshev, C. M., Vocke, C. D., and Waterston, R. H. (2000). Molecular profiling of clinical tissue specimens. *Am. J. Pathol.* **156,** 1109–1115.
48. Pasa-Tolic, L., Jensen, P. K., Anderson, G. A., Lipton, M. S., Peden, K. K., Martinovic, S., Tolic, N., Bruce, J. E., and Smith, R. D. (1999). High throughput proteome-wide precision measurements of protein expression using mass spectrometry. *J. Am. Chem. Soc.* **121,** 7949–7950.
49. Oda, Y., Huang, K., Cross, F. R., Cowburn, D., and Chait, B. T. (1999). Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96,** 6591–6596.

50. Wang, P., and Schellhorn, H. (1995). Induction of resistance to hydrogen peroxide and radiation in *Deinococcus radiodurans*. *Can. J. Microbiol.* **41,** 170–176.
51. Carbonneau, M., Melin, A., Perromat, A., and Clerc, M. (1989). The action of free radicals on *Deinococcus radiodurans* carotenoids. *Arch. Biochem. Biophys.* **275,** 244–251.
52. Goshe, M., Conrads, T., Panisko, E., Angell, N., Veenstra, T., and Smith, R. (2001). Phosphoprotein isotope-coded affinity tag approach for isolating and quantitating phosphopeptides in proteome-wide analyses. *Anal. Chem.* **73,** 2578–2586.
53. Oda, Y., Nagasu, T., and Chait, B. (2001). Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nat. Biotechnol.* **19,** 379–382.
54. Zhou, H., Watts, J., and Aebersold, R. (2000). A quantitative method for proteome-wide analysis of phosphorylation. *In* ''Proceedings of the 48th ASMS Conference on Mass Spectrometry and Allied Topics,'' Long Beach, CA, June 2000.
55. Tolmachev, A. V., Udseth, H. R., and Smith, R. D. (2000). Radial stratification of ions as a function of mass to charge ratio in collisional cooling radio frequency multipoles used as ion guides or ion traps. *Rapid Commun. Mass Spectrom.* **14,** 1907–1913.
56. Belov, M. E., Anderson, G. A., Angell, N. H., Shen, Y., Tolic, N., Udseth, H. R., and Smith, R. D. (2001). Dynamic range expansion applied to mass spectrometry based on data-dependent selective ion ejection in capillary liquid chromatography Fourier transform ion cyclotron resonance for enhanced proteome characterization. *Anal. Chem.* **73,** 5052–5060.
57. Pasa-Tolic, L., Harkewicz, R., Anderson, G. A., Tolic, N., Shen, Y., Zhao, R., Thrall, B. D., Masselon, C., and Smith, R. D. (2001). Increased proteome coverage based upon high performance separations and DREAMS FTICR mass spectrometry. *Anal. Chem.* (Submitted).
58. Gorshkov, M. V., Masselon, C., Anderson, G. A., Udseth, H. R., and Smith, R. D. (2001). Dynamically assisted gated trapping for FTICR mass spectrometry. *Rapid Commun. Mass Spectrom.* **15,** 1558–1561.
59. Davis, M. T., Stahl, D. C., Hefta, S. A., and Lee, T. D. (1995). A microscale electrospray interface for on-line, capillary liquid chromatography tandem mass spectrometry of complex peptide mixtures. *Anal. Chem.* **67,** 4549–4556.
60. Davis, M. T., and Lee, T. D. (1998). Rapid protein identification using a microscale electrospray LC/MS system on an ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **9,** 194–201.
61. Goodlett, D. R., Wahl, J. H., Udseth, H. R., and Smith, R. D. (1993). Reduced elution speed detection for capillary electrophoresis mass-spectrometry. *J. Microcol. Sep.* **5,** 57–62.
62. Martin, S. E., Shabanowitz, J., Hunt, D. F., and Marto, J. A. (2000). Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-EST Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **72,** 4266–4274.
63. Smith, R. D., Wahl, J. H., Goodlett, D. R., and Hofstadler, S. A. (1993). Capillary electrophoresis/mass spectrometry. *Anal. Chem.* **65,** A574–A584.
64. Masselon, C., Anderson, G. A., Harkewicz, R., Bruce, J. E., Pasa-Tolic, L., and Smith, R. D. (2000). Accurate mass multiplexed tandem mass spectrometry for high-throughput polypeptide identification from mixtures. *Anal. Chem.* **72,** 1918–1924.
65. Avery, M., and Fouda, H. (1991). ''Characterization of Impurities in Synthetic Peptides by Atmospheric Pressure Ionization MS.'' Paper presented at the 39th ASMS Conference on Mass Spectrometry and Allied Topics Nashville, TN.

CURRENT STRATEGIES FOR QUANTITATIVE PROTEOMICS

By THOMAS P. CONRADS, HALEEM J. ISSAQ, AND VAN M. HOANG

Biomedical Proteomics Program, SAIC-Frederick, National Cancer Institute at Frederick, Frederick, Maryland 21702

I. Introduction

A major trend underlying current biological research is the development and application of analytical methods capable of making global measurements of entire cellular systems. These advances have created unique opportunities in the field of medicine, where the results from gene expression studies are expected to help identify cellular alterations associated with disease etiology, progression, outcome, and response to therapy. This revolution, driven largely by the rapid and highly successful genome sequencing efforts, is quickly moving biological research into a ''postgenomic era'' involving the explicit study of complex biological ''systems.'' A major goal therein lies to obtain a greater understanding of the function of proteins in a cellular context, as well as their more conventionally delineated molecular function. This global view will provide a greater understanding of the cellular responses to events that include cell division, differentiation, respiration, hormonal signaling, and changes in homeostasis. In addition, an understanding is gained of where signaling and metabolic pathways converge to form networks. Such proteins are prime candidates whose activity can be targeted for drug development, gene therapy, genetic manipulations, and so on. The increased global understanding of cellular systems gleaned from studies made possible by applying these new technologies will offer an assortment

ADVANCES IN
PROTEIN CHEMISTRY, Vol. 65
0065-3233/03 $35.00

of information and opportunities to guide the ''single gene'' approaches presently dominating conventional biological research. The promise of moving to more global measurements of biological systems is to establish new diagnostic approaches and therapeutic targets for a host of maladies, including infectious diseases, behavioral disorders, developmental defects, neurodegenerative diseases, aging, and cancer.

Although the availability of complete genome sequences opens the door to important biological advances, much of the real understanding of cellular systems and the roles of its constituents will necessarily be based on proteomics, which is nothing more than the study of the entire complement of proteins, and their modifications, expressed by a given cell [1]. The capability to precisely measure changes in the expression of all proteins and their modifications would simultaneously provide for an understanding of the function of the proteins participating in manifold pathways and would provide insights into how cellular networks interrelate. When combined with the capability to model and simulate cellular systems (and ultimately predictive capabilities will be developed), proteomics should provide the basis for attacking some of the problems least understood in biology.

A necessary component to obtaining information that contributes to a global view of cellular processes is the ability to make high-throughput measurements. Because cellular proteomes are dynamic, proteomics must be high throughput to enable proteomic measurements of cells under many different conditions in a facile, yet precise, manner (i.e., minutes or hours). Until now, however, analytical techniques have fallen far short of the requirements needed to provide the means necessary for conducting high-throughput proteome measurements. The predominant proteomic methodology has historically been a two-step process involving two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) to fractionate proteins based on isoelectric point and molecular mass, followed by mass spectral identification of resolved spots on the gel [2]. Although 2D-PAGE separations have demonstrated the ability to provide a detailed view of potentially thousands of proteins expressed by a cell [3], it remains a relatively slow and labor-intensive technology. Even after many years of development, it remains doubtful whether 2D-PAGE technology can overcome its inherent limitations to meet the demands for conducting truly global and high-throughput proteomic measurements.

At the heart of proteomics is the characterization of proteins. Although this is not a new concept, the ability of proteomic technologies to characterize a large number of proteins in a complex mixture within a single experiment is what makes this area so novel. Past studies using classic protein characterization techniques such as Western blotting and

enzyme-linked immunosorbent assays (ELISAs) were designed to focus on a single protein at a time. In addition, these studies generally would focus on only a single aspect of that protein, whether it be, for example, its identify, quantity, modification state, localization, or structure. Of these, the identification and quantitation of a particular species are two of the most fundamental issues in the field of protein chemistry. For example, many cell biology experiments are designed to measure the quantity of a known protein as a result of some treatment to the cell in which it is expressed. This information is then used to develop a hypothesis as to which particular pathway or network is affected as a result of treating the cell. The initial discovery of an up- or downregulated protein is used as a guide to find other differentially regulated proteins. The discovery of other differentially regulated proteins is then accomplished using either a hypothesis- or a discovery-driven approach. When using a hypothesis-driven approach, classic approaches are employed to examine the expression levels of other proteins that may function within the same pathway as the initially characterized protein. This result provides the investigator evidence to ascertain precisely which pathway is affected and how the signal may be propagated. Using a discovery-driven approach, the identified protein can be used as the sole, or partial, basis for an affinity method to copurify associated proteins. This approach allows for the potential identification of novel proteins without any previous assumptions about the role of the initially characterized protein.

The promise of proteomics lies in its potential to identify and quantify large numbers of proteins in single experiments. The ultimate goal is a capability that allows for a broad characterization of the expression levels of proteins from two distinct cell populations so that the activated pathways can be readily determined. This vision is probably the major focus of proteomics today and, along with improving protein identification methods, has been the fastest developing areas in proteomics. Unlike techniques such as Western blotting and ELISAs that can use standards to generate a calibration curve, proteome technologies do not provide an absolute measurement of the abundance of a particular protein. In almost all quantitative proteome studies, the relative abundance of a protein extracted from two sources is what is reported. Therefore, the data reflect whether a protein is up- or downregulated and by how much. To establish an absolute quantitation would require the use of internal standards of known concentration, an impossible scenario considering the discovery-driven nature of proteomic studies and the complexity of the samples being analyzed.

There are two primary techniques to measure the relative abundances of proteins from distinct cell populations. These techniques make use

of two of the primary tools in proteomics today, 2D-PAGE and mass spectrometry (MS). The use of 2D-PAGE in quantitative proteomics is relatively straightforward. Protein extracts from two distinct cell populations are separated on two different 2D-PAGE gels. The gels are then stained with a visualization agent such as Coomassie blue or silver staining. Changes in relative abundances are then determined by comparing the spot intensity of a particular protein on the two gels. The second technique relies on MS to measure relative abundances of proteins. This measurement is commonly accomplished through the use of stable isotopes to differentially label proteins from different sources. When analyzed in a single experiment, a protein or peptide that is present in both samples will give rise to two distinct peaks representing the isotopically distinct forms of each species. The relative abundance of the protein between the two samples can then be measured by comparing the peak areas of the two signals. These methods represent the major focus of quantitative proteomics and although they are still in many ways being developed, they have already produced many exciting results.

II. Quantitative Two-Dimensional Polyacrylamide Gel Electrophoresis

Two-dimensional gel electrophoresis is an important separation technique for proteomics research, whereby qualitative as well as quantitative protein expressions can be investigated. An essential aspect of accurate quantitation of proteins using 2D-PAGE is the ability to resolve all the proteins in a mixture, to be able to detect them, and then to apply quantitative measurement procedures. Two different approaches are used for the optimization of 2D electrophoretic separations: (1) the use of narrow pH gradients with overlapping intervals; or (2) the generation of wide-range pH gradients with extended separation distance [4]. The error in quantitative measurements of protein separated by 2D-PAGE may be due to poor solubilization, incomplete resolution, and staining, in addition to other experimental errors such as human error, instrumental errors, and blotting. The following procedures have been used for the detection of proteins: organic dye, silver stain, Coomassie Brilliant Blue, radiolabeling, fluorescent stain, chemiluminescent stain, and mass spectrometry. To detect protein differences between two samples, normal and diseased states, two 2D gels are run and compared visually or instrumentally. Many procedures have been used to quantify proteins resolved by 2D-PAGE. Three main approaches to differential display proteomics are currently being pursued: difference gel electrophoresis,

multiplexed proteomics, and isotope-coded affinity tagging. These and other methods of protein quantitation are discussed in this review.

A. *Visualization*

Because proteins are colorless (in general), a staining procedure, as mentioned above, is used to visualize the protein spots, mostly abundant proteins and polypeptides, on the 2D gel. The intensity of the spots is then compared visually. This detection procedure can be used to quantitatively identify differences in the levels of individual proteins following modulation of a tissue or cell and would give an estimate, not an accurate measurement, of the concentration levels of two matched proteins on two different gels from two different protein mixtures.

B. *Densitometry*

Densitometry is an instrumental technique that is more accurate than visual inspection by the naked eye. In densitometry the resolved spots on the 2D gel are scanned and their densities are determined with a densitometer. The density of the spots is directly proportional to protein concentration. Alaiya *et al.* [5] used 2D electrophoresis and a laser densitometer for the analysis of protein expression in ovarian tumors. A sample set of approximately 400 gel spots was quantified from each gel. Peaks for the protein spots were located and counted. The individual polypeptide quantities were expressed as parts per million (ppm) of the total integrated optical density. The total spot counts and the total optical density are directly related to the protein concentration.

C. *Difference Gel Electrophoresis*

Difference gel electrophoresis (DIGE) is a modification of 2D-PAGE that requires only a single gel to reproducibly detect differences between two protein samples. DIGE circumvents the reproducibility problems associated with comparing two different 2D gels. The principle of the method is simple. Two different protein mixtures are labeled with two different fluorophores, 1-(5-carboxypentyl)-1′-propylindocarbocyanine halide (Cy3) *N*-hydroxysuccinimidyl ester and 1-(5-carboxypentyl)-1′-methylindodicarbocyanine halide (Cy5) *N*-hydroxysuccinimidyl ester fluorescent dyes, respectively. The labeled proteins are mixed and separated in the same 2D gel, thus enabling the analyst to run two different samples on the same gel in a 2D

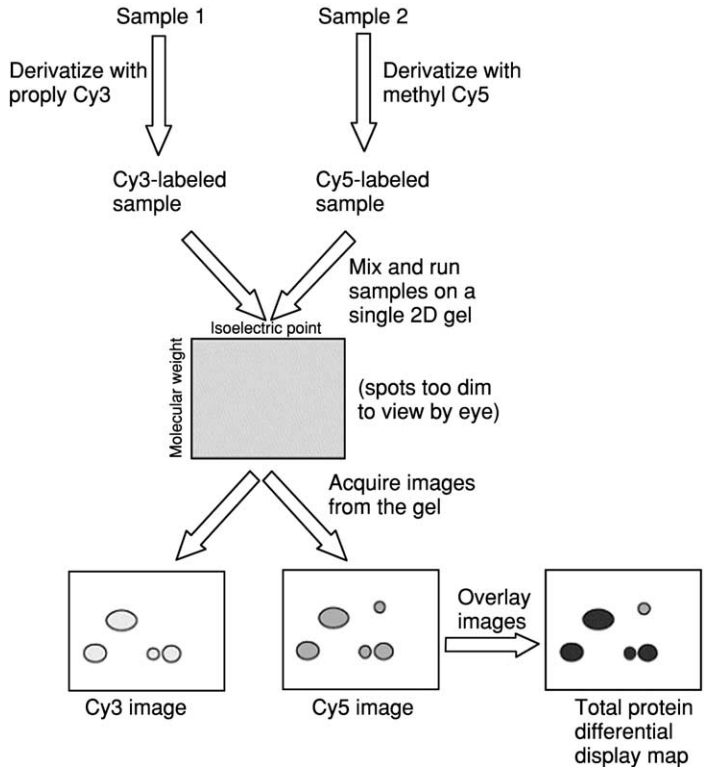


FIG. 1. Schematic illustration of the difference gel electrophoresis (DIGE) technology platform. [Reprinted with permission from reference 5.]

format, as shown in Fig. 1. Protein spots are then detected by fluorescence imaging immediately after electrophoresis, using a dual-laser scanning device or xenon arc-based instrument with different excitation/emission filters in order to generate two separate images. The images are then matched by a computer-assisted overlay method, signals are normalized, and spots are quantified. Differences in expression are identified by a pseudo-colored image and data spreadsheet. DIGE can maximally evaluate three different samples using three different fluorophores [6, 7] with a sensitivity equal to silver staining [8].

2D DIGE was applied to quantify the differences in protein expression between laser capture microdissection-procured esophageal carcinoma cells and normal epithelial cells and to define cancer-specific and normal-specific protein markers [9]. Analysis of the 2D images from protein lysates of ~250,000 cancer cells and normal cells identified 1038 protein spots in

cancer cell lysates and 1088 protein spots in normal cell lysates. Of the detected proteins, 58 spots were up regulated by >3-fold and 107 were down regulated by >3-fold in cancer cells. Global quantification of protein expression between laser capture-microdissected patient-matched cancer cells and normal cells, using 2D DIGE in combination with mass spectrometry, is a powerful tool for the molecular characterization of cancer progression and identification of cancer-specific protein markers [9].

D. Multiplexed Proteomics

The principle of multiplexed proteomics (MP) is easy to understand and the procedure is easy to perform (Fig. 2). It is designed to allow the parallel determination of protein expression levels as well as certain attributes of the proteins. Unlike DIGE, in which two different samples are run on the same 2D gel, in MP two different samples are run on two different gels and then stained for a particular functionality, for example glycosylation, after which the two gels are imaged. Gels are then stained for total protein expression, using SYPRO Ruby stain, and imaged again. The images resulting from both stains are matched by computer-assisted rubber sheeting and overlay methods. If needed, the signals are normalized and the spots are then quantified. Differences in glycosylation and protein expression are identified and recorded. MP technology can be used to evaluate an almost limitless number of 2D gels with respect to two or three different attributes, protein expression, posttranslational modification, drug-binding capability, and so on [8]. MP technology was used to assay for the presence of glycosylated proteins, followed by the detection of the total protein profile [10]. The procedure is as follows: Pro-Q Emerald 300 dye-labeled gels and blots are poststained with SYPRO Ruby dye, allowing sequential two-color detection of glycosylated and nonglycosylated proteins. Both fluorophores are excited with midrange ultraviolet (UV) illumination. Pro-Q Emerald 300 dye maximally emits at 530 nm (green) whereas SYPRO Ruby dye maximally emits at 610 nm (red). As little as 300 pg of α_1-acid glycoprotein (40% carbohydrate) and 1 ng of glucose oxidase (12% carbohydrate) or avidin (7% carbohydrate) are detectable in gels after staining with Pro-Q Emerald 300 dye. Besides glycoproteins, as little as 2–4 ng of lipopolysaccharide is detectable in gels with Pro-Q Emerald 300 dye whereas 250–1000 ng is required for detection by conventional silver staining. Detection of glycoproteins may be achieved in sodium dodecyl sulfate–polyacrylamide gels, 2D gels, and on polyvinylidene difluoride membranes [10].

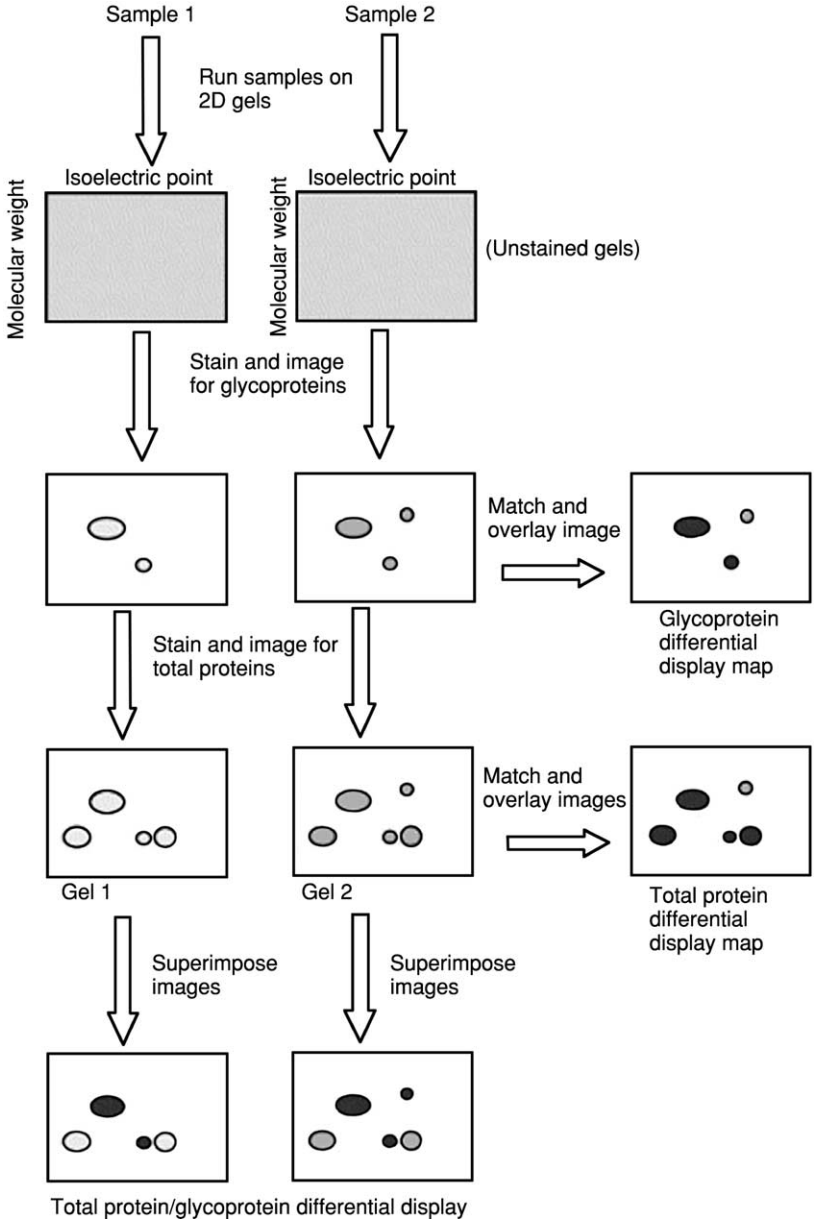


FIG. 2. Schematic illustration of the multiplexed proteomics (MP) technology platform. [Reprinted with permission from reference 5.]

E. Protein Mapping

Protein expression mapping may be defined as the quantitative study of global changes in protein expression in cells and tissue extracts by 2D gel electrophoresis image analysis. The advantage of this method is the direct determination of protein abundance and detection of posttranslational modifications that results in shifts in mobility. Because thousands of proteins are imaged on a 2D gel, the analyst can obtain a picture of a cell or tissue protein profile at a given point in time [11]. Goldfarb used 2D electrophoresis and computer imaging for the quantification of human milk casein. Milk samples from 20 mothers were run on 2D gels. A slide was taken of each silver-stained gel with a Kodak control strip, and the slide was scanned into a PowerMac running Photoshop 3. The NIH Image program was used to measure the area and integrated density of the spots. A Kodak control scale provided calibration and conversion to optical density units [12].

F. Chemiluminescence

Chemiluminescent probes offer highly sensitive quantitative analyses of proteins blotted from 2D electrophoretic gels onto nitrocellulose or polyvinylidene supporting matrix and further analyzed with specific antibodies to detect specific protein antigens. Secondary antibodies are used to amplify the signal via chemiluminescent signals. To ensure the quality of the data, and minimize artifacts, an internal standard should be used. Conrad *et al.* [13] described a protocol that uses a standard series of substrate-specific antigens that can be tailored for any chemiluminescent probe to calculate a regression line equation that normalizes differences in light emission that occur between experiments.

III. Metabolic Labeling Applications for Quantitative Proteomics

Although 2D-PAGE combined with MS has formed the basis of many proteomic investigations, there has been a significant shift away from slab gel techniques and toward capillary-based methods to conduct proteome measurements. This shift has been largely fueled by limitations inherent to gel-based methods in the areas of sensitivity, dynamic range, proteome coverage, and throughput. Other articles in this volume adequately cover these limitations. The potential of capillary-based separations coupled directly on-line with MS to provide broad proteome coverage was initially demonstrated by their ability to identify hundreds of proteins in

a high-throughput manner. For example, the MudPIT technology, developed by Yates and co-workers, has shown the capability of identifying hundreds if not thousands of proteins in a single analysis [14]. The broad proteomic identification capabilities of these solution-based techniques combined with MS have led to the development of methods to quantitate the proteins being detected.

Metabolic labeling of proteins with isotopes was developed previous to the proteomics era. For example, the development of the radioimmunoassay to measure the amount of insulin in plasma dates back to 1960 [15]. Radioisotopes have also been used to monitor cellular processes. For example,^{32}P is often used in nucleotide sequencing, phospholipid research, and studies that look for the site of protein phosphorylation [16] whereas ^{125}I can be used to label proteins in metabolic studies. Pulse–chase experiments using ^{35}S-labeled methionine are often done to study protein synthesis. For example, methionine incorporation was used to identify newly synthesized proteins in synaptosomes [17] and to examine the effects of 12-*O*-tetradecanoyl phorbol 13-acetate on the synthesis of lipoprotein lipase [18]. Heavy isotope labeling (i.e., ^{2}H, ^{13}C, or ^{15}N) of proteins has also been used frequently in structural elucidation by nuclear magnetic resonance (NMR) [19, 20].

Stable isotope labeling in quantitative mass spectrometry traditionally has been used in small molecule analyses such as in drug metabolism and pharmacokinetic studies. Application of this concept to protein mass spectrometry includes metabolic labeling of proteins in which stable isotopes are incorporated into proteins in cell culture systems during translation. Examples of this are given in this section and include ^{15}N labeling in cultured cells, isotope depletion in *Escherichia coli*, and specific amino acid labeling using auxotrophic organisms.

The basis of ^{15}N labeling of proteins of microorganisms is as outlined in Fig. 3. First, cells are grown in medium containing a natural abundance of the isotopes of nitrogen or in ^{15}N-enriched medium. The two populations of cells are then harvested and pooled. The proteins are extracted and evaluated by first separating proteins on 2D gels followed by an in-gel trypsin digest of each protein spot and analysis of the peptides by mass spectrometry, or the 2D gel procedure can be omitted. In this second case, the protein mix is digested with trypsin in solution and subjected to liquid chromatography to (LC)-MS. MS is able to distinguish between the two separate pools of proteins because one set contains the natural abundance of nitrogen, whereas proteins from the other set will have ^{15}N incorporated. This will increase the mass of ^{15}N-labeled peptides and, as a result, MS will show paired peaks of peptides. The ratio of this pair of peaks is used for relative quantitation between the two samples. Therefore, MS

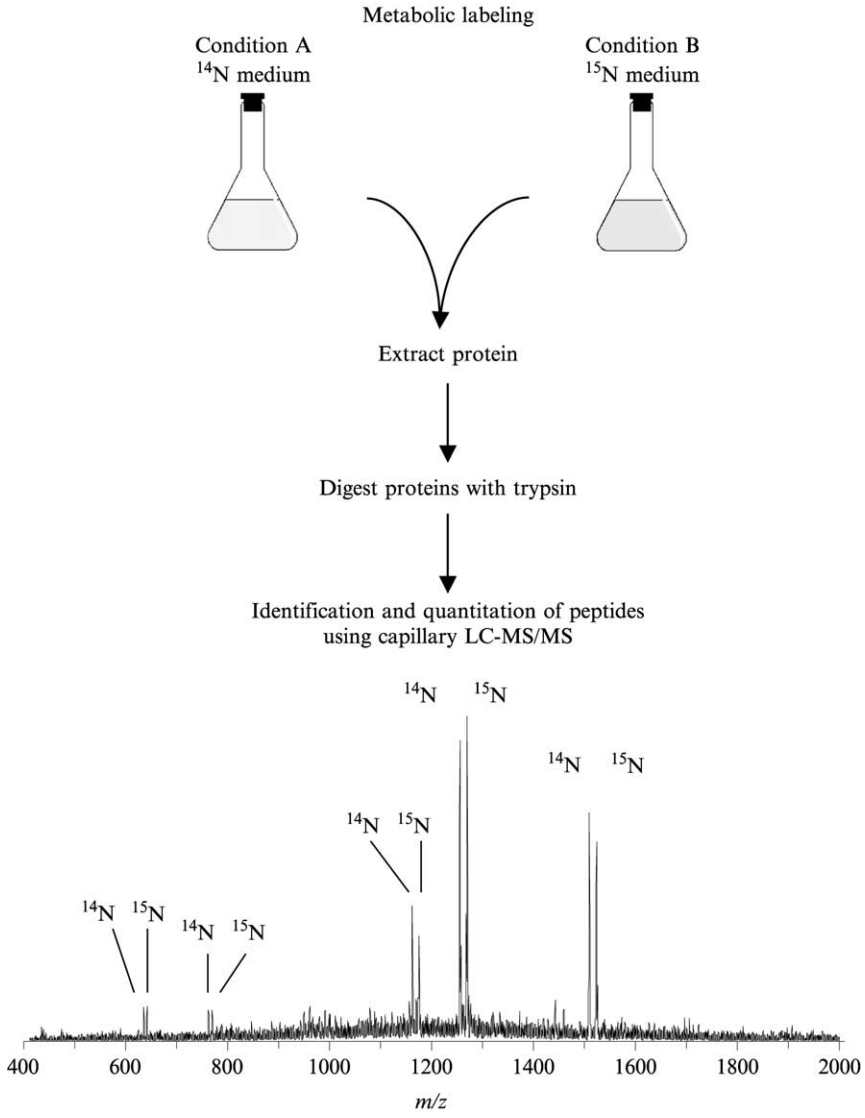


FIG. 3. Schematic representation of the combined ^{15}N metabolic labeling/cysteinyl residue affinity tag approach for quantitative proteome analyses. Equivalent numbers of cells grown under two different conditions are pooled and proteins are extracted from the cells. Iodoacetyl-PEO-biotin is used to label cysteine residues and the proteins are then digested with trypsin. The modified peptides are isolated by affinity chromatography and analyzed by capillary LC-tandem MS.

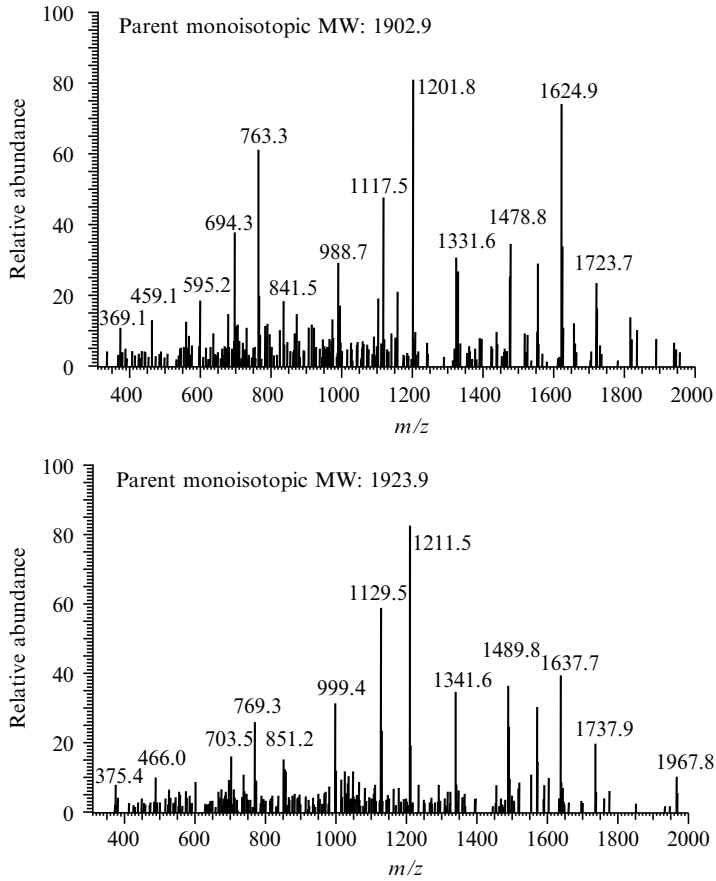


FIG. 4. Comparison of tandem MS of ^{14}N- and ^{15}N-labeled versions of an identical peptide. The tandem MS spectrum of a peptide from the yeast protein arginase extracted from cells grown in normal isotopic abundance medium (*top*) is compared with the tandem MS spectrum of the same peptide extracted from cells cultured in ^{15}N-enriched medium. The spectra are qualitatively similar and identified the same peptide.

analysis is used to compare the relative amounts of each protein in the two different samples and also to identify the protein. Importantly, the tandem MS spectra of ^{14}N and ^{15}N peptides are qualitatively similar. The two MS/MS spectra shown in Fig. 4 both identified the peptide as HETGGLVFFEPILDACR originating from the yeast protein arginase. Both spectra are similar, the most notable difference being the mass shift of the fragments ions originating from the ^{15}N-labeled version of the peptide.

This method has been successfully used in multiple studies. Oda *et al.* first examined the differences in protein expression between wild-type *Saccharomyces cerevisiae* and a mutant unable to express the G_1 cyclin CLN2 [21]. They were able to identify and relatively quantitate changes in the expression of the most abundant proteins between the mutant and wild-type yeast; they reported a quantifiable decrease in triosephosphate isomerase, *S*-adenosylmethionine synthetase 2, and open reading frame (ORF) YLR109w in the mutant.

Conrads *et al.* reported on a variation of this experiment, which utilized (+)-biotinyliodoacetamidyl-3,6-dioxaoctanediamine (iodoacetyl-PEO-biotin), a cysteine-reactive label containing a biotin moiety, in conjunction with ^{15}N labeling [22]. Analogous to the previously described experiment, *Deinococcus radiodurans* was grown in either normal or ^{15}N-enriched medium. The two cell populations were pooled and the proteins were extracted. Next, iodoacetyl-PEO-biotin was added to the protein extract to label cysteinyl residues. The mixture was digested with trypsin and iodoacetyl-PEO-biotin-labeled peptides were affinity purified by avidin affinity chromatography. The labeled cysteinyl peptides were eluted and subjected to analyses by Fourier transform ion cyclotron resonance (FTICR) mass spectrometry, which afforded high resolution and mass accuracy to the measurements. The addition of the cysteine derivatization step allows for the enrichment of cysteine-containing peptides. This both reduces the complexity of the mixture of peptides analyzed and introduces an additional constraint, that is, peptide contains cysteine, into the database searches and interpretation of the spectra to aid in protein identification. This method was also shown to be successful with mammalian cells by using B16 cells, a murine melanoma cell line.

^{15}N labeling has also been used in conjunction with MudPIT (multidimensional protein identification technology) [23]. MudPIT is the use of a biphasic capillary column coupled directly to a tandem mass spectrometer. The biphasic nature of the column is due to it containing both strong cation-exchange and reversed-phase resin. Peptides are applied to the column and are fractionated first by ion exchange and then by reversed-phase chromatography. By using MudPIT, an impressive 1484 proteins were identified from *S. cerevisiae*, including proteins that are traditionally more difficult to observe, such as membrane proteins and transcription factors [1]. By using ^{15}N labeling with MudPIT, the power of multidimensional separations is coupled with relative quantitation by metabolic labeling. The investigators were able to show accurate quantitation with a complex mixture of peptides derived from *S. cerevisiae*.

In contrast to labeling proteins with a heavy isotope, Jensen *et al.* chose to use isotopically depleted medium, that is, medium depleted of ^{13}C, ^{15}N, and so on [24]. This depletion resulted in higher mass accuracy because the broad isotopic envelope of the protein is reduced due to the shift in mass toward the monoisotopic species. Another benefit was the increase in sensitivity as a result of collapsing the peak toward the monoisotopic species. When proteins isolated from *E. coli* cultured on isotopically depleted medium were analyzed by capillary isoelectric focusing (CIEF) and FTICR mass spectrometry, the investigators were able to identify the monoisotopic mass of proteins. This result was in contrast to *E. coli* cultured on normal medium, where the monoisotopic peak of proteins was undetectable. As a demonstration of the increased sensitivity of the method, it was also shown that more proteins were detected from cultures derived from isotopically depleted medium than from normal isotopic abundance medium.

As a final example of metabolic labeling, Veenstra *et al.* grew a multiple auxotrophic strain of *E. coli* (*leuB*, *argH*, *thr*, *his*) on natural isotopic abundance minimal medium supplemented with leucine or heavy isotope-labeled leucine (leucine-d_{10}) [25]. There is a 10-Da mass difference between leucine and leucine-d_{10} as a result of the substitution of 10 hydrogens for 10 deuteriums. The investigators were able to determine the number of leucine residues present in a protein by comparing the mass difference between the protein extracted from the auxotroph grown on medium supplemented with either leucine or leucine-d_{10}. Also, by using CIEF-FTICR, they were able to obtain accurate molecular mass measurements of the protein. By using both of these pieces of information—accurate molecular weight and number of leucine residues—they were able to identify intact proteins from the databases. For example, they identified a protein pair that gave two monoisotopic masses of 7327.5 and 7346.8 Da shown in Fig. 5. They were able to determine that there were two leucines in the protein because of the mass difference of 19 Da. A search of the *E. coli* genome, using the MW 50 Da and leucine composition as search constraints, they were able to identify the protein as the cold shock protein Csp-E after accounting for a cleaved N-terminal methionine.

Using a similar strategy, Chen *et al.* used an *E. coli* auxotroph requiring essential amino acids including glycine and methionine and containing an isopropyl-β-D-thiogalacto pyranoside (IPTG)-inducible plasmid that expresses the ubiquitin-like protein UBL1 with a histidine tag [26]. This strain was grown on M9 medium supplemented with the required amino acids including glycine and methionine. In parallel, the strain was also grown on the same medium, but with glycine-d_2 (Gly-d_2) or with

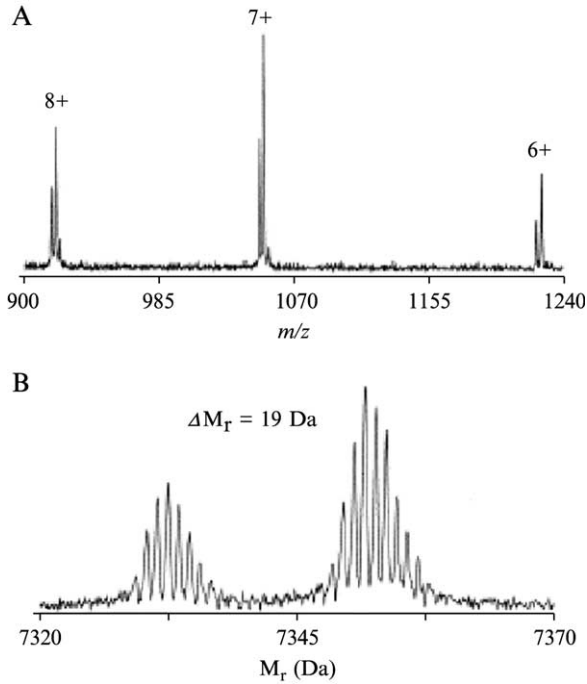


FIG. 5. (A) Mass spectrum and (B) deconvoluted mass spectrum of normal isotopic abundance and leucine-d_{10}-labeled forms of an *E. coli* protein identified as the cold shocklike protein Csp-E observed in the CIEF-FTICR analysis of a mixture of *E. coli* lysate grown in minimal medium containing leucine and minimal medium containing leucine-d_{10}. [Reprinted with permission from reference 25.]

methionine-d_3 (Met-d_3), or with both Gly-d_2 and Met-d_3. After induction, UBL1 was purified. This resulted in four separate populations of UBL1: UBL1 containing natural isotope abundance glycine and methionine, UBL1 with Gly-d_2 incorporated and natural isotope abundance methionine, UBL1 with Met-d_3 and natural isotope abundance glycine, and UBL1 containing Gly-d_2 and Met-d_3. The proteins were digested with trypsin and resulting peptides were mixed at various ratios and analyzed by matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) to demonstrate relative quantitation. Also, partial sequence information was obtained by postsource decay (PSD). Similar to Veenstra *et al.*, the numbers of glycines and methionines were used as an additional constraint along with PSD data and peptide monoisotopic mass to aid in identification.

IV. Chemical Modification Strategies for Quantitative Proteome Measurements

Although stable isotope labeling can be accomplished at the metabolic level during translation, this is necessarily the case only if the system is amenable to growth in a medium enriched for a given isotope. Fortunately, methods have been developed to differently label proteins on a global level after they have been extracted from the cell with distinctive isotopic tags. The most popular example of these methodologies utilizes a reagent referred to as isotope-coded affinity tags (ICAT), which allows global measurements of peptide relative abundances from virtually any source [27].

ICAT reagents can be thought of as possessing three vital chemical moieties: an iodoacetyl functionality that is used to covalently modify the sulfhydryls of reduced cysteinyl residues, an ethylene glycol linker region that can be synthesized to possess either eight protons (d_0) or eight deuterons (d_8), and a biotin moiety that allows for the selective isolation of ICAT-labeled peptides by avidin affinity chromatography. In the application of ICAT labeling, two cell extracts to be compared are each labeled either with the d_0 or the d_8 version of ICAT, combined, and digested, and ICAT-labeled peptides are isolated by avidin affinity chromatography as illustrated in Fig. 6. Hence, ICAT labeling results in differentially stable isotope-labeled ''sister'' cysteinyl-containing polypeptides (Cys-polypeptides) whose masses are separated exactly by the mass of the heavy ICAT label, or approximately 8 Da.

We have developed another strategy for comparing relative abundances between peptides from two proteome samples that combines ^{15}N metabolic labeling and postextraction cysteine affinity tagging to isolate and quantitate Cys-polypeptides analogous to the ICAT strategy [28]. To demonstrate this labeling strategy, proteome samples isolated from equal numbers of *D. radiodurans* or mouse B16 melanoma cells cultured in normal isotopic abundance and ^{15}N-enriched media were labeled with iodoacetyl-PEO-biotin. Iodoacetyl-PEO-biotin contains all of the elements of the light ICAT reagent: a cysteine-specific reactive group with a biotin functionality that can be used to isolate derivatized peptides using immobilized avidin. The samples were then analyzed by LC-FTICR. Pairs of differentially labeled Cys-polypeptides were observed whose mass-to-charge (m/z) ratio differed based on the number of nitrogen atoms in the peptide, as shown in Fig. 7. More than 600 pairs of Cys-polypeptides were observed in the analysis of the *D. radiodurans* proteome. The average ratio of the areas under the differentially isotopically labeled versions of each peptides was ~1.12. This corresponds to the expected ratio because

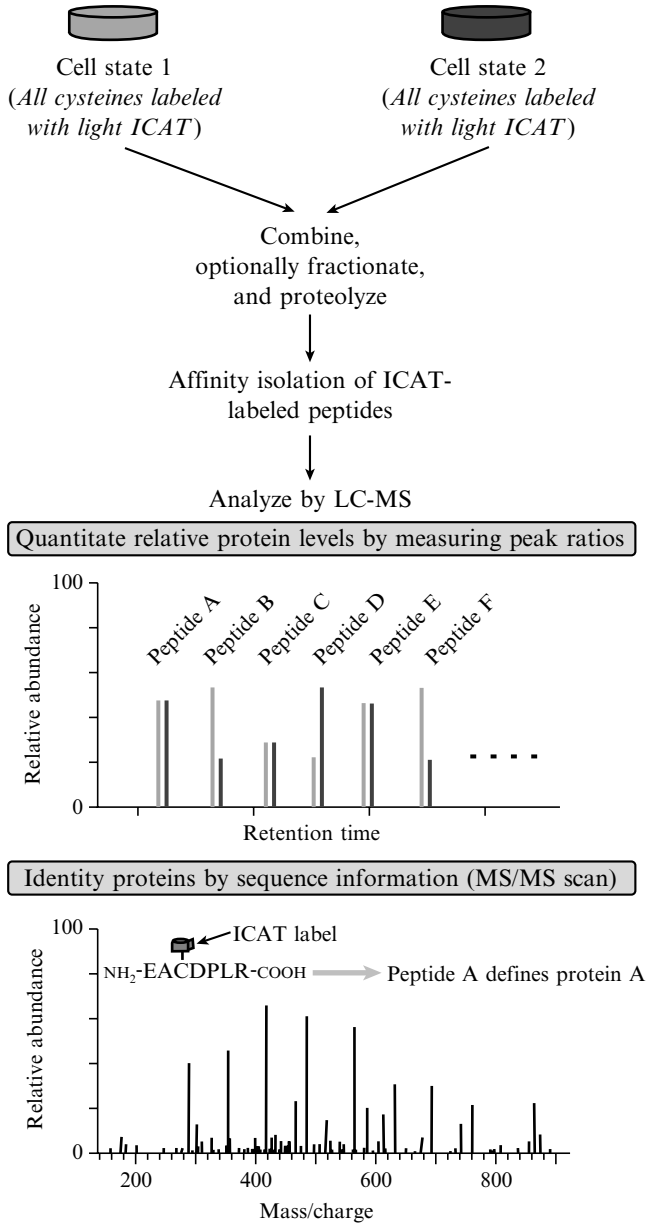


FIG. 6. The ICAT strategy for quantifying differential protein expression.

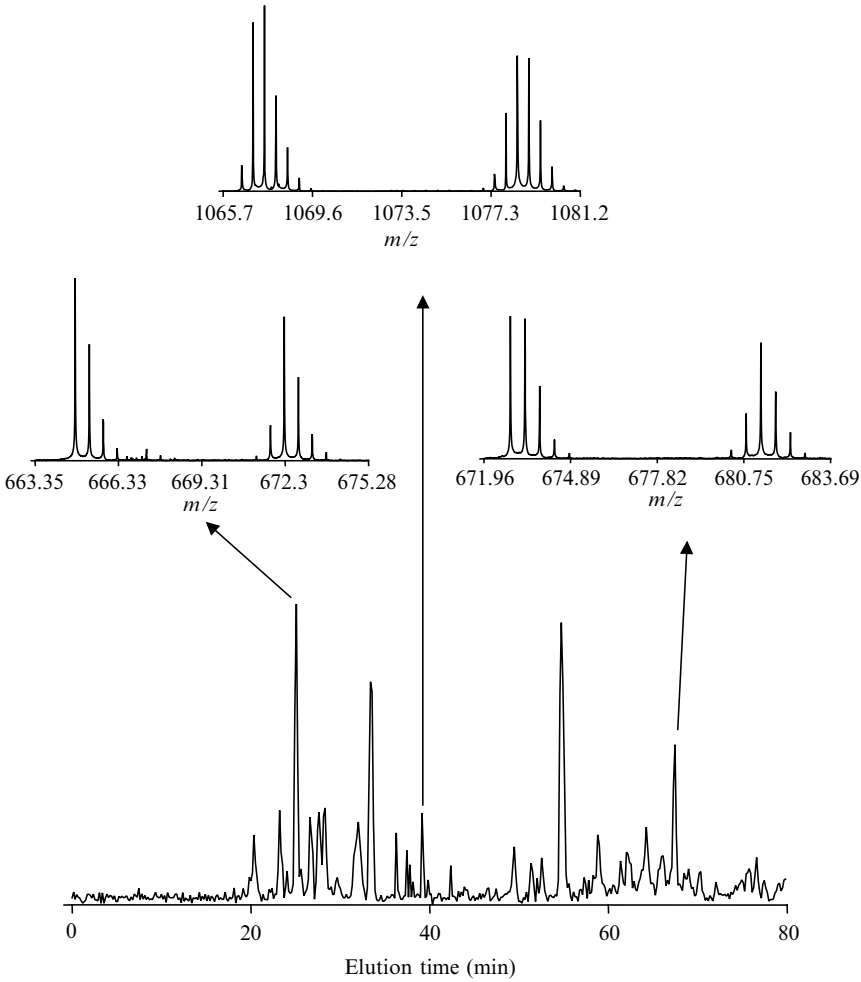


FIG. 7. Examples of Cys-polypeptides observed in the LC-FTICR analysis of peptides isolated from a combined culture of *D. radiodurans* grown in normal and ^{15}N-enriched media. Pairs of related peptides can be detected on the basis of the distinctive isotopic distributions of the ^{15}N-labeled peptide and its normal isotopic abundance partner.

the sample contained equal numbers of cells from the cultures grown in either normal isotopic abundance or ^{15}N-labeled medium. The standard deviation for these experiments was approximately 10% (0.10%). Although the use of the PEO-biotin affinity tag to isolate only

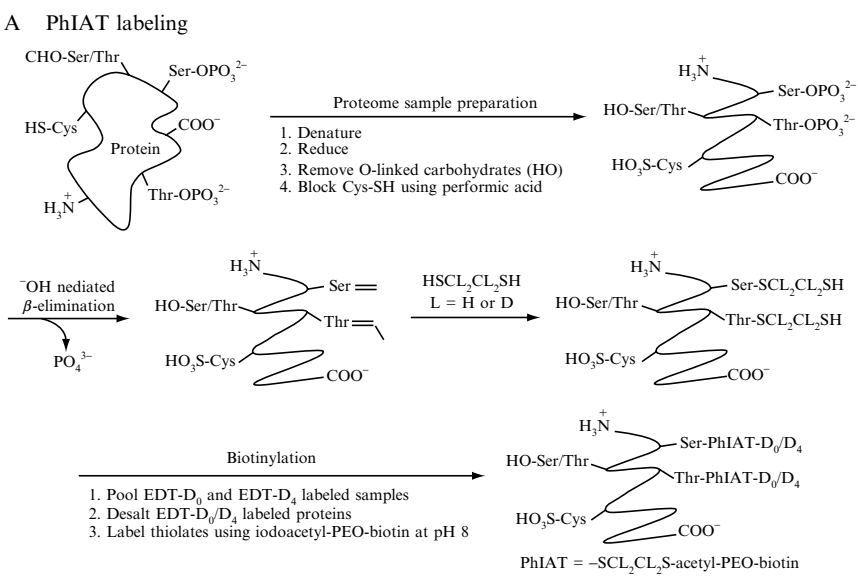
Cys-polypeptides significantly reduces the complexity of the mixture, the proteome samples still contain a formidable number of peptides.

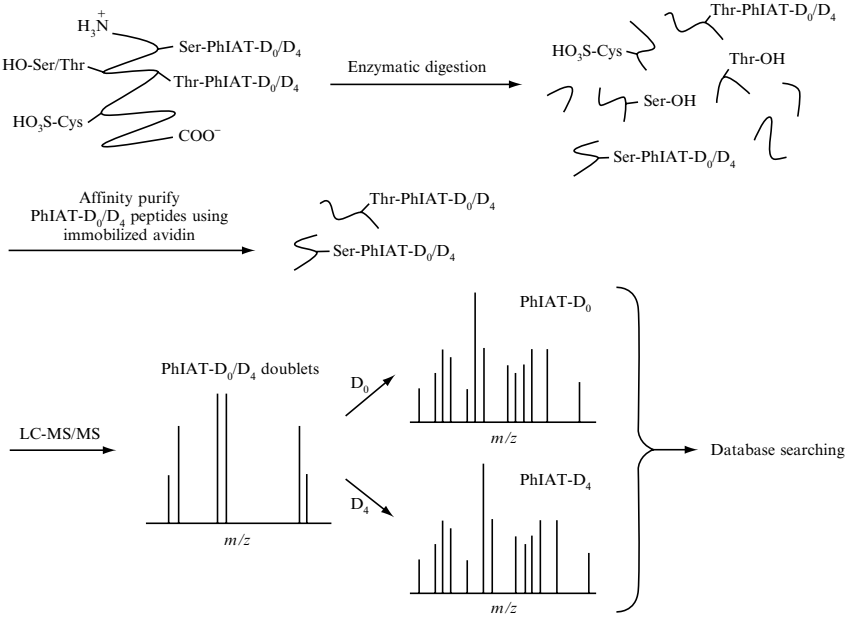
The labeling strategies that use isotopically defined medium offer the highest precision and the broadest proteome coverage. However, as mentioned, metabolic labeling is limited to cells that can be cultured in special medium, whereas postextraction isotopic labeling (ICAT) is universally applicable to proteins extracted from every conceivable source. In addition, postextraction labeling can provide a reduced mixture complexity in some schemes, and can aid identification by providing an additional sequence constraint. Postextraction isotopic labeling does, however, require additional sample processing and necessarily results in decreased protein coverage, and possibly decreases quantitative precision and/or accuracy.

Aside from quantifying protein abundances, proteomic studies also have the opportunity to target posttranslational modifications. Unfortunately, identifying sites of posttranslational modifications is more difficult than simple peptide identification because of the fragility of the protein modification bond and the low stoichiometry of the modification. Although more than 200 different protein modifications have been described [29], arguably the most important to cell function is phosphorylation. Phosphorylation modulates protein activity and propagates signals within cellular pathways and networks [30–32]. Studies estimate that one-third of all mammalian proteins may be phosphorylated [30]. Processes including protein kinase activation, cell cycle progression, cellular differentiation and transformation, development, peptide hormone response, and adaptation are all regulated by changes in the state of protein phosphorylation.

Characterizing phosphorylation events, even for a modest number of proteins, is an analytical challenge. The most prevalent method to study phosphorylation of proteins involves the use of ^{32}P-labeled inorganic phosphate ($^{32}P_i$). In these studies, $^{32}P_i$-labeled proteomes are analyzed by 2D-PAGE, and the relative spot intensities produced by autoradiography are compared to assess the relative extent of phosphorylation [33, 34]. However, this method of detection is not amenable to high-throughput proteomic studies because of the difficulties and regulations for handling radioactive isotopes. Thus, strategies have been developed using stable isotope labeling and mass spectrometric detection.

The use of $^{14}N/^{15}N$ metabolic labeling to study phosphorylation was reported by Chait and co-workers [35]. This method determines the relative quantitation of the phosphorylation state of proteins by measuring the intensity ratio (^{14}N:^{15}N) of the nonphosphorylated and phosphorylated peptide species. In addition to the loss of the phosphate moiety


A PhIAT labeling
CHO-Ser/Thr
Ser-OPO3 2–
HS-Cys
COO–
Protein
Thr-OPO3 2–
H3N+
Proteome sample preparation
1. Denature
2. Reduce
3. Remove O-linked carbohydrates (HO)
4. Block Cys-SH using performic acid
HO-Ser/Thr
HO3S-Cys
–OH nediated
β-elimination
PO4 3–
Ser
Thr
HSCL2CL2SH
L = H or D
Ser-SCL2CL2SH
Thr-SCL2CL2SH
Biotinylation
1. Pool EDT-D0 and EDT-D4 labeled samples
2. Desalt EDT-D0/D4 labeled proteins
3. Label thiolates using iodoacetyl-PEO-biotin at pH 8
Ser-PhIAT-D0/D4
Thr-PhIAT-D0/D4
PhIAT = –SCL2CL2S-acetyl-PEO-biotin


B Enrichment and analysis
Enzymatic digestion
Thr-OH
Ser-OH
Affinity purify
PhIAT-D0/D4 peptides using
immobilized avidin
LC-MS/MS
PhIAT-D0/D4 doublets
D0
D4
m/z
PhIAT-D0
PhIAT-D4
Database searching

during collision-induced dissociation (CID), the presence of other peptides makes analysis of only the phosphopeptides more difficult. Immuno- and metal-affinity columns have been used to enrich mixtures for phosphopeptides [36, 37], but these procedures often result in the isolation of many nonphosphorylated peptides mediated through nonspecific interactions that complicate the downstream analysis by introducing uncertainty about the nature of the sample.

A novel strategy to isolate and quantify phosphopeptides using a combination of stable isotope labeling and MS was developed concurrently, and independently, by two groups [38, 39]. The first step in the reaction involves blocking reactive thiolates via reductive alkylation or performic acid oxidation (Fig. 8). The phosphate groups on the protein are removed from the pSer and pThr residues via hydroxide ion-mediated β elimination, resulting in their conversion to dehydroalanyl and β-methyl dehydroalanyl residues, respectively. The next modification involves a Michael-type addition of the bifunctional reagent 1,2-ethanedithiol (EDT), creating a free thiolate group at the site of the former phosphate group. This thiolate is then covalently modified with iodoacetyl-PEO-biotin, a thiol-reactive molecule containing a free biotin group. The net result is the covalent modification of phosphoryl residues with a linker molecule that contains a terminal biotin group and a stable isotopic label that enables relative quantification of the phosphorylation state of the protein or peptide. This isotopic labeling and quantitation is achieved by using commercially available sources of either the light ($HSCH_2CH_2SH$ or EDT-D_0) or the heavy ($HSCD_2CD_2SH$ or EDT-D_4) isotopic version of EDT. After modification, the samples are combined and enzymatically digested and the modified peptides are specifically extracted by immobilized avidin chromatography. These extracted peptides are then ready to be analyzed by reversed-phase liquid chromatography (LC) coupled directly on-line with MS.

The modified peptides are identified and quantitated using both MS and MS/MS strategies. In the MS mode, the masses of the intact peptides are measured and the peak intensity of the individual species is used to quantitatively measure the phosphorylation status of the peptide. The

FIG. 8. Illustration of the use of phosphoprotein isotope-coded affinity tags for quantitative analysis of the phosphoproteome. In the initial steps (A) phosphoseryl and phosphothreonyl residues are labeled with the PhIAT reagents and in the second phase (B) the PhIAT-labeled peptides are enriched by avidin affinity chromatography. The enriched mixture is analyzed by capillary reversed-phase LC-tandem MS and peptides are identified by subsequent database searching of the CID spectra. [Reprinted with permission from reference 39.]

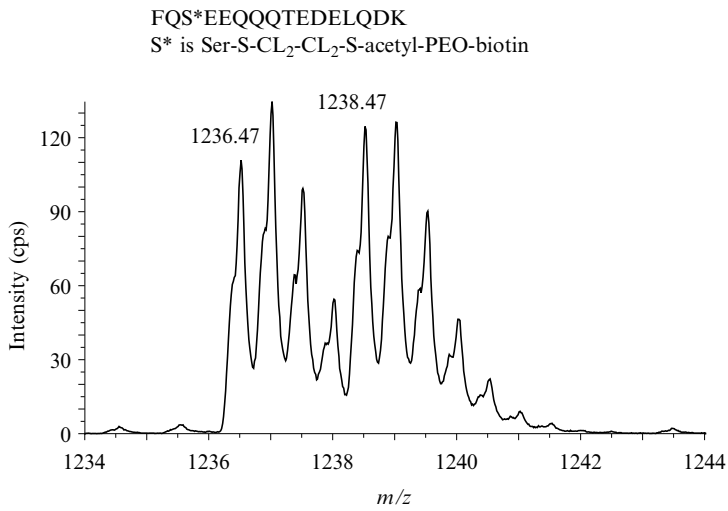


FIG. 9. Mass spectra of PhIAT-D_0/D_4-labeled β-casein peptides. The enriched mixture of biotinylated phosphopeptides was analyzed by capillary reversed-phase liquid chromatography coupled directly on-line to a PE Sciex API QStar Pulsar hybrid quadrupole-TOF mass spectrometer using approximately 1.0 and 0.5 $\mu g/\mu l$ of overall sample concentration per injection, respectively. The $[M+2H]^{2+}$ ion pair corresponds to the mass of the PhIAT-D_0/D_4-derivatized β-casein phosphopeptide FQS*EEQQQTEDELQDK, where S* has the modified side chain-CH_2-SCL_2CL_2S-acetyl-PEO-biotin and L contains either H (PhIAT-D_0 label) or D (PhIAT-D_4 label). [Reproduced with permission from reference 39.]

pairs of peaks originating from the modified versions of the phosphopeptides are easily recognized because they are separated by 4.0 Da (i.e., the mass difference between the EDT-D_0 and EDT-D_4 labels). For example, the mass spectrum of the modified phosphorylated peptide FQpSEEQQQTEDELQDK from β-casein is shown in Fig. 9 [38]. The 2.0-*m/z* difference between the $[M+2H]^{2+}$ ions at 1236.96 and 1238.97 *m/z* corresponds to the expected 4-Da mass difference. Although most phosphopeptide pairs can be anticipated to be separated by 4.0 Da, there are invariably cases in which multiple phosphorylation sites exist within a single peptide and these pairs will be separated by some integral value equal to the number of phosphorylation sites times 4.0 Da.

Although this method provides a means to quantitate changes in the phosphorylation state of a protein it also allows the specific phosphorylation site to be determined. In a typical MS/MS analysis of phosphopeptides, the phosphate group can readily dissociate from the peptide during MS/MS, preventing the site-specific assignment of the phosphate

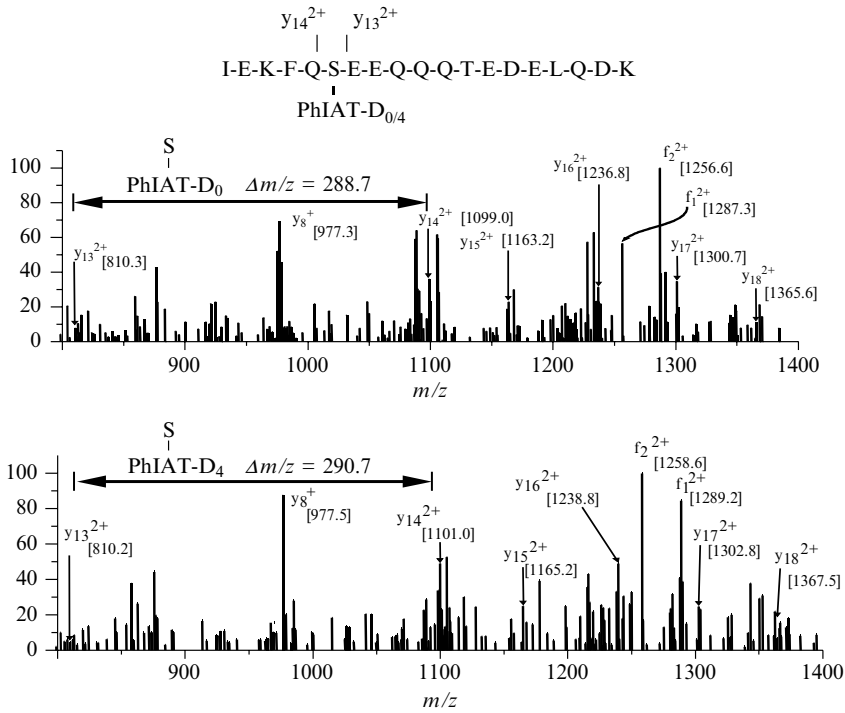


FIG. 10. Tandem mass spectrometry identification of PhIAT-modified phosphopeptides. Tandem MS/MS spectra of a phosphopeptide modified with the light (*top*) and heavy (*bottom*) isotopic versions of EDT and iodoacetyl-PEO-biotin are shown. Retention of the modification on the serine residue during MS/MS allows the specific site of phosphorylation to be determined. [Reproduced from reference 23 with permission.]

modification. During MS/MS the covalent modification remains attached to the residue as shown in Fig. 10 [40]. In this example the MS/MS spectra for the EDT-D_0- and EDT-D_4-modified versions of a peptide are shown. Within this peptide there are at least two possible sites of phosphorylation. The mass difference between the y_{13}^{2+} and y_{14}^{2+} daughter ions, however, is equal to the mass of a seryl residue modified as described above. This result unambiguously defines the phosphorylated residue as the serine and not the threonine.

Aebersold and co-workers have developed an alternative method to isolate and quantitate phosphopeptides from complex proteome mixtures [41]. This method potentially provides greater enrichment because the phosphopeptides are covalently linked to a solid support during processing at the expense of somewhat more complex chemistry.

In addition, whereas the previously described method in amenable only to pSer and pThr residues, this method is equally applicable to pTyr-containing peptides. The sequence of chemical reactions for selectively isolating phosphopeptides from a peptide mixture consists of six steps. In the first step the peptide amino groups are protected with *text*-butyl dicarbonate (tBoc). The carboxylate and phosphate groups are subsequently modified via a carbodiimide-catalyzed condensation reaction to form amide and phosphoramidate bonds. An acid wash is used to hydrolyze the phosphoramidate bond and deprotect the phosphate group. Cystamine is attached to the regenerated phosphate group via another carbodiimide-catalyzed condensation reaction and a free sulfhydryl group is generated at each phosphate group by reduction of the internal disulfide of cystamine. This reaction allows the peptides to be attached to iodoacetyl groups immobilized on glass beads. The covalent attachment of the peptides allows much more stringent washing conditions to be used than in a noncovalent coupling. The ability to use more stringent washing conditions should ultimately reduce the number of nonspecifically bound components being recovered. The phosphopeptides are recovered by cleavage of phosphoramidate bonds, using trifluoroacetic acid (TFA). The use of TFA also removes the tBoc protection group, allowing the peptides to be recovered with free amino and phosphate groups. The carboxylate groups, however, remain blocked. The strategy describe does not provide a direct method to quantitate the phosphorylation status of a protein; however, the blocking of the carboxylates using either normal isotopic abundance or deuterated ethanolamine (i.e., ethanolamine-d_4) allows for the peptides to be quantified by MS.

V. Conclusions

Rapid analytical developments in the field of proteomics are enabling a new tool set for biological investigation, allowing advancements in our understanding of protein structure, function, and organization in complex signaling and regulatory networks. One of the most exciting aspects of more recent proteomic developments is the development of both gel-based and solution-based labeling strategies that allow relative protein abundance changes to be measured in cell systems. Such gel-based technologies as DIGE and MP enable thousands of protein abundances to be compared across cell systems in a precise and relatively high-throughput manner. Developments in stable isotope labeling and bulk peptide chemistry have provided tools both to measure relative protein abundance changes between cell systems and to decrease the complexity

of the system, such as with the use of ICAT reagents. Finally, new methodologies are rapidly advancing our ability to conduct global investigations of posttranslational modifications, perhaps leading to the challenge of conducting truly global measurements. These advances have significant implications for our understanding of how cellular protein fluxes respond in the context of disease and toxicity.

Acknowledgments

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. NO1-CO-12400.

References

1. Wasinger, V. C., Cordwell, S. J., Cerpa-Poljak, A., Yan, J. X., Gooley, A. A., Wilkens, M. R., Duncan, M. W., Harris, R., Williams, K. L., and Humphery-Smith, I. (1995). Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis* **16,** 1090–1094.
2. O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250,** 4007.
3. Klose, J., and Kobalz, U. (1995). Two-dimensional electrophoresis of proteins: An updated protocol and implications for a functional analysis of the genome. *Electrophoresis* **16,** 1034–1059.
4. Gorg, A., Obermaier, C., Boguth, G., and Weiss, W. (1999). Recent developments in two-dimensional gel electrophoresis with immobilized pH gradients: Wide pH gradients up to pH 12, longer separation distance and simplified procedure. *Electrophoresis* **20,** 712–717.
5. Alaiya, A. A., Franzen, B., Moberger, B., Silversward, C., Linder, S., and Auer, G. (1999). Two-dimensional analysis of protein expression in ovarian tumors shows a low degree of intratumoral heterogeneity. *Electrophoresis* **20,** 1039–1046.
6. Unlu, M., Morgan, M. E., and Minden, J. S. (1997). Difference gel electrophoresis: A single gel method for detecting changes in protein extracts. *Electrophoresis* **18,** 2071–2077.
7. Tonge, R., Shaw, J., Middleton, B., Rowlinson, R., Rayner, S., Young, J., Pognan, F., Hawkins, E., Currie, I., and Davison, M. (2001). Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. *Proteomics* **1,** 377–396.
8. Patton, W. F. (2002). Detection technologies in proteome analysis. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **771,** 3–31.
9. Zhou, G., Li, H., DeCamp, D., Chen, S., Shu, H., Gong, Y., Flaig, M., Gillespie, J., Hu, N., Taylor, P., Emmert-Buck, M., Liotta, L., Petricoin, E., and Zhao, Y. (2002). 2D differential in-gel electrophoresis for the identification of

esophageal squamous cell cancer-specific protein markers. *Mol. Cell Proteomics* **1,** 117–123.

10. Steinberg, T. H., Pretty On Top, K., Berggren, K. N., Kemper, C., Jones, L., Diwu, Z., Haugland, R. P., and Patton, W. F. (2001). Rapid and simple single nanogram detection of glycoproteins in polyacrylamide gels and on electroblots. *Proteomics* **1,** 841–855.
11. Blackstock, W. P., and Weir, M. P. (1999). Proteomics: Quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* **15,** 121–127.
12. Goldfarb, M. (1999). Two-dimensional electrophoresis and computer imaging: Quantitation of human milk casein. *Electrophoresis* **20,** 870–874.
13. Conrad, C. C., Malkowsky, C. A., Talent, J., Rong, D., Lakdawala, S. W., and Gracy, R. W. (2001). Chemiluminescent standards for quantitative comparison of two-dimensional electrophoresis Western blots. *Proteomics* **1,** 365–369.
14. Washburn, M. P., Wolters, D., and Yates, J. R., III. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247.
15. Yalow, R. S., and Berson, S. A. (1960). Immunoassay of endogenous plasma insulin in man. *Obes. Res.* **4,** 583–600.
16. de Nadal, E., Fadden, R. P., Ruiz, A., Haystead, T., and Arino, J. (2001). A role for the Ppz Ser/Thr protein phosphatases in the regulation of translation elongation factor 1Bα. *J. Biol. Chem.* **276,** 14829–14834.
17. Jimenez, C. R., Eyman, M., Lavina, Z. S., Gioio, A., Li, K. W., van der Schors, R. C., Geraerts, W. P., Giuditta, A., Kaplan, B. B., and van Minnen, J. (2002). Protein synthesis in synaptosomes: A proteomics analysis. *J. Neurochem.* **81,** 735–744.
18. Ranganathan, G., Kaakaji, R., and Kern, P. A. (1999). Role of protein kinase C in the translational regulation of lipoprotein lipase in adipocytes. *J. Biol. Chem.* **274,** 9122–9127.
19. Xia, B., Pikus, J. D., Xia, W., McClay, K., Steffan, R. J., Chae, Y. K., Westler, W. M., Markley, J. L., and Fox, B. G. (1999). Detection and classification of hyperfine-shifted ^{1}H, ^{2}H, and ^{15}N resonances of the Rieske ferredoxin component of toluene 4-monooxygenase. *Biochemistry* **38,** 727–739.
20. Shindo, K., Masuda, K., Takahashi, H., Arata, Y., and Shimada, I. (2000). Backbone ^{1}H, ^{13}C, and ^{15}N resonance assignments of the anti-dansyl antibody Fv fragment. *J. Biomol. NMR* **17,** 357–358.
21. Oda, Y., Huang, K., Cross, F. R., Cowburn, D., and Chait, B. T. (1999). Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96,** 6591–6596.
22. Conrads, T. P., Alving, K., Veenstra, T. D., Belov, M. E., Anderson, G. A., Anderson, D. J., Lipton, M. S., Pasa-Tolic, L., Udseth, H. R., Chrisler, W. B., Thrall, B. D., and Smith, R. D. (2001). Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and ^{15}N-metabolic labeling. *Anal. Chem.* **73,** 2132–2139.
23. Washburn, M. P., Ulaszek, R., Deciu, C., Schieltz, D. M., and Yates, J. R., III. (2002). Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Anal. Chem.* **74,** 1650–1657.
24. Jensen, P. K., Pasa-Tolic, L., Anderson, G. A., Horner, J. A., Lipton, M. S., Bruce, J. E., and Smith, R. D. (1999). Probing proteomes using capillary isoelectric focusing-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **71,** 2076–2084.

25. Veenstra, T. D., Martinovic, S., Anderson, G. A., Pasa-Tolic, L., and Smith, R. D. (2000). Proteome analysis using selective incorporation of isotopically labeled amino acids. *J. Am. Soc. Mass Spectrom.* **11,** 78–82.
26. Chen, X., Smith, L. M., and Bradbury, E. M. (2000). Site-specific mass tagging with stable isotopes in proteins for accurate and efficient protein identification. *Anal. Chem.* **72,** 1134–1143.
27. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999.
28. Conrads, T. P., Alving, K., Veenstra, T. D., Belov, M. E., Anderson, G. A., Anderson, D. J., Lipton, M. S., Pasa-Tolic, L., Udseth, H. R., Chrisler, W. B., Thrall, B. D., and Smith, R. D. (2001). Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and ^{15}N-metabolic labeling. *Anal. Chem.* **73,** 2132–2139.
29. Krishna, R. G., and Wold, F. (1998). ''Proteins: Analysis and Design,'' p. 121. Academic Press, San Diego, California.
30. Pawson, T., and Scott, J. D. (1997). Signaling through scaffold, anchoring, and adaptor proteins. *Science* **278,** 2075.
31. Cohen, P. (1982). The role of phosphorylation in neural and hormonal control of cellular activity. *Nature* **296,** 613.
32. Cohen, P. (1992). Signal integration at the level of protein kinases, protein phosphatases and their substrates. *Trends Biochem. Sci.* **17,** 408.
33. van der Greer, P., and Hunter, T. (1994). Phosphopeptide mapping and phosphoamino acid analysis by electrophoresis and chromatography on thin-layer cellulose plates. *Electrophoresis* **15,** 544.
34. Mason, G. G., Murray, R. Z., Pappin, D., and Rivett, A. J. (1998). Phosphorylation of ATPase subunits of the 26S proteasome. *FEBS Lett.* **430,** 269.
35. Oda, Y., Huang, K., Cross, F. R., Crowburn, D., and Chait, B. T. (1999). Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96,** 6591.
36. Yaron, A., Hatzubai, A., Davis, M., Lavon, I., Amit, S., Manning, A. M., Andersen, J. S., Mann, M., Mercurio, F., and Ben-Neriah, Y. (1998). Identification of the receptor component of the IκBα-ubiquitin ligase. *Nature* **396,** 590–594.
37. Cao, P., and Stults, J. T. (2000). Mapping the phosphorylation sites of proteins using on-line immobilized metal affinity chromatography/capillary electrophoresis/electrospray ionization multiple stage tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **14,** 1600.
38. Oda, Y., Nagasu, T., and Chait, B. T. (2001). Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nat. Biotechnol.* **19,** 317–318.
39. Goshe, M. B., Conrads, T. P., Panisko, E. A., Angell, N. A., Veenstra, T. D., and Smith, R. D. (2001). Phosphoprotein isotope coded affinity tag approach for isolating and quantitating phosphopeptides in proteome-wide analyses. *Anal. Chem.* **73,** 2578–2586.
40. Goshe, M. B., Veenstra, T. D., Panisko, E. A., Conrads, T. P., Angell, N. H., and Smith, R. D. (2002). Phosphoprotein isotope coded affinity tag approach for isolating and quantitating phosphopeptides in proteome-wide analyses. *Anal. Chem.* **74,** 607–616.
41. Zhou, H., Watts, J. D., and Aebersold, R. (2001). A systematic approach to the analysis of protein phosphorylation. *Nat. Biotechnol.* **19,** 375–378.

PROTEOME ANALYSIS OF POSTTRANSLATIONAL MODIFICATIONS

By TIMOTHY D. VEENSTRA

SAIC-Frederick Inc., National Cancer Institute at Frederick, Frederick, Maryland 21702

I. Introduction

With the success of the various genome sequencing projects [1] and the development of techniques to measure differences in gene expression at the transcription level [2], considerable focus has shifted toward proteomics—the characterization of gene expression at the protein level [3]. Proteomic analysis represents a much more complex characterization than either genome sequencing or mRNA profiling. In genomics, the goal is to sequence the vast string of deoxynucleotide bases that comprise the genome of a particular species. In mRNA array analysis, the goal is to measure the relative abundance of the gene transcripts from two or more different cell types. Although each of these areas represents tremendous technical achievements, they both are called on to provide one basic characteristic: either the sequence of a genome or the abundance of the gene products. Proteomics, on the other hand, requires the analysis of a group of biomolecules with an almost undefined number of characteristics. Compared with DNA and RNA, which are composed of 4 different monomeric subunits, proteins are composed of at least 20 different amino acids. At present, no high-throughput instrument has been developed that

ADVANCES IN
PROTEIN CHEMISTRY, Vol. 65

0065-3233/03 $35.00

sequences proteins in a comparable fashion to DNA. Proteins come in a wide variety of sizes and structures and are localized throughout the cell. Proteins can be acidic or basic, soluble or insoluble in an aqueous environment, monomeric or oligomeric, and so on. The chemical heterogeneity of proteins represents one of the major factors that make proteomics a greater analytical challenge than genomic or transcriptomics.

Most of the initial efforts in global proteomics have been focused on methods to effectively identify a large number of proteins in a rapid fashion [4]. In global proteomic projects most proteins are identified by mass spectrometry (MS)-based methods in which only a single peptide may be used to infer the presence of a particular species. Unfortunately, this type of simple identification does not do justice to the overall description of a protein. To adequately describe a protein requires characterizing its expression level, its location within the cell, its interactions with other biomolecules, its function, and so on. One of the key descriptors of a protein, amenable to proteomics technology, is the delineation of posttranslation modifications (PTMs). Although significant advances have resulted in methods to effectively identify and determine relative protein abundances, delineation of protein function, or importance to cell phenotype, solely on the basis of abundance changes still provides only a limited view of the proteome because numerous vital activities of proteins are modulated by PTMs that may not be reflected by changes in protein abundance. For example, the treatment of tumor-promotion sensitive murine JB6 epidermal cells with either 12-*O*-tetradecanoylphorbol 13-acetate (TPA) or epidermal growth factor (EGF) results in an increase in the phosphorylation state of the mitogen-activated protein kinases (MAPKs) Erk1 and Erk2, when compared with control cells [5]. The protein abundances of Erk1 and Erk2 in control and TPA- and EGF-treated cells, however, showed no difference. A similar result has been shown for the activation of Akt and MEK1 by growth hormone in murine 32D leukemic cells [6]. These results show that protein activation, and hence cellular processes, can often be controlled by an alteration in phosphorylation state and not by abundance within the cell.

Techniques have been developed over the past few decades to determine whether a protein is posttranslationally modified. The predominant method has been the use of affinity reagents such as anti-phosphoamino acid-specific monoclonal antibodies (mAbs) to detect phosphoproteins or lectins to identify glycoproteins. Although these affinity-based detection methods can determine whether a protein is modified they may not necessarily identify the specific site of modification. Knowledge of the sites of modification is important because an identical modification at a

different site within the same protein can have widely different effects on the activity of the protein. In addition, several different enzymes may modify a single protein, each providing an indication into which cell pathway may be active. For example, the α-amino-3-hydroxy-5-methyl-4-isoxazolepropionate (AMPA) receptor is phosphorylated by protein kinase C, protein kinase A, and cAMP-dependent protein kinase II [7]. Each of these modifications is related to a different AMPA receptor function. Although site-specific mAbs can be produced, this is typically a laborious procedure and it is unlikely that a useful inventory of site-specific antibodies can be produced for all types of modifications.

Mass spectrometry provides the best available technology for the site-specific identification of posttranslational modifications. The present attributes of MS used to measure masses as well as to obtain sequence information about peptides are directly applicable to the site-specific identification of modifications. Although MS has primarily been used for identification, there is also an active research area devoted to quantifying the extent of modification through various stable isotope-labeling techniques. In addition to instrumental methods, software tools are being developed to aid in the identification of modified peptides. In this article we review current progress in the identification of protein modifications, with emphasis on phosphorylation and glycosylation.

II. Phosphorylation

One of the most important posttranslational protein modifications used to modulate protein activity and propagate signals within cellular pathways and networks is phosphorylation [8]. Cellular processes including cell cycle progression, differentiation, development, peptide hormone response, and adaptation are all regulated by protein phosphorylation. Often regulation of protein function by phosphorylation occurs without a change in the abundance of the protein. Classic methods used to study protein phosphorylation primarily employ protein radiolabeling with ^{32}P-labeled inorganic phosphate ($^{32}P_i$). The radioactive proteins are used in fractionation procedures such as two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) or high-performance liquid chromatography (HPLC). To determine the amino acid types that are modified, the phosphoproteins is completely hydrolyzed and the phosphoamino acid content determined. The specific site(s) of phosphorylation can be determined by proteolytic digestion of the radiolabeled protein, separation and detection of phosphorylated peptides (e.g., by two-dimensional peptide mapping), followed by Edman sequencing. To measure differences in

relative abundances of phosphorylation, ^{32}P-labeled proteomes are separated by 2D-PAGE and the relative spot intensities are compared [8]. The use of $^{32}P_i$ to label proteins does not lend itself to high-throughput proteome-wide analysis because of issues with handling radioactive compounds and the associated contamination of analytical instrumentation. Fortunately, MS-based methods have been developed that provide more effective methods to identify, and potentially quantify, specific sites of phosphorylation.

In its simplest form, MS can be used to provide an accurate mass measurement of an intact phosphorylated protein. Comparing this mass with the calculated mass of the unmodified protein or the mass of the protein after phosphatase treatment allows the number of bound phosphate groups to be calculated [9]. Unfortunately, analysis of the intact protein by this method does not provide any information related to the specific site of phosphorylation—a key piece of information that can directly affects the function of the protein. To identify the specific phosphorylated residues requires analysis of the protein at the peptide level. Peptides are generated by enzymatic or chemical digestion of the intact protein and are then analyzed by either MS or tandem MS [10]. Although MS measurements can confirm the presence of a phosphate group on a peptide, tandem MS is still necessary to establish the specific site of phosphorylation when two or more phosphorylatable residues are present. Because most phosphorylation sites are identified at the peptide level, this article focuses on MS analysis of phosphopeptides.

A. Identification of Phosphorylated Proteins

1. Enrichment of Phosphopeptides

To identify phosphorylated residues, a protein or mixture of proteins is digested into peptides either chemically (i.e., cyanogens, bromide) or enzymatically (i.e., trypsin). Although trypsin is the most commonly used enzyme to prepare peptides for MS analysis, alternative methods or enzymes may be required depending on the primary sequence neighboring the phosphorylated residue. If the phosphorylated residue is located within a lysyl (Lys)- or argininyl (Arg)-rich or Lys/Arg-poor region the peptides generated with trypsin may be too small or large to allow accurate identification of the modification. Obviously the choice of a proteolysis agent is simplified when the site(s) of phosphorylation is known. When searching for unknown phosphorylated residues a digestion strategy using many different enzymes may be required to produce peptides of sufficient size that allow broad coverage of the primary structure of the protein. A key

factor in identifying a phosphorylated site is the ability to identify as many of the peptides generated from a digest of the phosphoprotein as possible, so that the peptides of interest are not missed. Unfortunately, complete coverage of a protein by MS analysis is rarely achieved.

One of the major difficulties in the analysis of phosphopeptides is their relatively low abundance when compared with other peptides within the sample [11]. Because the presence of other peptides can suppress the ability to detect phosphopeptides by MS, phosphopeptide detection can be enhanced using methods that reduce the amount of nonphosphorylated peptides within the mixture. Generating a mixture enriched for phosphopeptides not only increases the ability to detect this class of peptides but also aids in the downstream analysis because the identification of a greater percentage of phosphopeptides can be anticipated. Several strategies have been developed to enrich the sample for phosphorylated peptides or phosphoproteins before MS analysis.

2. *Immunoaffinity Chromatography*

One of the earliest means to enrich a sample for phosphopeptides before MS analysis is the use of antibodies [12]. Antibodies can be used to immunoprecipitate a phosphorylated protein(s) from a mixture or to select phosphopeptides from a mixture containing all types of modified and unmodified peptides. The selection of the antibody to use depends on many factors including prior knowledge of the sample. If it is known that a protein of interest is phosphorylated and the question is at which site(s), an antibody directed toward a specific epitope within the protein may be used. If there is no prior knowledge about the sample and the goal is to identify unknown phosphorylation sites, a phospho-specific antibody is the best choice. A number of phosphoseryl (pSer), phosphothreonyl (pThr), and phosphotyrosyl (pTyr)-specific antibodies have been developed against specific phosphoproteins [13]. The potential specificity of such antibodies makes site-specific phosphorylation analysis feasible, providing the relevant antibodies can be prepared. In the case of phospho-specific antibodies, the phosphoamino acid is part of the recognition epitope but is often not the major recognition factor between the antibody and phosphoprotein. The contribution of the neighboring residues allows high-affinity antibodies to be prepared, so that the phosphorylation status of single proteins within complex mixtures can be monitored.

A common procedure is the use of antibodies whose affinity is more dependent on the phosphorylation state of a specific residue type (i.e., Ser, Thr, or Tyr) and not the neighboring primary sequence [14].

These antibodies are generally used to generate mixtures of proteins or peptides phosphorylated at a specific residue type or to indicate the presence of phosphorylated proteins within a complex mixture that has been fractionated, for example, using 2D-PAGE. Yanagida *et al.* used this strategy to study changes in the phosphorylation states of proteins extracted from murine fibroblasts L929 cells treated with tumor necrosis factor α (TNF-α) [15]. In this study, proteins were extracted from L929 cells at various time points after TNF-α treatment. The protein extracts were separated by 2D-PAGE and either electroblotted onto a polyvinylidene fluoride (PVDF) membrane or silver stained. The blot was immunostained with an anti-pTyr monoclonal antibody to enable the identification of phosphoproteins as well as to quantify any changes in the phosphorylation state of the proteins over time. The protein spots that immunostained with the anti-pTyr antibody were correlated with their position on the silver-stained gel. These gel spots were then excised from the 2D-PAGE gel, digested with trypsin, and analyzed by matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) MS to identify the phosphoprotein. Twenty-one different phosphoproteins were identified within the L929 cell lysates, including 8 that showed a time-dependent change in their phosphorylation state based on the immunostaining of the PVDF membrane.

Although the above describes the use of phosphoamino acid-specific antibodies to isolate the intact phosphoprotein before digestion and MS analysis, they may also be used postdigestion to enrich for phosphopeptides. This strategy was used to identify the phosphorylated residues within EphB2, a receptor tyrosine kinase involved in neuronal axon guidance, neural crest cell migration, the formation of blood vessels, and the development of facial structures and the inner ear [16]. In this study, EphB2 was first isolated with an anti-EphB2 antibody followed by tryptic digestion of the intact protein. An anti-pTyr antibody was then used to isolate the phosphopeptides from this mixture. Eight major peaks observed by MALDI-MS represented phosphopeptides containing nine different pTyr residues.

3. Immobilized Metal Affinity Chromatography

One of the most popular affinity-based methods to extract phosphoproteins or phosphopeptides from complex mixtures is immobilized metal affinity chromatography (IMAC) [17]. In IMAC, trivalent cations, such as Fe^{3+} or Ga^{3+}, are bound to a solid support such as iminodiacetic or nitrilotriacetic acid. Passing a complex mixture over an IMAC column results in enrichment for the phosphorylated species due to the affinity of the phosphate moiety for the metal ion. After washing the column, the

remaining bound species are eluted with high-pH or phosphate buffer. Cao and Stults developed a two-dimensional separation incorporating IMAC followed by capillary electrophoresis (CE) coupled directly on-line with electrospray ionization (ESI)-MS/MS to enable the preconcentration, separation, and identification of phosphopeptides [18]. The system was initially demonstrated with phosphorylated β- and α-casein and shows considerable promise for the analysis of complex mixtures of phosphoproteins. In addition, custom-made, nanoscale IMAC columns have been used to enrich phosphorylated peptides from proteins that had been previously separated by sodium dodecyl sulfate (SDS)–PAGE [19]. These miniaturized IMAC columns have been combined with MALDI-MS to characterize the *in vivo*-phosphorylated peptides from the human p47/phox phosphoprotein.

Although immuno and IMAC columns are effective in enriching mixtures of phosphopeptides they do not provide any direct method to quantify the degree of phosphorylation. It would be useful to measure the relative phosphorylation state of a peptide between two different samples, for instance, when one is treated with a growth factor, drug, and so on. The Chait laboratory was the first to show the ability to measure the relative abundance of a phosphopeptide from two distinct samples [20]. The general strategy is shown in Fig. 1. In this initial demonstration, a wild-type yeast strain and a G_1 cyclin Cln2-deficient yeast strain were grown in natural isotopic abundance medium and ^{15}N-enriched medium, respectively. After combining the cells and extracting the proteins, the Cln2-dependent protein STE20 was isolated and analyzed by ESI-MS. Mass spectral measurement of the intensity ratios of the isotopically labeled ($cln2^-$) and unlabeled ($CLN2^+$) phosphopeptides showed that at least four sites exhibit large increases in phosphorylation in the $CLN2^+$ cell pool. These Cln2-dependent sites appear to be consensus cyclin-dependent pSer/pThr, consistent with direct phosphorylation of Ste20 by Cln2-Cdc28 [21]. The two disadvantages of this approach are as follows: because the protein itself is isotopically labeled in heavy isotope-enriched medium, the method is amenable only to organisms that can be metabolically labeled. In addition, this approach does not provide any means to specifically enrich a mixture for phosphopeptides beyond the use of IMAC or antibodies described above.

4. Chemical Modification and Isotopic Labeling of Phosphopeptides

Fortunately, two new methods have been developed that provide for the specific enrichment and quantification of phosphopeptides. Both methods incorporate stable isotopes to differentially label the samples to be compared, and employ subsequent MS analysis for the identification and quantitation of the enriched phosphopeptide mixture.

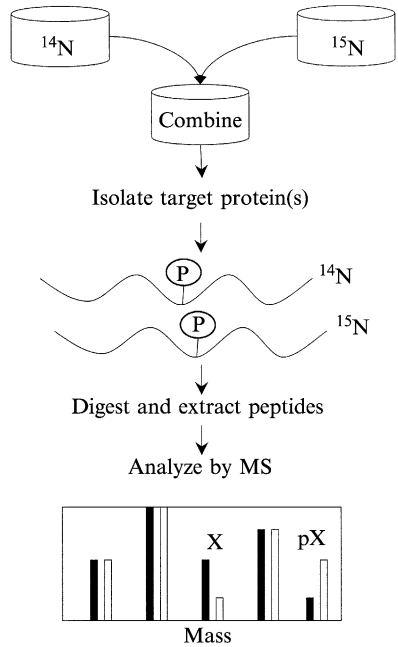


FIG. 1. Method for site-specific quantitation of changes in the level of phosphorylation on proteins. For illustration, peptides that remain unchanged in the two cell pools are assumed to be present in equal abundance, and the level of phosphorylation of the peptide is assumed to change from 30% (^{14}N) to 70% (^{15}N), leading to a decrease in the measured intensity ratio of unphosphorylated peptide X and an increase for phosphorylated peptide pX.

The first strategy to isolate and quantify phosphopeptides we discuss was developed concurrently, and independently, by two groups [22, 23]. Although there are subtle differences in the specific procedures, the overall approaches of both methods are similar. The reaction scheme illustrating the labeling of the phosphoseryl (pSer) and phosphothreonyl (pThr) residues of phosphoproteins is outlined in Fig. 2. Unfortunately, the chemistry involved in this procedure does not make it amenable to pTyr residues. The first step involves blocking reactive thiolates via reductive alkylation or performic acid oxidation. In the following step, phosphate groups are removed via hydroxide ion-mediated β elimination from the pSer and pThr residues, resulting in their conversion to dehydroalanyl and β-methyl dehydroalanyl residues, respectively. These newly formed α,β-unsaturated double bonds render the β carbon in each sensitive to nucleophilic attack, hence the next modification involves a Michael-type addition of the bifunctional reagent 1,2-ethanedithiol

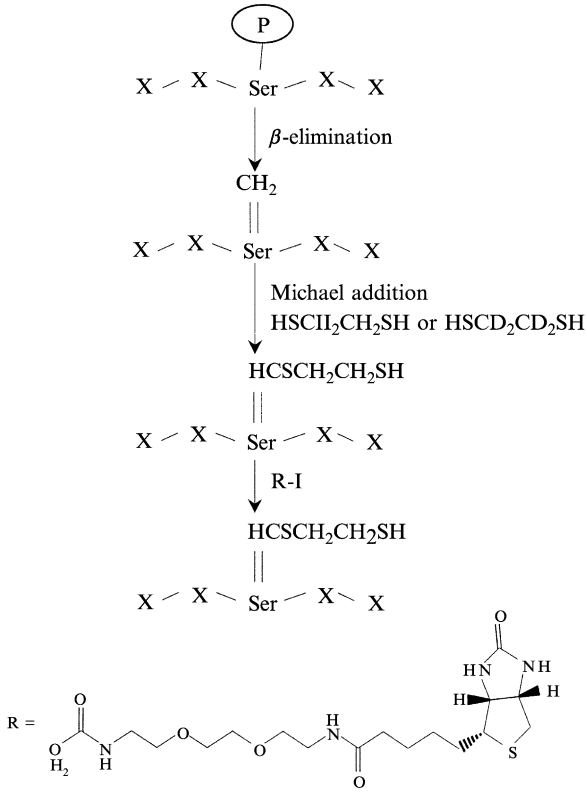


FIG. 2. Isotope affinity strategy for isolating and quantifying phosphopeptides. Proteins containing phosphoseryl (X = H) or phosphothreonyl (X = CH_3) residues are modified with reagents containing both an isotopically labeled linker and biotin group. After proteolytic digestion, these modified peptides are isolated by immobilized avidin affinity chromatography. Light (L = H, EDT-D_0) and heavy (L = D, EDT-D_4) isotopic versions of 1,2-ethanedithiol (EDT) are used to quantitate the relative phosphorylation state of phosphopeptides extracted from two different sources.

(EDT). The addition of EDT to either the dehydroalanyl or β-methyl dehydroalanyl residues results in the creation of a free thiolate in place of what was formerly a phosphate moiety. This thiolate can now serve as a reactive site that can be covalently modified with iodoacetyl-PEO-biotin, a thiol-reactive molecule containing a free biotin group. The end result is the covalent modification of phosphoryl residues with a linker molecule that contains a terminal biotin group. The stable isotopic labeling that enables relative quantification is achieved by using commercially available sources of either the light ($HSCH_2CH_2SH$ or EDT-D_0) or the heavy

($HSCD_2CD_2SH$ or EDT-D_4) isotopic version of EDT. Once the samples have been modified they are digested with trypsin (or another proteolytic enzyme of choice) and the modified peptides are specifically extracted by immobilized avidin chromatography, taking advantage of the high affinity of the biotin–avidin interaction. These extracted peptides are analyzed by reversed-phase liquid chromatography (LC) coupled directly on-line with MS.

The identification and quantitation of the phosphorylated peptides are achieved by two MS strategies. In the MS mode, the masses of the intact peptides are measured and the signal intensity of the individual species provides a direct measure of the peptides phosphorylation status. The signals originating from the modified versions of the phosphopeptides are recognized as pairs separated by the mass difference between the EDT-D_0 and EDT-D_4 labels (i.e., 4.0 Da). The mass spectrum of the phosphorylated peptide FQS^{p}EEQQQTEDELQDK from β-casein in which the pSer residue has been modified with EDT (D_0 or D_4) and iodoacetyl PEO-biotin is shown in Fig. 3A [23]. The 2.01-m/z difference between the $[M+2H]^{2+}$ ions at 1236.96 and 1238.97 m/z corresponds to the expected 4.02-Da mass difference.

Another key attribute of this strategy to identify phosphopeptides is that the modification remains attached to the residue during tandem MS (or MS/MS) fragmentation of the peptide. During MS/MS the intact peptide ion collides with an inert gas (usually nitrogen or helium), which causes it to fragment into smaller ions. These fragment, or daughter, ions provide partial sequence information that is used in conjunction with commercially available computer algorithms to identify the peptide. In a typical experiment studying unmodified phosphopeptides, the phosphate group can readily dissociate from the peptide during MS/MS (indeed, even in MS analyses as well), preventing the site-specific assignment of the phosphate modification. The modification strategy described above, however, allows the exact phosphorylation site to be determined by MS/MS as shown in Fig. 3B [23]. In this example the MS/MS spectra for the EDT-D_0- and EDT-D_4-modified versions of a peptide are shown. The mass difference between the y_{13}^{2+} and y_{14}^{2+} daughter ions is equal to the mass of a seryl residue modified as described above. Although there is at least one other possible site of phosphorylation within this peptide, the MS/MS spectrum clearly identifies it as the seryl residue.

One of the novel attributes of the labeling strategy described above is its ability to quantify the relative phosphorylation of a peptide from two different samples. As an illustration, several stoichiometric amounts of β-casein were processed as described in Fig. 2. As shown for a modified phosphopeptide in Fig. 4, the ratios integrated for each mass spectrum

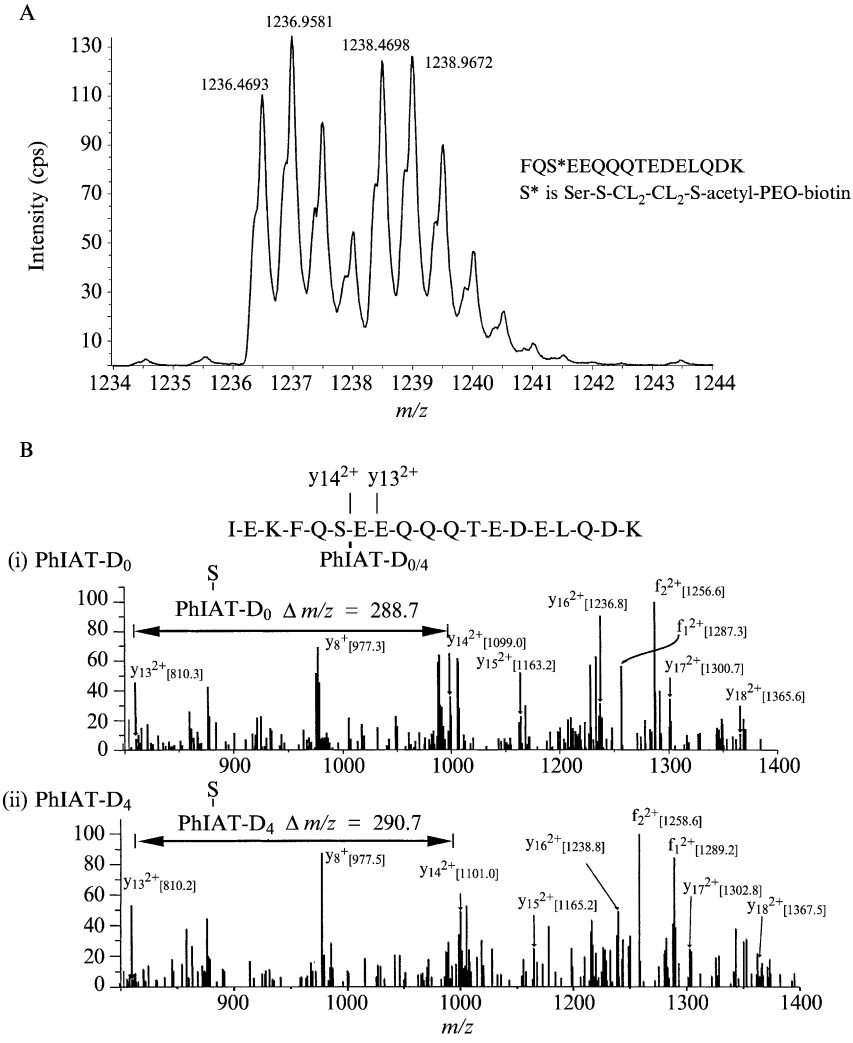


FIG. 3. (A) Mass spectra of EDT-D_0/D_4-labeled β-casein peptide. The enriched mixture of biotinylated phosphopeptides was analyzed by capillary reversed-phase liquid chromatography coupled directly on-line to a PE Sciex API QStar Pulsar hybrid quadrupole-TOF mass spectrometer. The $[M+2H]^{2+}$ ion pair corresponds to the mass of the derivatized β-casein phosphopeptide FQS*EEQQQTEDELQDK, where S* has the modified side chain -CH_2-SCL_2CL_2 S-acetyl-PEO-biotin and L is either H (EDT-D_0 label) or D (EDT-D_4 label). (B) Tandem mass spectrometry identification of β-casein phosphopeptides. Tandem MS/MS spectra of a phosphopeptide modified and affinity isolated, using the (i) light and (ii) heavy isotopic versions of EDT and iodoacetyl-PEO-biotin, are shown. Both labeled versions of the phosphopeptide were identified in a single LC data-dependent MS/MS analysis of the enriched mixture. [Reproduced from reference 23 with permission.]

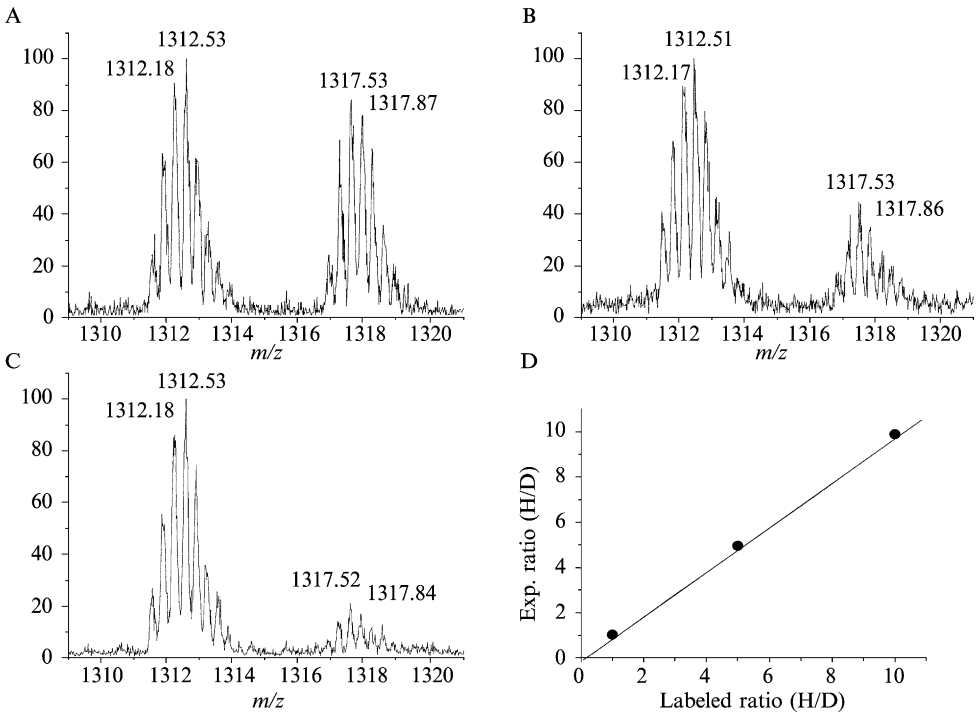


FIG. 4. Stoichiometric isotopic labeling of β-casein phosphopeptides. Samples of β-casein containing ratios of (A) 1:1, (B) 5:1, and (C) 10:1 were labeled with EDT-D_0:EDT-D_4, combined, biotinylated, affinity isolated with immobilized avidin, and analyzed by LC-TOF-MS. Integrated reconstructed ion chromatograms for each species were used to calculate the D_0:D_4 ratio. (D) These measured isotope ratios were plotted against the molar ratios of the β-casein sample labeled with either EDT-D_0 or EDT-D_4. [Reproduced from reference 23 with permission.]

correlate well with the stoichiometric concentration ratios of 1:1, 5:1, and 10:1 used in the labeling experiment. The two protein samples are pooled after the isotopic labeling step, eliminating any variations associated with sample handling.

A second labeling method to isolate and quantify phosphopeptides has been developed in the laboratory of R. Aebersold [24]. Although the chemistry involved is more complex, this method potentially provides greater enrichment because the phosphopeptides are covalently linked to a solid support during processing. The sequence of chemical reactions for selectively isolating phosphopeptides from a peptide mixture consists of six steps as shown in Fig. 5. To eliminate potential intra- and intermolecular condensation, the peptide amino groups are protected using *tert*-butyl dicarbonate (tBoc) chemistry [25]. After this the carboxylate and phosphate groups are modified via a carbodiimide-catalyzed condensation reaction to form amide and phosphoramidate bonds. The phosphoramidate bonds are then hydrolyzed via a brief acid wash to deprotect the phosphate group and cystamine is attached to the regenerated phosphate group via another carbodiimide-catalyzed condensation reaction. A free sulfhydryl group is generated at each phosphate group by reduction of the internal disulfide of cystamine, which allows the peptides to be attached to iodoacetyl groups immobilized on glass beads. The covalent attachment of the peptides allows stringent washing conditions to be used, thereby reducing the number of nonspecifically bound components being recovered with the phosphopeptides of interest. The phosphopeptides are recovered by cleavage of phosphoramidate bonds, using trifluoroacetic acid at a concentration that also removes the tBoc protection group, thus regenerating peptides with free amino and phosphate groups. The carboxylate groups, however, remain blocked from step 2.

Zhou *et al.* [24] noted that this method yielded mixtures highly enriched in phosphopeptides with minimal contamination from other peptides. Because this strategy does not require the removal of the phosphate group it is equally applicable to pSer-, pThr-, and pTyr-containing peptides. The MS/MS spectra of the modified phosphopeptides were of high-enough quality to allow the peptides to be identified by sequence database searching. The MS/MS spectra could discriminate between pSer/pThr- and pTyr-containing peptides because pSer and pThr lose an H_3PO_4 group on MS/MS, allowing these residues to be identified via a fragment ion corresponding to the loss of 98 Da [26]. Phosphotyrosyl residues are more stable and do not lose their phosphate group during fragmentation. Although this strategy does not provide a direct method to quantify changes in phosphorylation state between peptides from two different samples, the blocking of the carboxylates using either normal isotopic abundance or

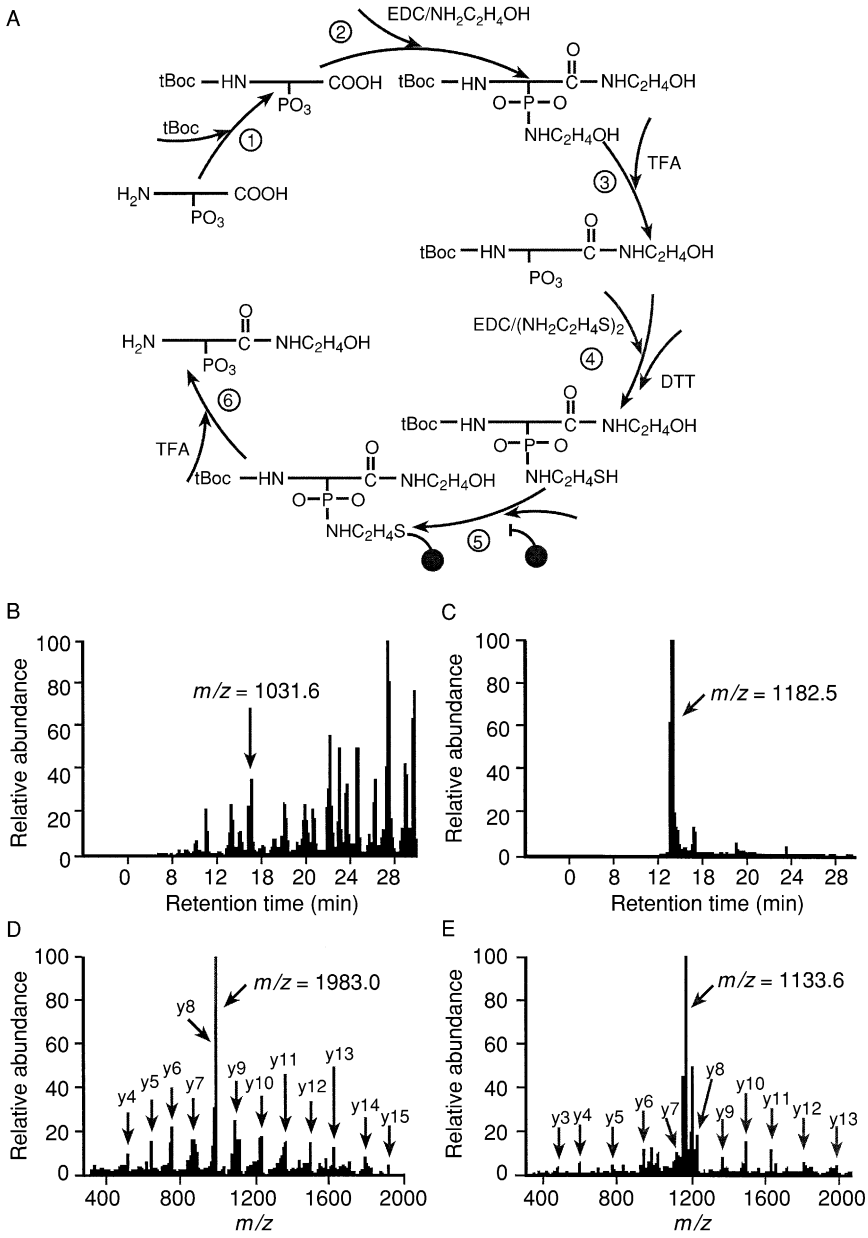


FIG. 5. Phosphopeptide isolation strategy and validation with phosphoprotein β-casein. (A) Schematic illustration of the chemistry involved in selective phosphopeptide isolation. Full details are given in text. (B–E) Phosphopeptide isolation from β-casein. A

deuterated ethanolamine (i.e., ethanolamine-d_4) would allow for incorporation of stable isotope tags that later can be differentiated and quantified by MS.

Although the labeling strategies described above were developed using model systems, they have also been assessed with cell lysates because their greatest utility will be in their application to proteome-wide identification and quantitation of phosphopeptides. In the case of the strategy proposed by Zhou *et al.* [24], phosphopeptides were isolated from a *Saccharomyces cerevisiae* cell lysate and analyzed by LC-MS/MS, with collision-induced dissociation (CID) spectra recorded and searched against the sequence database. Greater than 80% of the useful MS/MS spectra identified phosphopeptides.

III. Mass Spectral Identification of Phosphopeptides

A. *Phosphopeptide Mapping*

Once the sample has been prepared the next step is to identify the phosphopeptides and the specific sites of phosphorylation. A variety of MS-based methods have been designed that either identify a peptide as being phosphorylated and/or identify the specific site of phosphorylation. The most straightforward method of identifying phosphorylated peptides, termed phosphopeptide mapping, involves enzymatically digesting a purified phosphoprotein and analyzing the resulting fragments by MS [7]. If the protein being studied is known, the phosphopeptide(s) is identified as having an experimental mass shifted by a multiple of 80 Da compared with its predicted mass. If the identity of the protein is unknown, the complement

tryptic digest of β-casein was analyzed by LC-MS/MS both before (B and D) and after (C and E) phosphopeptide isolation according to the procedure in (A). The starting material for phosphopeptide isolation is 10 pmol. (B) Ion chromatogram of 1 pmol of β-casein digest before phosphopeptide isolation. The peak at $m/z = 1031.6$ represents the doubly charged form of the expected tryptic phosphopeptide from β-casein. (C) Ion chromatogram of the isolated phosphopeptides of β-casein digest. The peak at $m/z = 1182.5$ represents the doubly charged form of the same tryptic phosphopeptide from β-casein indicated in (B), additionally modified on its seven carboxylate groups with ethanolamine. (D) CID spectrum of β-casein digest in (B). The peak at $m/z = 983.0$ represents the doubly charged form of the selected parent ion ($m/z = 1031.6$) minus the H_3PO_4 group. (E) CID spectrum of isolated phosphopeptides of β-casein digest in (C). Again, the peak at $m/z = 1133.6$ represents the doubly charged form of the selected parent ion ($m/z = 1182.5$) minus H_3PO_4, and the *y*-ion series used for peptide identification is indicated. [Reproduced from reference 24 with permission.]

of measured masses can be used first to identify the protein and the phosphopeptides distinguished as described above. An additional means to identifying phosphopeptides via peptide mapping involves measuring the masses of the digested protein before and after phosphatase treatment. Because treatment with phosphatase enzymatically removes the phosphate groups, peaks representing phosphopeptides in the original mass spectrum will show a decrease of 80 Da after phosphatase treatment (Fig. 6).

B. Collision-Induced Dissociation

The most common method of identifying sites of phosphorylation is CID of peptides produced by ESI. Although phosphopeptide mapping is capable of identifying phosphorylated peptides it does not provide the specific site of phosphorylation, unless the peptide contains a single phosphorylatable residue. As described below, there are many different CID strategies; however, they all rely on the lability of the phosphoester bonds of the pTyr, pThr, and pSer residues. This bond can be easily fragmented within a collision cell, an ESI ion source, or during postsource decay (PSD) in an MALDI instrument. The net result is the loss of a phosphate group that can be identified by several phosphate-specific ion scans. Fragment ions can be measured in a triple quadrupole instrument [27], an ion trap instrument [28], or a hybrid quadrupole time-of-flight (Qq-TOF) instrument [29]. Loss of phosphate as HPO_3 or H_3PO_4 is a favored fragmentation event, particularly when the parent ion is singly charged, and usually dominates over the backbone cleavages that are useful for sequence determination (Fig. 7). As a general trend for low-energy CID of phosphopeptides, it has been observed that phosphate tends to be lost from pSer more readily than from pThr, and from pThr more readily than from pTyr. Phosphate is generally eliminated from shorter phosphopeptides more readily than from longer phosphopeptides because roughly the same amount of CID energy is dispersed across fewer bonds.

C. In-Source Collision-Induced Dissociation

Although ESI is normally per formed to produce molecular ions to provide molecular weight information about an intact molecule, the voltage on the sample cone can be increased relative to the skimmer voltage to induce some in-source fragmentation, referred to as in-source CID [30]. This change causes the ions to accelerate more quickly through the region between the sample cone and skimmer, which although under vacuum still

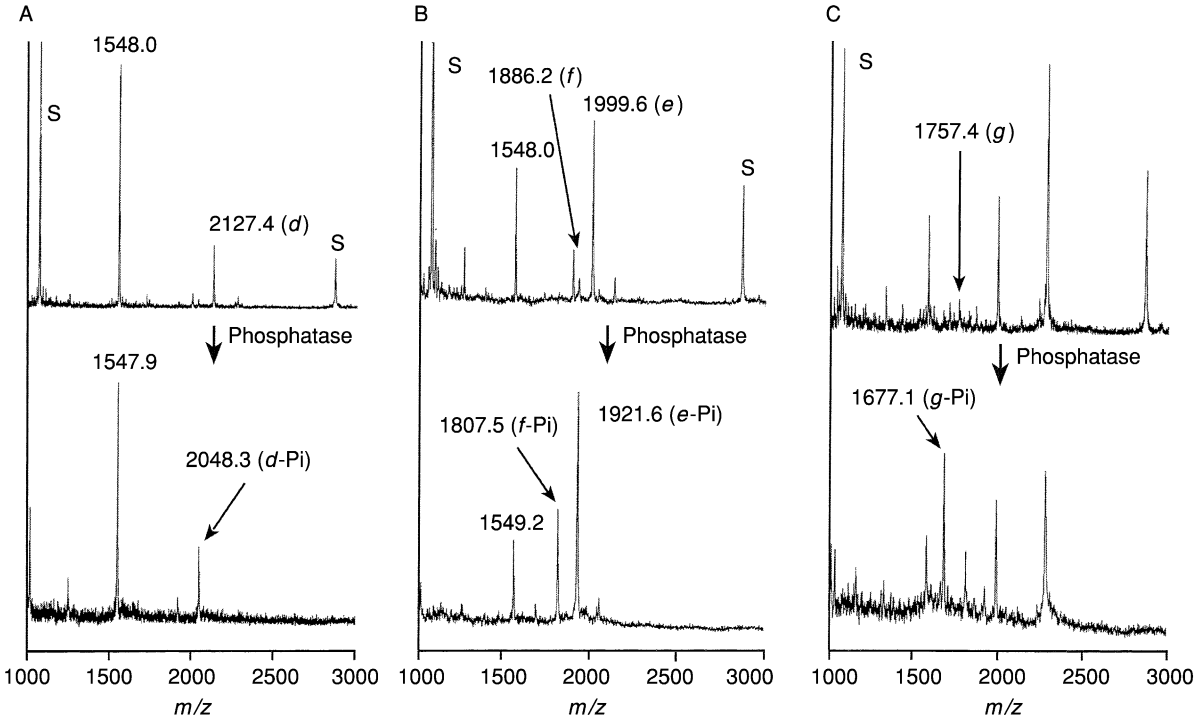


FIG. 6. MALDI-MS analysis of CD45 phosphopeptides before and after phosphatase treatment. In each of these mass spectra, the phosphorylated peaks were reduced by 80 Da after phosphatase treatment and were more prominent in dephosphorylated form. The other major peaks in the top panel were not altered by phosphatase treatment. [Reproduced from reference 58 with permission.]

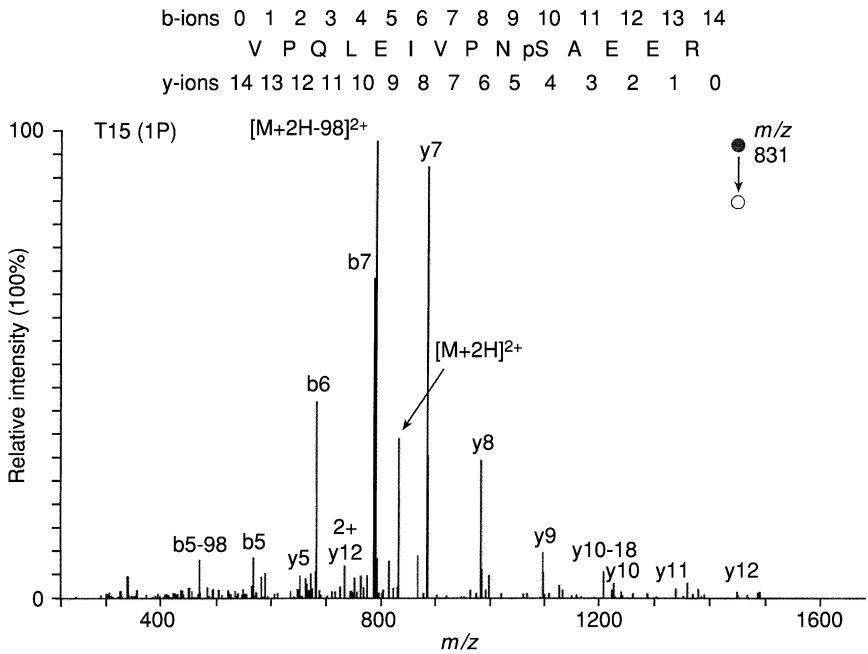


FIG. 7. Collision-induced dissociation of phosphopeptides. The loss of 98 Da is readily apparent in this tandem MS spectrum of a phosphorylated peptide from α-casein. [Reproduced from reference 18 with permission.]

possesses a relatively high pressure. Although the resulting collisions are not sufficiently energetic to provide extensive peptide fragmentation, they may be sufficiently energetic to fragment weaker covalent bonds such as C–N, C–O, and P–O.

In-source CID of phosphopeptides in negative ionization mode produces ions at m/z 63 (PO_2^-), m/z 79 (PO_3^-), and m/z 97 ($H_2PO_4^-$), which can be selectively monitored as a diagnostic of phosphopeptides [31]. This selective ion monitoring can be combined with LC/MS to establish the elution time of a phosphopeptide. This combination has been successfully implemented by Annan and co-workers in a method they refer to as multidimensional MS-based phosphopeptide mapping [32]. In this strategy, they analyze peptide mixtures by LC-MS under in-source CID conditions and monitor for the presence of the phosphopeptide diagnostic marker ions PO_2^- and PO_3^- at m/z 63 and 79, respectively. The measurement of the peptide masses is combined with precursor ion scanning for m/z 79 so that unmodified peptides in the collected fractions are not detected. Fractions containing peptides that produce

phosphopeptide-specific ions are collected, and the molecular weights of the phosphopeptides within each fraction are determined by nanoelectrospray, again in the negative ion mode. In addition, a further analysis to sequence the phosphopeptide by positive ion nanoelectrospray tandem MS is added.

D. *Neutral Loss Scanning*

Neutral loss scanning is carried out in positive ion mode and uses tandem MS with a triple quadrupole mass spectrometer to detect peptides that undergo a neutral loss of H_3PO_4 (98 Da) after CID [33]. This scanning strategy makes use of all three quadrupole regions within the instrument. The first quadrupole, Q1, scans the entire mass range. The second quadrupole region, Q2, acts as the collision cell, and Q3 scans in parallel to Q1 but lower, at m/z $98/n$ (where n is the charge on the ion). For example, neutral loss scanning for a loss of phosphate from an $[M+2H]^{2+}$ phosphopeptide ion requires an offset value of m/z 49. Only those peptides that undergo a neutral loss of 98 Da in the collisional cell can pass through Q3. This loss reflects the facile cleavage of phosphate from pSer and pThr due to a process known as β elimination. Only pSer and pThr, not pTyr, may undergo neutral loss of H_3PO_4 via β elimination. Although neutral loss scanning has not been widely used as in-source CID to detect phosphopeptides, it does have the advantage of being carried out in positive ion mode. This advantage makes it compatible with data-dependent tandem MS, enabling the phosphopeptide to be identified via the partial sequence information obtained by CID. The major disadvantage of neutral loss scanning is its propensity for false positives and the need to know the charge state of the phosphate-containing ion; therefore it has infrequently been used to characterize unknown samples.

E. *Precursor Ion Scanning*

Precursor ion scanning is conducted with an ESI triple quadrupole instrument and is carried out in the negative ion mode, and therefore suffers from the problem associated with sequencing in positive ion mode immediately after detecting the loss of phosphate [34]. In this mode Q1 and Q3 are continuously scanned at a fixed m/z ratio. As in neutral loss scanning Q2 is used as a collision cell, only ions at m/z 79 (PO_3^-) are allowed to pass through Q3. As a result the mass spectrum shows only ions that have lost 79 m/z. Precursor ion scanning is best done during direct

infusion with a nanospray source because the resulting ion discrimination simplifies mixture analysis.

The foregoing discussion has dealt only with CID of ions produced by ESI. The development of instruments in which an MALDI source is coupled to a Qq-TOF analyzer [35] has enabled effective CID fragmentation of peptide ions generated by MALDI. Characterization of phosphopeptides in such instruments has not yet been widely reported, although standard compounds have been investigated [36]. These reports show a significant prompt loss of HPO_3 or the elements of H_3PO_4 from phosphopeptides, resulting in a characteristic pair or series of peaks in the MALDI-MS spectrum. In the analysis of a multiply phosphorylated peptide, the fully dephosphorylated ion formed by this loss was observed to produce a more easily interpretable CID fragmentation pattern than that produced from the fully phosphorylated form [36]. In this case, the formerly phosphorylated residues were identified by the presence of dehydroalanine residues in place of pSer residues, as previously described for peptides undergoing loss of H_3PO_4 [37].

F. Electron Capture Dissociation

As mentioned above, one of the primary difficulties in identifying phosphorylated sites is the lability of the phosphoester linkage, which typically fragments before the amide bonds. A newly developed ion fragmentation technique, electron capture dissociation (ECD), appears to be a promising solution for this problem. ECD is a soft fragmentation technique that mainly induces fragmentation of the backbone of a peptide or protein, forming c and z ions [38]. Labile modifications such as γ-carboxylation, O-glycosylation, or phosphorylation, however, are retained. In comparison with the CID spectra of peptide and protein ions, ECD spectra show far less side-chain loss of posttranslational modifications such as carboxyl, sulfonate, and glycosyl groups, but usually with a higher proportion of backbone cleavage to provide sequence information.

The use of ECD to identify the phosphopeptides and sites of phosphorylation has been demonstrated for β-casein and many other proteins [39]. No phosphate loss is observed for any of the fragment ions generated in the ECD experiment of β-casein. The ECD-generated fragment ions clearly identify phosphorylation at Ser-15. They also indicate three more phosphorylation sites among the cluster of four seryl residues at positions 17, 18, 19, and 22; at Thr-24; and one more between Lys-32 and Val-59, among Ser-35 and Ser-57 and Thr-41 and Thr-55. In none of the ECD spectra of these phosphorylated peptides was a

corresponding nonphosphorylated peptide found; thus, ECD should be valuable for quantitative determination of the degree of phosphorylation. Although ECD is presently unique to Fourier transform ion cyclotron resonance (FTICR) [40], in future either ECD will become available on other types of spectrometers or FTICR may increase in popularity.

G. *Coupled ESI/Inductively Coupled Plasma MS*

A method that combines ESI and inductively coupled plasma (ICP) MS for the identification and quantitation of phosphopeptides has been developed by Wind *et al.* [41]. In this strategy, the eluent from an LC separation of a tryptic digest of a phosphoprotein is interfaced alternatively to ICP-MS and to ESI-MS. The ICP-MS is used to monitor for the presence of ^{31}P and ESI-MS measures the molecular masses of the corresponding peptides. Aligning the two separate LC runs allows phosphopeptides to be identified by aligning the two separate LC runs and determining the peptides that produce a ^{31}P signal. The two advantages of this strategy are its high selectivity and the fact that the signal intensity of the ^{31}P is directly proportional to the molar amount of ^{31}P in the LC eluate, unlike ESI or MALDI-MS, in which the phosphopeptide ionization efficiency (and hence its quantitation) is compound dependent. In addition, the detection limit is approximately 1 pmol of phosphopeptide injected. Although this method has not been widely used, promising results have been demonstrated for β-casein, activated human MAP kinase Erk1, and the catalytic subunit of protein kinase A [41]. Although combined ESI/ICP-MS has relied on peptide mapping to identify phosphopeptides, it is obvious that the strategy is amenable to tandem MS identification as well.

IV. Glycosylation

The attachment of carbohydrates to proteins is one of the most common posttranslational modifications that occur within the cell. Until more recently, the glycosylation of proteins was thought to occur only in eukaryotes; however, investigations have identified many glycoproteins within archaea and bacteria [42]. Although a significant percentage of proteins, particularly secreted and membrane proteins, have been shown to be glycosylated, the function of the carbohydrate attachment has no well-defined universal purpose. The biological activity of a glycoprotein is often not detectably different after deglycosylation. Although there is no universal purpose for this modification, the most relevant function of the glycosyl groups is to increase the solubility of a protein in aqueous medium,

and permit specific interactions with other molecules by affecting the surface properties of the protein [43]. Indeed, carbohydrate groups are commonly found on classes of proteins such as immunoglobulins, proteases, cytokines, hormones, and cell surface receptors that function primarily through their interactions with other biomolecules. The characterization of protein glycoforms is desirable because differences in glycosylation may be observed in disease states, such as the spongiform encephalopathies [44]. The structural elucidation of glycoproteins remains a significant challenge. Whereas unmodified proteins can often be studied by X-ray crystallography or nuclear magnetic resonance spectroscopy, those methods may not provide structural information about the glycolytic portion of glycoproteins because of their nonrigid structure.

The two main types of glycosylation are referred to as N-linked (covalently attached to the nitrogen atom of Asn side chains) or O-linked (attached to the oxygen atom of primarily Ser and Thr side chains) [45]. The Asn residue that is glycosylated generally occurs within the sequence -Asn-Xaa-Ser-, -Asn-Xaa-Thr-, or -Asn-Xaa-Cys-(where Xaa is any residue except Pro). The signal that determines the site of O-linked glycosylation is not readily apparent from the residues surrounding the modified residue. Although O-linked glycosylation occurs primarily within the Golgi as a posttranslational modification in eukaryotes, the attachment of N-linked glycosylations occurs cotranslationally as the growing protein emerges from the endoplasmic reticulum.

Structural determination of glycans is complicated by their branched character, the isomeric nature of the monosaccharides, and the variety of linkage positions possible between neighboring monosaccharides [46]. The mass spectral analysis of N-linked eucaryotic glycoproteins is somewhat simplified because these glycans contain a common pentasaccharide core consisting of three mannose and two *N*-acetylglucosamines (GlcNAc). Antennae of additional sugar residues are attached to this core via two outer mannose residues. The additional sugars are attached in a variety of different configurations. In the high-mannose type, additional mannose residues are linked to the core, whereas in the complex type, GlcNAc, galactose, sialic acid, and L-fucose residues are present. The antennae chains typically terminate in sialic acid residues. Hybrid-type glycoproteins incorporate features of both high-mannose and complex glycans. The structures of some of the common N-linked glycans found in eukaryotes are shown in Fig. 8. In contrast, O-linked glycans have no common core structure and range from monosaccharides to large sulfated polysaccharides. Accordingly, N-linked glycans have received the greatest attention and methods to study their structures have been developed to a greater extent than for O-linked glycans.

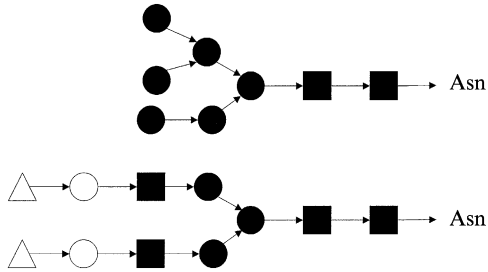


FIG. 8. Structures of some of the commonly found N-linked eukaryotic glycans. These glycans are most often found covalently bound to protein through asparaginyl residues. The monomeric saccharide groups are represented as follows: (■) *N*-acetylglucosamine; (●) mannose; (○) galactose; (△) sialic acid.

A. Identification of Glycosylated Proteins

Characterization of the glycosylated protein requires four fundamental pieces of information: (1) identification of the sites of glycosylation, (2) quantitation of the extent of glycosylation at each site, (3) identification of the number of different glycoforms of each protein, and (4) structural characterization of the glycolytic side chain. MS plays a significant role in both the identification of the sites of glycosylation and the structural characterization of the glycolytic side chain. The complete characterization of the different glycoforms is highly dependent on the ability to fractionate the various glycoforms and then identify the glycolytic composition of each species.

Several MS-based mapping strategies have been developed to identify the site of attachment of the N- or O-linked glycan [47]. The location of the modified site(s) is usually performed via peptide mapping of the glycoproteins before and after treatment with a deglycosylation agent [48]. Modified peptides can be ascertained by changes in the masses observed between the two analyses. O-linked carbohydrates are typically removed from the glycoproteins by base-catalyzed β elimination, whereas N-linked groups are cleaved with *N*-glycanase [47]. Both the native and deglycosylated proteins are proteolytically digested (usually with trypsin) before MS analysis. The appearance of new mass spectral signals, at lower m/z for O-linked carbohydrates and higher m/z for N-linked carbohydrates compared with the respective deglycosylated peptides, identify those peptides that were originally glycosylated. For N-linked sugars, the carbohydrate attached (i.e., complex, high mannose, or hybrid) can be

determined by first digesting the glycoprotein with endoglycosidase H, which releases the high-mannose and hybrid-type sugars [48]. The resulting glycopeptide contains an N-acetylglucosamine (GlcNAc) residue attached to an Asn residue. Such a peptide will have a mass 203 Da higher than that of its deglycosylated counterpart. As an illustration glycopeptides from endopolygalacturonase (EPG II) were treated with endoglycosidase H and compared with the same peptides, but undigested. EPG II has one potential site for N-linked glycosylation that had previously been shown to be occupied with eight different high-mannose structures, whose compositions range from ($Man_5GlcNAc_2$) to ($Man_{12}GlcNAc_2$) [49]. As shown in Fig. 9, the MALDI mass spectrum taken after digestion showed that the high-mannose structures were no longer present, and a new ion was observed at *m/z* 2061.2, which was representative of the mass of the peptide with one GlcNAc residue attached.

When performing comparative peptide mapping by fast atom bombordment (FAB), MALDI, or ESI-MS, *N*-glycanase can be used to convert the

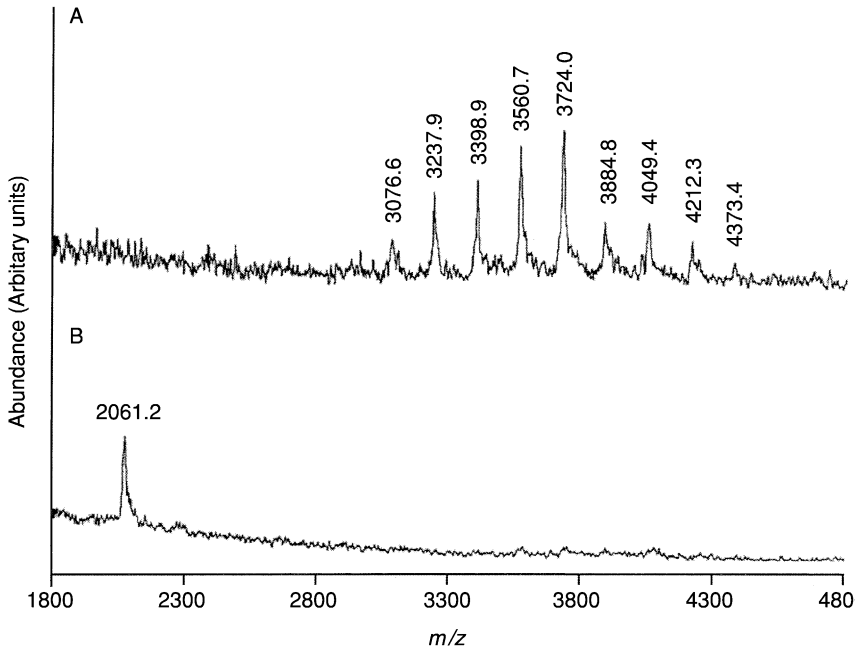


FIG. 9. MALDI mass spectra of the glycopeptides from EPG II before (A) and after (B) treatment with endoglycosidase H. [Reproduced from reference 49 with permission.]

Asn residue to which the oligosaccharide is attached to an Asp residue [50]. This conversion increases the mass of the deglycosylated peptide by 1 Da relative to the calculated mass of the primary sequence, assuming no residue conversion. The presence of low-mass ions that are diagnostic of carbohydrates can assist in the identification of glycopeptides being analyzed by LC-MS/MS. These low-mass ions can be produced by CID in either the source region of a collision cell or in a triple quadrupole mass spectrometer. To produce in-source fragments the potential difference between the orifice and skimmer is increased to cause fragmentation. This method is useful for producing sugar-specific oxonium ions corresponding to Hex^+ (m/z 162), $HexNAc^+$ (m/z 203), $NeuAc^+$ (m/z 274 and 292), Hex-$HexNAc^+$ (m/z 366), and NeuAc-Hex-$HexNAc^+$ (m/z 657) [51].

ESI-MS/MS presents a number of other significant limitations in the analysis of N-linked glycoprotein oligosaccharides, primarily when only a single stage of fragmentation is used. The complex branching present in many oligosaccharides makes structural elucidation by MS/MS extremely difficult. To determine intersaccharide linkages cross-ring fragmentation products must be identified; however, these are usually of low intensity in MS/MS spectra. In the case of glycans of the complex variety, these and other ions may be entirely absent from the collision spectrum because of the low favorability of multiple-bond cleavage processes when in the presence of highly facile pathways originating at glycosidic bonds adjacent to acetylhexosamines and neuraminic acids. The ultimate results are spectra of lower information content, requiring the researcher to employ additional derivatization methods or other analytical techniques, such as methylation analysis, to establish the missing structural details. For structural determination of complex carbohydrates, MS^n using an ion-trap mass spectrometer has proved particularly useful [52]. In particular, MS^3 and higher order MS^n experiments enable the detailed examination of common glycosylation features of these complex glycans: galactosylation, core fucosylation, sialylation of lactosamine antennae, and attachment of *N*-acetylglucosamine residues at both arm-type and bisecting-type mannoses of the conserved core structure (Fig. 10). The removal of the most labile saccharide substituents from complex glycans allows the attainment of fine characterization of underlying glycosidic linkages, branching patterns, sequences, and composition.

1. Electron Capture Dissociation

As with phosphoprotein analysis described above, ECD hold tremendous promise in the characterization of glycoproteins [53]. The ESI FTICR mass spectrum of the unfractionated tryptic digest of a 28-kDa

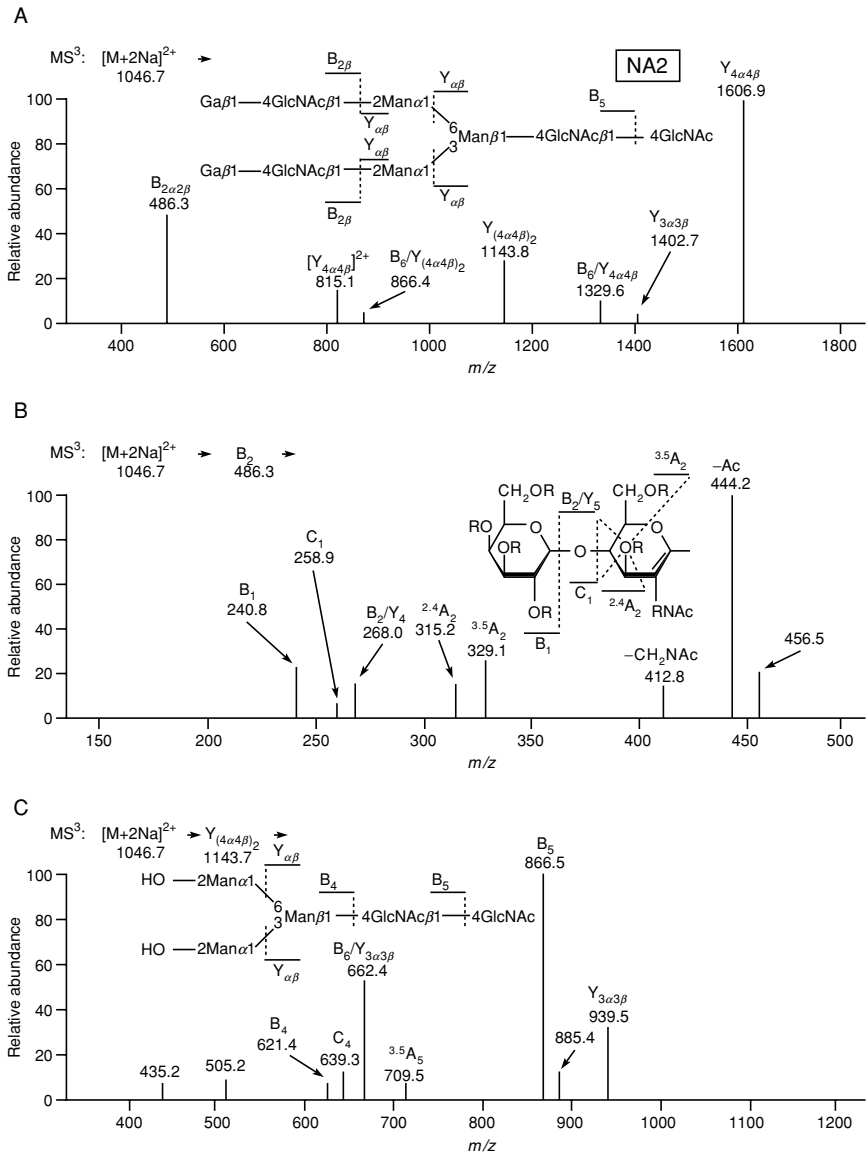


FIG. 10. Ion trap analysis of the permethylated asialo, galactosylated biantennary glycan, NA2, from human fibrinogen: (A) Tandem MS of the doubly sodiated parent, m/z 1046.7; (B) MS^3 of the Gal(1–4)GlcNAc antennae (R = CH_3); (C) MS/MS/MS of the isolated pentasaccharide core. [Reproduced from reference 52 with permission.]

lectin isolated from the coral tree, *Erythrina corallodendron*, was acquired. Two of the major peaks were accurately matched to known glycopeptide structures [40]. The ECD fragment spectrum obtained from a suspected N-glycosylated peptide of m/z 1005.5 (residues 100–116) in the lectin digest is shown in Fig. 11A. The ECD spectrum of this peptide is dominated by N-terminal c-type ions. The amino acid sequence and glycan structure of this peptide, and sites of dissociation in ECD and in frared multiphoton dissociation (IRMPD), are shown in Fig. 11B and C.

The observed fragment ions correspond to cleavages at 11 of 15 peptide backbone amide bonds. One of the bonds not cleaved by ECD is at the N-terminal side of a proline residue, a bond not routinely fragmented by this technique because of its cyclic nature. All sites that are not cleaved were located close to the glycosylated Asn residue. A possible explanation for the absence of cleavage at those sites is that the bulky glycan sterically hinders access to the backbone carbonyl oxygens of the corresponding amino acid residues. The carbonyl oxygens are involved in the proposed ECD cleavage mechanism in which, after electron capture, an energetic hydrogen atom is released and subsequently captured by a site of high hydrogen atom affinity, such as a carbonyl oxygen [54]. In the ECD spectrum of the lectin glycopeptide, three glycosylated fragment ions are observed. All three fragments contain the entire, complex glycan structure; no carbohydrate loss due to cleavage of glycosidic bonds is observed. The current result extends the applicability of ECD for glycopeptide analysis to N-glycosylated peptides and to peptides containing branched, highly substituted glycans. A peptide sequence tag containing six amino acid residues obtained from the achieved ECD fragment ion series was used to retrieve the protein from the database.

Because ECD does not result in the fragmentation of the carbohydrate, IRMPD [55] was employed to characterize the attached monosaccharides [40]. The IRMPD fragment ion spectrum obtained from the N-glycosylated peptide of m/z 1005.5 in the lectin digest is shown in Fig. 12. A complex fragmentation pattern is observed; however, several ions were identified as the parent glycopeptide with loss of one or more sugars. In the $900 < m/z < 1100$ range, both doubly and triply protonated ions are seen. Several correspond to the parent glycopeptide with loss of one or more sugars. Finally, in the $200 < m/z < 900$ range, triply protonated ions corresponding to the parent glycopeptide with loss of several sugars are seen. Also, singly protonated, dehydrated sugar ions are observed. The fragmentation sites are illustrated in Fig. 12 (bottom). Dissociation at each glycosidic bond was observed, as well as the loss of the entire glycan. The extensive monosaccharide losses are consistent with the presence of multiple branch points in the structure of the glycan.

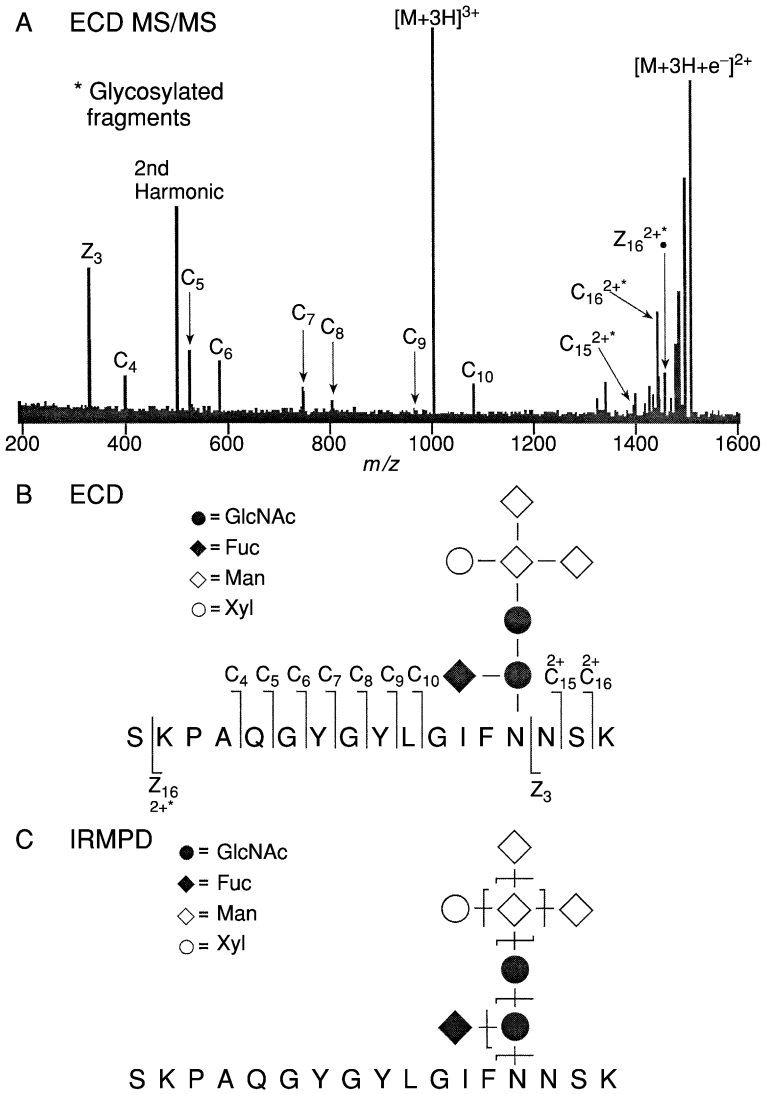


FIG. 11. (A) Electron capture dissociation (ECD) FTICR mass spectrum obtained from the triply protonated N-glycosylated peptide of *m/z* 1005.5 (peptide segment 100–116) from the tryptic digest of a 28-kDa lectin isolated from *Erythrina corallodendron*. The *y* axis is magnified ×10. Cleavages at 11 of 15 backbone amide bonds are observed. The observable c ions provide a peptide sequence tag of six amino acids. No fragmentation of the branched, N-linked heptasaccharide was observed. (B) Dissociation sites for ECD. (C) Major dissociation sites for infrared multiphoton dissociation (IRMPD). [Reproduced from reference 40 with permission.]

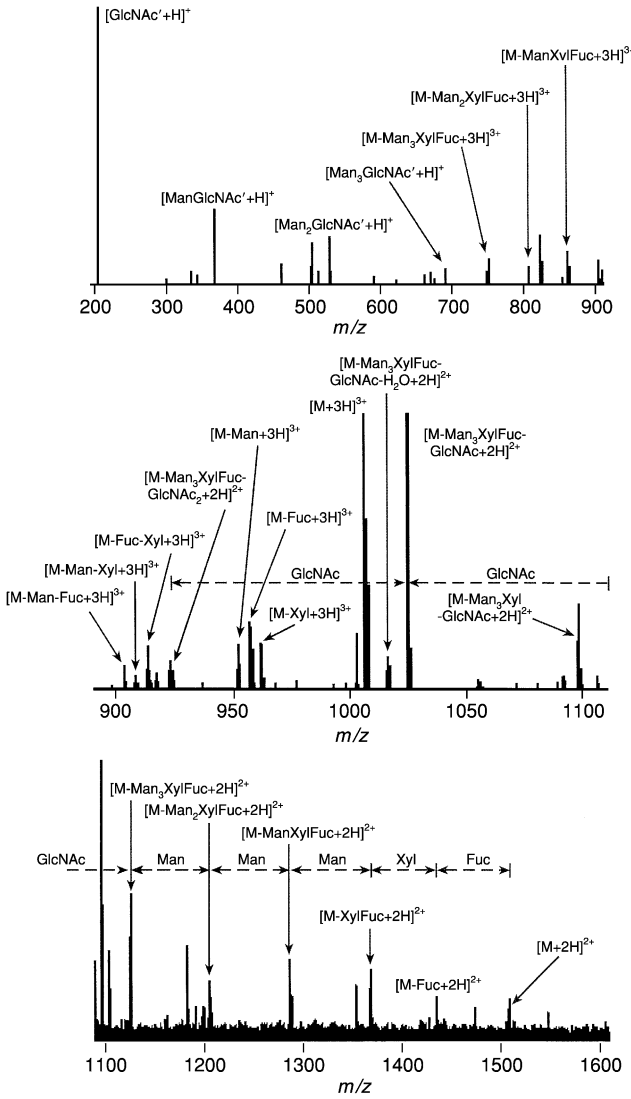


FIG. 12. Infrared multiphoton dissociation (IRMPD) FTICR mass spectrum (displayed in three segments) from the triply protonated N-glycosylated peptide of *m/z* 1005.5 in the lectin digest (Fig. 11).The *y* axis is magnified ×2 (low- and high-mass regions) and ×3 (intermediate-mass region). Extensive fragmentation of the glycan is seen. The achieved fragmentation enabled partial determination of the glycan structure, including the presence of three branching sites. No ions corresponding to cleavage at the peptide backbone are observed. [Reproduced from reference 40 with permission.]

Observation of the doubly protonated fragment at *m/z* 1097.0 ([M-$Man_3XylGlcNAc$+2H]$^{2+}$) made it possible to specify the site of fucosylation as the inner GlcNAc residue. Although the IRMPD spectrum provides no information about the peptide sequence, combining this information with that obtained by ECD provided a complete characterization of the glycopeptide.

2. *Proteome-Wide Identification of Glycoproteins*

Much of the focus in global proteomics has been on the proteome-wide identification and quantitation of proteins or peptides. There has been much attention directed to characterizing the phosphoproteome. Although not as aggressively pursued, the characterization of the glycome (i.e., all the glycopeptides in an organism) is also generating much-needed attention. One of the major obstacles in characterizing the glycome is the complexity and heterogeneity of the carbohydrate chains that make up the glycan. At present the best approaches involve separating a proteome mixture by 2D-PAGE and then using methods to specifically stain or detect the glycosylated proteins. One detection method developed uses the fluorescent hydrazide, Pro-Q Emerald 300 dye, which may be conjugated to glycoproteins by a periodic acid Schiff's mechanism [56]. The glycols present in glycoproteins are initially oxidized to aldehydes, using periodic acid. The dye then reacts with the aldehydes to generate a highly fluorescent conjugate. The glycoproteins can be detected directly in the gels within 2–4 h of electrophoresis. Gels labeled with the Pro-Q Emerald 300 dye can also be subsequently stained with SYPRO Ruby dye, which allows sequential two-color detection of glycosylated and nonglycosylated proteins. Detection of glycoproteins may be achieved in sodium dodecyl sulfate–polyacrylamide gels, two-dimensional gels, and on polyvinylidene difluoride membranes. Lectin affinity chromatography has also been used to isolate glycopeptides resulting from the tryptic digestion of complex proteome mixtures [57]. Although this method circumvents the use of 2D-PAGE, for complete glycoprotein coverage many different lectins are required because of the heterogeneity of the carbohydrate groups present within a glycome sample. Even when the glycoproteins are fractionated, however, a significant amount of work is still required to characterize both the modified residue and the carbohydrate attached.

V. Conclusions

The advances that have been made in the technology to identify proteins in complex biological mixtures have also benefited the

characterization of posttranslational modifications. Although characterization of phosphorylated proteins, and even more so glycoproteins, is far from routine, a much greater effort is being focused toward these two classes of proteins than ever before. Unfortunately we have only begun to scratch the surface, based on the number of known, and possibly yet discovered, posttranslational modifications. Fortunately there is also much room for improved developments in ways to analyze these modifications and these improvements can, and will be, made in many different areas ranging from sample preparation through to instrumentation.

Acknowledgments

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. NO1-CO-12400.

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U.S. Government.

References

1. Broder, S., and Venter, J. C. (2000). Sequencing the entire genomes of free-living organisms: The foundation of pharmacology in the new millennium. *Annu. Rev. Pharmacol. Toxicol.* **40,** 97–132.
2. Brown, P. O., and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* **22,** 33–37.
3. Dongre, A. R., Opiteck, G., Cosand, W. L., and Hefta, S. A. (2001). Proteomics in the post-genome age. *Biopolymers* **60,** 206–211.
4. Aebersold, R., and Goodlett, D. R. (2001). Mass spectrometry in proteomics. *Chem. Rev.* **101,** 269–295.
5. Huang, C., Ma, W. Y., Young, M. R., Colburn, N., and Dong, Z. (1998). Shortage of mitogen-activated protein kinase is responsible for resistance to AP-1 transactivation and transformation in mouse JB6 cells. *Proc. Natl. Acad. Sci. USA* **95,** 156–161.
6. Liang, L., Jiang, J., and Frank, S. J. (2000). Insulin receptor substrate-1-mediated enhancement of growth hormone-induced mitogen-activated protein kinase activation. *Endocrinology* **141,** 3328–3336.
7. McDonald, B. J., Chung, H. J., and Huganir, R. L. (2001). Identification of protein kinase C phosphorylation sites within the AMPA receptor GluR2 subunit. *Neuropharmacology* **41,** 672–679.
8. Cohen, P. (2000). The regulation of protein function by multisite phosphorylation—a 25 year update. *Trends Biochem. Sci.* **25,** 596–601.
9. Han, J. M., Kim, J. H., Lee, B. D., Lee, S. D., Kim, Y., Jung, Y. W., Lee, S., Cho, W., Ohba, M., Kuroki, T., Suh, P. G., and Ryu, S. H. (2002). Phosphorylation-dependent regulation of phospholipase D2 by protein kinase Cδ in rat pheochromocytoma PC12 cells. *J. Biol. Chem.* **277,** 8290–8297.

10. Molloy, M. P., and Andrews, P. C. (2001). Phosphopeptide derivatization signatures to identify serine and threonine phosphorylated peptides by mass spectrometry. *Anal. Chem.* **73,** 5387–5394.
11. Annan, R. S., Huddleston, M. J., Verma, R., Deshaies, R. J., and Carr, S. A. (2001). A multidimensional electrospray MS-based approach to phosphopeptide mapping. *Anal. Chem.* **73,** 393–404.
12. Sun, T., Campbell, M., Gordon, W., and Arlinghaus, R. B. (2001). Preparation and application of antibodies to phosphoamino acid sequences. *Biopolymers* **60,** 61–75.
13. Tsai, E. M., Wang, S. C., Lee, J. N., and Hung, M. C. (2001). Akt activation by estrogen in estrogen receptor-negative breast cancer cells. *Cancer Res.* **61,** 8390–8392.
14. Marcus, K., Immler, D., Sternberger, J., and Meyer, H. E. (2000). Identification of platelet proteins separated by two-dimensional gel electrophoresis and analyzed by matrix assisted laser desorption/ionization-time of flight-mass spectrometry and detection of tyrosine-phosphorylated proteins. *Electrophoresis* **21,** 2622–2636.
15. Yanagida, M., Miura, Y., Yagasaki, K., Taoka, M., Isobe, T., and Takahashi, N. (2000). Matrix assisted laser desorption/ionization-time of flight-mass spectrometry analysis of proteins detected by anti-phosphotyrosine antibody on two-dimensional-gels of fibrolast cell lysates after tumor necrosis factor-α stimulation. *Electrophoresis* **21,** 1890–1898.
16. Kalo, M. S., Yu, H. H., and Pasquale, E. B. (2001). In vivo tyrosine phosphorylation sites of activated ephrin-B1 and ephB2 from neural tissue. *J. Biol. Chem.* **276,** 38940–38948.
17. Gaberc-Porekar, V., and Menart, V. (2001). Perspectives of immobilized-metal affinity chromatography. *J. Biochem. Biophys. Methods* **49,** 335–360.
18. Cao, P., and Stults, J. T. (1999). Phosphopeptide analysis by on-line immobilized metal-ion affinity chromatography-capillary electrophoresis-electrospray ionization mass spectrometry. *J. Chromatogr. A.* **853,** 225–235.
19. Stensballe, A., Andersen, S., and Jensen, O. N. (2001). Characterization of phosphoproteins from electrophoretic gels by nanoscale Fe(III) affinity chromatography with off-line mass spectrometry analysis. *Proteomics* **1,** 207–222.
20. Oda, Y., Huang, K., Cross, F. R., Cowburn, D., and Chait, B. T. (1999). Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96,** 6591–6596.
21. Wu, C., Whiteway, M., Thomas, D. Y., and Leberer, E. (1995). Molecular characterization of Ste20p, a potential mitogen-activated protein or extracellular signal-regulated kinase kinase (MEK) kinase kinase from Saccharomyces cerevisiae. *J. Biol. Chem.* **270,** 15984–15992.
22. Oda, Y., Nagasu, T., and Chait, B. T. (2001). Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome. *Nat. Biotechnol.* **19,** 379–382.
23. Goshe, M. B., Conrads, T. P., Panisko, E. A., Angell, N. H., Veenstra, T. D., and Smith, R. D. (2001). Phosphoprotein isotope-coded affinity tag approach for isolating and quantitating phosphopeptides in proteome-wide analyses. *Anal. Chem.* **73,** 2578–2586.
24. Zhou, H., Watts, J. D., and Aebersold, R. (2001). A systematic approach to the analysis of protein phosphorylation. *Nat. Biotechnol.* **19,** 375–378.
25. Bodanszky, A. B. M., ed. (1984). ''*The Practice of Peptide Synthesis,*'' Vol. 21. Springer-Verlag, New York.
26. Bennett, K. L., Stensballe, A., Podtelejnikov, A. V., Moniatte, M., and Jensen, O. N. (2002). Phosphopeptide detection and sequencing by matrix-assisted laser

desorption/ionization quadrupole time-of-flight tandem mass spectrometry. *J. Mass Spectrom.* **37,** 179–190.
27. Annan, R. S., Huddleston, M. J., Verma, R., Deshaies, R. J., and Carr, S. A. (2001). A multidimensional electrospray MS-based approach to phosphopeptide mapping. *Anal. Chem.* **73,** 393–404.
28. Merrick, B. A., Zhou, W., Martin, K. J., Jeyarajah, S., Parker, C. E., Selkirk, J. K., Tomer, K. B., and Borchers, C. H. (2001). Site-specific phosphorylation of human p53 protein determined by mass spectrometry. *Biochemistry* **40,** 4053–4066.
29. Polson, A. G., Huang, L., Lukac, D. M., Blethrow, J. D., Morgan, D. O., Burlingame, A. L., and Ganem, D. (2001). Kaposi's sarcoma-associated herpesvirus K-bZIP protein is phosphorylated by cyclin-dependent kinases. *J. Virol.* **75,** 3175–3184.
30. Nemeth-Cawley, J. F., Karnik, S., and Rouse, J. C. (2001). Analysis of sulfated peptides using positive electrospray ionization tandem mass spectrometry. *J. Mass Spectrom.* **36,** 1301–1311.
31. Beck, A., Deeg, M., Moeschel, K., Schmidt, E. K., Schleicher, E. D., Voelter, W., Haring, H. U., and Lehmann, R. (2001). Alkaline liquid chromatography/electrospray ionization skimmer collision-induced dissociation mass spectrometry for phosphopeptide screening. *Rapid Commun. Mass Spectrom.* **15,** 2324–2333.
32. Annan, R. S., Huddleston, M. J., Verma, R., Deshaies, R. J., and Carr, S. A. (2001). A multidimensional electrospray MS-based approach to phosphopeptide mapping. *Anal. Chem.* **73,** 393–404.
33. Schlosser, A., Pipkorn, R., Bossemeyer, D., and Lehmann, W. D. (2001). Analysis of protein phosphorylation by a combination of elastase digestion and neutral loss tandem mass spectrometry. *Anal. Chem.* **73,** 170–176.
34. Pandey, A., Andersen, J. S., and Mann, M. (2000). Use of mass spectrometry to study signaling pathways. *Sci STKE* PL1.
35. Shevchenko, A., Loboda, A., Ens, W., and Standing, K. G. (2000). MALDI quadrupole time-of-flight mass spectrometry: A powerful tool for proteomic research. *Anal. Chem.* **72,** 2132–2141.
36. Lee, C. H., McComb, M. E., Bromirski, M., Jilkine, A., Ens, W., Standing, K. G., and Perreault, H. (2001). On-membrane digestion of β-casein for determination of phosphorylation sites by matrix-assisted laser desorption/ionization quadrupole/time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **15,** 191–202.
37. Tholey, A., Reed, J., and Lehmann, W. D. (1999). Electrospray tandem mass spectrometric studies of phosphopeptides and phosphopeptide analogues. *J. Mass Spectrom.* **34,** 117–123.
38. McLafferty, F. W., Horn, D. M., Breuker, K., Ge, Y., Lewis, M. A., Cerda, B., Zubarev, R. A., and Carpenter, B. K. (2001). Electron capture dissociation of gaseous multiply charged ions by Fourier-transform ion cyclotron resonance. *J. Am. Soc. Mass Spectrom.* **12,** 245–249.
39. Shi, S. D., Hemling, M. E., Carr, S. A., Horn, D. M., Lindh, I., and McLafferty, F. W. (2001). Phosphopeptide/phosphoprotein mapping by electron capture dissociation mass spectrometry. *Anal. Chem.* **73,** 19–22.
40. Hakansson, K., Cooper, H. J., Emmett, M. R., Costello, C. E., Marshall, A. G., and Nilsson, C. L. (2001). Electron capture dissociation and infrared multiphoton dissociation MS/MS of an N-glycosylated tryptic peptic to yield complementary sequence information. *Anal. Chem.* **73,** 4530–4536.
41. Wind, M., Edler, M., Jakubowski, N., Linscheidm, M., Wesch, H., and Lehmann, W. D. (2001). Analysis of protein phosphorylation by capillary liquid chromatography

coupled to element mass spectrometry with ^{31}P detection and to electrospray mass spectrometry. *Anal. Chem.* **73,** 29–35.
42. Schaffer, C., Graninger, M., and Messner, P. (2001). Prokaryotic glycosylation. *Proteomics* **1,** 248–261.
43. Dell, A., and Morris, H. R. (2001). Glycoprotein structure determination by mass spectrometry. *Science* **291,** 2351–2356.
44. Priola, S. A., and Lawson, V. A. (2001). Glycosylation influences cross-species formation of protease-resistant prion protein. *EMBO J.* **20,** 6692–6699.
45. Kobata, A. (2000). A journey to the world of glycobiology. *Glycoconj. J.* **17,** 443–464.
46. Gerwig, G. J., and Vliegenthart, J. F. (2000). Analysis of glycoprotein-derived glycopeptides. *EXS* **88,** 159–186.
47. Harvey, D. J. (2001). Identification of protein-bound carbohydrates by mass spectrometry. *Proteomics* **1,** 311–328.
48. Liu, T., Li, J. D., Zeng, R., Shao, X. X., Wang, K. Y., and Xia, Q. C. (2001). Capillary electrophoresis-electrospray mass spectrometry for the characterization of high-mannose-type N-glycosylation and differential oxidation in glycoproteins by charge reversal and protease/glycosidase digestion. *Anal. Chem.* **73,** 5875–5885.
49. Colangelo, J., and Orlando, R. (2001). On-target endoglycosidase digestion matrix-assisted laser desorption/ionization mass spectrometry of glycopeptides. *Rapid Commun. Mass Spectrom.* **15,** 2284–2289.
50. Plummer, T. H., Jr., and Tarentino, A. L. (1981). Facile cleavage of complex oligosaccharides from glycopeptides by almond emulsin peptide: *N*-Glycosidase. *J. Biol. Chem.* **256,** 10243–10246.
51. Yeboah, F. K., and Yaylayan, V. A. (2001). Analysis of glycated proteins by mass spectrometric techniques: Qualitative and quantitative aspects. *Nahrung* **45,** 164–171.
52. Weiskopf, A. S., Vouros, P., and Harvey, D. J. (1998). Electrospray ionization-ion trap mass spectrometry for structural analysis of complex N-linked glycoprotein oligosaccharides. *Anal. Chem.* **70,** 4441–4447.
53. Zubarev, R. A., Horn, D. M., Fridriksson, E. K., Kelleher, N. L., Kruger, N. A., Lewis, M. A., Carpenter, B. K., and McLafferty, F. W. (2000). Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal. Chem.* **72,** 563–573.
54. Mirgorodskaya, E., Roepstorff, P., and Zubarev, R. A. (1999). Localization of O-glycosylation sites in peptides by electron capture dissociation in a Fourier transform mass spectrometer. *Anal. Chem.* **71,** 4431–4436.
55. Little, D. P., Speir, J. P., Senko, M. W., O'Connor, P. B., and McLafferty, F. W. (1994). Infrared multiphoton dissociation of large multiply charged ions for biomolecule sequencing. *Anal. Chem.* **66,** 2809–2815.
56. Steinberg, T. H., Pretty On Top, K., Berggren, K. N., Kemper, C., Jones, L., Diwu, Z., Haugland, R. P., and Patton, W. F. (2001). Rapid and simple single nanogram detection of glycoproteins in polyacrylamide gels and on electroblots. *Proteomics* **1,** 841–855.
57. Hirabayashi, J., Arata, Y., and Kasai, K. (2001). Glycome project: Concept, strategy and preliminary application to *Caenorhabditis elegans*. *Proteomics* **1,** 295–303.

MAPPING PROTEIN MODIFICATIONS WITH LIQUID CHROMATOGRAPHY-MASS SPECTROMETRY AND THE SALSA ALGORITHM

By DANIEL C. LIEBLER, BEAU T. HANSEN, JULIET A. JONES, HAMID BADGHISI, AND DANIEL E. MASON

Southwest Environmental Health Sciences Center, College of Pharmacy, University of Arizona, Tucson, Arizona 85721

I. Introduction

There are many ways in which the completion of human and other genome sequences will change biology and medicine (Lander *et al.*, 2001; Rubin *et al.*, 2000; Venter *et al.*, 2001). Most fundamentally, the information in a well-annotated genome sequence represents a catalog of all possible gene products in cells. Genes ultimately encode proteins, which perform most of the functions of cells. A proteome is the collection of proteins in a cell or other biological compartment. Whereas the genome in any cell in an organism is essentially invariant over time, the proteome is highly variable. Different cells express different sets of genes at different stages of cell and life cycles, with development or aging, and in response to stress or disease. Thus organisms have one genome, but many proteomes.

Most widely practiced forms of proteomic analysis are focused on identifying the proteins that comprise proteomes and quantifying variations in their expression (Liebler, 2001). However, protein expression per see does not necessarily dictate function. Most proteins undergo posttranslational

ADVANCES IN PROTEIN CHEMISTRY, Vol. 65

0065-3233/03 $35.00

modifications that govern enzymatic activity, protein–protein interaction, or protein turnover. These include widely studied regulatory modifications (e.g., phosphorylation), modifications that govern turnover (e.g., ubiquitination), and modifications by xenobiotics or endogenous electrophiles (protein adducts). To adequately describe the biology of proteomes, proteomic analysis must characterize modified protein forms.

Modifications to biological macromolecules play important roles in the toxicity and carcinogenicity of many chemicals and related environmental stresses. Studies of DNA adducts have clearly established their essential roles in mutagenesis (Dipple, 1995; Garner, 1998; Marnett and Plastaras, 2001; Miller and Miller, 1981; Strickland *et al.*, 1993). Protein modifications also have been appreciated as important, but an understanding of their roles in toxicity and carcinogenesis has been hindered by a lack of analytical methods to characterize the protein targets (Cohen *et al.*, 1997; Nelson and Pearson, 1990). Advances in protein and peptide separation methods, in mass spectrometry (MS) instrumentation, and in computational tools for analysis of MS data have driven an rapid expansion of the proteomics field (Liebler, 2001; Pandey and Mann, 2000; Yates, 1998). These tools offer new opportunities to characterize protein modifications, including those produced by xenobiotics and their reactive metabolites. However, these tools are broadly applicable to characterization of proteomic diversity including xenobiotic adducts, posttranslational modifications, and sequence variants.

We have focused our more recent work on the application of LC-tandem MS (LC-MS-MS)-based approaches to identifying protein targets of modifications and mapping the modifications at the level of amino acid sequence. This article describes (1) the tandem MS (MS-MS) analysis of modified peptides and the impact of different adduct chemistries on peptide ion fragmentation, (2) a new MS-MS data analysis tool called SALSA (scoring algorithm for spectra analysis), which was developed in our laboratory, (3) the application of SALSA to sequence-specific mapping of modifications to proteins, and (4) the complementary relationship of SALSA to other data analysis algorithms (e.g., Sequest) for MS-MS data. The integration of Sequest and SALSA provides a powerful approach to proteome characterization by LC-MS-MS.

II. MS-MS Fragmentation of Modified Peptides

MS-MS fragmentation of a peptide in a quadrupole ion trap, triple quadrupole, or quadrupole-time of flight mass analyzer yields MS-MS spectra that encode all or part of the peptide sequence (Yates, 1998).

Specifically, fragmentation along the peptide backbone yields b- and y-series ions (as well as other fragment ions) indicative of sequence. In addition, peptides bearing modifications may also yield other distinct fragmentations derived from the modifying moieties themselves. We have studied the fragmentation of model peptides modified with several different types of chemical adducts or common biological modifications (Hansen *et al.*, 2001; Jones and Liebler, 2000; Mason and Liebler, 2000). These studies revealed that different peptide modifications confer three types of characteristics on MS-MS spectra: (1) product ions derived from adduct moieties, (2) neutral or charged losses arising from adduct-specific fragmentations, and (3) shifts in b- or y-ion signals along the m/z axis. Examples of each type of modification are presented below and in Table I, which lists examples of these characteristic features in MS-MS spectra of modified peptides.

A. Product Ions from Peptide Modifications

Facile fragmentation of some chemical modifications yields adduct-derived product ions, which appear at characteristic m/z values. MS-MS of peptides adducted at histidine with polycyclic aromatic hydrocarbon diol epoxides yielded characteristic fragment ions for the protonated diol epoxide moiety as well as product ions represent losses of H_2O and (H_2O + CO) from the protonated diol epoxide moiety (Harriman *et al.*, 1998). Pyrrole adducts derived from cysteinyl peptide adducts of the pyrrolizidine alkaloid metabolite dehydromonocrotaline also generated a protonated pyrrole fragment ion or its dehydration product (Hansen *et al.*, 2001). An *N,N*-dithiocarbamoyl adduct on a cysteine residue fragmented at the thiocarbamoyl C–S bond to yield a prominent m/z 100 fragment (Shen *et al.*, 2000). One characteristic of most adduct-derived ions is their relatively small m/z value (typically < 300). These ions may not be detected in ion trap MS-MS analyses of peptides of $m/z > \sim 1200$, because of the inability of the analyzer to resolve ions <25% of the precursor m/z (Jonscher and Yates, 1997). However, such ions would be detected with triple quadrupole or quadrupole-time of flight analyzers.

B. Neutral and Charged Losses from Peptide Modifications

Modified peptide ions can eliminate neutral or charged fragments derived from the modification. In the case of neutral losses, the resulting peptide ion is of lower m/z [Eq. (1)]. Charged losses result from elimination of a charged fragment from a multiply charged peptide

TABLE I
Representative Tandem Mass Spectrometry Spectral Characteristics of Endogenous and Xenobiotic-Derived Peptide Modifications

Modification (amino acid)[a]	Spectral characteristics[b]	Ref(s).
Phosphorylation (S, T)	NL 98; IP 167 (pS), 181 (pT)	Qin and Chait, 1997; Schroeder *et al.*, 1990; DeGnore and Qin, 1998
Phosphorylation (Y)	IP 243	DeGnore and Qin, 1998
Nitration (Y)	IP 208	Jiao *et al.*, 2001
Benzoquinone (C)	NL 142; IP 142, 211	Mason and Liebler, 2000
Pyrrolizidine dehydropyrrole (M+135 adduct) (C)	PI 118, 120, 136; NL 135; CL 135	Hansen *et al.*, 2001
Pyrrolizidine dehydropyrrole (M+117 adduct) (C)	PI 118, 120; NL 117; CL 117	Hansen *et al.*, 2001
N,N-Diethylthiocarbamoyl (C)	PI 100, 72; IP 202; CL 100	Shen *et al.*, 2000
Benzo [*a*]pyrene diol epoxide (H)	PI 303, 285, 257; NL 303; CL 303	Harriman *et al.*, 1998
Chrysene diol epoxide (U)	PI 279, 261, 233; NL 279; CL 279	Harriman *et al.*, 1998
5-Methylchrysene diol epoxide (U)	PI 293, 275, 247; NL 293; CL 293	Harriman *et al.*, 1998
Benzo[*g*]chrysene diol epoxide (U)	PI 328, 311, 283; NL 328; CL 328	Harriman *et al.*, 1998
S-[2-*S*-Cysteinyl)acetyl] glutathione (C)	PI 75, 129, 178, 274, 307, 349; NL 75, 129, 273, 307, 348; IP 450	Jones and Liebler, 2000
S-[2-*S*-Cysteinyl)ethyl]glutathione (C)	NL 75, 129; IP 436	Erve *et al.*, 1995a,b
S-Carboxymethyl (C)	IP 161; NL 60	Jones and Liebler, 2000
2-Chloroacetyl (C, K, H)	IP 179 (C), 213 (H), 204 (K); NL 36, 78, 96, 114	Jones and Liebler, 2000
4-Hydroxynonenal (H)[c]	IP 295	Bolgar *et al.*, 1996

[a]Standard one-letter codes for the modified amino acids are used. U denotes that an unspecified amino acid was adducted.
[b]Adduct-specific characteristics are those described in Fig. 1. Spectral characteristics apply only to positive ion tandem MS done by low-energy collision-induced dissociation on ion trap and triple quadrupole instruments. The abbreviations used are PI, product ion; NL, neutral loss, CL, charged loss; IP, ion pair.
[c]The adduct was reduced with sodium borohydride before analysis.

adduct ion to yield an adduct-derived product ion (see above) and a peptide ion of lower charge state [Eq. (2)]. In most such cases, the adduct is of lower mass than the peptide and the resulting peptide ion appears at higher m/z than the peptide adduct precursor ion.

$$\text{Peptide} - \text{X}^{2+} \rightarrow \text{peptide}^{2+} + \text{X} \quad (1)$$

$$\text{Peptide} - \text{X}^{2+} \rightarrow \text{peptide}^{+} + \text{X}^{+} \quad (2)$$

Perhaps the best-known example of a neutral loss from modified peptide ions is the β elimination of phosphoric acid (H_3PO_4) from phosphoserine and phosphothreonine residues (DeGnore and Qin, 1998; Qin and Chait, 1997; Schroeder *et al.*, 1990). However, a number of chemical adducts also result in neutral losses from peptide adduct ions in MS/MS. In some cases, elimination of the adduct moiety accounts for the neutral loss, as has been observed for polycyclic aromatic hydrocarbon diol epoxide adducts (Harriman *et al.*, 1998), pyrrole adducts from dehydromonocrotaline (Hansen *et al.*, 2001), styrene oxide adducts on N-terminal amines or cysteine (Badghisi and Liebler, 2002), and 4-hydroxynonenal Michael adducts on histidine or cysteine residues (Bolgar *et al.*, 1996; D. C. Liebler and A. J. L. Ham, unpublished observations). In other cases, the neutral loss results from fragmentation of the adduct moiety, as we and others have observed in fragmentation of adducts that incorporate the tripeptide glutathione as part of the adduct (Erve *et al.*, 1995a,b; Jones and Liebler, 2000). These glutathione-containing adducts fragment with characteristic neutral losses of glutamate, glycine, and other glutathione-derived fragments. In some other cases, the adducts fragment with loss of the adduct moiety together with a fragment derived from the adducted peptide. This has been observed in the fragmentation of benzoquinone adducts on peptide cysteine residues, which eliminate the hydroquinone moiety plus –SH (D. E. Mason and D. C. Liebler, unpublished observations).

Charged losses occur when multiply charged peptide adduct ions eliminate a charged fragment to yield an adduct-derived product ion (see above) and a peptide ion of lower charge state than the precursor. This has been observed with polycyclic aromatic hydrocarbon diol epoxides (Harriman *et al.*, 1998) and with pyrrole adducts derived from dehydromonocrotaline (Hansen *et al.*, 2001). It is notable that these species fragment either by neutral loss or charged loss pathways with the generation of adduct-derived product ions. Another characteristic of such adducts is the tendency of the neutral or charged loss pathways to dominate fragmentation to the extent that little or no fragmentation of the adducted peptide is observed (Hansen *et al.*, 2001; Harriman *et al.*, 1998). Thus, although the appearance of characteristic product ions or

losses could indicate that the MS/MS spectrum is from a peptide bearing one of these modifications, it would be difficult to determine the identity of the targeted peptide without further experimental work.

C. *Characteristic Ion Pair Signals from Peptide Modifications*

Modified amino acids can yield characteristic pairs of ions that appear in either b- or y-ion series on MS-MS of the modified peptide. For example, a cysteine has a residue mass of 103, which means that a b- or y-ion series of a peptide containing a cysteine residue will contain a pair of signals separated on the m/z axis by 103 units. Such pairs of signals can be characteristic for certain modified peptides. Phosphorylation of tyrosine gives rise to a pair of signals separated by 243 units (DeGnore and Qin, 1998), whereas nitration of tyrosine gives rise to a pair of signals separated by 208 units (Jiao *et al.*, 2001). Other examples of such ion pair signals include phosphorylation of serine (167 units) or threonine (181 units) (DeGnore and Qin, 1998), hydroquinone adducts of cysteine (211 units) (Mason and Liebler, 2000), 4-hydroxynonenal on histidine (295 units) (Bolgar *et al.*, 1996), 2-chloroacetylation of histidine (213 units) or lysine (204 units) (Jones and Liebler, 2000), and S-carboxymethylation of cysteine (161 units) (Jones and Liebler, 2000).

Ion pair signals can be a useful means of identifying peptides bearing modifications to specific amino acids, but the practical utility of these features depends on their intensity relative to other spectral features. Any given pair of signals may occur in MS-MS spectra as a result of noise or contamination. Indeed, examination of large data sets for MS-MS spectra containing ion pairs separated by 161 units (for *S*-carboxymethylcysteine) yields many ''hits'' that cannot be confirmed as S-carboxymethylated peptides, but instead reflect other structural features that result in signals separated by 161 units (Jones and Liebler, 2000). Nevertheless, the effect of a modification on a series of ion signals (as opposed to a single pair) can be highly diagnostic for the occurrence of the modification in a defined sequence motif (see below).

III. SALSA: A Pattern Recognition Algorithm to Identify MS-MS Spectra for Modified Peptides

A. *Rationale for SALSA*

The collection of a large data set of MS-MS spectra for all of the peptides in a sample should enable the detection of both modified and unmodified peptide forms. Automated acquisition of MS-MS spectra for large numbers

of peptides in complex mixtures is now possible. Analysis of such data sets with Sequest or other data mining tools allows the identification of proteins in the sample by correlating MS-MS data with database peptide sequences. When likely modifications can be anticipated (e.g., phosphorylation on tyrosine, serine, or threonine), these tools can enable assignment of MS-MS spectra for the modified peptides to the correct sequences. However, unanticipated modifications pose a problem for this approach, as neither the mass of the modifications nor their amino acid sequence locations are known. As described above, modifications may give rise to diverse combinations of product ions, characteristic losses, or ion pairs, but the appearance of any of these characteristics may vary somewhat with individual peptide sequence contexts. Thus, a data evaluation tool suited to detecting multiple adduct-specific features in MS-MS spectra is needed to selectively identify spectra of modified peptides.

We reported the development of a novel algorithm and software called SALSA (scoring algorithm for spectral analysis) (Hansen *et al.*, 2001). SALSA evaluates MS-MS spectra for specific, user-defined features, including product ions at specific m/z values, neutral or charged losses from singly or doubly charged precursors, and ion pairs or series (Fig. 1).

Essential features of SALSA are described briefly below and have been detailed elsewhere.

- Averaged CID spectra are preprocessed to subtract nonfragment ions, to distinguish spectra of singly charged precursors from those of doubly charged precursors, and to normalize ion intensities as a percentage of the total ion current (%TIC).
- Processed spectra then are evaluated for user-specified spectral characteristics. These include product ions, neutral losses, charged losses, and ion pairs. SALSA scores specific product ions by identifying the most abundant ion within a user-selectable window centered at the designated m/z value and recording the %TIC value for the selected ion. Neutral losses and charged losses from doubly charged precursors are scored in an analogous manner. Ion pairs are defined as two fragment ions that are a specified distance apart on the m/z axis. If a match exists, the ion pair is scored as the geometric mean of the %TIC values for the detected ions.
- To reduce background noise in scoring, each search criterion is designated either as primary or secondary. Whereas primary criteria are automatically scored when detected, a secondary criterion is

scored only when the linked primary criterion is detected in the same MS-MS spectrum. Thus, the scoring of secondary criteria is contingent on the presence of other primary indicators.

- The output of the SALSA algorithm is in the form of a list of MS-MS scans ranked in order of decreasing score. In addition, SALSA provides the precursor m/z, retention time, scan numbers, and ion signals that matched search criteria for each spectrum.

B. Application of SALSA to Detection of MS-MS Spectra of Peptide Adducts

We first illustrated the utility of data-dependent LC-MS-MS analysis and SALSA in the detection of peptide adducts spiked into a tryptic digest of bovine serum albumin, rat glutathione-*S*-transferase, and equine

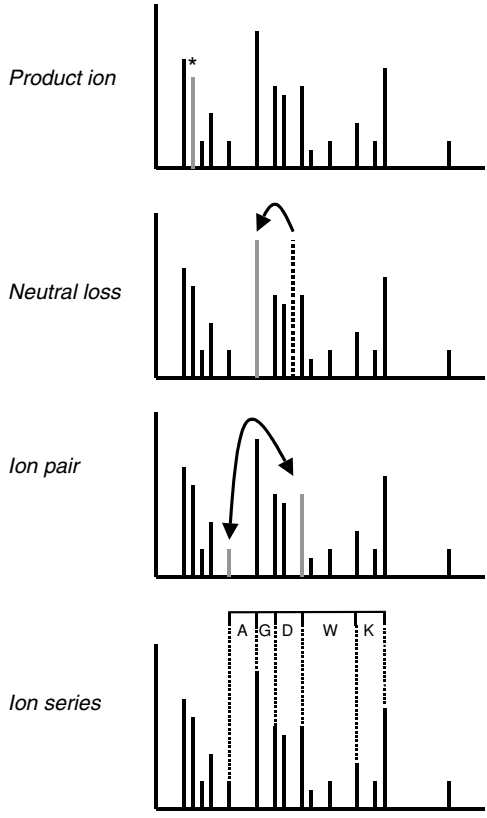


FIG. 1. MS-MS spectral features detected by the SALSA algorithm. See text for discussion.

apomyoglobin (Hansen *et al.*, 2001). The sample was analyzed by data-dependent LC-MS-MS on a quadrupole ion trap instrument. Previous studies with model peptide adducts had indicated specific fragmentation characteristics of dehydropyrrole (DHP), *S*-cysteinylhydroquinone (HQ), and *S*-carboxymethyl (CM) peptide adducts. For DHP adducts, neutral and charged losses (from singly and doubly charged precursors, respectively) of 117 and 135 amu were entered as primary search criteria with neutral loss of H_2O from either singly or doubly charged precursors as secondary to the neutral or charged loss of 135 amu. Neutral losses from doubly charged precursors of 117 and 135 amu also were entered as secondary criteria linked to the corresponding charged losses. SALSA search criteria for HQ adducts include the neutral loss of 142 amu as a primary criterion, with the neutral loss of 160 amu and ion pairs separated by 142 and 211 amu as secondary criteria linked to the primary neutral loss (Mason and Liebler, 2000). CM-cysteine peptides typically do not display adduct-specific product ions or neutral losses (Jones and Liebler, 2000), so an ion pair separated by 161 amu was used as the only search criterion.

Table II lists the distributions of SALSA scores for MS-MS scans from the three LC-MS-MS analyses of the peptide adduct-spiked digests. The great majority of the MS-MS scans received relatively low SALSA scores [log (SALSA score + 1) values < 0.5]. In contrast, the MS-MS scans of the peptide adducts all received log(SALSA score + 1) values above 1.0. Thus, SALSA clearly differentiated peptide adduct MS-MS scans from scans for other species. SALSA analyses thus serve as a means for quickly ranking all the MS-MS scans in the data file based on their correspondence to adduct-specific search criteria. These data underscore two other key points about SALSA. First, SALSA scores are highly dependent on search strategy and spectral characteristics and are useful only as relative measures of concordance between spectral features and search criteria. Second, the ability of SALSA to distinguish MS-MS spectra of specific peptide adduct depends on its fragmentation characteristics and the extent to which these characteristics distinguish its MS-MS spectrum from all others. In the case of the DHP and HQ adducts, there is a clear resolution of adduct MS-MS spectra from all others, as these adducts display multiple characteristic features. *S*-Cysteinyl-CM adducts, on the other hand, display only a single ion pair as an adduct-specific characteristic. The frequency with which this ion pair appears in MS-MS spectra of nontarget peptides is relatively high. Thus, SALSA scores for the CM adduct spectra are less effectively distinguished from those of other MS-MS spectra.

TABLE II
Distributions of SALSA Scores for MS-MS Spectra from a Tryptic Digest of Bovine Serum Albumin, Rat Glutathione-S-transferase, and Equine Myoglobin Spiked with Peptide Adducts[a]

log (SALSA score + 1)	LVACGAK-DHP-spiked digest (20 pmol/μg protein)	AVAGCAGAR-HQ-spiked digest (57 pmol/μg protein)	AVAGCAGAR-CM-spiked digest (12 pmol/μg protein)
0	963	662	431
0.01–0.2	7		
0.21–0.4	2	3	
0.41–0.6	2	5	96
0.61–0.8			233
0.81–1.0			79
1.01–1.2		4 (1.19, 1.16)	8 (1.02)
1.21–2.0	3 (1.74, 1.73, 1.69)		1

[a] Numbers of MS-MS scans with log(SALSA score + 1) in the indicated ranges are listed for each analysis. Values of log(SALSA score + 1) for actual peptide adduct spectra are listed in parentheses.

C. Sequence Motif Analysis of Peptide MS-MS Data with SALSA

Many peptide modifications are like *S*-cysteinyl-CM adducts in that they do not necessarily produce characteristic product ions and neutral or charged losses in MS-MS. We realized that SALSA-assisted identification of MS-MS spectra from such modified peptides would require a more sensitive, robust search approach than ion pairs. To meet this need, SALSA was modified to enable the detection of ion series, which consist of two or more signals in fixed relationship to each other (Liebler *et al.*, 2002). These series can correspond to b- or y-ion series that are indicative of specific peptide sequences. When the user inputs a peptide sequence motif (i.e., a string of amino acids corresponding to part of a peptide sequence), SALSA generates a ''virtual ruler'' for comparison with spectra (Fig. 2). The ruler is matched to each MS-MS spectrum in a data file in multiple alignments (Fig. 2A and B). Scans that match part of the ion series defined by the ruler may include those with missing ion signals (Fig. 2C). However, peptides bearing modifications or amino acid substitutions will generate MS-MS spectra with ion series that provide at least partial matches to the ruler. SALSA scores are calculated on the basis of the intensities of the MS-MS signals that match the ruler (Fig. 2D). This application of SALSA is particularly valuable for finding MS-MS spectra from specific peptide sequences and can distinguish highly similar peptides of the same m/z value, but with subtle sequence differences (Liebler *et al.*, 2002). A key

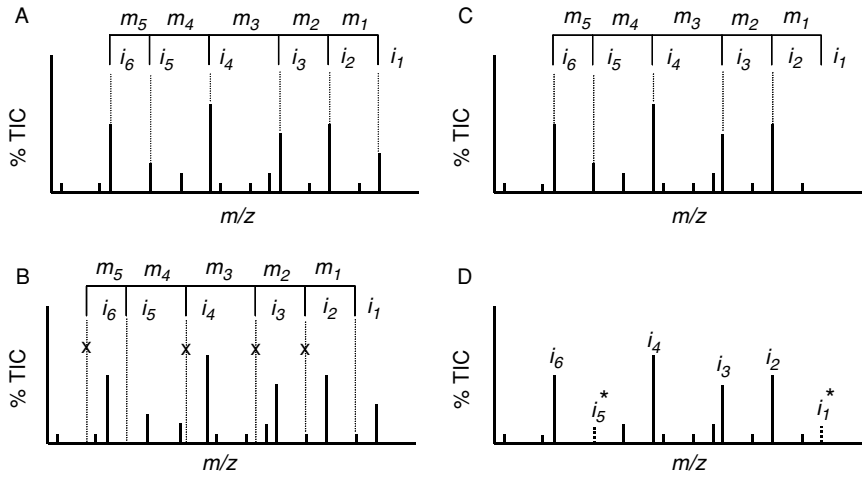


FIG. 2. (A–D) SALSA ion series detection and scoring scheme. See text for discussion. [Reproduced from Liebler *et al.* (2002), with permission from the American

advantage of detecting modifications by this approach is that it is not necessary to anticipate the mass or sequence specificity of the modification. SALSA is thus suited to the detection of MS-MS spectra displaying either posttranslational modifications or sequence variations.

As noted above, SALSA scores are determined by several factors including (1) the search strategy used, (2) the length of the search motif, (3) the number of ions that match the search series, and (4) the intensities of the scored ions (Liebler *et al.*, 2002). SALSA scores do not provide an absolute measure of spectral quality or the fidelity of the match between the search motif and the MS-MS spectrum. Thus, the absolute values of SALSA scores are less important than the relative values for ranking MS-MS scans in a data set. A ranking of the MS-MS scans by SALSA score quickly identifies those MS-MS scans originating from the target peptide or its modified or variant forms.

IV. Mapping Protein Modifications with SALSA

A. *Sequence Mapping Oxidative Modifications in Bovine Serum Albumin*

Initial work in our laboratory focused on the detection of modified forms of bovine serum albumin (BSA) (Liebler *et al.*, 2002). These studies analyzed commercially purchased BSA (Sigma, St. Louis, MO), which was not subjected to any treatment with oxidants or other modifiers before

analysis. Tryptic digests of BSA were analyzed by LC-MS-MS and the data then were analyzed with Sequest and SALSA. The results are summarized in Fig. 3. The highlighted sequences (is boldface and underlined) indicate peptides to which MS-MS spectra were assigned by Sequest and SALSA. The first number in parentheses below each highlighted peptide corresponds to the number of MS-MS scans assigned to the sequence by Sequest. Sequest analysis of the data identified MS-MS spectra corresponding to 37 BSA tryptic peptides and accounting for 66.2% coverage by amino acid sequence. A SALSA analysis of the same data file was performed with ion series searches corresponding to the central sequence of each peptide. The second number in parentheses corresponds to the number of MS-MS spectra assigned to the sequence by SALSA and the third number indicates the number of modified or variant forms of that peptide sequence assigned to MS-MS spectra by SALSA. SALSA detected the MS-MS spectra for unmodified peptides also detected by Sequest. However, SALSA also detected MS-MS spectra of several variant peptides not assigned by Sequest. All detected MS-MS scans assigned as variants of a target sequence

DTHK**SEIAHR** (1,1,0) FK**DLGEEHFK** (3,3,0) GLVLIAFSQYLQQCPFDEHVK**LVNELTEFAK** (2,2,0) **TCVADESHAGCEK** (1,2,1)

SLHTLFGDELCK (1,2,1) **VASLR** (1,1,0) **ETYGDMADCCEK** (1,1,0) QEPERNECFLSHKDDSPDLPK**LKPDPNTLCDEFK** (1,1,0)

ADEKKFWGK**YLYEIAR** (1,1,0) **RHPYFYAPELLYYANK** (1,2,1) **YNGVFQECCQAEDK** (2,3,1) **GACLLPK** (1,1,0) IETMREKVLASSAR

QRLRCASIQKFGERALKAWSVARLSQKFPK**AEFVEVTK** (2,2,0) **LVTDLTK** (1,2,1) **VHKECCHGDLLECADDR** (1,1,0) ADLAK

YICDNQDTISSK (2,2,1) **LKECCDKPLLEK** (2,2,0) **SHCIAEVEK** (1,1,0) DAIPENLPPLTADFAEDKDVCKNYQEAK

DAFLGSFLYEYSR (3,4,1) **RHPEYAVSVLLR** (3,3,0) LAK**EYEATLEECCAK** (1,1,0) **DDPHACYSTVFDK** (1,1,0) LK**HLVDEPQNLIK** (3,3,0)

QNCDQFEK (1,1,0) **LGEYGFQNALIVR** (5,6,1) YTR**KVPQVSTPTLVEVSR** (3,5,2) SLGKVGTRCCTKPESER**MPCTEDYLSLILNR** (2,6,4)

LCVLHEK (1,1,0) **TPVSEK** (1,1,0) VTK**CCTESLVNR** (1,4,4) **RPCFSALTPDETYVPK** (1,1,0) AFDEK**LFTFHADICTLPDTEK** (3,5,3) QIK

KQTALVELLK (2,2,0) HKPKATEEQLK**TVMENFVAFVDK** (3,6,3) **CCAADDKEACFAVEGPK** (1,7,4) **LVVSTQTALA** (1,1,0)

FIG. 3. BSA tryptic peptides for which MS-MS scans were detected by Sequest and SALSA. Detected peptides are indicated by sequences in bold face and underlined. Numbers in parentheses in the format (*a*,*b*,*c*) beneath each highlighted sequence indicate (*a*) the number of MS-MS scans for the indicated sequence detected by Sequest, (*b*) the number of MS-MS scans for the indicated sequence detected by SALSA, and (*c*) the number of MS-MS scans for variants of the indicated sequence detected by SALSA. [Reproduced from Liebler *et al.* (2002), with permission from the American Chemical Society.]

displayed strong y-ion series identity or homology with ion series for the unmodified peptides. MS-MS spectra corresponding to modified or variant peptide forms were found for 14 BSA peptides detected. Inspection of the MS-MS scans corresponding to variant forms of the peptides MPCTEDYL-SLILNR and CCAADDKEACFAVEGPK indicated that they were due primarily to M + 16 and M + 32 variants reflecting oxidative modification at the cysteine and cysteine / methionine, respectively.

B. Sequence Mapping Xenobiotic Adducts in Hemoglobin

Reactive intermediates derived from a variety of xenobiotics react with hemoglobin (Hb) to form adducts, which have been widely studied as markers of exposure to these chemicals (for reviews, see Ehrenberg *et al.*, 1996; Farmer, 1995). Previous work to identify and quantify adducts relied on indirect assays and a modified Edman degradation procedure (Pauwels *et al.*, 1997; Rappaport *et al.*, 1993; Sepai *et al.*, 1993; Tornqvist *et al.*, 1986; Yeowell-O'Connell *et al.*, 1996). MS had been used previously to identify a histidine adduct of styrene oxide in human Hb (Kaur *et al.*, 1989) and, more recently, Moll and colleagues (2000) employed electrospray MS to identify peptides adducted with butadiene monoxide in mouse erythrocytes *in vitro*. Although this was the first systematic application of MS to comprehensive mapping of Hb adducts, the analyses did not include tandem MS to provide sequence-specific mapping of adducts.

We applied LC-MS-MS and SALSA to the sequence-specific mapping of aliphatic epoxide adducts formed with human hemoglobin *in vitro* (Badghisi and Liebler, 2002). Human hemoglobin was incubated with 40 m*M* styrene oxide in isotonic saline for 6 h at 37°C and the hemoglobin then was recovered and digested with trypsin, and the peptides were analyzed by LC-MS-MS. Analysis of the data with Sequest indicated that MS-MS spectra assigned to expected tryptic peptides yielded 79% coverage (by amino acid sequence) for Hbα and 86% coverage for Hbβ. SALSA analyses of the data files for tryptic peptide sequence motifs identified MS-MS spectra of both unmodified and modified peptides. A SALSA search for MS-MS spectra corresponding to the N-terminal tryptic peptide VHLTPEEK (residues 1–8) of Hbβ employed the search motif HLTPEE to detect spectra containing ion series that matched the expected y-ion series for the peptide (Liebler *et al.*, 2002). SALSA awarded scores of >20 to 17 MS-MS spectra that displayed ion series matching the search motif. The highest scoring MS-MS scans resulted from collision-induced dissociation (CID) of a precursor ion of m/z 537.4, which corresponded to the doubly charged ion of the styrene oxide monoadduct. Inspection of

these MS-MS spectra indicated the expected y-ion series for the target peptide, whereas the b-series ions were shifted by +120 amu, which corresponds to the mass of the styrene oxide adduct. This information unambiguously established the location of the adduct at the peptide N terminus. In addition, the presence of a peak at m/z 853.3 corresponds to unmodified y_7-ion, thus providing additional support that the modification is at the N-terminal valine, rather than on the adjacent histidine residue. The MS-MS spectrum of the doubly charged ion of the styrene oxide adduct also displays a prominent neutral loss of 120 amu, which corresponds to the styrene oxide moiety. As noted above, adduct-specific neutral and charged losses as well as adduct-derived product ions are characteristic features of MS-MS fragmentation of peptide adduct ions (Liebler *et al.*, 2002; Hansen *et al.*, 2001; Harriman *et al.*, 1998). SALSA also detected MS-MS scans for the singly charged ion of the styrene oxide adduct and both the doubly and singly charged ions of the unmodified VHLTPEEK peptide.

SALSA also scored MS-MS scans corresponding to two unanticipated variants of the target peptide (Badghisi and Liebler, 2002). The corresponded to a modification of +28 amu, probably due to an N-terminal formylation. The second corresponded to a modification of +43 amu, probably due to an N-terminal carbamylation. The formylation may have been in the original protein sample or may be due to formic acid in the mobile phase, whereas the N-terminal carbamylation is frequently observed in analyses of peptides from enzymatic digests done in the presence of urea (Stark *et al.*, 1960; Stark, 1965).

We further extended this approach to map hemoglobin adducts formed by a mixture of aliphatic epoxides (Badghisi and Liebler, 2002). Hemoglobin incubated with a mixture of styrene oxide, ethylene oxide, and butadiene dioxide (all at 40 mM) in isotonic saline at 37°C for 6 h yielded a mixture of adducts. The epoxide-treated protein mixture then was digested with trypsin and analyzed by LC-MS-MS with data-dependent scanning.

A SALSA search with sequence motifs corresponding to all hemoglobin peptides was used to detect MS-MS scans for the native peptides and their adducts. This analysis detected MS-MS spectra for modified variants of five Hbβ and two Hbα tryptic peptides in datafile from LC-MS-MS analysis of the alkylated hemoglobin mixture. The results are summarized in the adduct map in Fig. 4. In the Hbβ chain, amino acid targets included histidine (His-77, His-97, and His-143), Cys-93, and the N-terminal amine. In the Hbα chain, modification was observed only at the N-terminal valine and at a single histidine residue (His-20). Incubation of hemoglobin with the epoxide mixture at 40 μM yielded only the Hbα and Hbβ N-terminal valine adducts and the Hbβ Cys-93 adducts of the epoxides. The position of modification in each case was verified by inspection of the MS-MS

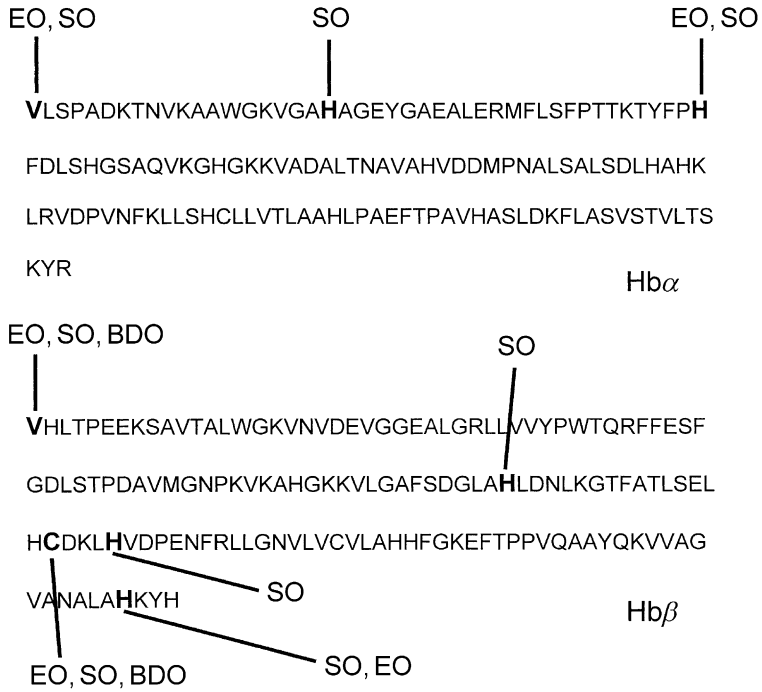


FIG. 4. Map of epoxide adducts detected in human hemoglobin by LC-MS-MS and SALSA analysis. The sequence locations of adducts from styrene oxide (SO), ethylene oxide (EO), and butadiene dioxide (BDO) are indicated.

spectra, in which the b- and/or y-series ions all were displaced by the mass of the adduct.

Modification by styrene oxide and ethylene oxide corresponded to addition of the epoxide mass to the target peptide via a nucleophilic attack at an epoxide carbon with ring opening to form a β-hydroxy-substituted adduct. Modification by butadiene dioxide produced two types of adducts that were distinguished by adduct mass. These included two possible M+86 adducts, which result from direct addition of the protein nucleophile to the diepoxide, and an M+104 adduct, which arises from nucleophilic attack at one epoxide with hydrolytic ring opening of the other.

This work with hemoglobin demonstrates that LC-tandem MS and SALSA can be applied to the identification and sequence-specific mapping of protein-xenobiotic adducts. An important advantage of employing LC-MS-MS and SALSA is that modifications to a protein can be detected without prior knowledge of the exact chemical nature of the modifying species. In previous studies of chemical modification of proteins, the

investigators knew the chemical nature of the modifying species and thus could directly search the MS data for MS-MS spectra of known peptides modified by chemicals of known mass. However, unanticipated modifications would not be detectable by this approach. Moreover, some modifications by known electrophiles may yield adducts that undergo adventitious oxidation, hydrolysis, or other modifications during sample workup. These unanticipated modifications would yield adducted peptides that would not be identified by directly searching the data for peptides of the expected mass modification. The advantage of SALSA in these situations is that it can identify MS-MS spectra for variant and modified peptides even when the exact nature and location of the modification are not known. This capability is possible because SALSA searching for spectra with ion series motifs will detect ion series patterns, which are at least partially conserved in spectra of modified peptide forms (Liebler *et al.*, 2002).

SALSA has many applications in proteomics research beyond the mapping of xenobiotic adducts. Endogenous posttranslational modifications, sequence variants, splice variants, and other modified protein forms all could be detected with LC-MS-MS and SALSA. Regardless of the modification type, the fundamental requirement for reliable detection by SALSA is that MS-MS spectra for the modified or variant peptide forms are actually acquired. The utility of SALSA is dependent on the quality of the data set; if the MS-MS spectra are not acquired, SALSA cannot find them. In many cases, modified or variant protein forms may be present at low abundance relative to unmodified forms. Improvements in MS hardware and in LC separation strategies (e.g., tandem LC) will increase the chances of obtaining MS-MS spectra of these low-abundance components of complex mixtures (Washburn *et al.*, 2001; Wolters *et al.*, 2001). In other cases, modifications may be on peptides that are either too long or too short to generate good MS-MS spectra. This probably will require digestion strategies employing multiple enzymes with complementary cleavage specificities (Gatlin *et al.*, 2000).

V. Comparison of SALSA with Other Software for Analysis of MS-MS Data

Other algorithms and software tools already are available for the analysis of tandem MS data including Sequest (Eng *et al.*, 1994), Mascot (Perkins *et al.*, 1999), Pep-Frag (Fenyo *et al.*, 1998), MS-Tag (Clauser *et al.*, 1999), and Pep-Sea (Mann and Wilm, 1994). These tools identify proteins from tandem MS data by comparing features of the spectra with those of

theoretical spectra of protein sequences in databases. SALSA differs from these other tools in two fundamental ways. First, SALSA finds MS-MS spectra that display characteristics of a sequence motif, rather than database sequences that match spectra. In other words, Sequest uses MS-MS scans to mine databases; SALSA uses sequences to mine MS-MS scans. Second, the SALSA algorithm searches for spectral features without regard either to a peptide ion precursor m/z or specific m/z values of product ions. This allows SALSA to identify MS-MS scans that contain m/z signals in some specified relation to each other, regardless of their absolute positions on the m/z axis. SALSA then can identify MS-MS scans that correspond to the target peptide and any variants, whether anticipated or not.

The ability of SALSA to identify unanticipated variants is a distinct advantage over other data analysis tools. Sequest and related programs can correlate MS-MS data of modified peptides with database sequences if the user specifies the nature of the modification and the amino acid modified (Yates *et al.*, 1995). Nevertheless, unanticipated modifications preclude correct correlation of precursor m/z or the MS-MS data with database protein sequences. SALSA can be used to best advantage when combined with Sequest and similar tools for mining proteomic diversity. Initial analysis of MS-MS data files with Sequest identifies proteins that are represented in the analyzed sample. The identified protein sequences then form the basis for generating SALSA motif searches to mine the data for MS-MS scans corresponding to modified and variant peptide forms. Inspection of these MS-MS scans allows confirmation of the sites and masses of modifications.

LC-MS-based approaches to proteome analysis will continue to improve as (1) instrumentation and techniques allow acquisition of more complete data sets for complex peptide mixtures and (2) data-mining algorithms and software allow more comprehensive extraction of the data contained therein. The large, complex data sets generated by increasingly comprehensive proteome analyses will contain information not only about the identities of diverse proteins, but also their many variant and modified forms. A great challenge in working with this expected wealth of data is the higher order assembly of the Sequest and SALSA search outputs in forms that are accessible to biologists. The DTASelect and Contrast programs (Tabb *et al.*, 2002), which assemble the results from multiple Sequest analyses to analyze complex protein mixtures, represent a new generation of tools that integrate complex data sets. Further development and elaboration of the approaches used in SALSA-based analyses will also be needed to accurately map modifications and variant forms in complex proteomes. This challenge is the object of continuing work in our laboratory.

ACKNOWLEDGMENTS

The authors acknowledge support from NIH Grants ES10056, ES06694, and ES07091 and from ThermoFinnigan, L.L.C.

REFERENCES

Badghisi, H., and Liebler, D. C. (2002). Sequence mapping of epoxide adducts in human hemoglobin with LC-tandem MS and the SALSA algorithm. *Chem. Res. Toxicol.* **15,** 799–805.

Bolgar, M. S., Yang, C. Y., and Gaskell, S. J. (1996). First direct evidence for lipid/protein conjugation in oxidized human low density lipoprotein. *J. Biol. Chem.* **271,** 27999–28001.

Clauser, K. R., Baker, P., and Burlingame, A. L. (1999). Role of accurate mass measurement (10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **71,** 2871–2882.

Cohen, S. D., Pumford, N. R., Khairallah, E. A., Boekelheide, K., Pohl, L. R., Amouzadeh, H. R., and Hinson, J. A. (1997). Selective protein covalent binding and target organ toxicity. *Toxicol. Appl. Pharmacol.* **143,** 1–12.

DeGnore, J. P., and Qin, J. (1998). Fragmentation of phosphopeptides in an ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **9,** 1175–1188.

Dipple, A. (1995). DNA adducts of chemical carcinogens. *Carcinogenesis* **16,** 437–441.

Ehrenberg, L., Granath, F., and Tornqvist, M. (1996). Macromolecule adducts as biomarkers of exposure to environmental mutagens in human populations. *Environ. Health Perspect.* **104** (Suppl. 3)**,** 423–428.

Eng, J. K., McCormack, A. L., and Yates, J. R. (1994). 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.

Erve, J. C., Barofsky, E., Barofsky, D. F., Deinzer, M. L., and Reed, D. J. (1995a). Alkylation of *Escherichia coli* thioredoxin by *S*-(2-chloroethyl)glutathione and identification of the adduct on the active site cysteine-32 by mass spectrometry. *Chem. Res. Toxicol.* **8,** 934–941.

Erve, J. C., Deinzer, M. L., and Reed, D. J. (1995b). Alkylation of oxytocin by *S*-(2-chloroethyl)glutathione and characterization of adducts by tandem mass spectrometry and Edman degradation. *Chem. Res. Toxicol.* **8,** 414–421.

Farmer, P. B. (1995). Monitoring of human exposure to carcinogens through DNA and protein adduct determination. *Toxicol. Lett.* **82–83,** 757–762.

Fenyo, D., Qin, J., and Chait, B. T. (1998). Protein identification using mass spectrometric information. *Electrophoresis* **19,** 998–1005.

Garner, R. C. (1998). The role of DNA adducts in chemical carcinogenesis. *Mutat. Res.* **402,** 67–75.

Gatlin, C. L., Eng, J. K., Cross, S. T., Detter, J. C., and Yates, J. R., III. (2000). Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Anal. Chem.* **72,** 757–763.

Hansen, B. T., Jones, J. A., Mason, D. E., and Liebler, D. C. (2001). SALSA: A pattern recognition algorithm to detect electrophile-adducted peptides by automated evaluation of CID spectra in LC-MS-MS analyses. *Anal. Chem.* **73,** 1676–1683.

Harriman, S. P., Hill, J. A., Tannenbaum, S. R., and Wishnok, J. S. (1998). Detection and identification of carcinogen-peptide adducts by nanoelectrospray tandem mass spectrometry. *J. Am. Soc. Mass Spectrom.* **9,** 202–207.

Jiao, K., Mandapati, S., Skipper, P. L., Tannenbaum, S. R., and Wishnok, J. S. (2001). Site-selective nitration of tyrosine in human serum albumin by peroxynitrite. *Anal. Biochem.* **293,** 43–52.

Jones, J. A., and Liebler, D. C. (2000). Tandem MS analysis of model peptide adducts from reactive metabolites of the hepatotoxin 1,1-dichloroethylene. *Chem. Res. Toxicol.* **13,** 1302–1312.

Jonscher, K. R., and Yates, J. R. (1997). The quadrupole ion trap mass spectrometer—a small solution to a big challenge. *Anal. Biochem.* **244,** 1–15.

Kaur, S., Hollander, D., Haas, R., and Burlingame, A. L. (1989). Characterization of structural xenobiotic modifications in proteins by high sensitivity tandem mass spectrometry: Human hemoglobin treated in vitro with styrene 7,8-oxide. *J. Biol. Chem.* **264,** 16981–16984.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P.,

Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., and Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.

Liebler, D. C. (2001). ''Introduction to Proteomics: Tools for the New Biology.'' Human Press, Totowa, NJ.

Liebler, D. C., Hansen, B. T., Davey, S. W., Tiscareno, L., and Mason, D. E. (2002). Peptide sequence motif analysis of tandem MS data with the SALSA algorithm. *Anal. Chem.* **74,** 203–210.

Mann, M., and Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66,** 4390–4399.

Marnett, L. J., and Plastaras, J. P. (2001). Endogenous DNA damage and mutation. *Trends Genet.* **17,** 214–221.

Mason, D. E., and Liebler, D. C. (2000). Characterization of benzoquinone-peptide adducts by electrospray mass spectrometry. *Chem. Res. Toxicol.* **13,** 976–982.

Miller, E. C., and Miller, J. A. (1981). Searches for ultimate chemical carcinogens and their reactions with cellular macromolecules. *Cancer* **47,** 2327–2345.

Moll, T. S., Harms, A. C., and Elfarra, A. A. (2000). A comprehensive structural analysis of hemoglobin adducts formed after in vitro exposure of erythrocytes to butadiene monoxide. *Chem. Res. Toxicol.* **13,** 1103–1113.

Nelson, S. D., and Pearson, P. G. (1990). Covalent and noncovalent interactions in acute lethal cell injury caused by chemicals. *Annu. Rev. Pharmacol. Toxicol.* **30,** 169–195.

Pandey, A., and Mann, M. (2000). Proteomics to study genes and genomes. *Nature* **405,** 837–846.

Pauwels, W., Farmer, P. B., Osterman-Golkar, S., Severi, M., Cordero, R., Bailey, E., and Veulemans, H. (1997). Ring test for the determination of N-terminal valine adducts of styrene 7,8-oxide with haemoglobin by the modified Edman degradation technique. *J. Chromatogr. B Biomed. Sci. Appl.* **702,** 77–83.

Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567.

Qin, J., and Chait, B. T. (1997). Identification and characterization of posttranslational modifications of proteins by MALDI ion trap mass spectrometry. *Anal. Chem.* **69,** 4002–4009.

Rappaport, S. M., Ting, D., Jin, Z., Yeowell-O'Connell, K., Waidyanatha, S., and McDonald, T. (1993). Application of Raney nickel to measure adducts of styrene oxide with hemoglobin and albumin. *Chem. Res. Toxicol.* **6,** 238–244.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., Cherry, J. M., Henikoff, S., Skupski, M. P., Misra, S., Ashburner, M., Birney, E., Boguski, M. S., Brody, T., Brokstein, P., Celniker, S. E., Chervitz, S. A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R. F., Gelbart, W. M., George, R. A., Goldstein, L. S., Gong, F., Guan, P., Harris, N. L., Hay, B. A., Hoskins, R. A., Li, J., Li, Z., Hynes, R. O., Jones, S. J., Kuehl, P. M., Lemaitre, B., Littleton, J. T., Morrison, D. K., Mungall, C., O'Farrell, P. H., Pickeral, O. K., Shue, C., Vosshall, L. B., Zhang, J., Zhao, Q., Zheng, X. H., Zhong, F., Zhong, W., Gibbs, R., Venter, J. C., Adams, M. D., and Lewis, S. (2000). Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215.

Schroeder, W., Covey, T., and Hucho, F. (1990). Identification of phosphopeptides by mass spectrometry. *FEBS Lett.* **273,** 31–35.

Sepai, O., Anderson, D., Street, B., Bird, I., Farmer, P. B., and Bailey, E. (1993). Monitoring of exposure to styrene oxide by GC-MS analysis of phenylhydroxyethyl esters in hemoglobin. *Arch. Toxicol.* **67,** 28–33.

Shen, M. L., Johnson, K. L., Mays, D. C., Lipsky, J. J., and Naylor, S. (2000). Identification of the protein–drug adduct formed between aldehyde dehydrogenase and *S*-methyl-*N,N*-diethylthiocarbamoyl sulfoxide by on-line proteolytic digestion high performance liquid chromatography electrospray ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **14,** 918–923.

Stark, G. R. (1965). Reactions of cyanate with functional groups of proteins. III. Reactions with amino and carboxyl groups. *Biochemistry* **4,** 1030–1036.

Stark, G. R., Stein, W. H., and Moore, S. (1960). Reactions of cyanate present in aqueous urea with amino acids and proteins. *J. Biol. Chem.* **235,** 3177–3181.

Strickland, P. T., Routledge, M. N., and Dipple, A. (1993). Methodologies for measuring carcinogen adducts in humans. *Cancer Epidemiol. Biomarkers Prev.* **2,** 607–619.

Tabb, D. L., McDonald, W. H., and Yates, J. R. (2002). DTASelect and Contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1,** 21–26.

Tornqvist, M., Mowrer, J., Jensen, S., and Ehrenberg, L. (1986). Monitoring of environmental cancer initiators through hemoglobin adducts by a modified Edman degradation method. *Anal. Biochem.* **154,** 255–266.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di, F. V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri,

J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., and Nodell, M. (2001). The sequence of the human genome. *Science* **291,** 1304–1351.

Washburn, M. P., Wolters, D., and Yates, J. R. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19,** 242–247.

Wolters, D. A., Washburn, M. P., and Yates, J. R. (2001). An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73,** 5683–5690.

Yates, J. R. (1998). Mass spectrometry and the age of the proteome. *J. Mass Spectrom.* **33,** 1–19.

Yates, J. R., Eng, J. K., McCormack, A. L., and Schieltz, D. (1995). Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **67,** 1426–1436.

Yeowell-O'Connell, K., Jin, Z., and Rappaport, S. M. (1996). Determination of albumin and hemoglobin adducts in workers exposed to styrene and styrene oxide. *Cancer Epidemiol. Biomarkers Prev.* **5,** 205–215.

EMERGING ROLE OF MASS SPECTROMETRY IN STRUCTURAL AND FUNCTIONAL PROTEOMICS

By STEPHEN NAYLOR[*,†] AND RAJIV KUMAR[†,‡]

[*]Beyond Genomics, Inc., Waltham, Massachusetts 02451,[†]Department of Biochemistry and Molecular Biology, and [‡]Nephrology Research Unit and Department of Medicine, Mayo Clinic/Foundation, Rochester, Minnesota 55905

I. Introduction

Many cellular processes are controlled and regulated by interactions between metal ions, hormones, endogenous ligands, proteins, and nucleic acids with receptors, enzymes, other proteins, and nucleic acids [1–11]. Furthermore, therapeutic drugs, toxins, and infectious agents frequently act by associating with cellular targets [12–19]. These biomolecular processes principally involve noncovalent interactions of component molecules held together primarily by hydrophobic and electrostatic forces. Hence, such complexes can be transient and the interaction between components is often characterized by low-affinity (micromolar to nanomolar) binding constants. Thus, techniques that delineate the structure and composition of such biological complexes, as well as the stoichiometry and rate of interactions of such components in a physiological environment, are fundamental to a comprehensive understanding of cellular processes. Thus knowledge gained by applying such techniques can be used not only to understand the nature of such interactions but also to devise methods, agents, or drugs that regulate, enhance, or block such interactions.

ADVANCES IN
PROTEIN CHEMISTRY, Vol. 65

0065-3233/03 $35.00

An understanding of this latter set of processes is essential to the overall success of the biotechnology industry. Hence, an ideal technique would be one that allows the examination of biological interactions within the organism itself, does not perturb the system, is nondestructive, or, at least, sparing of material, specific, sensitive, reproducible, and rapid. To this end, numerous methods have already been developed that allow the examination of biological interactions [20–22]. These methods have different measured end points and also have inherent advantages and disadvantages. They include spectroscopic methods such as fluorescence [23], circular dichroism [24], light scattering [25] and nuclear magnetic resonance spectroscopy [26], surface plasmon resonance analysis [27], differential scanning and titration calorimetry [28], analytical ultracentrifugation [29]; electrophoretic methods [30], radioactive tracer binding experiments [31], and yeast two-hybrid methods [32, 33].

More recently, an emerging technology that appears to offer some significant advantages over most other techniques is mass spectrometry. The purpose of this review is to highlight the use of mass spectrometry in the real-time analysis of biomolecular noncovalent complexes, including protein–metal ion, protein–ligand or therapeutic drug, protein–protein, protein–DNA, and protein–RNA interactions. We have not attempted to describe in detail the large number of examples already available because there are a number of excellent survey reviews on the subject [34–43]. We attempt to provide an overview of this approach and describe appr opriate examples highlighting the power of this technology.

A. Definition of Structural and Functional Proteomics

The word ''proteome'' was first coined in the mid-1990s to describe the protein complement of an organism's genome [44, 45]. Furthermore, as Blackstock and Weir, [40] have noted, the ''proteome was imperceptibly transmuted into a new discipline, proteomics.'' This new discipline has induced a flood of activity and has resulted in publication of two books in this area [46, 47], and has been the subject of a number of reviews [40, 43, 48–54]. As with any new and growing field of investigation, nascent nomenclature is often confusing or contradictory. Proteomics has been described as functional genomics at the protein level [40], whereas large-scale protein structure has been designated as ''structural genomics.'' [55] Finally, a further subdivision of proteomics into ''expression proteomics'' and ''cell map proteomics'' has been proposed [40]. The former relies on the determination of quantitative maps of protein expression from organisms, organs, tissue, or cells. This approach allows

for the measurement of perturbations on biological systems by comparing a control versus diseased/challenged protein complement. We would propose that this is more clearly defined by the term differential proteomics (see Table I).

Cell map proteomics involves the systematic identification of individual proteins participating in protein–protein multicomponent complexes [40]. Characterization of such protein complexes affords the potential of determining specific protein(s) function in the cell. Hence we would suggest that this is more aptly defined by the term functional proteomics. Furthermore, the function of individual proteins is also mediated by a myriad of interactions that also include protein–DNA, protein–RNA, and protein–protein–DNA/RNA. Hence we would also suggest that the investigation of all such interactions come under the descriptor ''functional proteomics'' (see Table I). Finally, aspects that contribute to an understanding of the structure of a protein, such as determination of metal ion:protein binding stoichiometries, or identification of a specific ligand-binding site, can be described as structural proteomics (Table I). In summary, the use of the descriptors differential, structural, or functional proteomics is a simple attenuation of such terms commonly used in protein biochemistry to describe differing protein concentrations, structure, and function in the cell.

TABLE I
Definitions in Proteomics

Term	Definition
Proteome	The total protein profile of an organism, organ, tissue, cell, or organelle at a specified time
Proteomics	Systematic analysis and identification of the entire protein complement of an organism, organ, tissue, cell, or organelle at a specified time
Differential proteomics	Quantitative determination of the differences in protein complement of a normal/control organism, organ, tissue, cell, or organelle versus challenged/diseased counterpart
Structural proteomics	Identification of all interactions by metal ions, drugs, ligands, peptides, and other proteins that affect protein structure
Functional proteomics	Identification of all protein–protein interactions within a specified cell system, as well as protein–DNA and protein–RNA interactions occurring in an organism or cells and the modulation effects of ligands, toxins, drug, and metal ions on all those protein interactions that affect function

B. *Mass Spectrometry-Structural and Functional Proteomics*

Proteomic and differential proteomic analyses have historically relied on two-dimensional polyacrylamide gel electrophoresis (2D PAGE) as the platform for the separation of proteins derived from organisms, tissue, and cells [45–48, 51]. Indeed, workers in the 1970s had demonstrated the power of 2D PAGE in separating proteins on the basis of their different isoelectric points and molecular weights. They started to construct 2D gel databases even though the identities of the proteins were not known [56, 57]. However, with the advent of modern biological mass spectrometry (MS), the rapid identification of such proteins became possible and opened the door to present-day proteomic investigations. However, in the case of structural and functional proteomic analyses, the presence of sodium dodecyl sulfate (SDS) in the gel disrupts the noncovalent interactions one is actually attempting to investigate. Furthermore, real-time analysis of the modulation effects of ligands, drugs, and metal ions is often necessary and the time constraints imposed by gel separations necessitate consideration of other options. Hence, most investigations in structural and functional proteomics have involved direct infusion of analytes into the MS, affinity chromatography coupled with MS, or other on-line chromatography-MS approaches.

The mass spectrometer is an analytical instrument capable of determining the molecular mass of biological analytes to a high degree of accuracy (typically, 0.01–0.001%). The essence of a modern mass spectrometer consists of a source region, the mass spectrometer itself that separates ions on a mass to charge (m/z) basis, and a detection system all under workstation control. For the interested reader, more extensive and detailed reviews are available [58, 59]. However, it should be noted that a limitation of MS is that compounds can be analyzed only in the gas phase, either as positively or negatively charged ions. Hence the source region serves as both the sample inlet and ionization chamber. Today, two commonly used ''soft'' ionization techniques are preferred in the analyses of peptides and proteins, namely matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI).

The ESI process generates charged microdroplets containing analytes. Gentle evaporation of the droplets in the source results ultimately in a charge transfer from the water droplet to the analyte, leading to the creation of gas-phase ions [60, 61]. Such ions are detected as a series of multiply charged ion species with differing m/z ratios, as shown in Fig. 1A for the ESI-MS analysis of the recombinantly expressed protein calbindin D_{28K}. To determine the relative molecular weight (M_r) of the protein, a simple software algorithm ''transforms'' the multiply charged ion series

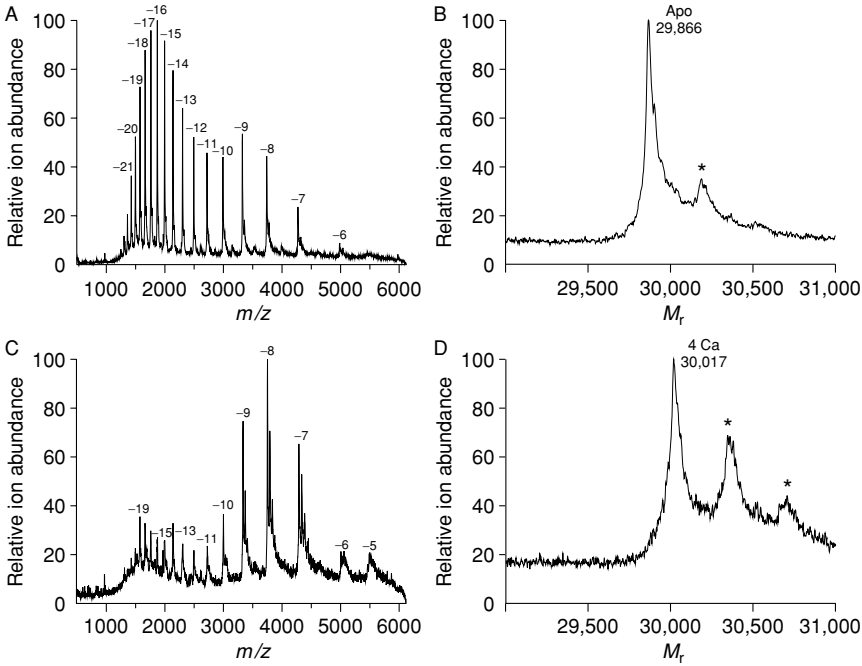


FIG. 1. Negative ion ESI-MS analysis of calbindin D_{28K} by direct infusion of the 60 μM protein solution in 4 mM NH_4HCO_3 at pH 8.0. (A) Multiply charged raw data file of calbindin D_{28K} after addition of 20 scans between mass range 600 and 6000 Da. (B) The transformed spectrum affording M_r = 29,866, indicating that the protein is in the apo form. The peak marked by the asterisk (*) corresponds to EDTA/Na adduct, bound nonspecifically to the protein. (C) Multiply charged raw data file of calbindin D_{28K} after 20 scans between mass range 600 and 6000 Da in the presence of 1 mM calcium acetate. (D) Transformed spectrum, indicating the addition of four Ca^{2+} ions per mole equivalent of protein. The peaks marked by asterisks (*) correspond to nonspecifically bound EDTA/Na adducts. (Published with permission.)

into a single value. This is shown in Fig. 1B, for calbindin D_{28K}, where the measured signal at M_r = 29,866 indicates that the protein is in the apo form. In the case of MALDI, the analyte is mixed and cocrystallized with a photoactive organic acid, which readily absorbs energy from laser irradiation [62]. Hence when the target containing the analyte and organic matrix is placed in the source region and subjected to laser bombardment analytes are projected into the gas phase, typically as singly charged ions. However, MALDI-MS requires mixing of analytes with an organic acid that may significantly disrupt any noncovalent interactions that are present. Hence, although there are some exceptions, the principal

method of choice in structural and functional proteomic analyses has, to date, been ESI-MS [39].

In the case of proteins derived from protein complexes in functional proteomics analyses, it is often necessary to obtain sequence information about these proteins [40, 63]. MS has almost completely replaced classic Edman sequencing in this regard. The former offers significantly enhanced sensitivity, and faster throughput, and can also deal with protein mixtures. It is typically carried out on proteolytic digests (such as trypsin) of proteins, and utilizes the approach commonly referred to as tandem mass spectrometry (MS/MS) [64]. Typically, the proteolytically derived peptides are individually ionized in the ESI source and separated on the basis of their m/z values in mass spectrometer 1. Subsequently, one peptide of specific m/z value is subjected to collision-induced dissociation by colliding only these ions with an inert gas such as xenon. The collisions between ions and gas cause fragmentation, and these ions (commonly referred to as product ions) are separated on the basis of their m/z values in mass spectrometer 2. These sequential losses of amino acids provide valuable sequence information and when interrogated against expressed sequence tag (EST) databases readily afford protein identification [65, 66].

Loo has discussed the advantages conferred by ESI-MS in the analysis of noncovalent complexes compared with other techniques [36]. Loo describes in some detail advantages of mass spectrometry in solving problems, first outlined by McLafferty in 1981 [67], namely specificity, sensitivity, and speed (defined as ''S'' advantages). Loo added a fourth advantage in the analysis of noncovalent complexes—stoichiometry. This last addition is obviously an important parameter when considering protein–metal ion, or protein–ligand interactions. We would also add a fifth ''S'' advantage, namely, selectivity. In ESI-MS analysis of complexes it is possible in a single experiment to selectivity measure and determine the mass of the intact complex itself, the M_r of the individual proteins making up the multimer, or the molecular weight of the specific metal ion, ligand, or drug metabolite bound to the protein complex.

C. *Important Parameters in Structural and Functional Proteomics*

Typically, cellular processes involving proteins exist under conditions of close to neutral pH (6.8–7.2), osmolality of 300 mmol/kg, and ionic salt concentrations of 150–200 mM [68]. In some cellular compartments this may deviate from intracytoplasmic conditions. For example, lysosomes are acidic with a pH approaching 4.0 [69], whereas in mitochondria the pH is closer to 7.3–7.7 [70], and the renal medulla is hyperosmolar [71].

In conventional ESI-MS such conditions are not conducive to sensitive and reproducible mass spectrometric analysis [61]. Frequently, the pH must be modified to either acidic (positive ion MS) or basic (negative ion MS) pH, in order to ensure efficient ionization of analytes occurs. Furthermore, high concentrations of involatile salts are also deleterious to MS performance as they precipitate out in the source region. Finally, source temperature conditions often must be kept high in order to ensure efficient droplet evaporation occurs in order to desolvate protein complexes [61]. Also, oftentimes the presence of organic solvents such as methanol or acetonitrile in protein solutions can enhance sensitivity of analyte detection and signal stability. Obviously, the acquisition of ESI-MS data under more physiologically relevant conditions would be preferable.

More recently, we, as well as others, have demonstrated that it is possible to detect noncovalent complexes by ESI-MS, spraying from simple aqueous solutions [72–74]. Furthermore, the use of high (millimolar) concentrations of volatile ammonium salts, such as acetate, formate, and bicarbonate in the ESI-MS analysis of protein complexes allows preparation of aqueous solutions with comparable to ionic strengths to *in vivo* cellular conditions, as well as spraying complexes in the pH range of approximately 6.5–7.8. However, it is the advent of nanoelectrospray (nESI) that has allowed the direct analysis of aqueous solutions at physiological pH levels with relatively high salt concentrations akin to intracellular conditions [75–77]. A number of studies have now demonstrated the power and potential of such an approach in the analyses of noncovalent protein complexes [63, 78].

II. Applications of ESI-MS in Structural and Functional Proteomics

ESI-MS was first demonstrated in the early 1990s to be a technique that can detect intact noncovalent complexes. Independently, Katta and Chait [79] were able to detect heme associated with myoglobin, whereas Ganem *et al.* [80] detected a receptor–ligand complex. Since that time there have been numerous reports describing protein–metal ion, protein–ligand, protein–protein (also protein–peptide), protein–DNA, and protein–RNA interactions. These results have been summarized in the excellent and detailed review by Loo [36] and more recently by other workers [39, 42]. We do not attempt to duplicate their thoroughness but provide appropriate examples that highlight the fact that ESI-MS is a powerful

tool with which to analyze biomolecular interactions and is the emerging central platform technology in structural and functional proteomics.

A. Protein–Metal Ion Interactions

Metal ion interactions with proteins have been widely reported to modulate function of proteins [36, 81]. For example, EF-hands and zinc fingers take up Ca^{2+} and Zn^{2+}, respectively, which leads to a wide range of subsequent protein activity including protein–protein complexation, transcription complex formation, and protein–ligand (peptide) binding. Typically, metal ion:protein binding stoichiometries have been determined by indirect optical spectroscopic methods such as fluorescence spectroscopy [82]. However, with the advent of ESI-MS it has been possible to directly and accurately determine the metal ion-binding stoichiometry to the protein [36]. A specific example in which ESI-MS afforded accurate metal ion-binding stoichiometries was in the analysis of the Ca^{2+}-binding protein, calbindin D_{28K} [83]. Previous optical spectroscopic and dialysis methods had yielded Ca^{2+}-binding stoichiometries ranging from three to six per mole of calbindin D_{28K}. The protein consists of six EF-hand domains that could possibly bind Ca^{2+}. Initially, the apoprotein, in the presence of 2 m*M* EDTA, was subjected to microelectrospray ionization-MS (μESI-MS). The multiply charged ion series is shown in Fig. 1A. On transformation a dominant ion at $M_r = 29{,}866$ was observed, corresponding to the molecular mass of the protein (Fig. 1B). Titration of Ca^{2+} into the apoprotein solution resulted in uptake of calcium by the protein. Addition of 1 m*M* Ca^{2+} resulted in a multiply charged ion series being detected as shown in Fig. 1C. On transformation, this afforded an $M_r = 30{,}017$, corresponding to the uptake of four Ca^{2+} ions (Fig. 1D), leaving two EF-hands free. More recently Troxler *et al.* [84] have used the same approach to demonstrate that α-parvalbumin binds two Ca^{2+} ions per mole of protein whereas nine mutant proteins had significantly altered function and bound either zero Ca^{2+} or one Ca^{2+} per mole of protein. Finally, Chazin and Veenstra [85] have demonstrated that calbindin D_{9K} binds two Ca^{2+} per mole of protein. In all cases ESI-MS was able to rapidly and accurately determine metal ion-binding stoichiometries of functionally active proteins.

An even more useful example of the power of ESI-MS in the study of metal ion–protein binding is the analysis of metalloproteins containing two distinct types of metal–binding sites. For example, the metalloprotein matrilysin contains both zinc- and calcium-binding sites. In such a situation conventional methods such as optical spectroscopy would not

be able to determine individual metal ion-binding stoichiometries. However, Yuan *et al.* [86] were able to clearly demonstrate on titrating both Zn^{2+} and Ca^{2+} at pH 7.4, using ESI-MS, that matrilysin bound two Zn^{2+} and two Ca^{2+} per mole of protein simultaneously. Finally, ESI-MS has also been used to determine the oxidation states of metal centers in iron–sulfur proteins [87], as well as to monitor metal ion fluxes in and out of metallothionein and drug-modified metallothionein [88].

Because transformed ESI-MS data provide discernible ion responses based on m/z values, it is possible to differentiate between the sequential binding of metal ions by a protein. Chazin and Veenstra [85] utilized this property of ESI-MS to demonstrate that calbindin D_{9K}, which contains two EF-hands, bound the second Ca^{2+} ion cooperatively, whereas the binding of the first Ca^{2+} ion to the protein had no effect on the metal ion affinity of the second EF-hand. More recently Gehrig *et al.* [89] have used the same approach to demonstrate that a number of metallothionein proteins also bound zinc, cadmium, or copper in a cooperative manner.

The quaternary structure of proteins can be a key factor in their functional role within cells. Hence, any technique that can rapidly determine gross quaternary structure, as well as detect changes on the basis of metal ion modulation, affords a powerful new approach. In that regard, Strupat *et al.* [90] investigated a series of myeloid-related proteins (MRPs), which belong to the S100 family. MRPs contain two EF-hands, with one high- and one low-affinity binding site. In a series of elegant experiments they first demonstrated that MRP8, MRP14, and its isoform MRP14* bound 2 mol of Ca^{2+} per mole of protein. They went on to show that MRP8, MRP14, and MRP14*, when mixed together in the absence of calcium, exist as heterodimers MRP8/MRP14 and MRP8/MRP14*. However, in the presence of Ca^{2+}, the two heterodimers tetramerize to form the heterotetramers (MRP8/MRP14 + MRP8/MRP14*), (MRP8/MRP14 × 2), and (MRP8/MRP14* × 2). They further demonstrated that eight Ca^{2+} were bound to individual nonphosphorylated tetramers.

Finally, conformational change in protein structure is integrally linked to protein function. The ESI-MS analysis of proteins allows the detection of gross conformational changes of proteins on metal ion uptake. Previous studies in the early 1990s had shown that the ESI-MS multiply charged spectrum of a protein was different in the native state versus denatured state. On the basis of hydrogen/deuterium (H/D) exchange and circular dichroism (CD) measurements it was concluded that this change was due to a difference in protein conformation [91]. Hence, if the multiply charged spectrum of apocalbindin D_{28K} is inspected (Fig. 1A), the predominant charge states are centered around −16 to −17. However, on uptake of all four Ca^{2+} ions, the multiply charged spectrum of calbindin

D_{28K} is now significantly different, as seen in Fig. 1C. The charge state distribution is now centered around the −7, −8, and −9 charges, reflecting a change in the availability of negatively charged side chains at the protein surface, implying a change in tertiary structure of the protein. Changes in the multiply charged spectra of other proteins have also been reported on uptake of various metal ions [92].

Two studies from our laboratory have compared changes in the multiply charged ESI-MS spectra of proteins calbindin D_{28K} [including two mutant calbindins lacking 1 (Δ [2]) and 2 (Δ [2, 6]), EF-hands] [93] and vitamin D_3 receptor DNA-binding domain (VDR-DBD) by optical spectroscopic methods [94]. In both cases, a clear correlation between multiply charged ESI-MS spectra versus CD and fluorescence spectroscopy was observed, reflecting the fact that changes in the secondary and tertiary structures of the proteins were being detected on uptake of the metal ions. Because the ESI process produces a series of multiply charged protein ions, the distribution of charge states is related to the number of acidic or basic amino acid residues at or near the protein surface [60]. Hence the probability that the charge state distribution is related to secondary or tertiary structure is not an unreasonable one.

More recently, a number of workers have demonstrated that monitoring H/D exchange rates as a function of metal ion uptake by proteins also affords information about conformational change. For example, Nemirovskiy and co-workers [95] have used such an approach to determine gross tertiary structural changes in calmodulin on uptake of Ca^{2+}. They were able to show that apocalmodulin exchanged approximately 115 protons for deuteriums over 60 min (Fig. 2A and B). However, on titration with Ca^{2+}, the extent of exchange decreases significantly by approximately 24 protons (Fig. 2C), indicating that the ''Ca^{2+}-induced folding of the protein to a tighter, less solvent accessible form'' had occurred. No such differences were observed on uptake of Mg^{2+}. The authors conclude from this study that the H/D exchange differences reflect tertiary structural change of the protein on uptake of Ca^{2+}.

B. Protein–Ligand/Drug Interactions

Numerous cellular functions are triggered by noncovalent interactions of an endogenous ligand with a specific protein such as an enzyme or receptor. Furthermore, most therapeutic drugs exert their effects through interaction with specific protein receptors. However, it is interesting to note that of the approximately 3000 available therapeutics agents, approximately 25% do not have an identified target. Also, as the discovery of new orphan

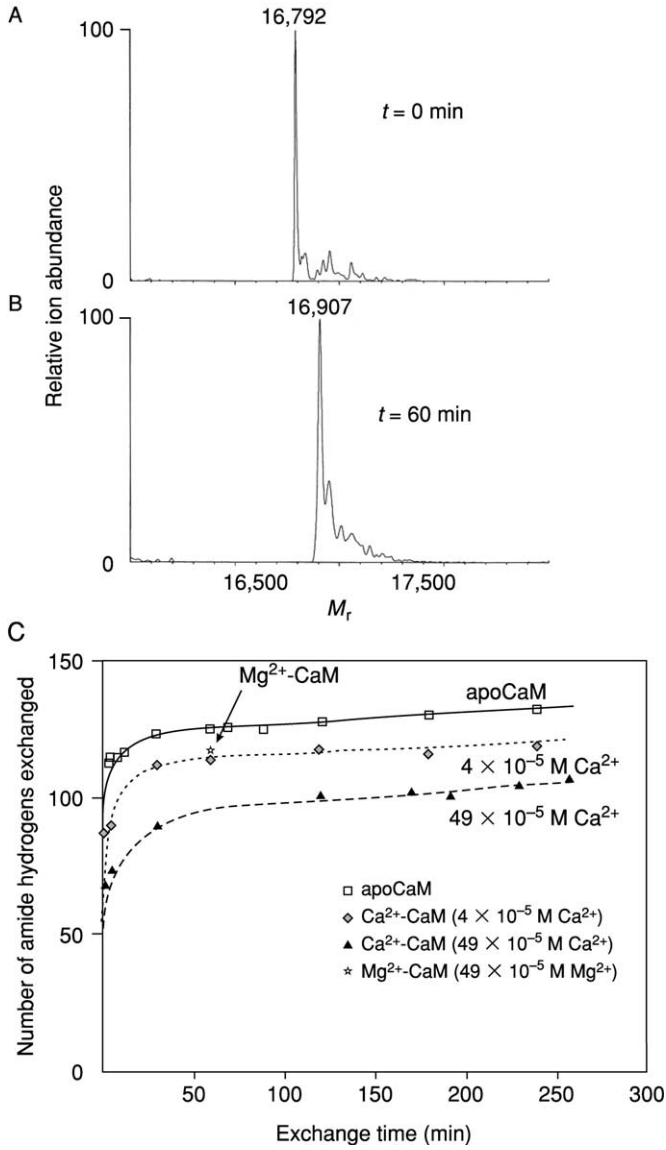


FIG. 2. Positive ion ESI-MS analysis of H/D exchange experiments on 15 μM calmodulin analyzed at pH 7 in NH_4OH buffer. (A) Apocalmodulin at time = 0 min in the presence of D_2O at pH 7. (B) Apocalmodulin at time = 60 min in the presence of D_2O at pH 7, and calcium acetate. (C) Time dependence for the H/D exchange of calmodulin with and without added Ca^{2+}. The extent of H/D exchange at one time and concentration of Mg^{2+} (control) is given as a reference. (Published with permission.)

receptors continues to significantly increase, the need for rapid analysis and identification of protein–ligand, and protein–drug, interactions is vital [96]. In this regard mass spectrometry is beginning to play an ever-increasing role. For reviews of this area see Loo [36] and others [97, 98].

Smith and co-workers have described the use of ESI-Fourier transform-MS in the detection of the interactions of the peptide ligand periplasmic oligopeptide-binding protein (OppA) with the protein SecB [99, 100]. This cytosolic chaperone protein facilitates the transport of polypeptides to the periplasmic space and the outer membrane. Initially, they demonstrated that SecB exists as a homotetramer, and that OppA binds to SecB in a 1:1 stoichiometry [99]. Subsequently, they demonstrated that SecB binds to both native and nonnative polypeptides, indicating two different, distinct binding sites [100]. In a further refinement of the approach, Ayed *et al.* used an ESI-time of flight (TOF)-MS instrument to actually determine K_a values for the dimer–hexamer equilibrium complexes of *Escherichia coli* citrate synthase ($K_a = 6.9 \times 10$ [10]M^{-2}) [101]. They also determined a K_d value (1.1 μM) for binding of the endogenous ligand NADH to the protein. They concluded that this approach would be widely used in ''obtaining fundamental physicochemical information about macromolecular interactions.''

Finally, in an elegant study Loo and co-workers [41] have demonstrated the power of this approach in detecting a therapeutic drug–protein–RNA complex. They used ESI-MS to study the ability of the aminoglycoside antibiotic neomycin to recognize HIV-1 Tat protein and *trans*-activation responsive (TAR) element RNA. Replication of HIV requires that Tat recognize the TAR element located at the 5′ end of the mRNA. Hence, binding of drugs to such a complex may interfere with the Tat–TAR RNA interaction and prevent HIV replication. Initially Loo *et al.* demonstrated that they could readily detect the interaction of neomycin with TAR, as shown in Fig. 3A, affording a drug:TAR stoichiometry of 3:1. However, on addition of Tat a 1:1:2 complex of Tat:TAR:neomycin was observed (Fig. 3B). The authors conclude that the molecular details of small molecule binding events are typically difficult to determine by most other biophysical techniques. However, ESI-MS affords high mass resolution and accuracy, in a rapid time frame requiring little material.

C. *Protein–Protein Interactions*

Historically, the modulation of protein–protein interactions by metal ions, ligands, or drugs or other regulatory proteins has been investigated by optical spectroscopic methods such as UV, near and far CD,

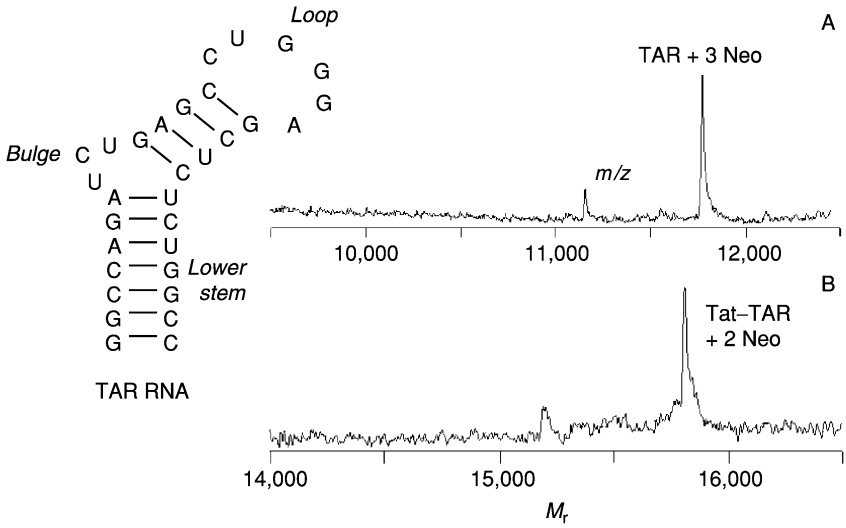


FIG. 3. Structure of TAR RNA and negative ion ESI mass spectra (transformed to the molecular weight domain) of the noncovalent binding between neomycin and TAR RNA (A) and the Tat peptide–TAR RNA complex (B). A concentration ratio of greater than 5:1 neomycin:TAR was used (in 10 m*M* ammonium acetate, pH 6.9). The maximum neomycin-binding stoichiometry is 3 and 2 to TAR and Tat–TAR, respectively. (Published with permission.)

fluorescence spectroscopy, as well as other techniques such as infrared (IR) and Raman spectroscopy [20–22]. Furthermore, the actual demonstration of the existence of protein heteromeric complexes has principally been based on nondenaturing gels and other chromatographic techniques [102]. In the former case, stoichiometries are difficult to quantitate and in the latter case, the identity of other proteins interacting with the target protein has been difficult to determine. However, more recently ESI-MS has found an ever-increasing role in directly measuring the stoichiometry of homomultimeric and heteromultimeric protein complexes, as well as in the rapid determination of the modulation effects of metal ions, drugs, and ligands on such complexes [36, 39]. An even more exciting development has been the use of MS and MS/MS in the rapid identification of constituent components of cellularly derived multimeric protein complexes. Typically a specific target protein of interest is immunoprecipitated out of a cell lysate or tissue extract, along with all the associated proteins of the complex. Subsequently these ''unknown'' proteins are identified by tandem mass spectrometry. This latter approach has been pioneered by Mann [43, 50, 54, 66, 77] and others [51, 52, 63, 103, 104]

and offers a powerful and direct approach for the analysis of such complexes.

The use of ESI-MS in studying multimeric protein complexes continues to increase. For example, Green *et al.* used this approach to investigate the assembly of hexagonal bilayer hemoglobins [105]. They demonstrated that these ~3.6-MDa complexes are composed of subassembly units consisting of monomeric Hb-disulfide-linked trimer Hb proteins. In another example, Dobson and Robinson monitored the formation of type II DNA-binding protein heterodimers from various *Bacillius* species [106]. They further developed this approach to analyze both the thermodynamic and kinetic attributes of the complexes and showed that their studies correlated well with the kinetics of protein dissociation and free energy differences between homo- and heterodimeric species, using other more conventional approaches. In an intriguing piece of work, Rostom and Robinson used ESI-MS to detect the intact GroEL-14mer complex at ~800,000 Da [38]. Furthermore, they were able to show by increasing collision energies (1–2 eV) that they could begin to dissociate specific component parts of this large complex into a single ring and monomeric species, thus demonstrating the stochiometry and subunit topology of the 14 subunits arranged in 2 heptameric rings.

A further example of the efficacious use of ESI-MS is in the study of the modulation effects of metal ions or ligands/drugs on protein complex stability and formation. For example, Fabris and Fenselau demonstrated that insulin present in β cells of the pancreas exists as a hexamer in the presence of zinc ions, but principally as the monomer in the absence of zinc [107]. They concluded that ESI-MS is a rapid method by which to characterize metal ion allosteric conformation of multimeric proteins. Along the same lines, two different groups simultaneously reported on the use of ESI-MS to determine the modulatory effects of Ca^{2+} on calmodulin in the binding of peptides [73, 108]. In both cases, only in the presence of calcium was it possible to detect the presence of calmodulin: calmodulin-dependent protein kinase II peptide complex. The measured mass of the protein–peptide complex was consistent with calmodulin bound to 4 mol equivalents of calcium. More recently, Shen *et al.* [109] have used μESI-MS to investigate the ability of the active drug metabolite of disulfiram to affect the stability of homotetramer complex formation of the target enzyme aldehyde dehydrogenase (ALDH). This work went on to extend the focus to look at the ability of a series of inhibitors such as prunetin, *N*-tosyl L-phenylalanine chloromethyl ketone (TCPK), and benomyl to affect ALDH homotetramer formation. Finally, Nettleton *et al.* [110] have used the same approach in evaluating the effect of mutant variants of the blood protein transthyretin on formation of the normal wild-type homotetramer. They demonstrated

that a strong correlation existed between the instability of the tetramer mutant, as determined in the mass spectrometer, with the propensity of this variant to form amyloid. They suggest that this approach can be useful in determining the mechanism of fibril formation in amyloidosis.

A key question regarding the function of a specific protein concerns how it interacts within the cellular environment. Knowledge of such interactions allows a ready understanding of protein function as well as opens up the possibility for development of new targets in therapeutic treatments of disease. One powerful approach, pioneered by Mann, is to purify an entire protein complex by some type of targeted affinity chromatography based on glutathione-*S*-transferase (GST) fusion protein, histidine-tagged protein, antibody, RNA, DNA, or peptide affinity [43, 50–54, 66, 103, 104]. A generic way is to tag a specific protein, immunoprecipitate the protein complex, and then subject the entire protein complex to enzymatic digestion followed by MS/MS analysis of individual peptides. For example, the 25S[U4/U6U5]tri-small nuclear ribonucleoprotein (tri-snRNP) is a central constituent of the nuclear pre-mRNA splicing machinery in yeast and delineating its functional dynamics is critical to an understanding of the spliceosomal cycle. Gottschalk *et al.* [77] were able to rapidly characterize the known proteins of the [U4/U6U5]tri-snRNP from *Saccharomyces cerevisiae*, as well as identify eight new proteins. The tri-snRNP was purified from a yeast extract, using an anti-m_3G-immunoaffinity column and Ni^{2+}-NTA chromatography. Gradient fractions obtained by 10–30% glycerol gradient centrifugation were then subjected to SDS-gel electrophoresis and visualized with silver stain, as shown in Fig. 4A. Subsequently, protein bands were excised, in-gel digested with trypsin, and analyzed by a combination of MALDI-TOF-MS and nano-ESI-MS and MS/MS. The mass spectrometric analysis of the protein mixture found in band 14 is shown in Fig. 4B. Product ion spectra of constituent peptides were obtained as shown in Fig. 4C and an EST was constructed and a database search performed. In this particular case, the protein was identified as Spp381p. Furthermore, the authors also reported the identification of Lsm2p, Lsm5p, Lsm6p, Snu13p, Dib1p, Snu23p, and Snu66p. Clearly, this example demonstrates the power and potential of this approach in determining the identity of protein constituents from noncovalently bound multimeric protein complexes.

D. Protein–DNA and Protein–RNA Interactions

Proteins serve as key regulators of genetic information by interacting with DNA at the transcriptional level and with RNA at the translational level. Typically, such interactions have been examined in electrophoretic

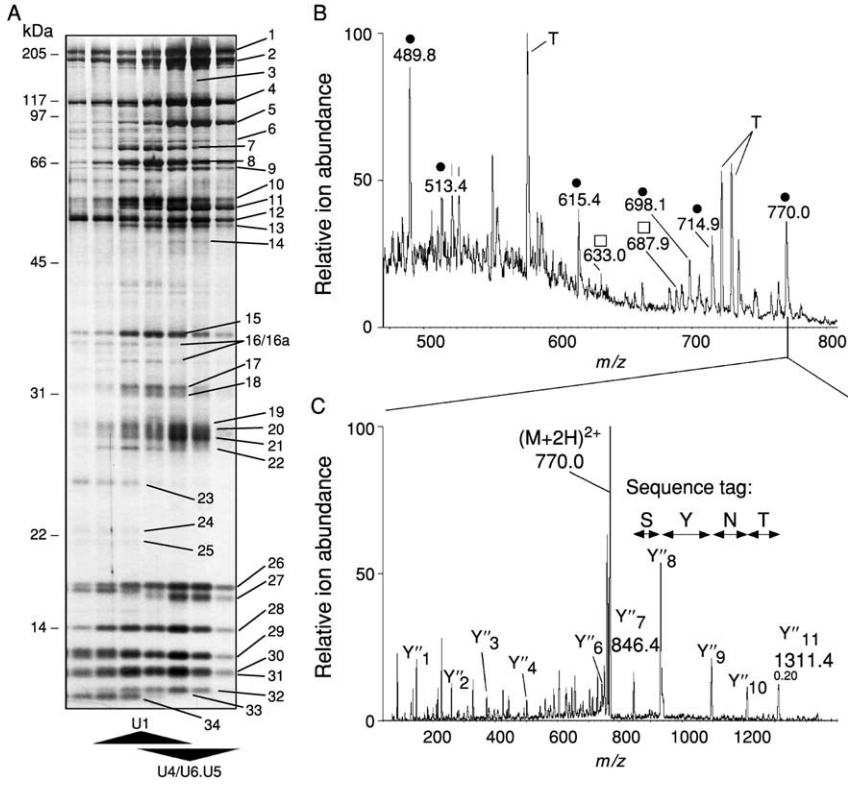


FIG. 4. Biochemical purification of the yeast [U4/U6·U5]tri-snRNP and identification of its protein components by mass spectrometry. The tri-snRNP was purified from whole-cell extract by anti-m_3G-cap immunoaffinity and Ni^{2+}-NTA chromatography followed by 10–30% glycerol gradient centrifugation. Gradient fractions were extracted with phenol–chloroform–isoamyl alcohol and analyzed for their protein content. (A) Gels were stained with silver. Proteins containing the majority of the tri-snRNP were pooled, separated on a 60-cm SDS gel consisting of 11.5% acrylamide (*top*) and 13% acrylamide (*bottom*), and Coomassie blue stained. The protein bands were excised and in-gel digested, and peptides were then eluted from the gel and analyzed by a combination of MALDI and nanoelectrospray tandem mass spectrometry. (B) Mass spectrometric analysis of the protein mixture found in band 14. Shown is part of the Q1 scan (normal mass spectrum) of the peptide mixture obtained after tryptic digestion of the protein band and micropurification of the resulting peptide mixture. The labeled peaks were selected in the first part of the mass spectrometer in turn and fragmented in its collision chamber. Mass spectra of the resulting fragments (tandem mass spectra) contain peptide ions that were subjected to tandem mass spectrometry, and the partial sequences derived can be used to identify the proteins in a sequence database using peptide sequence tags [65]. Two different proteins were identified by tandem mass spectrometry: marked with bullets are the peptides derived from Spp381p; the squares

gel mobility shift assays (EMSAs) [21]. However, this gel-based assay can often take several days, and it is not possible to determine metal ion; endogenous ligand, or drug-binding stoichiometries to such complexes. Hence, the development of techniques to determine and understand the structure, stoichiometry, and function of protein–DNA and protein–RNA interactions in real time would have broad utility and should be of widespread interest. Cheng and co-workers were the first to demonstrate the potential of ESI-MS to detect such complexes [111, 112]. They were able to detect the interaction of homodimer gene V protein with various single-stranded DNAs in a 1:1 complex of DNA to homodimer. Further, by increasing the length of the DNA, a second homodimer of gene V proteins bound to the oligonucleotide. Loo has described in some detail the early pioneering work in this area [36]. More recently, Liu and co-workers described the interactions of bacteriophage T4 RegA protein, a ''unique translational regulator,'' with a number of mRNA species [113]. They demonstrated that 1:1 protein:RNA stoichiometries were observed, but that loss of a single nucleotide from the mRNA could result in almost complete loss of binding. In a tour de force, Rostom *et al.* demonstrated that they could actually detect intact *E. coli* ribosomes [114]. Furthermore, by decreasing the Mg^{2+} concentration, they could dissociate and detect the 30S and 50S subunits, and by inducing gas-phase collisions could further induce the subunits to fragment into even smaller macromolecular assemblies.

The regulation of transcription by protein transcription factors involves the direct interaction of proteins with DNA elements. As noted above, EMSA is the most widely used technique to examine such interactions [21]. However, a good example of the ability of ESI-MS to measure in real time, as well as provide information not available from EMSA, is the study of 1α,25-dihydroxy vitamin D receptor DNA-binding domain protein (VDR-DBD) with the double-stranded DNA vitamin D response element (VDRE) transcription complex [74]. The vitamin D receptor is a member of the nuclear hormone receptor family that includes retinoic acid, thyroid hormone, glucocorticoid, estrogen, and progesterone receptors. The VDR-DBD contains two zinc finger-binding domains. However, it was

indicate peptides originating from a degradation product of Prp31p, and the peaks labeled with T are autolysis products of trypsin. (C) Product ion scan of the doubly charged peptide ion at m/z 770.0. From the fragment spectrum, a partial sequence of this peptide was determined, the sequence tag (846.4)SYNT(1311.4) was constructed, and a database search identified uniquely Spp381p in the database. The C-terminal or Y″ion series confirms the sequence of this peptide (LNTNYSTNEELIK).

not known how Zn^{2+} ions might influence the binding of the protein to DNA or what the precise stoichiometry of binding of the protein to DNA was. Hence, ESI-MS was used to probe these issues in a real-time scenario. When VDR-DBD was incubated with VDRE in the presence of 100 μM EDTA (to chelate any freely available zinc ions), and then subjected to negative ion ESI-MS, only VDRE DNA was detected (Fig. 5A). Concomitant analysis in positive ion ESI-MS, revealed a single ion series that on transformation revealed an ion corresponding to apo-VDR-DBD (M_r = 12,819) (shown in Fig. 5B). However, on addition of 100 μM Zn^{2+}, two new additional ion series were observed, labeled M and D in Fig. 5C. On transformation the M_r values revealed that they corresponded to 1:1 (labeled M) and 2:1 (labeled D) VDR-DBD:VDRE, respectively. Analysis of the same incubation in positive ion ESI-MS revealed that VDR-DBD contained predominantly 2 mol equivalents of Zn^{2+} per mole of VDR-DBD, affording M_r = 12,945, as shown in Fig. 5D. Presumably these Zn^{2+} ions were occupying the two high-affinity zinc finger-binding sites, and that binding of these two metal ions is necessary for binding to the VDRE to occur. Further addition of Zn^{2+} (200 μM) to the VDR-DBD:VDRE solution resulted in disassociation of the transcription complex as shown in Fig. 5E. Interestingly enough, analysis in positive ion ESI-MS revealed that the protein had bound additional Zn^{2+} ions (Fig. 5F). It can also be seen, on comparing the multiply charged ion series of apo-VDR-DBD (Fig. 5B), that this is different from the protein containing two Zn^{2+} ions (Fig. 5D). Further addition of Zn^{2+} bound to the protein (Fig. 5F) affords yet another, subtler, change in charge state distribution. All these data indicate that Zn^{2+} ion uptake onto VDR-DBD induces a change in protein conformation, which is responsible for association and ultimately dissociation of the transcription complex [74].

In a further extension to this work we have used ESI-MS to investigate the effect of both Zn^{2+} and endogenous ligands on the formation of transcription complexes containing full-length receptors [115, 116]. In particular, we used μESI-MS to analyze the binding of the full-length vitamin D receptor (VDR) and the retinoid X-receptor α (RXR) to VDRE. Initially we demonstrated that in the presence of 10 μM Zn^{2+}, VDR and RXR form a complex with VDRE [115]. This ion series on transformation affords a molecular mass M_r = 118,200, corresponding to a 4:1:1:1 Zn^{2+}:VDR:RXR:VDRE transcription complex. On increasing the Zn^{2+} concentration to 200 μM, the amount of transcription complex decreased considerably. However, in this case it was not possible to completely dissociate the complex. Finally, we investigated the effect of a variety of ligands on the stability of the transcription complex, including evaluating the endogenous ligands of VDR and RXR, namely 1α,25-dihydroxy vitamin

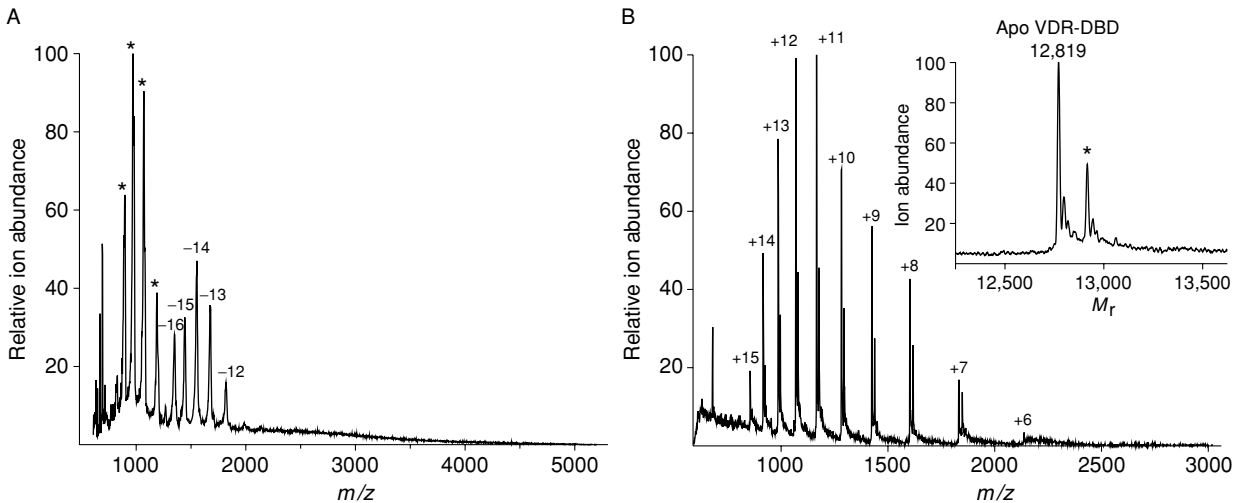
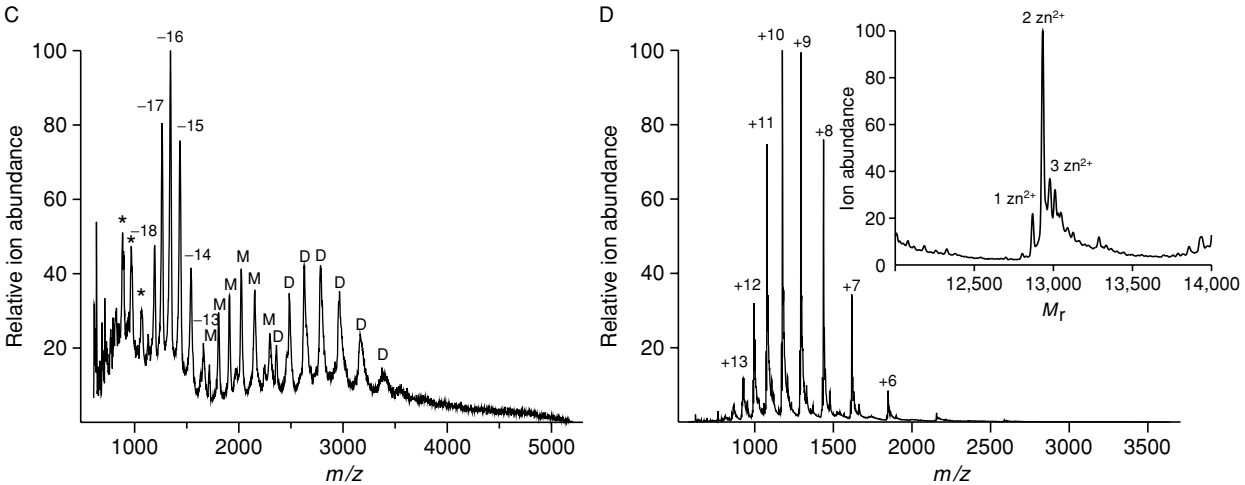


FIG. 5. (*Continued*)


C
Relative ion abundance
m/z
−16
−17
−15
−18
−14
−13
M
D
D
D
D
D
D
*
*
*
1000
2000
3000
4000
5000
100
80
60
40
20
D
Relative ion abundance
m/z
+10
+9
+11
+8
+12
+7
+13
+6
2 zn2+
3 zn2+
1 zn2+
Ion abundance
Mr
12,500
13,000
13,500
14,000
1000
1500
2000
2500
3000
3500

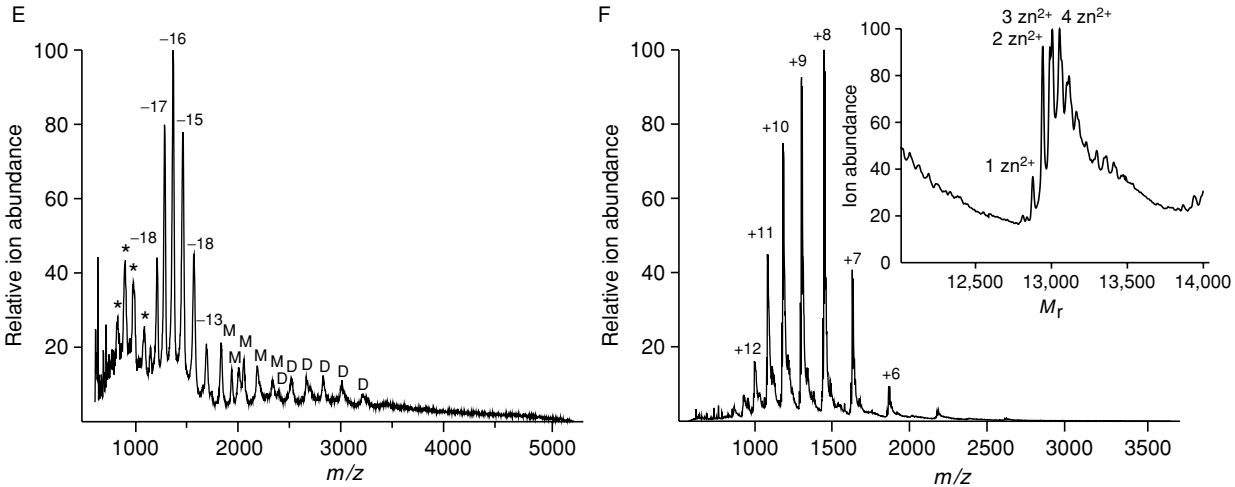


FIG. 5. Positive and negative ion ESI-MS analysis of the Zn^{2+}-dependent binding of VDR-DBD protein to VDRE to form a transcription complex. (A) Negative ion ESI-MS of 20 μM VDR-DBD plus 40 μM VDRE in the presence of 100 μM EDTA. Note that the charge states representing double-stranded DNA are numerically labeled (e.g., −15); single-stranded DNA is labeled (*). (B) Positive ion ESI-MS of the same analyte solution described in (A). *Inset*: Transformed spectrum affording M_r = 12,819, denoting apo-VDR-DBD. Note that the response labeled (*) is a nonspecific EDTA/Na adduct. (C) Negative ion ESI-MS of the same analyte solution described in (A), however, now with the addition of 100 μM Zn^{2+}. Note that M denotes a 2:1:1 ratio and D denotes a 4:2:1 ratio of Zn^{2+}:VDR-DBD:VDR, respectively. (D) Positive ion ESI-MS of the same analyte solution described in (C). *Inset*: Transformed spectrum affording a dominant ion M_r = 12,945 corresponding to VDR-DBD bound to two Zn^{2+}. (E) Negative ion ESI-MS of the same analyte solution described in (A), however, now with the addition of 200 μM Zn^{2+}. (F) Positive ion ESI-MS of the same analyte solution described in (E). *Inset*: Transformed spectrum denoting that VDR-DBD is fully loaded with Zn^{2+} ions. (Published with permission.)

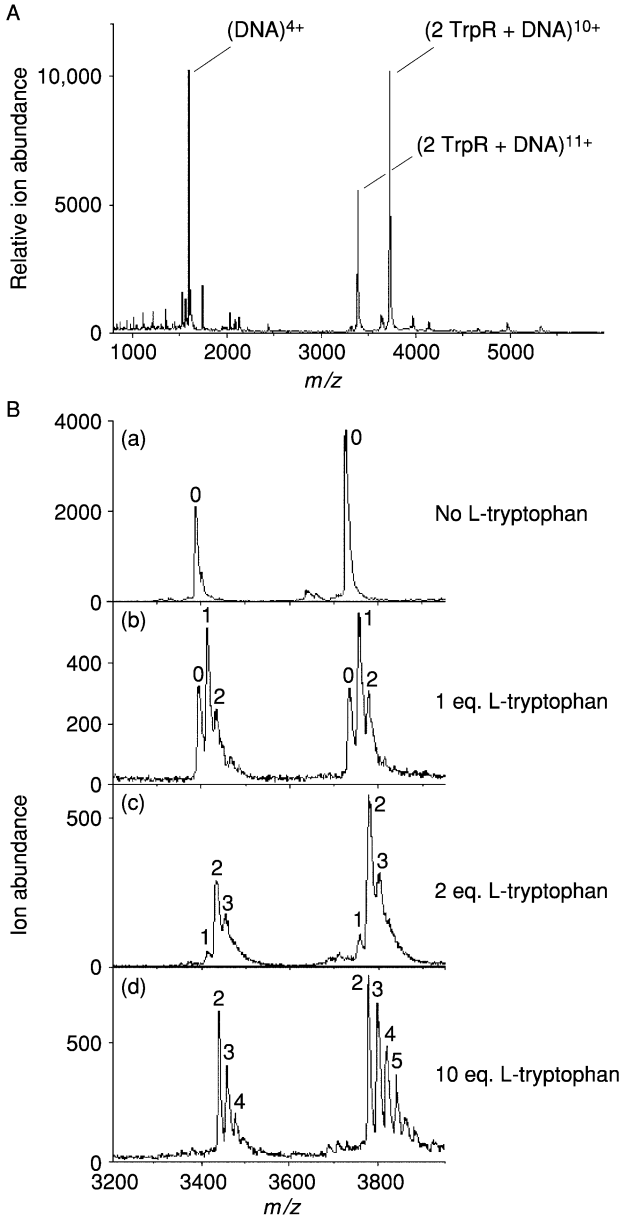


FIG. 6. (A) Positive ion ESI-MS of 20 μM TrpR plus 30 μM double-stranded operator DNA, sprayed in 5 mM ammonium acetate at pH 6.0. A molecular mass of 37,266 8 Da was measured for the TrpR dimer–DNA1 complex, which has the expected stoichiometry of [protein dimer:ds DNA] = [1:1]. (B) Study of the affinity of

D_3 [1α,25$(OH)_2D_3$] and 9-*cis*-retinoic acid (9-c-RA), respectively [116]. In the case of incubating 1α,25$(OH)_2D_3$ with VDR, RXR, and VDRE, an increase in mass of approximately 410 Da was observed in the transcription complex, corresponding to uptake of the ligand by VDR. However, no detectable change in the relative amount of transcription complex was observed. In contrast, incubation of the two receptors and VDRE with 9-c-RA resulted in a marked decrease in the VDR:RXR:VDRE complex and an appearance of an ion series, which on transformation was identifiable as the homodimer RXR complex with VDRE.

Other groups have also investigated the effects of ligands on transcription complex formation. For instance, Potier *et al.* have looked at the interaction of the tryptophan apo-repressor (TrpR), its corepressor L-tryptophan, and the specific operator DNA response element [117]. They demonstrated, using a purpose-built ESI-TOF-MS instrument, that TrpR bound to the operator double-stranded DNA as a protein homodimer: DNA complex, in the absence of L-tryptophan, as shown in Fig. 6A. No major responses corresponding to TrpR with single-stranded DNA, or complexes of other stoichiometries, were observed. Subsequently they showed that in the presence of L-tryptophan, a stable complex containing 2:1:1 L-tryptophan:TrpR:DNA was detectable (Fig. 6B). Hence the protein was now binding to the DNA as a monomeric unit. Increasing amounts of L-tryptophan led to further uptake of the repressor (Fig. 6B). However, when D-tryptophan was used in place of L-tryptophan no significant specific binding of the former ligand was detectable. Clearly all these results demonstrate the power of MS to examine in real time the effects of both metal ions and ligands on the amount and nature of receptor transcription complexes and to afford a rapid new method to assess the stability of such complexes and modulation effects.

Nordhoff *et al.* employed a different role for MS in looking at transcription complexes [118]. They took specific DNA probes immobilized onto Dynabeads and incubated them with either *E. coli* crude cell extracts or with a yeast strain of *S. cerevisiae* containing (or lacking) rat RXR-α cDNA. In either case, after incubation for 15 min, the magnetic beads were separated and washed. The beads were then directly analyzed by MALDI-TOF-MS for proteins associated with the DNA probes. In the

---

L-tryptophan for the TrpR–DNA complex. ESI mass spectra of the TrpR–DNA1 complex (10 μM) in presence of (a) 0, (b) 1 μM, (c) 20 μM, and (d) 100 μM L-tryptophan. Two charge states (10+ and 11+) of the ternary TrpR–DNA1–L-tryptophan complexes are shown. N = 0, 1, 2, …, corresponds to the number of L-tryptophans bound to the protein–DNA complex. All spectra were acquired in 5 mM ammonium acetate, pH 6.0, at V_c = 60 V. (Published with permission.)

case of the *E. coli* extract, employing a binding motif for cyclic adenosine monophosphate receptor protein (CRP), they were indeed able to identify CRP from the crude extract. In the case of the yeast strain they also confirmed the presence of rat RXR-αprotein, and clearly this approach has tremendous potential for high-throughput screening of transcription complex constituents and the effects of drug-induced or ligand-induced changes on such complexes, as well as any protein–DNA interaction.

III. Future Directions

Yates, in the review ''Mass Spectrometry and the Age of the Proteome,'' [51] has pointed out that proteins are pleiotropic. One consequence of pleiotropy is that ''protein function will have spatial, temporal and tissue specificity.'' Hence, protein function is complex and will be dependent on a number of factors, including protein structure, that necessitate ''systematic functional studies.'' In that regard, Mendelsohn and Brent [119] have described a number of approaches being used to study protein function, in particular, to identify protein–protein interactions. These methods included modified two-hybrid systems, *in vivo* reconstitution systems, fluorescence resonance energy transfer, and evanescent wave measurements such as surface plasma resonance (SPR). They also briefly describe the potential role of mass spectrometry in such studies and conclude that ''···mass spectrometric methods have tremendous potential. As their sensitivity and ease of use improves, mass spectrometry will come to complement biological methods for detecting and analyzing protein interactions, and *may eventually supplant them* [italics added].'' It is certainly clear from perusal of the current literature that mass spectrometry will play an ever-increasing role in the determination and identification of protein–protein interactions. However, it is also apparent that mass spectrometry is much more powerful in that it allows the study of all noncovalent interactions, including protein–metal ion, protein–drug, protein–RNA, and protein–DNA complexes, as described in this review.

Although we have focused in this review on published examples, it is clear that a number of exciting new developments will significantly contribute to the expanding role of mass spectrometry in this arena. For example, new developments in mass spectrometry instrumentation continue to occur at a rapid pace. Of particular significance in structural and functional proteomics is the emergence of ESI (micro- or nanospray) on time-of-flight (TOF) instruments, as well as the Fourier-transform ion cyclotron resonance (FT-ICR) mass spectrometers [53]. Such instruments

offer the capability of even greater sensitivities and higher mass resolution. The use of FT-ICR in particular offers some considerable advantage when used in conjunction with electron-capture dissociation (ECD) to fragment intact proteins in order to obtain sequence information data without the need to subject the protein to proteolytic digestion beforehand [120].

There is considerable activity in the development of protein microchip and microarray technologies [121]. In this approach typically a specific protein is immobilized in an array format on a chemically modified microchip/microwell surface. Subsequently a tissue extract or cell lysate is incubated with the target protein to obtain specific protein–protein(s) complexes. At present this approach has found some use in conjunction with MALDI-TOF-MS [122]. However, there are now efforts to couple such microchips directly to ESI-MS [123, 124]. The advent of such devices will allow the sample processing, isolation, and identification of proteins present in a protein complex, all on a microfabricated microfluidics chip coupled to an nESI-MS.

It is clear that the future of these exciting fields of structural and functional proteomics will rely on the innovative coupling of various protein processing and isolation techniques with mass spectrometry. An excellent example of this is the integration of SPR biosensors with MS [125, 126]. This powerful integrated approach couples the benefits of a sensitive, affinity capture device, and its ability to quantitate binding events with the ability to rapidly identify the bound substrate(s). Such a powerful combination approach offers significant advantages over either technique alone, and SPR-MS should find widespread use in future.

Finally, Fields [127] noted that ''the number one need in proteomics may be new technology.'' We certainly agree that in proteomics, as well as in structural and functional proteomics, innovation and creativity will be needed to fully exploit these rapidly developing and exciting fields. However, the five ''S'' advantages of mass spectrometry (speed, sensitivity, specificity, stoichiometry, and selectivity) make it an indispensable and powerful platform technology in structural and functional proteomics, and this is only the beginning.

Acknowledgments

We would like to thank Mrs. Diana Ayerhart and Mrs. Sharon Heppelmann for help in preparing this manuscript. We also acknowledge funding from Mayo Foundation (S.N. and R.K.), NIH (R.K.), and Finnigan MAT (S.N.) for work carried out in our laboratories and described in this review.

References

1. Gilman, A. G. (1995). Nobel Lecture: G proteins and regulation of adenylyl cyclase. *Biosci. Rep.* **15,** 65–97.
2. Berman, D. M., and Gilman, A. G. (1998). Mammalian RGS proteins: Barbarians at the gate. *J. Biol. Chem.* **273,** 1269–1272.
3. Tsai, M. J., and O'Malley, B. W. (1994). Molecular mechanisms of action of steroid/thyroid receptor superfamily members. *Annu. Rev. Biochem.* **63,** 451–486.
4. McKenna, N. J., Lanz, R. B., and O'Malley, B. W. (1999). Nuclear receptor coregulators: Cellular and molecular biology. *Endocr. Rev.* **20,** 321–344.
5. Lin, R. J., Kao, H. Y., Ordentlich, P., and Evans, R. M. (1998). The transcriptional basis of steroid physiology. *Cold Spring Harb. Symp. Quant. Biol.* **63,** 577–585.
6. Le Douarin, B., *et al.* (1996). Ligand-dependent interaction of nuclear receptors with potential transcriptional intermediary factors (mediators). *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **351,** 569–578.
7. Mark, M., Rijli, F. M., and Chambon, P. (1997). Homeobox genes in embryogenesis and pathogenesis. *Pediatr. Res.* **42,** 421–429.
8. Fischer, E. H. (1997). Cellular regulation by protein phosphorylation: A historical overview. *Biofactors* **6,** 367–374.
9. Fischer, E. H. (1999). Cell signaling by protein tyrosine phosphorylation. *Adv. Enzyme Regul.* **39,** 359–369.
10. Bramhill, D., and Kornberg, A. (1988). A model for initiation at origins of DNA replication. *Cell* **54,** 915–918.
11. Kornberg, R. D. (1999). Eukaryotic transcriptional control. *Trends Cell Biol.* **9,** M46–M49.
12. Waxman, D. J. (1999). P450 gene induction by structurally diverse xenochemicals: Central role of nuclear receptors CAR, PXR, and PPAR. *Arch. Biochem. Biophys.* **369,** 11–23.
13. Weinshilboum, R. M., Otterness, D. M., and Szumlanski, C. L. (1999). Methylation pharmacogenetics: Catechol *O*-methyltransferase, thiopurine methyltransferase, and histamine *N*-methyltransferase. *Annu. Rev. Pharmacol. Toxicol.* **39,** 19–52.
14. Ingelman-Sundberg, M., Oscarson, M., and McLellan, R. A. (1999). Polymorphic human cytochrome P450 enzymes: An opportunity for individualized drug treatment. *Trends Pharmacol. Sci.* **20,** 342–349.
15. Lewis, D. F. (2000). Structural characteristics of human P450s involved in drug metabolism: QSARs and lipophilicity profiles. *Toxicology* **144,** 197–203.
16. Abernethy, D. R., and Flockhart, D. A. (2000). Molecular basis of cardiovascular drug metabolism: Implications for predicting clinically important drug interactions. *Circulation* **101,** 1749–1753.
17. Kaslow, R. A., and McNicholl, J. M. (1999). Genetic determinants of HIV-1 infection and its manifestations. *Proc. Assoc. Am. Physicians* **111,** 299–307.
18. Miller, A. D. (1996). Cell-surface receptors for retroviruses and implications for gene transfer. *Proc. Natl. Acad. Sci. USA* **93,** 11407–11413.
19. Pereira, L. (1994). Function of glycoprotein B homologues of the family herpesviridae. *Infect. Agents Dis.* **3,** 9–28.
20. Hensley, P. (1996). Defining the structure and stability of macromolecular assemblies in solution: The re-emergence of analytical ultracentrifugation as a practical tool. *Structure* **4,** 367–373.
21. Kneale, G. G., Ed. (1994). ''Methods in Molecular Biology,'' ''DNA–Protein Interactions: Principles and Protocols.'' Vol. 30, Humana Press, Totowa, NJ.

22. Lakey, J. H., and Raggett, E. M. (1998). Measuring protein–protein interactions. *Curr. Opin. Struct. Biol.* **8,** 119–123.
23. Hovius, R., Vallotton, P., Wohland, T., and Vogel, H. (2000). Fluorescence techniques: Shedding light on ligand–receptor interactions. *Trends Pharmacol. Sci.* **21,** 266–273.
24. Siligardi, G., and Hussain, R. (1998). Biomolecules: Interactions and competitions by nonimmobilized ligand interaction assay by circular dichroism. *Enantiomer* **3,** 77–87.
25. Wyatt, P. J. (1993). Light scattering and the absolute characterization of macromolecules. *Anal. Chim. Acta* **272,** 1–40.
26. Song, J., and Ni, F. (1998). NMR for the design of functional mimetics of protein–protein interactions: One key is in the building of bridges. *Biochem. Cell Biol.* **76,** 177–188.
27. Satpaev, D. K., and Slepak, V. Z. (2000). Analysis of protein–protein interactions in phototransduction cascade using surface plasmon resonance. *Methods Enzymol.* **316,** 20–40.
28. Fisher, H. F., and Singh, N. (1995). Calorimetric methods for interpreting protein–ligand interactions. *Methods Enzymol.* **259,** 194–221.
29. Rivas, G., Stafford, W., and Minton, A. P. (1999). Characterization of heterologous protein–protein interactions using analytical ultracentrifugation. *Methods* **19,** 194–212.
30. Colton, I. J., Carbeck, J. D., Rao, J., and Whitesides, G. M. (1998). Affinity capillary electrophoresis: A physical–organic tool for studying interactions in biomolecular recognition. *Electrophoresis* **19,** 367–382.
31. Hainzl, T., and Boehm, T. (1998). A versatile expression vector for the in vitro study of protein–protein interactions: Characterization of E47 mutant proteins. *Oncogene* **9,** 885–891.
32. Bartel, P. L., and Fields, S. (1997). ''*The Yeast Two-Hybrid System.*'' Oxford University Press, New York.
33. Topcu, Z., and Borden, K. L. (2000). The yeast two-hybrid system and its pharmaceutical significance. *Pharm. Res.* **17,** 1049–1055.
34. Henion, J., Li, Y. T., Hsieh, Y. L., and Ganem, B. (1993). Mass spectrometric investigations of drug–receptor interactions. *Ther. Drug Monit.* **15,** 563–569.
35. Anderegg, R. J., Wagner, D. S., Blackburn, R. K., Opiteck, G. J., and Jorgenson, J. W. (1997). A multidimensional approach to protein characterization. *J. Protein Chem.* **16,** 523–526.
36. Loo, J. A. (1997). Studying noncovalent protein complexes by electrospray ionization mass spectrometry. *Mass Spectrom. Rev.* **16,** 1–23.
37. Winston, R. L., and Fitzgerald, M. C. (1997). Mass spectrometry as a readout of protein structure and function. *Mass Spectrom. Rev.* **16,** 165–179.
38. Rostom, A. A., and Robinson, C. V. (1999). Disassembly of intact multiprotein complexes in the gas phase. *Curr. Opin. Struct. Biol.* **9,** 135–141.
39. Veenstra, T. D. (1999). Electrospray ionization mass spectrometry: A promising new technique in the study of protein/DNA noncovalent complexes. *Biochem. Biophys. Res. Commun.* **257,** 1–5.
40. Blackstock, W. P., and Weir, M. P. (1999). Proteomics: Quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* **17,** 121–127.
41. Loo, J. A., DeJohn, D. E., Du, P., Stevenson, T. I., and Ogorzalek Loo, R. R. (1999). Application of mass spectrometry for target identification and characterization. *Med. Res. Rev.* **19,** 307–319.

42. Last, A. M., and Robinson, C. V. (1999). Protein folding and interactions revealed by mass spectrometry. *Curr. Opin. Chem. Biol.* **3,** 564–570.
43. Pandey, A., and Mann, M. (2000). Proteomics to study genes and genomes. *Nature* **405,** 837–846.
44. Wasinger, V. C., *et al.* (1995). Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium. Electrophoresis* **16,** 1090–1094.
45. Wilkins, M. R., *et al.* (1996). From proteins to proteomes: Large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology* **14,** 61–65.
46. Wilkins, M. R., Williams, K. L., Appel, R. D. and Hochstrasser, D. F., Eds. (1997). ''*Proteome Research: New Frontiers in Functional Genomics.*'' Springer-Verlag, Berlin.
47. Jolles, P., and Jornvall, H., Eds. (2000).''*Proteomics in Functional Genomics: Protein Structure Analysis.*'' Birkhauser Verlag, Basel, Switzerland.
48. Dunn, M. J. (1997). Quantitative two-dimensional gel electrophoresis: From proteins to proteomes. *Biochem. Soc. Trans.* **25,** 248–254.
49. Fenselau, C. (1997). MALDI MS and strategies for protein analysis. *Anal. Chem.* **69,** 661A–665A.
50. Kuster, B., and Mann, M. (1998). Identifying proteins and post-translational modifications by mass spectrometry. *Curr. Opin. Struct. Biol.* **8,** 393–400.
51. Yates, J. R. (1998). Mass spectrometry and the age of the proteome. *J. Mass Spectrom* **33,** 1–19.
52. Yates, J. R. (2000). Mass spectrometry: From genomics to proteomics. *Trends Genet.* **16,** 5–8.
53. Chalmers, M. J., and Gaskell, S. J. (2000). Advances in mass spectrometry for proteome analysis. *Curr. Opin. Biotechnol.* **11,** 384–390.
54. Andersen, J. S., and Mann, M. (2000). Functional genomics by mass spectrometry. *FEBS Lett.* **480,** 25–31.
55. Burley, S. K., *et al.* (1999). Structural genomics: Beyond the human genome project. *Nat. Genet.* **23,** 151–157.
56. Klose, J. (1975). Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues: A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26,** 231–243.
57. O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250,** 4007–4021.
58. Siuzdak, G. (1996). ''*Mass Spectrometry for Biotechnology.*''Academic Press, San Diego, CA.
59. Chapman, J. R., Ed.(1996).''*Protein and Peptide Analysis by Mass Spectrometry.*'' Humana Press, Totowa, NJ.
60. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71.
61. Cole, R. B., Ed. (1997). ''*Electrospray Ionization Mass Spectrometry: Fundamentals, Instrumentation and Applications.*'' John Wiley & Sons, New York.
62. Karas, M., and Hillenkamp, F. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **60,** 2299–2301.
63. Link, A. J., *et al.* (1999). Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676–682.
64. Busch, K. L., Glish, G. L., and McLuckey, S. A. (1988). ''*Mass Spectrometry/Mass Spectrometry: Techniques and Applications of Tandem Mass Spectrometry.*'' *VCH Publishers, New York.*

65. Mann, M. (1996). A shortcut to interesting human genes: Peptide sequence tags, expressed-sequence tags and computers. *Trends Biochem. Sci.* **21,** 494–495.
66. Neubauer, G., *et al.* (1998). Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat. Genet.* **20,** 46–50.
67. McLafferty, F. W. (1981). Tandem mass spectrometry. *Science* **214,** 280–287.
68. Smith, H. W. (1953). From ''Fish to Philosopher: The Story of Duo Internal Environment.'' Little Brown, Boston.
69. Ohkuma, S., and Poole, B. (1978). Fluorescence probe measurement of the intralysosomal pH in living sells and the perturbation of pH by various agents. *Proc. Natl. Acad. Sci. USA* **75,** 3327–3331.
70. Addanki, S., Cahill, F. D., and Sotos, J. F. (1967). Intramitochondrial pH and intra-extramitochondrial pH gradient of beef heart mitochondria in various functional states. *Nature* **214,** 400–402.
71. Kuhn, W., and Ryffel, K. (1942). Herstellung Konzentrierter Losungen aug verdunnten durch blosse membranwirkung: Ein Modellversuch zur Funktion der Niere. *Z. Physiol. Chem.* **276,** 145.
72. Fitzgerald, M. C., Chernushevich, I., Standing, K. G., Whitman, C. P., and Kent, S. B. (1996). Probing the oligomeric structure of an enzyme by electrospray ionization time-of-flight mass spectrometry. *Proc. Natl. Acad. Sci. USA* **93,** 6851–6856.
73. Veenstra, T. D., Tomlinson, A. J., Benson, L., Kumar, R., and Naylor, S. (1998). Low temperature aqueous electrospray ionization mass spectrometry of noncovalent complexes. *J. Am. Soc. Mass Spectrom.* **9,** 580–584.
74. Veenstra, T. D., *et al.* (1998). Metal mediated sterol receptor–DNA complex association and dissociation determined by electrospray ionization mass spectrometry. *Nat. Biotechnol.* **16,** 262–266.
75. Wilm, M., and Mann, M. (1996). Analytical properties of the nanoelectrospray ion source. *Anal. Chem.* **68,** 1–8.
76. Wilm, M., *et al.* (1996). Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379,** 466–469.
77. Gottschalk, A., *et al.* (1999). Identification by mass spectrometry and functional analysis of novel proteins of the yeast [U4/U6·U5]tri-snRNP. *EMBO J.* **18,** 4535–4548.
78. Rappsilber, J., Siniossoglou, S., Hurt, E. C., and Mann, M. (2000). A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal. Chem.* **72,** 267–275.
79. Katta, V., and Chait, B. T. (1991). Observation of the heme–globin complex in native myoglobin by electrospray-ionization mass spectrometry. *J. Am. Chem. Soc.* **113,** 8534–8535.
80. Ganem, B., Li, Y.-T., and Henion, J. D. (1991). Detection of noncovalent receptor–ligand complexes by mass spectrometry. *J. Am. Chem. Soc.* **113,** 6794–6796.
81. Feinberg, H., Greenblatt, H. M., and Shoham, G. (1993). Structural studies of the role of the active site metal in metalloenzymes. *J. Chem. Inf. Comput. Sci.* **33,** 501–516.
82. Veenstra, T. D., Gross, M. D., Hunziker, W., and Kumar, R. (1995). Identification of metal-binding sites in rat brain calcium-binding protein. *J. Biol. Chem.* **270,** 30353–30358.
83. Veenstra, T. D., Johnson, K. L., Tomlinson, A. J., Naylor, S., and Kumar, R. (1997). Determination of calcium-binding sites in rat brain calbindin D28K by electrospray ionization mass spectrometry. *Biochemistry* **36,** 3535–3542.

84. Troxler, H., Kuster, T., Rhyner, J. A., Gehrig, P., and Heizmann, C. W. (1999). Electrospray ionization mass spectrometry: Analysis of the Ca^{2+}-binding properties of human recombinant α-parvalbumin and nine mutant proteins. *Anal. Biochem.* **268,** 64–71.
85. Chazin, W., and Veenstra, T. D. (1999). Determination of the metal-binding cooperativity of wild-type and mutant calbindin D_{9K} by electrospray ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **13,** 548–555.
86. Yuan, Z., Feng, R., Castelhano, A., and Billedeau, R. (1994). Electrospray mass spectrometry study of metal ions in matrilysin: Evidence that two zincs and two calciums are required for inhibitor binding. *Ann. N. Y. Acad. Sci.* **732,** 489–492.
87. Johnson, K. A., Verhagen, M. F., Brereton, P. S., Adams, M. W., and Amster, I. J. (2000). Probing the stoichiometry and oxidation states of metal centers in iron–sulfur proteins using electrospray FTICR mass spectrometry. *Anal. Chem.* **72,** 1410–1418.
88. Zaia, J., Fabris, D., Wei, D., Karpel, R. L., and Fenselau, C. (1998). Monitoring metal ion flux in reactions of metallothionein and drug-modified metallothionein by electrospray mass spectrometry. *Protein Sci.* **7,** 2398–2404.
89. Gehrig, P. M., *et al.* (2000). Electrospray ionization mass spectrometry of zinc, cadmium, and copper metallothioneins: Evidence for metal-binding cooperativity. *Protein Sci.* **9,** 395–402.
90. Strupat, K., Rogniaux, H., Van Dorsselaer, A., Roth, J., and Vogl, T. (2000). Calcium-induced noncovalently linked tetramers of MRP8 and MRP14 are confirmed by electrospray ionization-mass analysis. *J. Am. Soc. Mass Spectrom* **11,** 780–788.
91. Katta, V., and Chait, B. T. (1991). Conformational changes in proteins probed by hydrogen-exchange electrospray-ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **5,** 214–217.
92. Lafitte, D., Capony, J. P., Grassy, G., Haiech, J., and Calas, B. (1995). Analysis of the ion binding sites of calmodulin by electrospray ionization mass spectrometry. *Biochemistry* **34,** 13825–13832.
93. Veenstra, T. D., Johnson, K. L., Tomlinson, A. J., Kumar, R., and Naylor, S. (1998). Correlation of fluorescence and circular dichroism spectroscopy with electrospray ionization mass spectrometry in the determination of tertiary conformational changes in calcium-binding proteins. *Rapid Commun. Mass Spectrom.* **12,** 613–619.
94. Veenstra, T. D., *et al.* (1998). Zinc-induced conformational changes in the DNA-binding domain of the vitamin D receptor determined by electrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom.* **9,** 8–14.
95. Nemirovskiy, O., Giblin, D. E., and Gross, M. L. (1999). Electrospray ionization mass spectrometry and hydrogen/deuterium exchange for probing the interaction of calmodulin with calcium. *J. Am. Soc. Mass Spectrom.* **10,** 711–718.
96. Williams, C. (2000). Biotechnology match making: Screening orphan ligands and receptors. *Curr. Opin. Biotechnol.* **11,** 42–46.
97. Lightstone, F. C., *et al.* (2000). Identification of novel small molecule ligands that bind to tetanus toxin. *Chem. Res. Toxicol.* **13,** 356–362.
98. Heller, M., Goodlett, D. R., Watts, J. D., and Aebersold, R. (2000). A comprehensive characterization of the T-cell antigen receptor complex composition by microcapillary liquid chromatography-tandem mass spectrometry. *Electrophoresis* **21,** 2180–2195.

99. Bruce, J. E., Smith, V. F., Liu, C., Randal, L. L., and Smith, R. D. (1998). The observation of chaperone–ligand noncovalent complexes with electrospray ionization mass spectrometry. *Protein Sci.* **7,** 1180–1185.
100. Randall, L. L., *et al.* (1998). The interaction between the chaperone SecB and its ligands: Evidence for multiple subsites for binding. *Protein Sci.* **7,** 2384–2390.
101. Ayed, A., Krutchinsky, A. N., Ens, W., Standing, K. G., and Duckworth, H. W. (1998). Quantitative evaluation of protein–protein and ligand–protein equilibria of a large allosteric enzyme by electrospray ionization time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **12,** 339–344.
102. Manabe, T. (2000). Combination of electrophoretic techniques for comprehensive analysis of complex protein systems. *Electrophoresis* **21,** 1116–1122.
103. Houry, W. A., Frishman, D., Eckerskorn, C., Lottspeich, F., and Hartl, F. U. (1999). Identification of in vivo substrates of the chaperonin GroEL. *Nature* **402,** 147–154.
104. Rout, M. P., *et al.* (2000). The yeast nuclear pore complex: Composition, architecture, and transport mechanism. *J. Cell Biol.* **148,** 635–651.
105. Green, B. N., *et al.* (1999). Electrospray ionization mass spectrometric determination of the molecular mass of the approximately 200-kDa globin dodecamer subassemblies in hexagonal bilayer hemoglobins. *J. Biol. Chem.* **274,** 28206–28212.
106. Vis, H., Dobson, C. M., and Robinson, C. V. (1999). Selective association of protein molecules followed by mass spectrometry. *Protein Sci.* **8,** 1368–1370.
107. Fabris, D., and Fenselau, C. (1999). Characterization of allosteric insulin hexamers by electrospray ionization mass spectrometry. *Anal. Chem.* **71,** 384–387.
108. Nemirovskiy, O. V., Ramanathan, R., and Gross, M. L. (1997). Investigation of calcium-induced, noncovalent association of calmodulin with melittin by electrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom.* **8,** 809–812.
109. Shen, M. L., Benson, L. M., Johnson, K. L., Lipsky, J. J., and Naylor, S. (2001). Effect of enzyme inhibitors on protein quaternary structure determined by on-line size exclusion chromatography-microelectrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom.* **12,** 97–104.
110. Nettleton, E. J., *et al.* (1998). Protein subunit interactions and structural integrity of amyloidogenic transthyretins: Evidence from electrospray mass spectrometry. *J. Mol. Biol.* **281,** 553–564.
111. Cheng, X., *et al.* (1996). Mass spectrometric characterization of sequence-specific complexes of DNA and transcription factor PU.1 DNA binding domain. *Anal. Biochem.* **239,** 35–40.
112. Cheng, X., Harms, A. C., Goudreau, P. N., Terwilliger, T. C., and Smith, R. D. (1996). Direct measurement of oligonucleotide binding stoichiometry of gene V protein by mass spectrometry. *Proc. Natl. Acad. Sci. USA* **93,** 7022–7027.
113. Liu, C., *et al.* (1998). Probing RegA/RNA interactions using electrospray ionization-Fourier transform ion cyclotron resonance-mass spectrometry. *Anal. Biochem.* **262,** 67–76.
114. Rostom, A. A., *et al.* (2000). Detection and selective dissociation of intact ribosomes in a mass spectrometer. *Proc. Natl. Acad. Sci. USA* **97,** 5185–5190.
115. Craig, T. A. Benson, L. M. Naylor, S. Kumar, R. (2000). Modulation effects of zinc on the formation of vitamin D receptor and retinoid X receptor α–DNA transcription complexes: Analysis by microelectrospray ionization mass spectrometry. Rapid *Commun. Mass Spectrom.* **15,** 1011–1016.
116. Craig, T. A., *et al.* (1999). Analysis of transcription complexes and effects of ligands by microelectrospray ionization mass spectrometry. *Nat. Biotechnol.* **17,** 1214–1218.

117. Potier, N., *et al.* (1998). Study of a noncovalent Trp repressor:DNA operator complex by electrospray ionization time-of-flight mass spectrometry. *Protein Sci.* **7,** 1388–1395.
118. Nordhoff, E., *et al.* (1999). Rapid identification of DNA-binding proteins by mass spectrometry. *Nat. Biotechnol.* **17,** 884–888.
119. Mendelsohn, A. R., and Brent, R. (1999). Protein interaction methods–toward an endgame. *Science* **284,** 1948–1950.
120. Zubarev, R. A., *et al.* (2000). Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal. Chem.* **72,** 563–573.
121. MacBeath, G., and Schreiber, S. L. (2000). Printing proteins as microarrays for high-throughput function determination. *Science* **289,** 1760–1763.
122. Davies, H., Lomas, L., and Austen, B. (1999). Profiling of amyloid β peptide variants using SELDI protein chip arrays. *Biotechniques* **27,** 1258–1261.
123. Li, J., *et al.* (1999). Integration of microfabricated devices to capillary electrophoresis-electrospray mass spectrometry using a low dead volume connection: Application to rapid analyses of proteolytic digests. *Anal. Chem.* **71,** 3036–3045.
124. Licklider, L., Wang, X. O., Desai, A., Tai, Y. C., and Lee, T. D. (2000). A micromachined chip-based electrospray source for mass spectrometry. *Anal. Chem.* **72,** 367–375.
125. Nelson, R. W., Nedelkov, D., and Tubbs, K. A. (2000). Biomolecular interaction analysis mass spectometry: BIA/MS can detect and characterize proteins in complex biological fluids at the low-to subfemtomole level. *Anal. Chem.* **72,** 404A–411A.
126. Williams, C., and Addona, T. A. (2000). The integration of SPR biosensors with mass spectrometry: Possible applications for proteome analysis. *Trends Biotechnol.* **18,** 45–48.
127. Fields, S. (2001). Proteomics: Proteomics in Genomeland. *Science* **291,** 1221–1224.

APPLICATION OF SEPARATION TECHNOLOGIES TO PROTEOMICS RESEARCH

By HALEEM J. ISSAQ

Separation Technology Group, Analytical Chemistry Laboratory, SAIC-Frederick, National Cancer Institute at Frederick, Frederick, Maryland 21702

I. Introduction

Separation science has played an active, and critical, role in many fields. These fields include organic synthesis for determination of the product purity; pharmaceutical science for assaying a product's purity, stability, and activity (i.e., pharmacokinetics); clinical chemistry and toxicology, where specimens of blood, urine, or tissue are analyzed for toxicological agents; forensic science for analysis of blood, hair, skin, or semen; and biology for analysis of amino acids, peptides, proteins, and nucleic acids. It is doubtful whether, if not for the innovative mind of the analytical chemist and the invention of modern capillary electrophoresis (CE) by Jorgenson and Lukacs [1], the Human Genome Project would have been completed so rapidly.

The next challenging analytical project for the separation scientist is in the field of proteomics. Although proteomics has more recently taken on many different meanings, in it was originally defined as the characterization of the entire protein complement within a cell, tissue, or organism. Although the analysis of the genome was a great accomplishment, it is not

ADVANCES IN
PROTEIN CHEMISTRY, Vol. 65
0065-3233/03 $35.00

as difficult as the characterization of all the proteins in a cell or tissue sample. Whereas the genome of a multicellular organism is static, its proteome varies depending on the cell type. In addition, the proteome of any given cell will dynamically change depending on alterations in its environment [2]. Although obviously only a small percentage of the estimated 30,000 to 35,000 open reading frames (ORFs) in the human genome will be expressed at any one time, the variety of different splice variants and isoforms that make up the ensemble of mature proteins represents an extremely complex milieu. In addition, these proteins are present at concentration levels spanning at least five orders of magnitude [3,4]. Presently, two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) is the method of choice for the separation of complex mixtures of proteins. In 2D-PAGE proteins are separated in two sequential steps based on their charge and mass. Quantitation of the separated proteins is based on their intensity after staining with Coomassie blue or silver stain, or SYPRO Ruby. Quantitation may also be carried out by 2D differential gel electrophoresis (DIGE), in which two proteome samples are covalently labeled with two different fluors and the fluorescence levels of each protein within an individual spot are compared. After quantitation, the protein spots are identified by mass spectrometry (MS) (Fig. 1). Unfortunately, 2D-PAGE has several limitations including its ability to visualize only the highest abundant proteins within a proteomic mixture [5]. Therefore, other sample preparation, concentration, separation, and detection methods need to be developed that would allow higher resolution than 2D-PAGE to separate a larger number of proteins, and sensitive enough to detect the low-abundance proteins. The advantages and limitations of 2D-PAGE in comparison are other separation methods are discussed later.

II. What Is Proteomics?

The burgeoning field of proteomics has given rise to a gamut of different definitions of what proteomics means. Yates [4] defined proteomics as the scientific discipline of characterizing and analyzing the proteins, protein interactions, and protein modifications of an organism. Gygi and Aebersold [5] have defined proteomics as the ability to systematically identify every protein expressed in a cell or tissue as well as to determine the salient properties of each protein such as abundance, state of modification, and involvement in multiprotein complexes. Others define it as the study of protein expression from cells, tissue, or organism, and the ability to study cellular processes at the molecular level. The goal in structural proteomics is to develop and apply experimental approaches to define the primary,

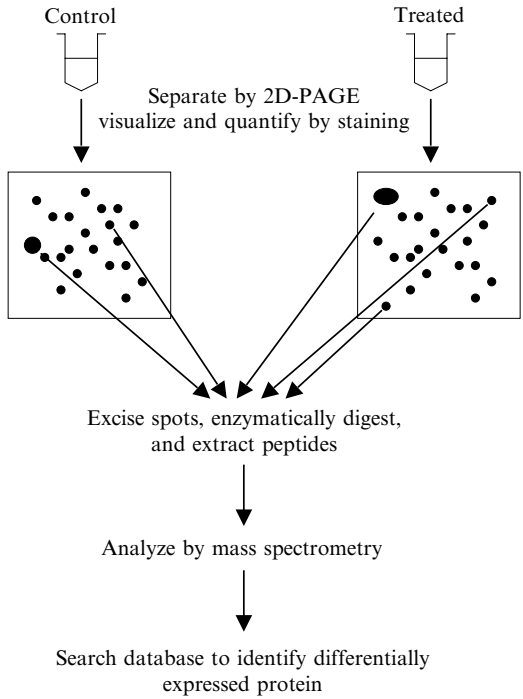


FIG. 1. Analysis of proteomes via 2D-PAGE-MS analysis. In a majority of comparative proteomic analyses, proteome samples extracted from separate cell systems are separated by 2D-PAGE. After staining the separated proteins, spots that show a difference in intensity are cored from the gel and enzymatically digested, and the resultant peptides are extracted from the gel matrix. The protein is then typically identified either by peptide mapping or tandem mass spectrometry.

secondary, and tertiary structure of proteins, whereas the goal of functional proteomics is to make use of the information provided by structural genomics to identify the function of each protein. Ultimately, the term proteomics has grown to encompass almost any experimental approach designed to characterize a protein or proteins at any level. This review, however, focuses on the role of the separation sciences in resolving complex mixtures of proteins for proteome analysis.

III. What Needs to Be Separated?

The type of analysis being undertaken best determines the optimal separation technique for any proteomic study. Does the study require the separation and analysis of a single specific protein, a group of proteins,

or all of the proteins in a cell or tissue? The problem, therefore, defines which technique [high-performance liquid chromatogry (HPLC), capillary electrophoresis (CE), slab gel electrophoresis], combination of techniques (multidimensional), or selective separation [affinity, isoelectric focusing (IEF), ion exchange, size exclusion, etc.] should be used. In general, the more complex the sample, the higher resolution the separation needs to be.

In a human cell, one can anticipate approximately 10,000–20,000 proteins, which when digested could produce upward of 10^6 tryptic peptides. No single chromatographic or electrophoretic method is capable of effectively separating such a complex mixture, so that MS can be used to identify a reasonable percentage of the peptides. It is abundantly clear that multidimensional separations of cell protein digests before MS analysis provide a much greater level of proteome characterization. The more effectively that species within a proteomic mixture are resolved simplifies the identification of the proteins, allows a higher effective dynamic range in the overall measurements, and ultimately provides more accurate results.

IV. Slab Gel Electrophoresis, High-Performance Liquid Chromatography, or Capillary Electrophoresis?

Slab gel electrophoresis (SGE) is an established technique for protein separation, whereby many samples can be analyzed simultaneously in a 1D format or an extremely complex mixture of proteins can be resolved in a 2D format. Compared with HPLC and CE, which require special instrumentation, SGE is economical. In spite of these advantages, however, SGE has the following limitations: poor solubility of hydrophobic and membrane proteins, limited dynamic range, difficulty in focusing highly basic and acidic proteins, poor sensitivity, and poor quantitation; in addition, it is highly laborious. On the other hand, CE and HPLC give higher sensitivity and faster analysis time, possess a large number of separation mechanisms, allow for the use of smaller sample sizes, and are more amenable to automation. A summary of the properties of each technique is given in Table I.

A. Selective Separation of a Specific Protein(s)

Selective separation of a specific protein(s) is achieved by affinity HPLC or affinity CE [6]. The principle of affinity is based on the ability of biologically active substances to bind specifically and reversibly to complementary substances, as shown in Fig. 2. These affinity reagents may include antibodies, lectins, aptamers, RNA or DNA fragments, and so

TABLE I
Comparison of High-Performance Liquid Chromatography Capillary, Electrophoresis, and Slab Gel Electrophoresis

Function	CE	SGE	HPLC
Automation	Yes	No	Yes
Speed	Seconds–minutes	Minutes–hours	Minutes
Sensitivity	n*M*–f*M*	μM	μM
Sample size	nl	μl	μl
Detection	On-column	Staining, fluorescence	Off-column
Quantitation	Peak area	Possible, but not simple	Peak area
Multiple samples	Yes	Yes	No
Multidimensional	Yes	Yes	Yes

on. The selection of a specific protein differs from the selection of a specific class of proteins in that the affinity reagent targets something unique to a single protein, such as an amino acid sequence (i.e., an epitope), and not to a functional group, such as a phosphorylation site, that may be present on potentially thousands of proteins. The binding sites of the immobilized substances must be sterically accessible even after they are coupled to the solid support and should not be deformed by immobilization. In the case of specific proteins an affinant is attached to the active surface of the column packing material (HPLC) or column surface (CE). The sample is injected onto the column and the protein(s) of interest is captured by the affinant, and the compounds that have no complementary binding site for the affinity-bound ligand will pass directly through the column. The captured protein is then eluted off the column by changing the properties of the buffer, pH, ionic strength, temperature, and so on.

B. Selective Separation of a Group of Proteins

Selective separation of a group of proteins from a cell lysate by HPLC or CE is achieved by affinity HPLC or affinity CE as mentioned above, or by stepwise changes in the experimental conditions, which are determined by the properties of the proteins of interest, such as hydrophobicity, charge, polarity, size, binding characteristic, or some specific modification. These selections are achieved by manipulations of either (1) the mobile-phase properties such as slope (time) of the gradient, percentage organic modifier, pH of the buffer, salt concentration, and ampholyte pH range,

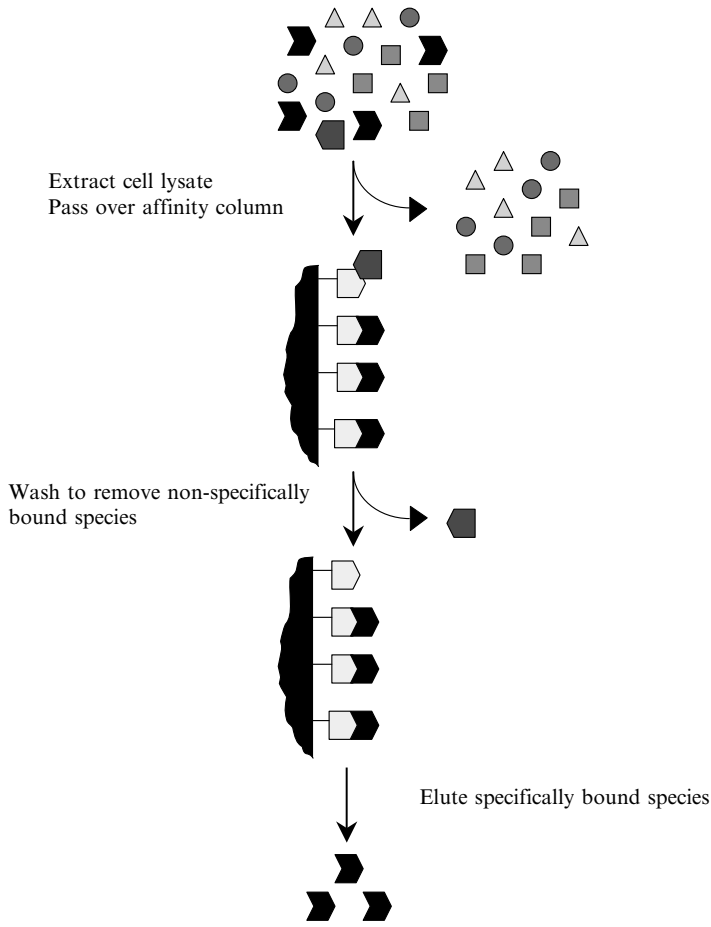


FIG. 2. Immunoaffinity chromatography isolation of a specific protein. In immunoaffinity chromatography, a cell extract is passed over a column containing an affinant covalently bound to a solid surface. Components within the extract with high affinity will noncovalently bind to the affinant. The column is then washed to remove species that may nonspecifically bind to the affinant. The specifically bound proteins are eluted off the column and analyzed.

or (2) column type, such as reversed-phase (C_{18}), ion exchange, size exclusion, gel filled, normal phase (silica), or affinity (type of antibody, lectin, aptamers, metals, DNA, etc.). Immobilized metal affinity ligands, such as copper and gallium, have been used to capture histidine-containing peptides and phosphopeptides [7, 8]. Methods to select peptides

containing cysteine, tryptophan, or methionine from pure proteins have also been reported [9, 10]. Other examples include the use of lectins to selectively isolate glycoproteins [11, 12].

C. *Separation of All Proteins in a Cell or Tissue*

No single chromatographic or electrophoretic procedure is likely to completely resolve a complex mixture of proteins extracted from a cell or tissue. A multidimensional method that employs orthogonal separation techniques or separation methods with different mechanisms of separation will significantly improve the chances of resolving such a complex mixture into its individual proteins. HPLC and CE allow for the combination of many different separations to increase the resolution of a complex mixture of proteins or peptides thereby increasing their chance of identification by MS. According to Giddings [13], the peak capacity of the multidimensional method is the product of the peak capacities of its component 1D methods. Therefore, if low- and high-resolution separation methods are combined in a multidimensional fractionation, the overall peak capacity will be greater than if only the high-resolution separation was used. Although this statement suggests that the coupling of many separations continues to make for more effective fractionation, factors such as complexity, robustness, and time required for the overall method need to be weighed against the required results. In addition, with the present pace of proteomics analysis, a truly effective multidimensional separation should be amenable to automation.

Although 2D-PAGE remains the most popular, many multidimensional approaches employing HPLC and/or CE have been developed. For example, any of the different slab gel, CE, or HPLC separation mechanisms listed in Table II can be combined to form a multidimensional separation. The advantages of HPLC-CE multidimensional methods are that they are amenable to automation, sensitive, reproducible, fast, and quantitative. Detection can be carried out by ultraviolet (UV) absorbance, MS, or laser-induced fluorescence (LIF). Two-dimensional gel electrophoresis is a good multidimensional separation method; however, it is manual, time consuming, and has limited sensitivity, poor quantitation, and limited dynamic range. Also, it has limited solubility of membrane and hydrophobic proteins. Slab gel electrophoresis, however, is a good sample preparation method. It can be used in the first dimension for resolving the proteins, from a cell lysate, into groups according to size, charge, or isoelectric point (p*I*).

TABLE II
Modes of Separation by Slab Gel Electrophoresis, High-Performance Liquid Chromatography, Capillary Electrochromatography, and Capillary Electrophoresis

	HIC		
Slab gel mode	HPLC mode	CEC mode	CE mode
Size	Reversed phase	Reversed phase	CZE
Charge	Ion exchange	Ion exchange	IEF
IEF	Size exclusion	Normal phase	SDS-PAGE
	Affinity		Affinity
	Normal phase		MECC

Abbreviations: CZE, capillary zone electrophoresis; HIC, hydrophobic interaction chromatography; IEF, isoelectric focusing; MECC, micellar electrokinetic chromatography.

V. Multidimensional Separations of Proteins

A. *Two-Dimensional Gel Electrophoresis*

The separation of a large number of proteins is routinely achieved by 2D-PAGE, a technique considered essential for proteomics research. In one study, however, Gygi *et al.* [14] showed that 2D-PAGE combined with MS detects only the most abundant proteins. Despite its limitations it is still the most used multidimensional separation technique in proteomics. Although the concept of 2D gel separation of complex protein mixtures has been around since the 1950s, the introduction by O'Farrell [15] of the separation of cellular proteins under denaturing conditions allowed hundreds of proteins to be resolved and serves as the basis of modern 2D-PAGE. Its relatively recent combination with MS analysis to identify the separated proteins has made 2D-PAGE a true phenomenon in proteomics. The list of samples that have been analyzed by 2D-PAGE/MS includes almost every conceivable life form ranging from human tissue to virus particles. After separation and staining, the spots are excised from the gel and digested with trypsin, and the peptide fragments are extracted and characterized by MS. Proteins are then identified by comparing the masses of the peptide fragments, or sequence information obtained by tandem MS, with predictions from either DNA or protein databases [16]. The wealth of data generated by 2D-PAGE separations has resulted in the construction of a number of different Web-based repositories for 2D-PAGE images. Many of the images deposited within the databases can be accessed at http://www.expasy.ch/ch2d/2d-index.html. Although a

majority of the images represent the analysis of human samples [17–19], results obtained with a variety of different eukaryotic and prokaryotic organisms can be accessed as well.

Proteomics studies are currently limited by the resolution capacity of 2D gels. In a standard 2D gel format 3000–10,000 proteins can be visualized depending on the method of spot detection. According to Vuong *et al.* [19] these spots are protein species with more than 10,000 molecules per cell. Detection methods such as fluorescence, imidazole/zinc, or silver staining cannot detect spots representing proteins with fewer than 1000 molecules per cell. It can be assumed that cells contain 10^9 protein molecules per cell and between 5×10^4 and 5×10^5 mRNA molecules per cell [19]. Thus one mRNA could conceivably produce many proteins. ''If we assume an average of five spots per gene per mammalian cell, and 10000 active genes per cell, we are faced with 50000 potential protein spots per cell. The 3000–10000 spots on 2D gels would then represent between 7–24% of the most abundant proteins'' [19]. Therefore, based on this argument 76% of the cell proteins fall below the detection limit of 2D-PAGE.

Investigators have devoted much intellectual thought and effort to increasing the number of protein species detectable by 2D-PAGE. The ultimate proteome analysis would allow all of the protein species in a biological sample to be visualized and characterized. Hoving *et al.* [20] developed a 2D-PAGE method in which they apply narrow-range immobilized pH gradient (IPG) strips in the first dimension. The IPG strips are typically 1–3 pH units wide and overlap with one another. In the initial study six IPG strips covering the pH ranges 3.5–5, 4.5–5.5, 5–6, 5.5–6.7, 6.2–8.2, and 7–10 were used. One to 2 mg of protein from a B lymphoma cell line was applied to each strip, and separated by IEF. The strips were then applied to individual sodium dodecyl sulfate (SDS)-PAGE gels and separated in the second dimension on the basis of molecular weight. The same sample was also run on a single standard 2D-PAGE gel with a single IPG strip with a pH range of 3–10. Using a single gel, the group was able to detect approximately 1500 spots; however, using the six ultrazoom gels they were able to detect about 5000 unique spots. In addition, whereas only 0.8 mg of protein could be loaded onto a single gel, the ultrazoom gels allowed them to load up to 11 mg of total protein. They estimate that the increase in resolution allowed for the detection of proteins present at only about 300 copies per cell. Wildgruber *et al.* compared the collective use of IPG strips with pH ranges of 4–5, 5–6, and 5.5–6.7 against gels run with IPG strips with pH ranges 3–10 and 4–7 [21]. They found that they were able to detect 2.3 and 1.6 times more protein spots with the narrow range gels than with the IPG 3–10 and 4–7 gels, respectively. Although this approach of using narrow range IPG strips will not be useful when sample is limited, it does provide for the detection of a

greater number of lower abundance proteins when sample quantity is not an issue.

B. Multidimensional HPLC, CE, and HPLC-CE Protein/Peptide Separations

The separation of all cell or tissue proteins is not a simple matter. In a human cell there are thousands of proteins that, when digested, would increase the complexity of the sample between one and two orders of magnitude. The separation of such a large number of compounds is not possible by only one chromatographic or electrophoretic run. To be able to separate a large number of proteins/peptides, a multidimensional approach is needed. In 2D gel electrophoresis proteins are first resolved and digested with trypsin, and the resultant peptides are analyzed by MS. In chromatographic and electrophoretic multidimensional approaches the proteins are generally digested into peptides and then separated. The advantage is that peptides are more soluble and easier to separate than intact proteins, especially hydrophobic and membrane proteins. The disadvantage is the increase in the number of species that must be resolved. This increase in number of species requires the combination of effective HPLC/CE methods into a single multidimensional method so that large numbers of peptides can be identified by MS analysis.

An important consideration when devising a multidimensional separation for proteomic study is whether it needs to be on-line or off-line with the analyzer (generally MS). Many factors need to be considered when making this decision. Assuming the sample is complex (or a single dimension, on-line separation would suffice), one of the major factors is sample amount. If only a small amount of sample is available, an on-line separation would minimize the potential for sample loss due to excess handling. Another important factor is time. On-line approaches are generally faster, possess higher throughput, and are generally automated; however, they also present significant technical challenges. For example when using an on-line method the second dimension must be much faster than the first dimension. In addition, it is sometimes difficult to make the buffer conditions needed for two separations compatible. Although more laborious and time-consuming, an off-line approach does have its advantages. It is simple and easy to perform, any column type and size can be used, any sample size can be analyzed, the equipment is generally commercially available, and it allows the reanalysis of collected fractions by multiple techniques [i.e., electrospray ionization-mass spectrometry (ESI-MS), mass-assisted laser desorption ionization (MALDI)-MS, LIF, etc.].

Because most liquid-phase separation methods used in proteomics today digest proteins into peptides before analysis, these studies are generally reliant on a single peptide species to represent an entire protein. This dependency requires that whatever has happened to the expression level protein within the cell is reflected within the measurement of one of its peptides. Because 2D-PAGE separates intact proteins, it is assumed that it would present a true representation of the protein's relative abundance. Regnier *et al.* [2] discussed in detail the subject of signature peptides, which are unique peptides associated with a single protein from which they were derived. This group used lectin affinity chromatography selection of peptides to reduce the complexity of the peptide mixture and show that identified peptides can act as an analytical surrogate of their protein of origin.

Many different LC and CE techniques have been combined in an attempt to increase the overall resolution of a proteomic separation. Moore and Jorgenson used a combination of online reversed-phase HPLC (RP-HPLC)/CE [22] and size-exclusion chromatography (SEC)/RP-HPLC/capillary zone electrophoresis (CZE) [23] to resolve a mixture of peptides. Issaq *et al.* [24, 25] used simple off-line RP-HPLC/CE for the separation of a mixture of protein digests, whereby fractions were collected every 30 automatically with the aid of a fraction collector into polypropylene 96-well microtiter plates. The fractions were then concentrated and analyzed by single-capillary CE [24] and by 96-array CE [25] with LIF detection. The CE separation mode can be either CZE [24, 25] or capillary gel electrophoresis. Such a 2D format can be used for the separation of intact proteins or peptides.

Opiteck *et al.* [26] described a 2D LC system that used SEC followed by RP-HPLC to separate a mixture of proteins extracted from *Escherichia coli* cells. One advantage of SEC is that it can be conducted under either denaturing or nondenaturing conditions. In this study, peaks eluting from the first dimension were automatically subjected to RP-HPLC to separate similarly sized proteins on the basis of their various hydrophobicities. The RP-HPLC also serves to desalt the analytes so that they can be detected by UV at 215 nm regardless of the SEC mobile phase used. The 2D chromatograms produced were displayed on an x–y axis, where spot intensities represent quantity of proteins. Selected collected fractions from RP-HPLC were then analyzed by MALDI-time of light (TOF)/MS or ESI/MS depending on sample concentration. Wall *et al.* [27] developed a 2D liquid-phase separation method that is capable of resolving a large number of cellular proteins. The proteins were separated into 20 fractions by pI, using IEF in the first dimension, after which the fractions were each analyzed by hydrophobicity, using nonporous RP-HPLC in the second

dimension. Peaks eluting from the HPLC column were collected and identified by proteolytic digestion and MALDI-TOF/MS. The authors reported improved resolution of low-mass and basic proteins compared with 2D-PAGE analysis.

One of the more successful multidimensional separations for complex proteome mixtures has been developed by Link and co-workers [28]. This two-dimensional separation method, termed MudPIT, is an on-line 2D ion-exchange/reversed-phase HPLC method, and was initially used to separate a mixture comprised of a tryptic digest of 80S ribosomes isolated from yeast. In the MudPIT approach, peptides are systematically separated, depending on charge in the first dimension and on hydrophobicity in the second. The acidified peptide mixture is loaded onto a strong cation-exchange (SCX) column. A discrete fraction of peptides is displaced from the SCX column directly onto a RP column, and is then separated and eluted from the RP column into the mass spectrometer (Fig. 3). This iterative process was repeated 12 times, using increasing salt gradient elution for the SCX column and an increasing organic concentration for the RP column [28]. The SCX and RP columns are packed at opposing ends of a single capillary column, minimizing the amount of sample loss between the two separation dimensions. A major advantage of this separation technology is that the entire system is coupled directly on-line with MS, enabling a large number of peptides to be directly identified in a high-throughput manner. The capabilities of MudPIT to characterize a complex cell lysate were then demonstrated with yeast. In this study, a tryptic digest of a yeast lysate was analyzed by 15 iterations of the MudPIT separation cycle. The resulting total ion chromatograms (TICs) from each of the iterations are shown in Fig. 4. The important characteristic to recognize is that each TIC is variable (showing different peptides are being analyzed by the MS in each iteration) and each has many peaks (showing a large number of peptides being separated). A wide variety of protein classes including proteins with extremes in p*I*, molecular weight, abundance, and hydrophobicity (including 131 proteins with three or more predicted transmembrane domains) were identified by the MudPIT separation method. In the MudPIT analysis of the *S. cerevisiae* proteome, 53% of the proteins detected and identified had codon adaptation index (CAI) values of less than 0.2 [28]. This unbiased protein coverage is essential for the detection of low-abundance proteins and is in stark contrast to studies that have shown that 2D-PAGE analysis makes it difficult to detect yeast proteins with CAI values of less than 0.2. A total of 2114 unique peptides were identified by the MudPIT-MS combination [28]. They have combined the MudPIT technology with a three-tiered protein digestion strategy to characterize 270 proteins from lens tissues, including

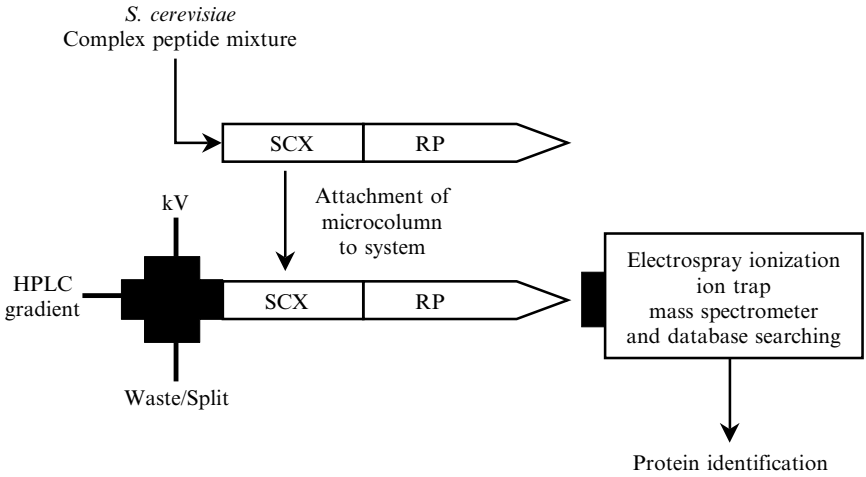
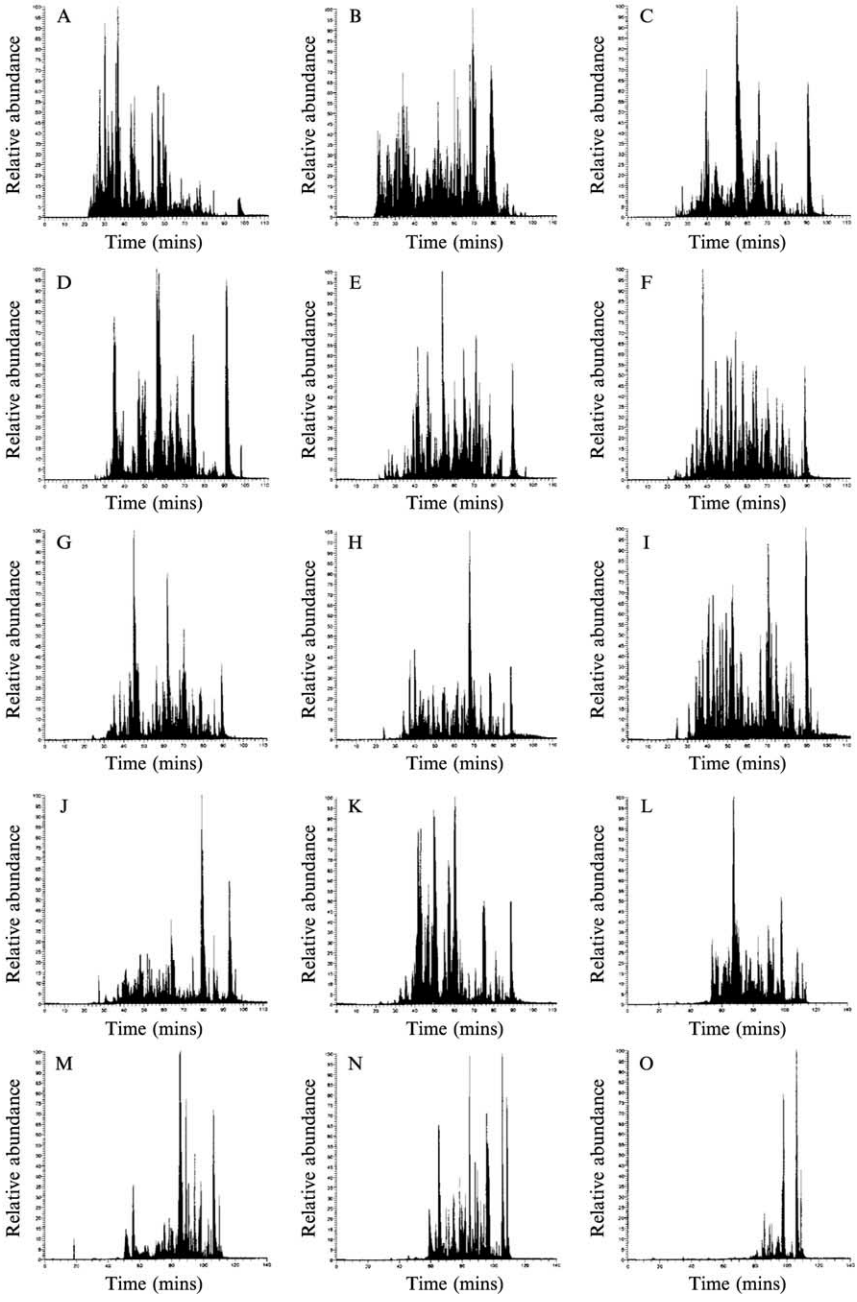


FIG. 3. Multidimensional protein identification technology (MudPIT). In the MudPIT multidimensional method, complex peptide mixtures are loaded onto a biphasic microcapillary column packed with strong cation-exchange (SCX) and reversed-phase (RP) materials. The column is then inserted on-line with an electrospray ionization ion-trap mass spectrometer (MS). Peptides are first displaced from the SCX to the RP by a salt gradient and eluted off the RP into the MS. In an iterative process, the microcolumn is reequilibrated and an additional salt step of higher concentration displaces peptides from the SCX to the RP. Peptides are again eluted by an RP gradient into the MS, and the process is repeated. The tandem mass spectra generated are correlated to theoretical mass spectra generated from protein or DNA databases.

11 different crystallins that were found to contain a total of 73 sites of modification. Their method was able to identify modifications ranging from Ser, Thr, and Tyr phosphorylation, to Arg and Lys methylation, Lys acetylation, and Met, Tyr, and Trp oxidations [29].

Although most multidimensional fractionation techniques have focused on the development of methods to fraction the entire peptide population obtained from a proteome sample, methods have also been developed to interrogate a specific subset of peptides. Hunt's group reported a multidimensional separation technology that enables the characterization of phosphopeptides from whole-cell lysate in a single experiment [30]. The proteins were digested with trypsin and the resulting peptides were converted to methyl esters. Immobilized metal-affinity chromatography (IMAC) was used in the first separation dimension to enrich for phosphopeptides, which were subsequently analyzed by nanoflow HPLC/ESI-MS. They detected more than 1000 phosphopeptides from an analysis of


A
B
C
D
E
F
G
H
I
J
K
L
M
N
O
Relative abundance
Time (mins)

a whole-cell lysate from *Saccharomyces cerevisiae*. A total of 216 peptide sequences defining 383 sites of phosphorylation were determined.

VI. On-Column HPLC and CE Protein Concentration

High sensitivity is an important and crucial aspect of proteomics research. The advantages of HPLC and CE over SGE are that dilute samples can be concentrated on-column before separation. In HPLC dilute solutions can be concentrated at the head of the column by manipulations of the mobile phase before the separation step. In CE different preconcentration approaches have been used. These methods include the use of antibodies [31], plugs of C_{18} [32], or electrophoretic techniques such as stacking [33, 34], isoelectric focusing (IEF) [35, 36], or isotachophoresis (ITP) [37, 38]. Figeys *et al.* [39] coupled a solid-phase extraction (SPE) device (C_{18}, 5-μm particles), CE-microelectrospray-MS/MS system to achieve limits of detection in the attomole to micromole range.

It is a fact that the more a protein mixture is manipulated the greater is the loss. An advantage of CE is that the analyst can introduce a cell(s) into the capillary or a microfluidic channel, lyse the cells, and then separate the proteins. Zhang *et al.* [40] used such a method to obtain a simple proteome map of a single cancer cell. An HT29 human colon adenocarcinoma cell was introduced into a fused-silica capillary and lysed, and the protein content was fluorescently labeled, separated by CZE, and detected by LIF. Several dozen components were readily resolved and detected. In another study,

FIG. 4. Chromatograms of a 15-cycle MudPIT analysis of a complex peptide mixture extracted from *S. cerevisiae*. The four buffer solutions used for the chromatography were 5% acetonitrile (ACN)–0.02% HFBA (buffer A), 80% ACN–0.02% HFBA (buffer B), 250 m*M* ammonium acetate–5% ACN–0.02% HFBA (buffer C), and 500 m*M* ammonium acetate–5% ACN–0.02% HFBA (buffer D). Cycle 1 (A) consisted of a 70-min gradient from 0 to 80% buffer B and a 10-min hold at 80% buffer B. Each of the next 12 cycles were 110 min with the following profile: 5 min of 100% buffer A, 2 min of *X*% buffer C, 3 min of 100% buffer A, a 10-min gradient from 0 to 10% buffer B, and a 90-min gradient from 10 to 45% buffer B. The 2-min buffer C in cycles 2–13 was as follows: cycle 2, 10% (B); cycle 3, 20% (C); cycle 4, 30% (D); cycle 5, 40% (E); cycle 6, 50% (F); cycle 7, 60% (G); cycle 8, 70% (H); cycle 9, 80% (I); cycle 10, 90% (J); cycle 11, 90% (K); cycle 12, 100% (L); and cycle 13, 100% (M). Cycle 14 (N) consisted of a 5-min 100% buffer A wash followed by a 20-min 100% buffer C wash, a 5-min 100% buffer A wash, a 10-min gradient from 0 to 10% buffer B, and a 90-min gradient from 10 to 45% buffer B. The chromatograms shown are representative of those obtained from samples of a complexity comparable to the one described. (Reproduced from reference 28.)

Chen and Lillard [41] developed an instrumental set-up for the continuous introduction and analysis of single cells. A flow-based interface that uses electro-osmotic flow provided continuous injection of intact cells through a capillary into a cell lysis junction and migration of the resulting cell lysate through a separation capillary for analysis. Hemoglobin and carbonic anhydrase were detected at a level of 37 and 1.6 amol by native fluorescence, using an excitation wavelength of 275 nm. A distinct advantage of single-cell analysis is that the information obtained presents a direct characterization of a specific cell and not a weighted average of a large cell population. Also, protein losses due to manipulations outside the capillary are eliminated. The disadvantage is that only abundant proteins are detected, and present technology does not afford a simple way to identify unknown species.

Microfluidics offers the possibility and has the potential for multitask manipulation of cells on one platform such as cell lysing, enzymatic digestion, and peptide labeling, followed by separation, LIF detection, and/or introduction into a mass spectrometer for peptide/protein identification. Gottsclich and co-workers [42] reported on an integrated microchip device for the enzymatic digestion, electrophoretic separation, and postcolumn labeling of proteins and peptides with naphthalene-2,3-dicarboxaldehyde and detection by LIF. To accomplish this, a reactor, an injector, a separator, and a second reactor were all integrated on a monolithic device. Bousse *et al.* [43] developed a microfabricated analytical device on a glass chip that performs a protein-sizing assay by integrating the required separation, staining, virtual destaining, and detection steps. Others also reported on a protein-sizing assay on a microchip using precolumn covalently labeled proteins [44]. Agilent Technologies have introduced the 2100 Bioanalyzer instrument equipped with a Protein 200 LabChip kit for the analysis of 10 protein samples in 45 min. Possible applications are comparison of protein expression levels, immobilized metal ion affinity chromatography, purity check, and so on. Proteins with molecular weights between 14,000 and 210,000 can be separated and sized.

VII. Detection

Unlike 2D gel electrophoresis, in which a stain or other procedure is needed for the detection and quantification of the resolved proteins, in HPLC, capillary electrochromatography (CEC), and CE proteins and peptides can be detected on-line by UV absorption, fluorescence, LIF, or mass spectrometry (MS). The strong absorption of the CO–NH peptide bond in the 185- to 220-nm wavelength region would allow the detection of the proteins. Detection sensitivity is increased at the lower end of this

wavelength region; however, the analyst should be careful in selecting a buffer that will not absorb at the selected detection wavelength. Proteins and peptides that contain an aromatic amino acid residue in their structure can be detected by UV absorption at 275–280 nm, and by fluorescence in the 200- to 300-nm wavelength range. However, CE fluorescence detection did not give better sensitivity than UV absorption detection. The detection of peptides and proteins by fluorescence and LIF depends on the quantum yield of the three aromatic amino acids. The quantum yield is greatest for tryptophan and lowest for phenylalanine, and the detection limits for these peptides when using a Kr-F UV laser operating at 248 nm were at least two orders of magnitude lower when compared with UV absorption at 214 nm [45]. Other lasers have been used in the range of 200–300 nm for the detection of native proteins. The optimum excitation was found to be 275 nm [46, 47].

It is almost universally agreed that MS is the instrumental analyzer that has made global proteomics a possibility. It is an ideal detection technique for peptides and proteins because of its universality, sensitivity, and selectivity. The success of MS as a detection technique for proteins and peptides is due to three sample introduction techniques: continuous-flow fast atom bombardment (FAB) [48], electrospray ionization (ESI) [49], and mass-assisted laser desorption ionization (MALDI) [50]. These three MS sample introduction techniques solved the problem of introduction into the mass spectrometer source of polar, nonvolatile compounds, such as peptides and proteins. According to Thomas *et al.* [51], ''ESI-MS is useful for probing a wide range of biological problems as a detector for HPLC and CE, in the study of noncovalent complexes, and for obtaining structural information. ESI does have limitations in that it is not very tolerant of the presence of salts (>1.0 mM), nor is it practical for the analysis of multicomponent samples. MALDI-MS has proven to be very effective where ESI is not very useful.'' Today, MS, especially when coupled to HPLC or CE, is one of the most important tools in proteomics research for peptide mapping, for the confirmation of protein sequences, and for the investigation of posttranslation modification. It is also used for quantification of proteins by employing isotope-labeling techniques [52, 53].

Although the major focus in maximizing the number of species that can be detected in a complex proteome sample has been on the development of multidimensional separation methods, other groups have focused more on developing analyzers with greater resolution, sensitivity, and dynamic range. Pasa-Tolic and co-workers developed an interesting approach for analyzing global proteolytic digests by ultrahigh-resolution capillary HPLC combined with a powerful Fourier transform ion cyclotron resonance

(FTICR) MS [54]. Whereas the focus of many multidimensional separations is to increase the dynamic range of the measurements, the superior dynamic range of FTICR compared with more conventional MS technologies lessens the demands on the separation technique. The high mass accuracy of FTICR makes possible a new concept of ''accurate mass tags'' that could largely eliminate the need for time-consuming MS/MS measurements [55]. Such strategy allows the measurement of many proteins and provides a methodology that could make possible large dynamic range measurements of proteomes in a high-throughput manner.

VIII. Proteome Quantitation Strategies

One of the primary goals of proteomics, beyond identification of the proteins presents, is the measurement of their relative abundance in two distinct samples. Two-dimensional polyacrylamide gel electrophoresis is the conventional method of measuring changes in protein expression levels by comparing spot intensities. Liquid-phase separation methods, however, have also been employed to measure changes in protein expression. One of the earliest demonstrations of an on-line separation coupled with MS to measure the relative abundances of proteins was conducted in the Smith laboratory [54, 56]. In this study, cultures of *Escherichia coli* were grown in both normal isotopic abundance and rare isotope (i.e., ^{15}N, ^{13}C)-depleted media. The culture grown in isotope-depleted medium was treated with cadmium (Cd^{2+}) and combined with an equal number of cells from the untreated culture. After extracting the soluble proteins, they were separated and analyzed by capillary IEF coupled directly on-line with FTICR-MS. The origin of the proteins observed in the mass spectra could be deduced from their isotope envelope; proteins harvested from cells grown in normal medium had a natural isotopic distribution whereas those harvested from cells grown in depleted medium contained a single major peak representing the monoisotopic mass of the protein. A number of proteins were observed that showed a significant change in their abundance when treated with Cd^{2+} [56].

One of the most popular methods for comparing the relative abundance of proteins from two distinct samples is the use of isotope-coded affinity tags (ICATs), initially demonstrated by Gygi and co-workers [52] to generate protein expression profiles of yeast utilizing galactose or ethanol as a carbon source. Although thought of foremost as a quantitative strategy, ICAT is inherently also a multidimensional fractionation method, based on its peptide selectivity. In this strategy, cysteinyl residues within proteins are modified with a thiol-reactive reagent that

contains a biotin moiety. The proteins are enzymatically digested and the modified peptides are recovered by immobilized avidin chromatography. The main purpose of affinity isolating only the cysteinyl-containing peptides is to reduce the complexity of the sample; however, because ICAT technology is designed for global proteomic studies, the postaffinity chromatography sample is still quite complex. Therefore, most investigators using ICAT technology have resorted to using a 2D separation technique after this initial fractionation step [57]. The most popular technique uses a combination of SCX and RP-HPLC coupled on-line with ESI-MS for identification and quantitation of the modified peptides. This approach was used to identify and quantify proteins contained in the microsomal fractions of naive and *in vitro*-differentiated human myeloid leukemia cells [57]. After ICAT labeling and affinity isolation of the modified cysteinyl-containing peptides, the peptide mixture was separated into 30 fractions via SCX chromatography (Fig. 5; see Color Insert). Each of these fractions was subsequently analyzed by RP-HPLC/MS, resulting in the identification and quantitation of 491 unique proteins.

IX. Conclusion

Many instrumental strategies have been developed that are based on efficient separation followed by MS identification and quantitation. Two-dimensional polyacrylamide gel electrophoresis is the method that biologists prefer. However, this 2-D procedure suffers from many limitations, of which sensitivity and precision of quantification are the main concerns. Many separation methods have been reported in the scientific literature employing a single high-resolution procedure as well as a multidimensional approach of HPLC/HPLC or HPLC/CE on-line or off-line with MS for protein identification. The separation method selected obviously depends primarily on the nature and complexity of the sample to be characterized. In addition, the focus of the proteome analysis also dictates the nature of the separation selected, as many separation techniques have been developed to extract specific classes of proteins. Regardless of the project focus, the continuing development of effective fractionation methods, both gel based and solution based, will be crucial to the continuing success of proteomics.

Acknowledgments

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. NO1-CO-12400.

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

References

1. Jorgenson, J., and Lukacs, K. D. (1981). *Anal. Chem.* **53,** 1298–1302.
2. Regnier, F., Amini, A., Chakraborty, A., Geng, M., Ji, J., Riggs, L., Sioma, C., Wang, S., and Zhang, X. (2001). *LC-GC* **19,** 200–213. and references therein.
3. Celis, E., and Gromov, P. (1999). *Electrophoresis* **10,** 16–21.
4. Yates, J. (2001). Proteomics Workshop at HPCE 2001, Boston, MA.
5. Gygi, S. P., and Aebersold, R. (2000). *Curr. Opin. Chem. Biol.* **4,** 489–494.
6. Turkova, J. (1999). *In* ''Analytical and Preparative Separation Methods of Biomolecules'' (H. Y. Aboul-Enein, Ed.), pp. 99–165. Marcel Dekker, New York.
7. Belew, M., Yip, T. T., Andersson, L., and Ehrnstrom, R. (1987). *Anal. Biochem.* **164,** 457–465.
8. Ji, J., Chakraborty, A., Geng, M., Zhang, X., Amini, A., Bina, M., and Regnier, F. (2000). *J. Chromatogr. B.* **745,** 197–210.
9. Barnard, G., Bayer, E., Wilchek, M., and Amir-Zaltsman, Y. (1986). *Methods Enzymol.* **133,** 284–288.
10. Coligan, J. E. (1999). *In* ''Current Protocols in Protein Science'' (J. E. Coligan, B. Dunn, H. Ploegh, D. Speicher, and P. Wingfield, Eds.), pp. A.1A. John Wiley & Sons, New York.
11. Posewitz, M. C., and Tempst, P. J. (1999). *Anal. Chem.* **71,** 2883–2892.
12. Schachter, H. (1991). *Glycobiology* **1,** 453–461.
13. Giddings, J. C. (1987). *J. High Resolut. Chromatogr. Chromatogr. Commun.* **10,** 319–323.
14. Gygi, S. G., Corthalas, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000). *Proc. Natl. Acad. Sci. USA* **97,** 9390–9395.
15. O'Farrell, J. (1974). *Biol. Chem.* **250,** 4007–4021.
16. Yates, J. (1998). *J. Mass Spectrom* **33,** 1–19.
17. Ueno, I., Sakal, T., Yamaoka, M., Yoshida, R., and Tsugita, A. (2000). *Electrophoresis* **21,** 1832–1845.
18. Simpson, R. J., Connolly, L. M., Eddes, J. S., Pereira, J. J., Moritz, R. L., and Reid, G. E. (2000). *Electrophoresis* **21,** 1707–1732.
19. Vuong, G. L., Weiss, S. M., Kammer, W., Priemer, M., Vingron, M., Nordheim, A., and Cahill, M. A. (2000). *Electrophoresis* **21,** 2594–2605.
20. Hoving, S., Voshol, H., and Oostrum, J. V. (2000). *Electrophoresis* **21,** 2617–2621.
21. Wildgruber, R., Reil, G., Drews, O., Parlar, H., and Gorg, A. (2002). *Proteomics* **2,** 727–732.
22. Moore, A. V., and Jorgenson, J. W. (1995). *Anal. Chem.* **67,** 3448–3455.
23. Moore, A. V., and Jorgenson, J. W. (1995). *Anal. Chem.* **67,** 3456–3463.
24. Issaq, H. J., Chan, K. C., Janini, G. M., and Muschik, G. M. (1999). *Electrophoresis* **20,** 1533–1537.
25. Issaq, H. J., Chan, K. C., Cheng, S. L., and Qingbo, L. (2001). *Electrophoresis* **22,** 1133–1135.
26. Opiteck, G. J., Ramirez, S. M., Jorgenson, J. W., and Moseley, M. A. (1998). *Anal. Biochem.* **258,** 349–361.
27. Wall, D. B., Kachman, M. T., Gong, S., Hinderer, R., Parus, S., Misek, D. E., Hanash, S. M., and Lubman, D. M. (2000). *Anal. Chem.* **72,** 1099–1111.

28. Wolters, D. A., Washburn, M. P., and Yates, J. R. 3rd (2001). *Anal. Chem.* **73***(*23*),* 5683–5690.
29. MacCoss, M. J., McDonald, W. H., Saraf, A., Sadygov, R., Clark, J. M., Tasto, J. J., Gould, K. L., Wolters, D., Washburn, M., Weiss, A., Clark, J. I., and Yates, J. R., III. (2000). *Proc. Natl. Acad. Sci. USA* **99,** 7900–7905.
30. Ficarro, S. B., McCleland, M. L., Stukenberg, P. T., Burke, D. J., Ross, M. M., Shabanowitz, J., Hunt, D. F., and White, F. M. (2002). *Nat. Biotechnol.* **20,** 301–305.
31. Guzman, N. (1999). *LC-GC.* **17,** 16–27.
32. Tomlinson, A. J., Guzman, N. A., and Naylor, S. (1995). *J. Capillary Electrophoresis* **2,** 247–266.
33. Aebersold, R., and Morrison, H. D. (1990). *J. Chromatogr.* **516,** 79–88.
34. Burgi, D. S., and Chien, R.-L. (1991). *Anal. Chem.* **63,** 2042–2047.
35. Lamoree, M. H., Tjaden, U. R., and van der Greef, J. (1997). *J. Chromatogr. A* **777,** 31–39.
36. Rodriguez-Diaz, R., Wehr, T., and Zhu, M. (1997). *Electrophoresis* **18,** 2134–2144.
37. Usdeth, H. R., Loo, J. A., and Smith, R. D. (1989). *Anal. Chem.* **61,** 228–232.
38. Stegehuis, D. S., Irth, H., Tjaden, U. R., and van der Greef, J. (1991). *J. Chromatogr.* **538,** 393–402.
39. Figeys, D., Ducret, A., and Aebersold, R. (1997). *J. Chromatogr. A.* **763,** 295–306.
40. Zhang, Z., Krylov, S., Arriaga, E. A., Polakowski, R., and Dovichi, N. J. (2000). *Anal. Chem.* **72,** 318–322.
41. Chen, S., and Lillard, S. J. (2001). *Anal. Chem.* **73,** 111–118.
42. Gottsclich, N., Culbertson, C. T., McKnight, T. E., Jacobson, C. J., and Ramsey, J. M. (2000). *J. Chromatogr. B.* **745,** 243–249.
43. Bousse, L., Mouradian, S., Minalla, A., Yee, H., Williams, K., and Dubrow, R. (2001). *Anal. Chem.* **73,** 1207–1212.
44. Yao, S., Anex, D. S., Caldwell, W. B., Arnold, D. W., Smith, K. B., and Schultz, P. G. (1999). *Proc. Natl. Acad. Sci. USA* **96,** 5372–5377.
45. Chan, K. C., Janini, G. M., Muschik, G. M., and Issaq, H. J. (1993). *J. Liq. Chromatogr.* **16,** 1877–1890.
46. Leo, T. T., and Yeung, E. S. (1992). *J. Chromatogr.* **595,** 319–325.
47. Issaq, H. J., and Chan, K. C. (1995). *Electrophoresis* **16,** 467–480.
48. Suter, M.J.F., and Caprioli, R. M. (1992). *J. Am. Soc. Mass Spectrom* **3,** 198–206.
49. Loo, J. A., Udesth, H. R., and Smith, R. D. (1989). *Anal. Biochem.* **179,** 404–412.
50. Walker, K. L., Chiu, R. W., Moning, C. A., and Wilkins, C. L. (1995). *Anal. Chem.* **67,** 4197–4204.
51. Thomas, J. J., Bakhtiar, R., and Siuzdak, G. (2000). *Acc. Chem. Res.* **33,** 179–187.
52. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999). *Nat. Biotechnol.* **17,** 249–307.
53. Regnier, F. (2001). Global quantification strategies for proteomics. *In*: ''HPCE 2001,'' Boston, MA [abstract L 1102].
54. Pasa-Tolic, L., Jensen, P. K., Anderson, G. A., Lipton, M. S., Peden, K. K., Martinovic, S., Tolic, N., Bruce, J. E., and Smith, R. D. (1999). *J. Am. Chem. Soc.* **121,** 7949–7950.
55. Conrads, T. P., Anderson, G. A., Veenstra, T. D., Pasa-Tolic, L., and Smith, R. D. (2000). *Anal. Chem.* **72,** 3349–3354.
56. Jensen, P. K., Pasa-Tolic, L., Peden, K. K., Martinovic, S., Lipton, M. S., Anderson, G. A., Tolic, N., Wong, K. K., and Smith, R. D. (2000). *Electrophoresis* **21,** 1372–1380.
57. Han, D. K., Eng, J., Zhou, H., and Aebersold, R. (2001). *Nat. Biotechnol.* **19,** 946–951.

PROTEOMICS OF MEMBRANE PROTEINS

By JULIAN P. WHITELEGGE, STEPHEN M GÓMEZ, AND KYM F. FAULL

Pasarow Mass Spectrometry Laboratory, Department of Psychiatry and Biobehavioral Sciences, Department of Chemistry and Biochemistry, and the Neuropsychiatric Institute, University of California Los Angeles, Los Angeles, California 90095

I. Introduction

A. *Why Are Membrane Proteins Important?*

Within the genomes sequenced thus far, about 30–35% of all open reading frames (ORFs) are known or predicted to encode polytopic transmembrane proteins. Of these the majority (20–25% of all ORFs) belong to the helix bundle class of membrane proteins (von Heijne, 1999). These proteins catalyze a multitude of essential functions including oxidative phosphorylation and electron transport, translocation of molecules into and out of cells, and signal transduction across membranes. Many of these proteins are crucial for the fundamental processes of life, including, as an illustrative example, the balance of the global ecosystem, where their role in photosynthesis is essential for maintenance of the carbon cycle that underlies the looming global warming environmental crisis. Later in this article we discuss in some detail our work on the proteomics of the photosynthetic machinery from spinach, pea, and tobacco chloroplasts.

ADVANCES IN
PROTEIN CHEMISTRY, Vol. 65
0065-3233/03 $35.00

With regard to the human condition there are numerous examples in which the origin of a disease has been traced to a defective membrane protein. Ion channels represent an interesting case. These constitute a major class of gateways for ion transport across membranes. In essence they are water-filled pores that extend across the cellular membrane (Hille, 1992). Ions pass through ion channels at a rate that approximates diffusion following the electrochemical gradient. Thus ion channels are passive transporters, and this is in contrast to active transporters such as Na^+/K^+ pumps (ATPases) that utilize metabolic energy from ATP hydrolysis to move ions across the membrane against electrochemical gradients. Cystic fibrosis, the most common lethal genetic disease among Caucasians, afflicting about 1 of every 2000–2500 newborns in northern Europe and the United States (Welsh *et al.*, 1995), is a consequence of a defective transmembrane conductance regulator protein, a chloride ion channel (Sheppard and Welsh, 1999; Gadsby and Nairn, 1999).

Among the active ion transporters are the sodium cotransport proteins. These are a major class of integral membrane proteins responsible for the accumulation of ions and nutrients in cells. The energy for transport is derived from the sodium electrochemical potential gradient across the cell plasma membrane. There are more than 55 known members of the sodium/glucose cotransporter (SGLT1) gene family and examples include the *Escherichia coli* Na^+/proline, the *Vibrio* Na^+/galactose, and the human Na^+/glucose and Na^+/iodide cotransporters (Turk and Wright, 1997). The disease called glucose-galactose-malabsorption is caused by mutations in the gene encoding the human intestinal Na^+/glucose contransporter protein, and it is this protein that is responsible for the success of oral rehydration therapy (ORT) in the treatment of infectious diarrhea. In 1980 infectious diarrhea was the leading cause of global child mortality. ORT was introduced in 1979 and rapidly became the most successful treatment for diarrheal diseases. The World Health Organization reports that the annual death toll attributable to diarrhea among children under 5 years of age fell from the estimated 4.6 million in 1980 to about 1.5 million today, primarily as a consequence of the success of ORT (Victora *et al.*, 2000). The mechanisms of ORT have been studied in *Xenopus* oocyte expression systems, where the Na^+/glucose cotransporter was found to transport at least 250 water molecules for each sugar molecule transported (Loo *et al.*, 1996). Extrapolation of the results from the test systems suggested that in the human intestine the Na^+/glucose cotransporter accounts for 5 liters of water absorption per day. Other animal and plant cotransporters such as the $Na^+/Cl^-/\gamma$-aminobutyric acid, Na^+/iodide, and H^+/amino acid transporters are also able to transport water and this suggests that cotransporters play an important role in water homeostasis.

Membrane proteins are of particular relevance to those medical disciplines involved with function of the central nervous system (CNS). The CNS is rich in neurotransmitter receptors, ion channels, and transporters, all of which are membrane proteins. Defects in the functioning of these proteins are implicated in many neuropsychiatric diseases. Therefore it is not too surprising that a focus in CNS drug development is for agents that interact with these membrane proteins. Fluoexetine hydrochloride (Prozac), for example, a member of the so-called selective serotonin reuptake inhibitor (SSRI) drug class, is now one of the most widely prescribed drugs in the United States. This drug has experienced enormous popularity in the mental health field for its antidepressant, antiobsessive–compulsive, and antibulimic properties. The drug has also attracted widespread attention in the lay press for a variety of reasons including its possible side effects (Appleton, 2000). The efficacy of this drug is linked to its selective inhibition of the reuptake of serotonin from the synaptic cleft in the CNS (Stark *et al.*, 1985; Fuller and Wong, 1990; Wong *et al.*, 1995). This effect is thought to be mediated via effects on a variety of membrane-bound ion channels, including voltage-activated K^+ channels. Another widely prescribed drug is omeprazole (Prilosec or Losec), which has experienced popularity for the treatment of ulcers because of its selective effect on the suppression of gastric acid secretion. This effect is thought to be due to specific inhibition of the H^+/K^+ ATPase membrane protein enzyme system located on the secretory surface of the gastric parietal cell. Because this enzyme system is regarded as the acid (proton) pump within the gastric mucosa, omeprazole has been characterized as a gastric acid pump inhibitor in that it blocks the final step of acid production. The reported annual sales for Prozac and Prilosec of $2.2 billion (www.lilly.com) and $6.3 billion (www.astrazeneca-us.com) for the year 2000, respectively, reflect the commercial and clinical potential of agents that affect the functioning of important membrane proteins.

B. What Are Membrane Proteins?

Biological membranes are composed of lipid and protein molecules in approximately equal proportions based on mass. The lipid molecules are arranged in a bilayer with the fatty acyl chains making up a 25- to 30-Å-wide hydrophobic core that is flanked by the relatively polar head group regions measuring about 10–15 Å each. The proteins associated with biological membranes fall into two general categories. Some are associated only with the surface of the membrane (extrinsic or peripheral membrane proteins) whereas others reside with a significant proportion of their mass

within the membrane (intrinsic or integral membrane proteins, IMPs). The peripheral membrane proteins, although essential components of membranes, are actually water soluble with a high affinity for a specific component of the membrane surface. The IMPs on the other hand, have different physicochemical properties, particularly solubility characteristics. Those members of this class that span the lipid bilayer and present a portion of their sequence on both sides of the membrane provide a conduit for communication between the interior and exterior of the compartment bounded by the membrane. The compartment can be a cell, in which case the membrane is the cellular membrane, or an organelle within a cell. Two general secondary structures characterize nearly all IMPs: transmembrane α helices and transmembrane β barrels. A unique feature of α-helical IMPs is the presence of one or several stretches of contiguous apolar amino acid residues. The environment created by the stretches of apolar amino acids thermodynamically favors the retention of these stretches within the nonpolar membrane lipid bilayer (Eisenberg, 1984). The hydrophobic segments, which are in many cases 20–25 amino acid residues long, impart characteristics to these proteins that have made their analysis difficult. The transmembrane β barrels of the porins present alternate apolar amino acid residues in the transmembrane portion of the primary sequence.

The factors that contribute to the topological character and the structure and stability of membrane proteins have been extensively reviewed elsewhere (Popot and Engelman, 2000; von Heinje, 1999; White and Wimley, 1999; Stowell and Rees, 1995; Haltia and Freire, 1995). Various types of hydropathy profile and hydrophobic moment analyses are available for predicting the number of transmembrane regions within amino acid sequences (Eisenberg, 1984). Two prominent structural motifs that have emerged from these analyses are proteins that contain 7 transmembrane-spanning segments, exemplified by bacteriorhodopsin and many mammalian G protein-coupled receptors, and proteins containing 12 transmembrane segments, as exemplified by many mammalian transporters. However, hydrophobicity analysis of bacterial, archaeal, and eukaryotic genomes found no obvious domination of the encoding of any particular number of transmembrane-spanning regions (Arkin *et al.*, 1997), suggesting there are many major groups of membrane proteins still to be characterized.

C. Why Are Membrane Proteins Difficult to Study?

IMPs are amphipathic in nature, often with several long stretches of apolar amino acids dispersed along their sequence. The result is a molecule with split personalities. Parts of the molecule are water soluble

and parts are not water soluble. Such molecules preferentially reside at the interface between polar and nonpolar environments. Although advantage can be taken of this property for sample preparation before mass spectrometry (see later), a consequence of this characteristic is that these proteins are difficult to crystallize or to study by other methods aimed at high-resolution structure determination. Only a handful of membrane proteins have been crystallized in a form that allows structure determination at atomic resolution, which is essential for understanding their mechanism of action (Kühlbrandt *et al.*, 1994; Picot *et al.*, 1994). Most topological information about these proteins has been derived from investigations employing site-directed mutagenesis and biochemical assays combined with sequence analysis and structure prediction. Furthermore, many membrane proteins require conformational flexibility in order to function, making it imperative to obtain dynamic information to fully understand their function. As an increasing number of genomes are sequenced and a growing number of membrane proteins are identified, the gap in knowledge about the structure and function of these proteins will likely increase. Moreover, in the postgenomic era, with proteomics emerging as a new field directed at the structural and functional assessment of the entire complement of proteins in a given organism, the limitations in working with large membrane proteins will present a formidable challenge.

D. Fast Atom Bombardment

In principle, fast atom bombardment (FAB) ionization (Barber *et al.*, 1981, 1982) could be useful for the analysis of membrane peptides and proteins because compounds that accumulate at the surface of the matrix/sample mixture are preferentially ionized by this technique. Because of their hydrophobicity and propensity for interface environments, membrane peptides and proteins would be expected to behave in this manner when the commonly used matrices of glycerol, thioglycerol, and *m*-nitrobenzyl alcohol are employed. However, the rapidly diminishing ion currents recorded with increasing mass from FAB sources (Green and Bordoli, 1986) have meant that there are relatively few examples in which this ionization technique has been successfully used above 6000 Da. This limitation in the practical mass range has restricted the application of FAB to peptide and protein analysis. The use of FAB for the analysis of the amino-terminal tryptic fragments derived from D1, D2, and CPa-2 proteins from photosystem II of spinach (Michel *et al.*, 1988) represented an application at the modest mass range of typical tryptic fragments. More

significant in this regard was the application of FAB to the measurement of the full-length membrane protein cytochrome *c* oxidase subunits VIIIa and VIIIb at 5438 and 4962 Da, respectively, and the ATP synthase subunit A6L at 7956 Da (Boyot *et al.*, 1988). This application revealed the production of stable and long-lived ion currents from purified hydrophobic proteins dissolved in aqueous dimethyl sulfoxide (DMSO) with thioglycerol containing 1% trifluoroacetic acid (TFA) as matrix. The reported quantities of these proteolipid-derived protein subunits used for the measurements of 1–2 nmol generally reflect the quantities of material often required for FAB, which is the other major limitation of this technique, presumably a consequence of the low ionization efficiency of the FAB source. Limited success was achieved when plasma desorption was used for analysis of lung surfactant proteins (Curstedt *et al.*, 1990).

II. Techniques

A. *Two-Dimensional Tricks for Intrinsic Membrane Proteins*

Although two-dimensional (2D) gel separations represent the most common technique used in proteomics, it has been known for some years that IMPs are poorly represented in such analyses primarily because of solubility problems within the primary isoelectric focusing (IEF) separation. Considerable efforts have been directed toward this problem and undoubtedly progress has been made (for reviews, see Molloy, 2000; Santoni *et al.*, 2000), although it is generally agreed that gel analysis excludes some membrane proteins either partially or completely. It was reported that some classes of Golgi proteins were never detected in 2D gel experiments but have been identified through chromatographic separations (K. Howell, personal communication; and see Lin *et al.*, 2001). Experiments that compare membrane protein recovery in 2D gel systems versus alternative techniques will be revealing.

B. *Peptide Mass Tags and Sequence Tags*

Most IMPs have loop regions that normally reside outside of the bilayer, providing a source of hydrophilic peptides suitable for traditional extraction and mass/sequence tag analysis, although sequence coverage may be restricted. Thus identification may be achieved while sacrificing the ability to detect posttranslational modifications to hydrophobic transmembrane segments. Recovery of peptides derived from transmembrane regions

from gels is unlikely, and alternatives to 2D gel analysis are recommended for rigorous analysis of membrane protein primary structure. Treatment of IMPs with proteases can be problematic because partially digested peptides often come out of solution and thus chemical cleavage is attractive. Cyanogen bromide (CNBr) treatment in high concentrations of formic acid is ideal from the point of view of solubility, although variable formylation results in the appearance of +28-Da adducts. Formic acid eliminates methionine oxidation during CNBr treatment (Joppich-Kuhn *et al.*, 1982) and preserves disulfide bonds. It was previously reported that formic acid treatment leads to disulfide cleavage (Villa *et al.*, 1989) but these authors relied on changes in retention time that are probably due to formylation. Repetition and expansion of these experiments, using mass spectrometry to more fully characterize the products, provided no evidence for disulfide cleavage by formic acid (Tjon *et al.*, 2000). If unformylated peptides are required, high concentrations of acetic acid may be substituted. Hydrophobic peptides derived from the transmembrane regions of membrane proteins may be separated by chromatographic methods described later (Section II.C.3). In addition, nonspecific cleavage is an approach that has been used with success; limited *in gel* acid hydrolysis with 0.1 *N* HCl allowed Shevchenko and co-workers to recover sufficient peptides from bacteriorhodopsin for identification (Shevchenko *et al.*, 2000).

An important goal in proteomics is to monitor the plasticity of posttranslational modifications during physiological changes. Because fragmentation raises issues concerning peptide recovery and thus the ability to monitor the full complement of modifications, it is desirable to include a mass spectrum of an intact protein within the framework of a proteomic experiment (Whitelegge *et al.*, 1998a). Thus the first methionine oxidation to afflict the structure of a protein can be conveniently monitored, for example, before localization of such a modification (Whitelegge *et al.*, 2000a). So-called top-down proteomics as described by McLafferty and co-workers (Kelleher *et al.*, 1999) embraces a comparable vision through the application of Fourier transform mass spectrometry (FT-MS). Smith and co-workers combined capillary IEF with FT-MS, demonstrating the power of such a combination for intact mass proteomics (Jensen *et al.*, 1999).

A useful approach to analyzing membrane, and other, proteomes including characterization of posttranslational modifications has been termed LC-MS+ (Whitelegge *et al.*, 1999b; Zhang *et al.*, 2001) (Section II.C.5). Typically, the proteins in isolated membranes are precipitated with acetone and resuspended in formic acid before separation by high-performance liquid chromatography (HPLC). A flow splitter inserted between the HPLC and electrospray ionization (ESI)-MS allows collection

of fractions concomitant with the collection of intact mass data (described in the next section). Each fraction is cleaved with either CNBr or trypsin to obtain peptides for subsequent generation of peptide mass or sequence tags and for identification of posttranslational modifications, usually by tandem mass spectrometry (MS/MS). CNBr is especially suited to intrinsic membrane proteins because cleavage can be performed in high concentrations of formic or acetic acid, thereby retaining the solubility of the proteins/peptides. Since the report that trypsin can cleave effectively at high organic solvent concentrations (Russell *et al.*, 2001) we have been investigating its suitability for generation of peptides from LC-MS+ fractions after titration to pH 8.0.

C. Intact Proteins: Intact Mass Tags

1. Sample Purification

As a general rule it is preferable to maintain a protein in the native conformation during purification. Taking advantage of native structure minimizes opportunities for undesirable chemistry, such as isoaspartate formation, and helps to avoid irreversible aggregation. Thus sample purification falls into two general categories; membrane subfractionation, in which membrane proteins remain surrounded by native bilayer lipids, and membrane solubilization, in which membrane proteins are extracted into detergent micelles. The specific methods used for individual membrane systems vary widely and a full discussion is beyond the scope of this review. The level of purity required for individual proteins depends on the nature of the downstream separation system to be employed [e.g., reversed-phase (RP) chromatography versus size-exclusion chromatography (SEC)] and the complexity of mass spectra that can be tolerated. Unfortunately, only limited sample heterogeneity can be handled by either ESI or matrix-assisted laser desorption ionization (MALDI), demanding a clear need for subfractionation. Moreover, it is becoming increasingly clear that minor subpopulations with functional significance may be observed only after increased purification (Whitelegge *et al.*, 1999a; Gómez *et al.*, 2002).

2. Sample Preparation

The techniques used for sample preparation depend largely on the identity and quantity of detergent present. If large amounts of a molecularly homogeneous detergent, such as 3-[(3-cholamidopropyl)

dimethylammonio]-1-propanesulfonate (CHAPS), or smaller amounts of a molecularly heterogeneous detergent, such as Triton X-100, are present, attempts to perform LC-MS will be unsuccessful for three reasons. First, chromatographic resolution will be dramatically reduced; second, the sample may not adsorb to the stationary phase; third, the chances of eluting the protein free of detergent will approach zero and small molecule ions will dominate the mass spectrum to the point at which protein ions become invisible. When most or all of the detergent must be removed before chromatography a number of strategies are available. Nondenaturing techniques such as dialysis or ultrafiltration may be effective for lowering detergent concentrations; however, as the detergent concentration is lowered membrane proteins have a tendency to precipitate and associate with the dialysis membranes, leading to loss of sample. Centrifugal ultrafiltration devices that move filtrate against the centrifugal field are attractive; as proteins come out of solution they are immediately moved away from the dialysis/filtration membrane. Although precipitation may lead to loss of activity and create problems for resuspension, it is feasible to recover such protein for analysis. However, despite removing sufficient detergent to precipitate the protein (i.e., below the critical micellar concentration) it is common to find significant amounts of detergent still remaining associated with the protein. Thus precipitation with organic solvents becomes attractive for MS analysis of the sample, although some covalent modifications occasionally result.

The identification of the optimal sample preparation method for the MS analysis of an IMP is illustrated by the analysis of bacteriorhodopsin. Bacteriorhodopsin is a seven-transmembrane IMP that binds a retinal cofactor via a Schiff base. The retinal cofactor plays a central role in the light-driven proton-pumping function of the molecule in the native purple membrane. Our earliest protocols used acetone precipitation to strip the protein of detergent and subsequent LC-MS analysis revealed two prominent mass measurements: one consistent with the apoprotein (26,784.2 Da) and a heavier component that agreed with the calculated average mass of the holoprotein (267 Da heavier). This result showed that acetone precipitation either resulted in the hydrolysis of the Schiff base or converted the protein to a state in which the Schiff base is more easily hydrolyzed during chromatography. To avoid hydrolysis of the Schiff base during sample preparation, purple membrane preparations were first treated with 1 m*M* CHAPS to disperse the membranes (the preparation is not fully solubilized at this point and centrifugation rapidly pellets the membranes) and then solubilized by addition of three volumes of formic acid–isopropanol (1:1, v/v) immediately before LC-MS analysis in the formic acid–isopropanol system (Section II.C.3). A reversible

semidenatured state of bacteriorhodopsin induced by addition of formic acid and isopropanol was described in the 1970s (Oesterhelt *et al.*, 1973). Mass measurements indicate almost complete recovery of the holoprotein with intact Schiff base-linked retinal (27,052.0 versus 27,050.06 Da calculated average mass). Clearly, in this example maintenance of the native covalent state of the protein was favored by the use of a small amount of detergent and the avoidance of acetone precipitation (see full discussion in Whitelegge *et al.*, 1998a). CHAPS is attractive in this respect because in contrast to most other detergents it is molecularly homogeneous and elutes at a defined retention time. If the same sample preparation procedure is combined with the (SEC) method described (Section II.C.3) most protein is recovered as the apoprotein, indicating that the chromatographic technique also plays a role in stabilizing the native covalent state of bacteriorhodopsin. Although bacteriorhodopsin is undoubtedly a special example, and the covalent structure of most proteins is not altered by organic solvent precipitation, this does provide a good example of the special treatment often required for IMPs.

When moderate concentrations of detergent inhibit precipitation in 80% acetone, chloroform–methanol precipitation (Wessel and Flugge, 1984; Whitelegge *et al.*, 1999b) should be tried as an alternative. Furthermore, the later procedure removes tightly bound lipids that may interfere with mass spectrometric analysis (J. P. Whitelegge and D. R. Blanco, unpublished data). In some cases displacement of tightly bound lipids is enhanced through detergent treatments before precipitation. Samples precipitated with organic solvents are dried for 2 min at atmospheric pressure and immediately solubilized, typically in 60% formic acid, although some proteins such as the lactose permease (12 transmembrane helices) of *E. coli* require 90% formic acid or TFA for solubilization (Whitelegge *et al.*, 1999b). Complete drying of the sample can lead to irreversible aggregation. When samples are analyzed directly from solutions with low detergent concentrations, they are typically acidified with 2 volumes of 60% formic acid immediately before LC-MS to ensure compatibility with the mobile phase.

The most hydrophobic membrane proteins, the proteolipids, behave more like lipids in that they partition into the chloroform phase of an aqueous chloroform–methanol phase separation. Although this can be a useful procedure for sample enrichment before LC-MS, problems can arise if a large amount of lipid and detergent has also entered this phase. Ether can be used to precipitate the proteolipids (Findlay, 1987) although it is not presently clear how effective this technique is for removing lipids and detergents.

3. *Chromatographic Separations*

The basis for our currently preferred chromatographic separations of membrane proteins in aqueous/organic solvent mixtures dates back many years. Khorana *et al.* used elevated concentrations (5%, v/v) of formic acid to perform RP separations of peptides derived from bacteriorhodopsin for sequence and MS analysis that revealed a pyroglutamate modification at the N terminus (Gerber *et al.*, 1979). Findlay and co-workers used SEC with LH60 resin and mobile phases of formic acid–ethanol and formic acid–acetic acid–chloroform–ethanol for separation of peptides derived from IMPs, such as rhodopsin, for Edman sequence analysis (Findlay *et al.*, 1981; Brett and Findlay, 1983; Pappin and Findlay, 1984). These solvent systems allowed solubilization and disaggregation of the most hydrophobic peptides, allowing separations of the monomeric molecules. The use of volatile organic acids, organic solvents, and low water content gave compatibility with Edman chemistry and, of course, MS analysis, although it was some years before the latter was realized. Many other chromatographic separations of membrane proteins have been described, typically incorporating mild detergents and salt to preserve function, which although useful for enrichment and purification are not directly compatible with mass spectrometry. Examples of both SEC and ion-exchange chromatography abound, providing a paradigm for simplification of the complex mixtures of membrane proteins in crude membrane extracts of cells. Once samples of limited complexity are prepared the strategy is dependent on the method of ionization to be employed.

a. Electrospray and Matrix-Assisted Laser Desorption Ionization. Peptide and protein analysis by mass spectrometry is now virtually the exclusive domain of ESI and MALDI. The older ionization techniques of FAB and plasma desorption (Macfarlane and Torgerson, 1976; Macfarlane, 1990) have been rendered obsolete, principally by the lower limits of detection and extraordinary mass range that can be obtained by ESI and MALDI. On the basis of our experience with membrane proteins isolated from bacterial, plant, and animal sources, we present an overview of the most successful methods currently available for sample preparation and purification of IMPs, and complete this review with a discussion of our preliminary analysis of the thylakoid membrane proteome.

ESI has been our ionization method of choice for intact proteins because of superior mass accuracy and resolution over MALDI for intact proteins greater than 15 kDa and because of the ability for direct HPLC coupling (LC-MS). Mass accuracy of small peptides from MALDI instruments is now comparable to, or exceeding, that obtained from

ESI-quadrupole or ion trap instruments. However, MALDI mass measurements of larger intact proteins typically achieve inferior accuracies (0.1% of the molecular weight), at least one order of magnitude worse than ESI. Nevertheless, there are times when MALDI is necessary, particularly when protein heterogeneity precludes ESI. In these cases purified membrane subfractions are either directly spotted with matrix solution or precipitated (Section II.C.2), dissolved in 60% formic acid, and then spotted with matrix solution. Sinapinic acid (20 mg/ml in 0.1% TFA–70% acetonitrile) or dihydroxybenzoic acid (DHB, 20 mg/ml in formic acid–isopropanol, 1:1) are useful matrix solutions, with the latter especially useful where samples are resistant to dissolution. The rare outer membrane protein 1 (Tromp1) from *Treponema pallidum* provides an interesting example of the analysis of a membrane protein by MALDI. After isolation by 2D gel electrophoresis, the protein was transferred to a polyvinylidene difluoride (PVDF) membrane, and eluted with 2% sodium dodecyl salt (SDS) containing 1% reduced Triton X-100 and 50 m*M* Tris, pH 9.0. After chloroform–methanol precipitation, the protein was dissolved in 4 μl of 60% formic acid and the MALDI mass spectrum was recorded with a sinapinic acid matrix (0.3 μl of sample with 0.5 μl of sinapinic acid [10 mg/ml in 0.1% TFA–70% acetonitrile]). An intense signal for the protonated molecule was seen in the MALDI-MS spectrum (Fig. 1). The measured molecular mass was more than 2 kDa, less than that calculated for the translated gene, indicating the likelihood of proteolytic trimming. Peptide mass fingerprinting provided data supporting N-terminal modification and outer membrane localization (Blanco *et al.*, 1999). Other matrix solutions reported to be useful for membrane proteins include 50 m*M* sinapinic acid in 70% formic acid (Schey *et al.*, 1992; Sharma *et al.*, 1997a), DHB mixed with succinic acid for ultraviolet (UV) MALDI or urea for infrared (IR) MALDI, as well as ferulic acid or nicotinic acid in tetrahydrofuran (Rosinke *et al.*, 1995) and α-cyano-4-hydroxycinnamic acid (Cadene and Chait, 2000). The ultrathin layer technique described by Cadene and Chait resulted in excellent signal to noise and is applicable to a wide range of IMPs.

The first spectra demonstrating the potential application of ESI to IMPs were published in 1993. In these studies, bacterio-opsin (bacteriorhodopsin apoprotein) was either recovered from an SDS gel (le Maire *et al.*, 1993) or bleached and dissolved in hexafluoroisopropanol (Schindler *et al.*, 1993). Acidified aqueous chloroform–methanol solutions have also been used for the analysis of inner mitochondrial membrane proteins (Fearnley and Walker, 1996) and bacterio-opsin (Hufnagel *et al.*, 1996). Undiluted formic acid has also been used as a solvent for the ESI-MS analysis of mannose transporter subunits and bacterio-opsin (Schaller *et al.*,

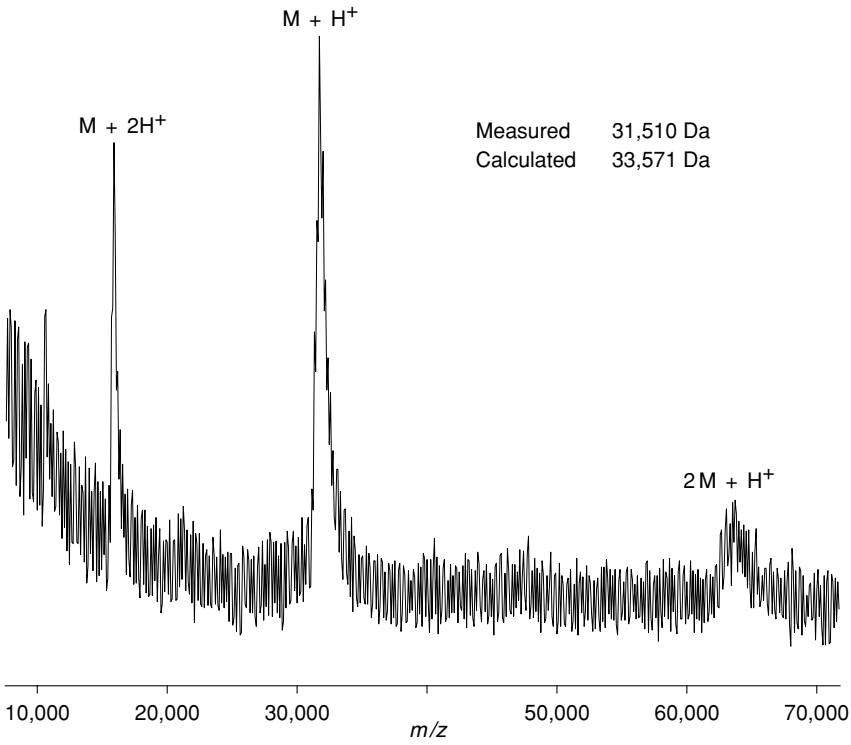


FIG. 1. Membrane proteins are amenable to MALDI: *Treponema* rare outer membrane protein 1 (Tromp1). Native Tromp1 was isolated from *Treponema pallidum* outer membrane preparations by 2D gel electrophoresis and the intact protein was recovered after transfer to PVDF membrane (Blanco *et al.*, 1999). Precipitated protein (1–2 μg) was dissolved in 2 μl of 60% formic acid and 0.3 μl was mixed with 0.5 μl of sinapinic acid matrix solution (10 mg/ml in 0.1% TFA–70% acetonitrile) on the sample plate. The mass spectrum revealed Tromp1 to be more than 2 kDa less than the molecular mass calculated from the gene sequence, presumably as a result of proteolytic trimming. An alternative matrix solution for difficult-to-solubilize membrane proteins is DHB (20 mg/ml) in formic acid–isopropanol (1:1, v/v).

1997). An alternative approach, in which detergent-solubilized bacteriorhodopsin was extracted into a nonpolar solvent phase by adding a chloroform–methanol–water solvent mixture to the aqueous detergent solution, was used to analyze bacterio-opsin by ESI-MS. In this method, an electrophoretic separation of the protein from the lipids and detergent occurs within the fused silica capillary that delivers the nonpolar, chloroform-rich phase to the ESI source, preventing suppression of

protein–ion formation by the lipid and detergent components (Barnidge *et al.*, 1999). In a variant of RP chromatography, photosynthetic reaction-center proteins were purified in 5% acetic acid with a propan-1-ol gradient before flow injection analysis (Sharma *et al.*, 1997b). Clearly, a variety of approaches have yielded success although the accuracy and resolution of some of the resulting spectra are not always comparable to those of globular proteins.

Whereas the above described studies typically rely on aqueous–acidic organic solvent mixtures, a more recent study showed the efficacy of analyzing membrane-bound peptides incorporated in large unilamellar vesicles of 1,2-dimyristoyl-*sn*-glycero-3-phosphocholine by ESI-MS. In this study, the hydrogen/deuterium (H/D) exchange properties of synthetic transmembrane peptides of varying length and composition were measured by ESI-MS. The peptide ions were resolved from the lipids by virtue of their *m/z* differences (Demmers *et al.*, 2000, 2001a). The ability to deliver the peptides to the ESI source allows H/D exchange experiments to be performed on the membrane bound peptides *in situ*. Attempts to perform similar experiments with intact transmembrane proteins have so far yielded spectra only after addition of 50% trifluoroethanol to the sample, which disrupts the membrane and potentially denatures the protein, and any existing noncovalent interactions (Demmers *et al.*, 2001b).

b. Reversed Phase High-Performance Liquid Chromatography. A number of alternatives for RP-HPLC of intact IMPs exist but, unfortunately, finding the most suitable protocols for a particular protein or mixture must generally be determined empirically. Some IMPs elute from RP stationary phases under standard conditions; however, these proteins generally have molecular weights of less than 30,000. The major intrinsic protein from bovine lens fibers (MIP or Aqp0), has six transmembrane helices, and elutes efficiently under these conditions (Fig. 2). Bacteriorhodopsin, however, which has seven transmembrane helices, does not. Many thylakoid membrane proteins also elute efficiently under these conditions, allowing high-performance separations (see Section III), although others remain column bound. If a protein does not elute under these conditions, however, it may be conveniently stripped of lipids and detergents before elution with a secondary gradient under conditions that promote elution, such as the aqueous formic acid–isopropanol system described later. Many proteins exhibit intermediate elution characteristics such that a second 0.1% TFA–acetonitrile gradient will elute ghost peaks, complicating MS analysis. As mentioned above, the primary gradient can be advantageously used to remove small intrinsic proteins, lipids, and detergents, providing purification and enrichment of the retained, typically larger, proteins.

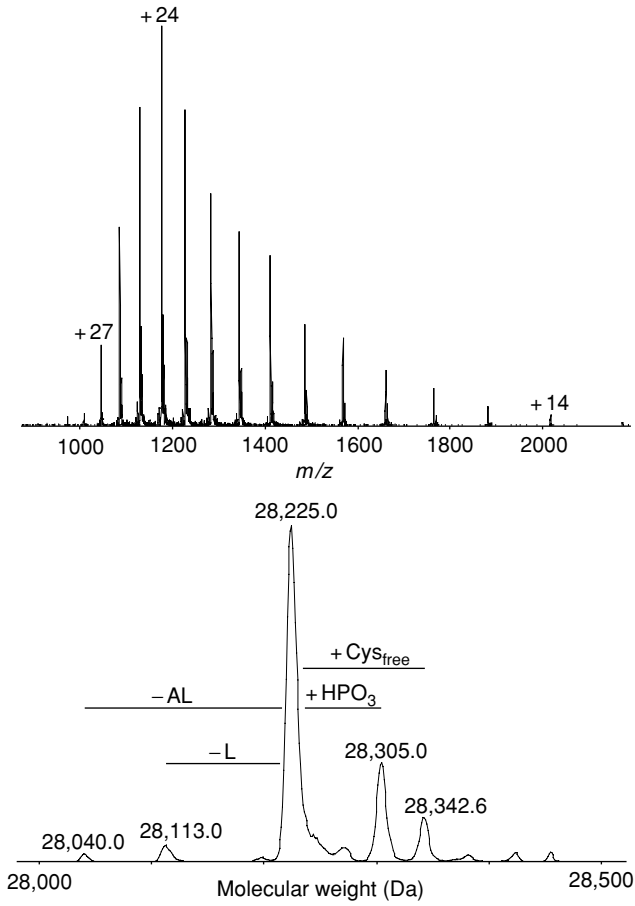


FIG. 2. Electrospray-ionization mass spectrometry of the bovine lens major intrinsic protein (MIP) after RP chromatography. *Top*: Mass spectrum. *Bottom*: Reconstructed molecular weight spectrum. MIP is unusual for two reasons. First, it elutes from a polymeric RP column with high efficiency in a standard 0.1% TFA–aqueous acetonitrile system; many IMPs fail to elute or elute with low efficiency under these conditions. Second, the measured MIP molecular mass matches that calculated without any adjustments (28,225.2 Da observed, 28,223.1 Da calculated); an unusual observation for eukaryotic proteins, which tend to be modified to various extents. The purified protein was loaded onto the RP column (PLRP/S, 300 Å, 5 μm, 2 $\times$ 150 mm: in Polymer Laboratories detergent and eluted with different retention to the detergent directly to the Ionspray source of a Sciex API III+ triple quadrupole mass spectrometer. Because many IMPs elute with poor efficiency in standard aqueous TFA–acetonitrile systems, a number of alternatives have been developed (see Section II.C.3).

Several different stationary phases have been investigated for separations of IMPs. A CN-bonded phase was reported to be more suitable for IMPs (Tarr and Crabb, 1983) whereas others have enjoyed success with C_4 or C_{18} aliphatic-bonded phases (Sharma *et al.*, 1997a,b). Our experience is that such columns tend to accumulate noneluted proteins and fail to regenerate efficiently such that resolution is soon lost before complete failure. Polymeric stationary phases may be rigorously regenerated and represent the columns of choice for IMP work. Several manufacturers provide these polystyrene–divinylbenzene copolymer phases and PLRP/S (Polymer Laboratories, Amherst, MA) is favored, whereas the PRP products (Hamilton, Reno, NV) provide less expensive alternatives. Lifetimes of these columns can be good provided the manufacturers' instructions concerning the use of entirely aqueous and especially entirely organic mobile phases are followed closely. Resolution declines with age and virgin columns provide ultimate performance. Polymeric stationary phase columns may also be heated (maximum, 80°C), allowing separations to be conducted over a wider range of temperatures than possible with silica-based resins.

Proteins that prove resistant to elution in standard TFA–acetonitrile gradients can often be eluted by the addition of isopropanol to the acetonitrile buffer (25–50%). This was first reported some years ago for RP chromatography of rhodopsin (Tarr and Crabb, 1983), and elution efficiency of bacteriorhodopsin rises from close to zero in the standard TFA–acetonitrile system to about 50% with addition of isopropanol to the elution buffer (i.e., 0.05% TFA–50% acetonitrile–50% isopropanol). These isopropanol-containing systems provide perhaps the ultimate in chromatographic resolution for IMPs although larger sample quantities must often be loaded to achieve acceptable signal to noise in LC-MS. Although we have no evidence to support differential elution efficiencies of different isoforms of a particular protein, this potential should be appreciated. The aqueous formic acid–isopropanol system should ideally be used to provide further evidence for relative abundances of isoforms.

Maximal elution efficiency for most membrane proteins is achieved by the aqueous formic acid–isopropanol system described (Whitelegge *et al.*, 1998a). This technique was developed for separation of hydrophobic tryptic peptides derived from IMPs for Edman sequence analysis (Whitelegge *et al.*, 1992) and combined the benefits of a heated polymeric column (40°C) with a 60% aqueous formic acid–isopropanol mobile phase first reported for elution of a membrane protein from a C_{18} column (Heukeshoven and Dernick, 1982; Wildner *et al.*, 1987). As described above, the aqueous formic acid–isopropanol system can be used as a secondary column eluant to recover membrane proteins retained

following application of the standard aqueous–acetonitrile buffer system. Alternatively, the system may be used for the primary separation. In such a case the column (PLRP/S, 300 Å, 2 × 150 mm, 100 μl/min, 40°C) is equilibrated in 95% solvent A (solvent A, 60% formic acid), 5% solvent B (solvent B, isopropanol) before loading of sample. Injection initiates a gradient through 100% solvent B over 40 min with eluent being passed through a UV (280 nm) detector before passage to the ESI source. Note that the time spent in 60% formic acid is minimized to avoid potential covalent formylation of the sample. Full safety precautions should be observed because of the dangers associated with pressurized concentrated acid. Line splitting between the UV and MS detectors allows fraction collection for downstream experiments (LC-MS+; Section II.C.5).

The lactose permease protein of *E. coli* is an example of an IMP that does not elute in the 60% aqueous formic acid–isopropanol solvent system. After chloroform–methanol precipitation it is necessary to solubilize the 12-transmembrane permease in 90% formic acid (undiluted TFA was also successful) and only then will it adsorb to the stationary phase that has been equilibrated in 90% formic acid in the absence of isopropanol. The protein elutes efficiently with a gradient of increasing isopropanol, allowing the protein mass spectrum to be recorded by ESI-MS. However, the high concentration of formic acid leads to multiple sample formylations, four or five being the most predominant, limiting the value of the result. Thus, an alternative technique was developed for satisfactory analysis of the permease.

c. Size-Exclusion High-Performance Liquid Chromatography. The described HPLC-based technique for SEC of hydrophobic peptides suitable for Edman sequencing (Whitelegge *et al.*, 1992) was largely based on the previous descriptions of SEC on Sephadex LH60, using formic acid–ethanol and formic acid–acetic acid–chloroform–ethanol solvents (Findlay, 1987). The deactivated silica TSK SW beads from TosoHaas (Montgomeryville, PA) provided excellent performance in SEC and are compatible with aqueous–organic solvent mixtures. High concentrations of formic acid are not compatible with SEC LC-MS, however, because of the abundant contaminant ions in the formic acid that give a large background for this solvent system. In RP chromatography, these ions are highly retained by reversed-phase matrices, so they do not interfere in the part of the chromatogram where proteins elute. Fortunately, a chloroform–methanol–1% aqueous formic acid (4:4:1, v/v/v) solvent system had been reported to be compatible with ESI of IMPs (Fearnley and Walker, 1996). This was used with a TosoHaas resin (G2000SW) for SEC-MS of the lactose permease. It was some surprise that the permease did not

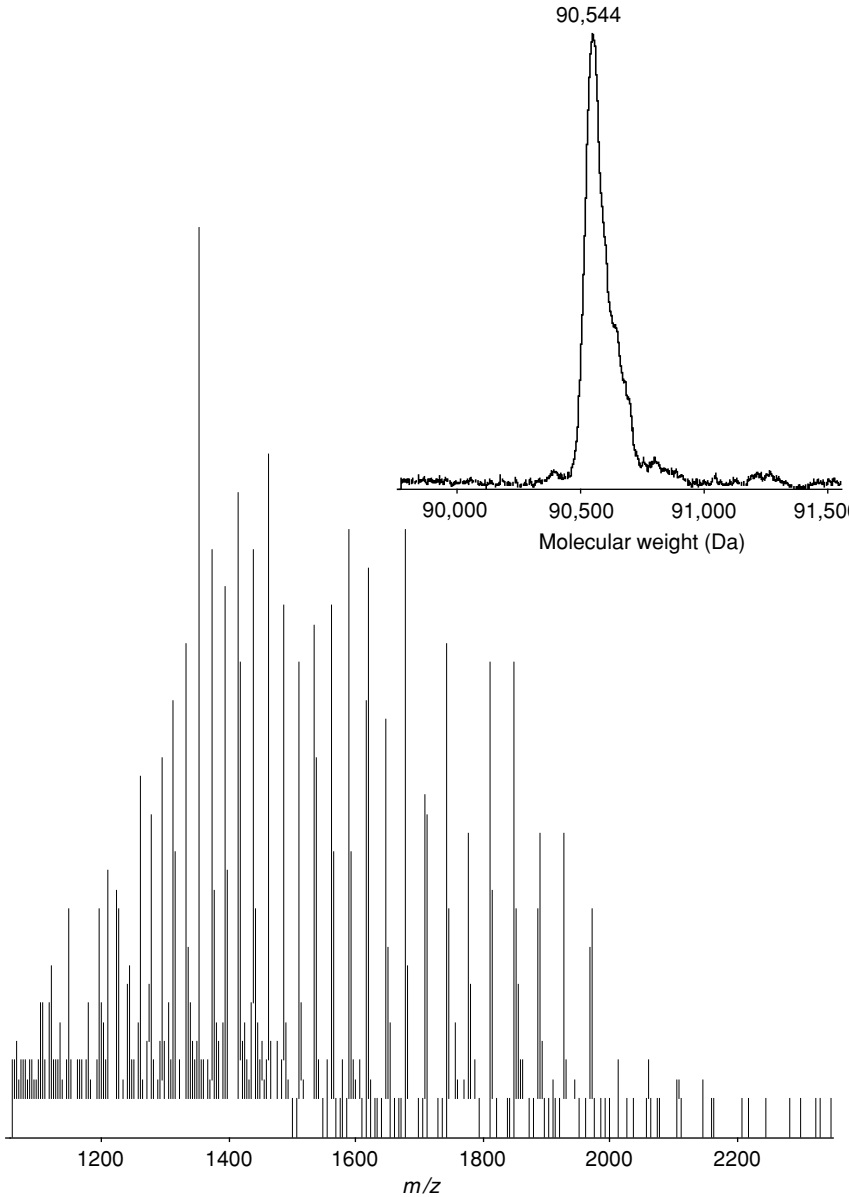


FIG. 3. Size-exclusion chromatography mass spectrometry (SEC-MS) provides a versatile system for analysis of IMPs. Studies with the lactose permease of *E. coli* demonstrated that the RP systems described (Section II.C.3) are not always adequate for larger polytopic membrane proteins. Here a 15-transmembrane helix sodium/galactose

precipitate on contact with the running solvent, instead transferring seamlessly to the mobile phase. The 47-kDa permease eluted close to the void volume of the G2000SW and was separated from numerous small molecule contaminants that had coprecipitated with the protein and introduced in the 90% formic acid used to solubilize the protein. Ionization efficiency from the aqueous–organic running solvent was good and the permease gave multiply charged ions ranging from more than 50 to fewer than 20 protons (this was the lowest detectable charge state on the Sciex API III, which can scan to 2400 m/z) (Whitelegge *et al.*, 1999b). A number of other proteins have since been shown to be compatible with this SEC-MS system including the potassium channel KscA (le Coutre *et al.*, 2000) and the 14-transmembrane domain Na^+/glucose cotransporter (Turk *et al.*, 2000), as well as many of the IMPs previously analyzed by RP chromatography. Latest protocols incorporate the Super SW2000 (4.6 × 300 mm; TosoHaas) that can be eluted at flow rates of 250 μl/min or lower, providing improved sensitivity (~1 μg of protein can often be detected with minimal inflection of the UV trace). The running temperature is elevated to 40°C in order to lower back pressure resulting from solvent viscosity. Many other size-exclusion columns have been tried, including some with specifications claiming equivalent performance to TosoHaas, but none of those investigated provided comparable performance, suggesting that the chemical deactivation procedure used by TosoHaas is superior to that used by other manufacturers. Figure 3 shows the mass spectrum of a chimera of the *Vibrio* Na^+/galactose cotransporter and green fluorescent protein recorded by SEC-MS. The measured mass of 90,544 Da is within 0.01% of the mass calculated for the protein either before chromophore assembly or after its hydrolysis. Provided samples are of limited heterogeneity; proteins of up to and exceeding 100-kDa can be profiled by ESI-MS with no loss of accuracy.

The lower resolution of SEC separations, compared with RP, limits the complexity of the samples that can be analyzed by SEC. The cytochrome b_6/f complex from thylakoid membrane (Zhang *et al.*, 2001; Whitelegge *et al.*, 2003) represents a good example of the use of SEC separations

cotransporter from *Vibrio parahaemoliticus*, in this case a chimera with green fluorescent protein, was analyzed by the SEC-MS system, revealing the covalent status of the protein. *Top*: Mass spectrum. *Bottom*: Reconstructed molecular weight spectrum. The measured mass of 90,544 Da suggests that either the protein had not yet constructed the fluorophore (90,544.6 Da, calculated) or if constructed it was subsequently hydrolyzed during sample preparation (90,542.5 Da, calculated). After fluorophore formation the mass is calculated to be 90,524.5 Da.

combined with MS analysis. Five larger subunits that coeluted (17–35 kDa) could be resolved only by their mass spectra. Unfortunately, the ionization efficiency of cytochrome *b* was considerably less than that of cytochrome *f*, necessitating the need for further experiments to confirm the mass of the most hydrophobic subunit. Four small subunits (3–8 kDa) were chromatographically resolved from the larger ones and could be resolved in the mass spectrum. Increased sample complexity would render the mass spectra of this sample too complex for deconvolution and the minor species would most likely be undetectable. For more complex mixtures, SEC separations are effective only if the Stoke's radii of the components vary over a useful range. Improvements in chromatography by using longer and/or multidimensional SEC separations are providing significant benefits in this respect.

d. Intact Proteins from Gels. The popularity of SDS–polyacrylamide gel electrophoresis (PAGE) and 2D gels has generated a demand for techniques that couple the excellent separation power of these methods with MS analysis of the intact proteins. Despite the potential for methionine oxidation and cysteine adduction with acrylamide, early results using electroelution to recover bacterio-opsin from the gel followed by ESI-MS analysis were promising (le Maire *et al.*, 1993). These and other techniques for direct elution from gels (Feick and Shiozawa, 1990), however, require relatively large amounts of protein for success and more sensitive procedures are desirable. An alternative approach involves electrophoretic transfer of proteins from the gel to a PVDF membrane followed by elution of the protein from the membrane with detergents followed by precipitation. This strategy typically allows 1–2 μg of the intact protein to be recovered. A low-abundance outer membrane protein from *Treponema pallidum* (Blanco *et al.*, 1999) and carbonic anhydrase were analyzed by ESI-MS after recovery from gels, using this technique (Whitelegge *et al.*, 2000b), although some proteins retained detergent after precipitation. Infrared multiphoton dissociation (IRMPD) in combination with an FT-ICR (ion cyclotron resonance) mass spectrometer has been used by McLafferty and co-workers to remove these SDS adducts (Fridriksson *et al.*, 1999) whereas the SEC-MS system (Section II.C.3.) has proved successful for removal of these adducts before the MS analysis (Fig. 4; Schroda *et al.*, 2001). A recurring observation is that intact proteins from gels have more covalent adducts than do proteins that have not been exposed to gel systems. Consequently the techniques described in detail here are the LC-MS protocols that result in spectra most closely allied to the *in vivo* covalent state of the protein.

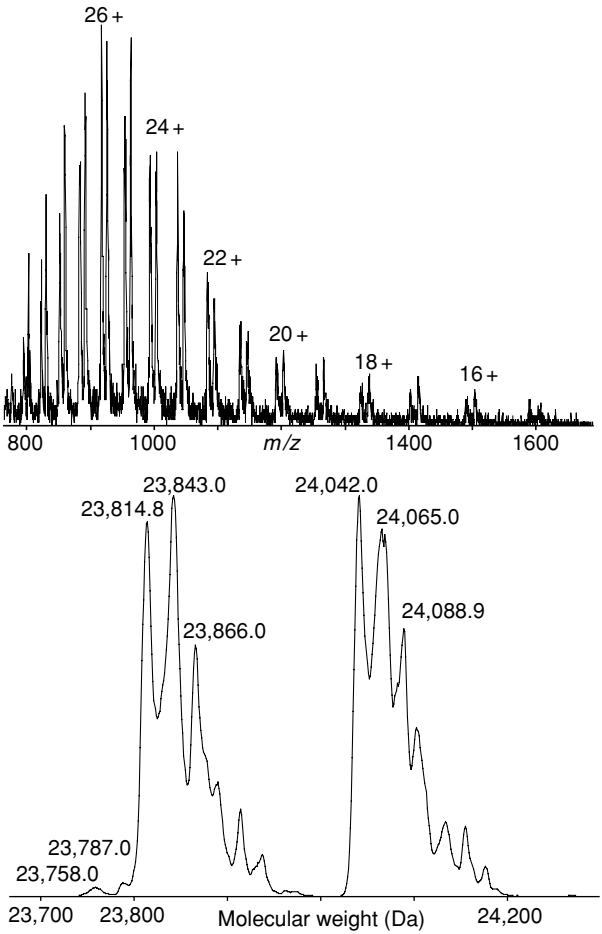


FIG. 4. Electrospray-ionization mass spectrometry of protein recovered by SDS–PAGE after immunoprecipitation. A chloroplast homolog of GrpE was recovered by immunoprecipitation and separated from antibody by SDS–PAGE. The gel was blotted to PVDF and stained briefly with Coomassie blue and the band of interest was excised. Protein was eluted with 2% SDS, 1% Triton X-100, 50 m*M* Tris (pH 9.0) for 16 h at room temperature and finally precipitated with chloroform–methanol. The protein recovered (2–3 μg) was dissolved in 60% formic acid for MALDI-TOF, but ESI indicated the presence of persistent detergent contaminants. Thus the SEC-MS technique was used to record the ESI spectra shown (*top*). Reconstruction of the molecular weight spectrum (*bottom*) revealed a pair of molecules differing by 227.2 Da. EST analysis had indicated two possible N termini for the mature protein caused by alternative splicing, one with an extra VQ (227.13 Da, calculated), and the intact mass measurements supported approximately equal expression of these two variants (23,812.5 and 24,039.6 Da, calculated). [From Schroda *et al.* (2001) with permission.]

4. Identification by Intact Mass

If the mass of a full-length protein is measured with sufficient accuracy, and the measurement matches the mass calculated from the gene sequence, it is possible to identify proteins by their intact mass. OmpA was separated from other *E. coli* membrane proteins by LC-MS and could be identified on the basis of its measured mass alone (le Coutre *et al.*, 2000). The major intrinsic protein from the bovine lens (Fig. 2) is another unusual example of a eukaryotic protein concerning which the measured mass agrees with that calculated from the sequence. The widespread occurrence of posttranscriptional and posttranslational modification in eukaryotes and especially humans makes the use of intact mass tags (IMTs) for identification generally unattractive, although future advances in software and databases may bring increased focus to this area. To illustrate how intact mass measurements are reconciled with available genomic data we have included a section describing studies of higher plant thylakoid membranes (Section III).

5. LC-MS+

Although much will be gleaned from 2D gel proteomics, a more complete picture of the membrane proteome will result from analytical techniques optimized for this subfraction. Furthermore, incorporation of intact mass measurements will provide the opportunity to visualize subtle alterations, such as methionine oxidation, that may modulate protein function without affecting mobility in 2D analysis (Whitelegge *et al.*, 1998a, 2000a). Liquid chromatography is far more amenable to downstream processing as opposed to gel technology, where extraction is problematic. The solvent line can be split to simultaneously acquire MS spectra and collect fractions. We use the acronym LC-MS+ to describe the MS analysis/fraction collection strategy (Fig. 5; Whitelegge, 1998; Whitelegge *et al.*, 1998b). Fractions can be stored for subsequent MALDI analysis if the ESI spectrum is too complex to interpret or for digestion/cleavage to generate peptide fragments for identification against databases and localization of posttranslational modifications. Our experience to date suggests that this approach will be important for the in-depth profiling of the membrane proteome. The ninth subunit of the chloroplast cytochrome b_6/f complex was identified as ferridoxin:$NADP^+$ oxidoreductase, using LC-MS+ with subsequent confirmation by cross-reaction with antibodies and by activity measurements (Zhang *et al.*, 2001).

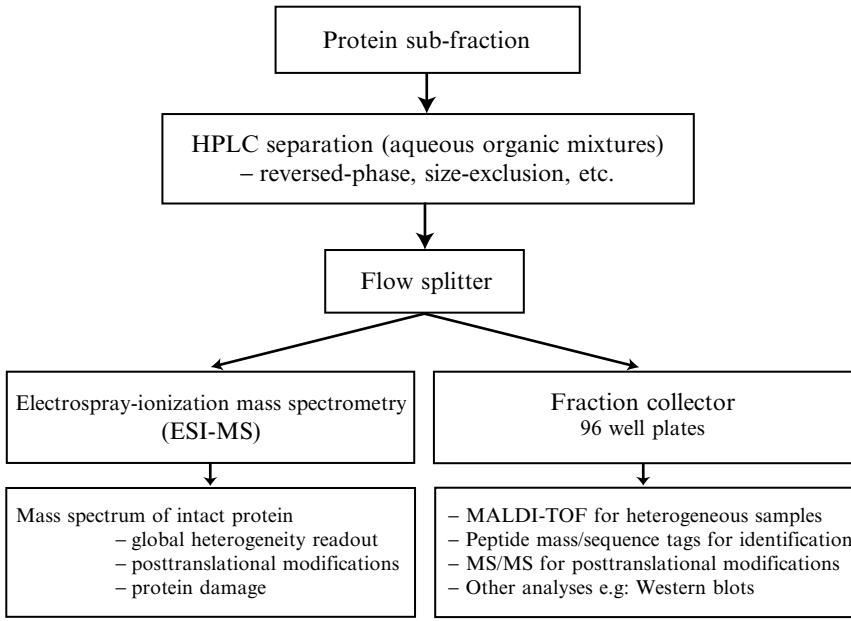


FIG. 5. Schematic of LC-MS+. Through the use of a ''T'' flow splitter, fractions are collected simultaneous to LC-MS. Mass and/or sequence tags are then used to increase the confidence of identifications when the intact mass measurement is ambiguous. Where measured intact mass deviates from mass calculated according to translated genome sequence, the masses of the fragments are used to localize the source of deviations. Tandem mass spectrometry is then used to characterize the primary structure of such fragments and define modifications. By combining intact mass measurements with downstream mass spectrometry experiments it is possible to monitor a proteome subfraction for the earliest events in physiological processes. [Adapted from Whitelegge (1998) with permission.]

III. Applications: The Thylakoid Membrane Proteome

A. *Generalized HPLC Elution Profiles*

The application of LC-MS in proteomic studies using intact mass as the means of identification may be possible because of the high conservation of protein sequence across species (Gómez *et al.*, 2002). Orthologs, in general, have similar chromatographic behavior (HPLC retention times) and similar masses. This information makes it possible to build a generalized HPLC elution profile for any proteome (tissue, membrane, cell type, etc.) by comparing the same proteome from different species.

The feasibility of this type of analysis does not demand availability of the genome sequence, but the analysis is facilitated if the genome sequence is available. It is necessary that some genomic information (i.e., protein-coding sequence) be known for each of the species being used to build the elution map, and that these sequence sets overlap to maximize coverage of the proteome being studied. Once generalized HPLC elution maps are available it is possible to perform intact mass proteomic studies on organisms for which there is no available DNA sequence information.

Such a proteomic survey of the thylakoid membranes [photosystem II (PS II) enriched] from the chloroplasts of spinach and pea has now been completed (Gómez *et al.*, 2002). The data set includes seven replicates of the spinach, three replicates of the pea, and four replicates of the tobacco profiles, in addition to two replicates of total thylakoids from *Arabidopsis thaliana*. A frequent observation was that the PS II preparations were often contaminated by the stromal extrinsic PS I proteins, and that the ESI-MS system was sufficiently sensitive to detect these contaminants even in the absence of appreciable peaks in the UV absorption profile. The generalized HPLC elution profile from the pea, spinach, and tobacco PS II preparations is shown in Fig. 6. Several of the peaks identified in pea but not in spinach, and vice versa, have been confirmed by comparison with tobacco or *A. thaliana* (not shown).

B. Assignment of Intact Mass Tags to Gene Sequences

A total of 177 IMTs were identified from the pea and spinach experiments. Noteworthy is the fact that none of the IMTs matched any of the predicted masses generated by simply translating the published gene sequences. In general, IMTs usually fail to correlate with the mass calculated from translations of published nucleic acid sequences. There are several mechanisms to explain this, including retention of *N*-formylmethionine (chloroplasts are reduced prokaryotes), secondary N- and C-terminal processing of the protein, covalent posttranslational modification of the protein (phosphorylation, glycosylation, oxidation, etc.), mRNA editing, the presence of proteins encoded by a different allele in a strain/race/cultivar other than the one from which the DNA sequence was obtained, paralogs other than the published sequence, and protein folding that traps endogenous ligands and other noncovalent adducts within the protein, in which case the measured mass is a composite of the two. There are also several nonphysiological reasons for IMTs not matching the predicted mass. These artifactual modifications include buffer (particularly sulfate and phosphate), detergent (i.e., SDS,

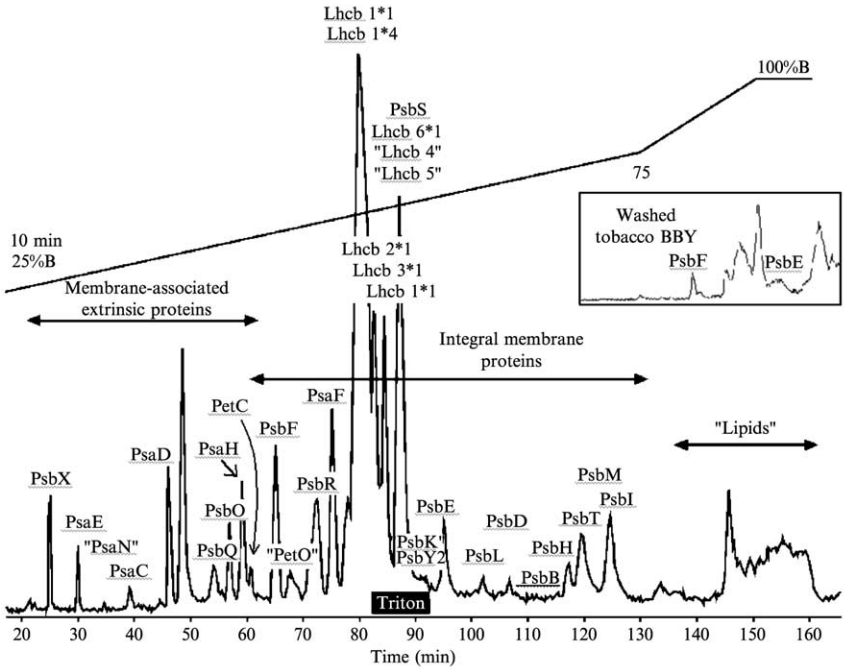


FIG. 6. RP-HPLC elution profile of pea PS II preparations. A polymeric stationary phase (PLRP/S, 2.1 × 150 mm, 5 μm × 300 Å, 40°C; Polymer Labs) is equilibrated in 0.1% TFA–5% acetonitrile before loading of the sample in 100 μl of 60% formic acid and gradient elution through 0.1% TFA–100% acetonitrile. The same profiles have been generated for pea, spinach and tobacco PS II preparations. The tobacco profile (*inset*) is from NaBr-washed membranes that have had the extrinsic proteins removed, as can be seen by the lack of signal in the corresponding region of the elution profile. Underlined names indicate the elution position for proteins that were not detected in this particular experiment. Names in quotes indicate proteins for which there is no pea, spinach, or tobacco sequence information, but on the basis of size and predicted hydrophobicity have been tentatively assigned. Triton X-100 contamination is indicated by the black box. [From Gómez *et al.* (2002) with permission.]

Triton X-100), and/or solvent (i.e., TFA) adducts on the protein, methionine oxidation during sample preparation, mutations of the gene during cloning, and errors in DNA sequencing. Noncovalent phosphate and sulfate adducts from buffers are extraordinarily persistent (Mirza and Chait, 1994) and can remain associated with the protein even after extensive washing.

Covalent modifications are generally simple to identify from the magnitude of the mass shift. For example, by adding 28 Da we matched four IMTs to four chloroplast genes whose products retain the *N*-formylmethionine (fMet). Removal of fMet from the predicted chloroplast masses allowed identification of a further 12 IMTs. Application of known protein modifications in addition to fMet removal allowed identification of five more IMTs encoded by the chloroplast. Nuclear-encoded proteins begin with *N*-methionine, but have a transit/targeting peptide necessary for entry to the chloroplast that is then removed to form the mature protein. Removal of published targeting peptides allowed the identification of 16 nuclear-encoded IMTs. Addition of 42 Da (acetylation) for blocked proteins permitted six more IMTs to be assigned. Other common covalent modifications were simple to identify because they result in minimal or predictable changes to the chromatographic retention time and have easily identified mass shifts (i.e., +16 Da for oxidation, +80 Da for phosphorylation). Nineteen IMTs were identified as either of these modifications of a known gene translation product. Finally, comparison of retention times of similar IMTs from one species with the other allowed the identification of 13 other gene products for which there was no available gene sequence. Incorporation of all these modifications simplified assignment of many IMTs to translated gene sequences such that it was possible to assign 75 of the 177 IMTs (42%) from the spinach and pea data sets (Gómez *et al.*, 2002).

In some cases a series of noncovalent adducts can be present (+1*n*, +2*n*, +3*n*, etc.) that is readily recognized in the molecular weight reconstruction of the mass spectra. Allelic differences, paralogs, mRNA editing, mutations arising during cloning, and DNA sequencing errors (collectively referred to here as sequence errors) present essentially the same type of difficulty when matching the calculated mass to the observed mass. It is likely that sequence-altered mass calculations may account for a significant proportion of the unassigned masses observed in an intact proteomic sample. In the thylakoid proteome example, 26 IMTs were encountered that had retention times and masses similar to those of assigned IMTs in another species, but did not agree with the masses calculated from any of the modified translations of the published gene sequences. It is likely that such differences are not due to further covalent modification of the protein, and in the absence of any other reasonable explanation they were therefore placed in the general category of sequence errors. The described process represents a methodology using phylogenetic comparison for reconciling a sequence error mass calculation with the observed mass in order to assign an identity. This methodology is illustrated in four examples described below.

1. Different Alleles/Paralogs: Tobacco PsaE

The gene *psaE* encodes a protein of 130–143 amino acids, of which the amino-proximal amino acids 38–51 are a chloroplast transit/targeting peptide that is removed from the mature protein (Anandan *et al.*, 1989; Obokata *et al.*, 1994). In both *Nicotiana tabacum* and *Nicotiana sylvestris* mature PsaE is isolated as a mixture of isoforms that differ by the presence or absence of N-terminal A^{52} (*N. sylvestris* PsaEb, numbering starts with ^{i}Met). The Ala^{+} and Ala^{-} isoforms of PsaE in *Nicotania* spp. are present in equimolar ratios and their presence has been shown not to be an artifact of thylakoid/protein isolation (Obokata *et al.*, 1993). We have confirmed the observations of Obokata *et al.* for *N. tabacum* and show that the differential processing of the N terminus of PsaE is conserved in pea, spinach, and *A. thaliana* (Gómez *et al.*, in preparation). Comparison of the published PsaE sequences suggests that the secondary processing of the N-terminal Ala may be conserved among the dicotyledons, but may not be in the monocotyledons (not shown).

Nicotiana tabacum is an amphidiploid species believed to have originated from natural hybridization of two autogamous diploid species, *N. tomentosiformis* and *N. sylvestris* (Goodspeed and Clausen, 1928; Smith, 1968; Gray *et al.*, 1974; Kung *et al.*, 1975). No *psaE* genes have been sequenced from *N. tabacum* or *N. tomentosiformis*, but two *psaE* genes have been sequenced from *N. sylvestris* (Obokata *et al.*, 1994). The tobacco PS II preparations from which the data reported here are taken were less contaminated by extrinsic membrane proteins than were the pea and spinach PS II preparations (~10-fold less), and had significantly less than the *Arabidopsis* thylakoid preparation (~100-fold less). Nevertheless, the PsaE proteins in the tobacco preparations were still detectable, although because of the small amount of material and resulting weak signals, the error in the mass measurements is understandably higher.

Three pairs of masses were recognized as differing by 71 Da, indicative of the presence or absence of the N-terminal Ala (Fig. 7). The predicted masses for the *N. sylvestris* PsaE proteins are as follows: PsaEa, 9928.3 Da; PsaEa(Ala^{-}), 9857.2 Da; PsaEb, 9882.2 Da; PsaEb(Ala^{-}), 9811.1 Da. The major observed pair (9882.1 2.6 and 9809.5 1.1 Da) were assigned as the orthologs to the PsaEb protein derived from the ancestral *N. sylvestris* gene. The observed mass pair with the second highest abundance (9927.2 1.8 and 9857.2 1.4 Da) are the orthologs to the *N. sylvestris* PsaEa pair. The third observed mass pair (9894.8 3.9 and 9822.2 1.8 Da) have IMTs that are 13 Da larger than the corresponding PsaEb masses. We attribute this to the presence of either a third PsaE paralog in *N. tabacum* (*psaEc*) or an allele of the other two genes. Alternatively, the error in the

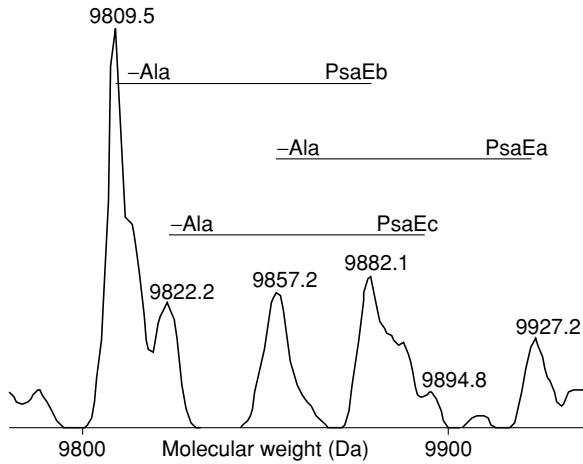


FIG. 7. Reconstructed mass spectrum of *Nicotiana tabacum* PsaE isoforms. Data were collected and analyzed as described in Fig. 6. The reconstructed mass spectrum is from a single representative LC-MS experiment. The masses shown are the averages of four experiments.

mass measurements is large enough that the third pair of masses could possibly be due to a single oxidation (+16 Da) of the PsaEb isoforms. The absence of +16-Da adducts in any of the PsaE isoforms from pea, spinach, or *A. thaliana* is supportive of the proposed third gene/allele; however, the tobacco preparations were made from older leaves that may be more susceptible to oxidative damage.

2. *Point Mutations: Spinach PsbH*

The 10-kDa phosphoprotein of PS II (PsbH) also has characteristic posttranslational modifications that simplify MS identification. PsbH is the only PS II core complex protein that is doubly phosphorylated (Gómez *et al.*, 1999; Vener *et al.*, 2001). Our initial results for spinach PsbH did not match the published sequence and the mass difference was erroneously attributed to secondary N-terminal processing of A^2 and to the acetylation of T^3 (Gómez *et al.*, 1999). Subsequent analysis of tobacco PsbH showed, as previously reported (Farchaus and Dilley, 1986; Webber *et al.*, 1989; Ikeuchi *et al.*, 1989), that A^2 is indeed the N-terminal amino acid and that it is not blocked. Alignment of spinach PsbH with published translations of PsbH orthologs showed residues 29 and 36 were at variance with conserved amino acids at these positions (Fig. 8A; see Color Insert). This homology alignment shows Glu at position 29 (instead of Lys) in 26 of the

29 proteins examined, and a uniformly conserved Gly at position 36 (instead of Arg). The calculated mass of spinach $^{29}E^{36}G$-PsbH (7598.9 Da) is within 0.008% of the observed mass (7599.5 1.5 Da).

The entire spinach chloroplast genome was sequenced (Schmitz-Linneweber *et al.*, 2001). The mature translated PsbH product from this accession agrees with and confirms our proposed changes to the PsbH protein sequence (Fig. 8B). Both changes can be attributed to guanine-to-adenine transitions in the original spinach *psbH* sequence (Westhoff *et al.*, 1986), but it is not possible to retrospectively determine whether the changes were due to mutations in cloning or errors in DNA sequencing. The most current GenBank *psbH* sequence also encodes an additional hexapeptide not present in other higher plant *psbH* sequences (Fig. 8B). It seems likely that this N-terminal hexapeptide may be an artifact of the gene finder used to annotate the spinach chloroplast genome, although a similar propeptide was predicted in an alternative form of PsbH from the cyanobacteria *Prochlorothrix hollandica* (Greer and Golden, 1992). Further work is required to clarify the existence of a propeptide at the N terminus of newly translated PsbH.

3. *Insertions and Deletions: Pea PsbT*

The single membrane-spanning protein encoded by the chloroplast gene *psbT* has been localized to the core complex of PS II, and MS analysis of the spinach protein has shown that it retains *N*-formylmethionine (Zheleva *et al.*, 1998). We confirmed the MS characterization of spinach PsbT (Gómez *et al.*, 2002), but found that the pea protein, which has a retention time similar to that of the spinach protein, did not match the mass predicted from the published sequence (Lehmbeck *et al.*, 1989). The alignment of all published PsbT orthologs shows an obvious difference in the pea sequence compared with the others (Fig. 9; see Color Insert). The pea PsbT protein has the sequence ^{13}KELV^{16} instead of the highly conserved sequence ^{13}TLGII^{17} found in almost all the published orthologs. Comparison of the several *psbT* gene sequences shows that there has been a four-nucleotide deletion in the two codons for ^{13}TL^{14}, resulting in a codon for Lys and a single nucleotide insertion in the codon for I^{17} (Fig. 9B). Replacing the pea-specific KELV sequence with the conserved TLGII sequence results in a calculated mass of 4060.1 Da that is within 0.02% of the observed mass (4060.9 Da). This and the previous example demonstrate the usefulness of comparing the amino acid sequences of orthologs to identify errors in the published sequences. Searching for nonconservative changes in highly conserved residues and changing them to match the consensus has been an effective tool for

reconciling IMTs with a published gene sequence. The following example demonstrates that errors can also be caused by phylogenetic comparisons that are not carefully reasoned.

4. *Protein Sequence Adjustments Necessitated by Error Correction of DNA Sequences: Arabidopsis AtpB*

The *atpB* gene has been used for several phylogenetic studies of chloroplast evolution (Ikeda *et al.*, 1992; Hoot *et al.*, 1999; Savolainen *et al.*, 2000) and as a result there are hundreds of *atpB* sequences in the databases that can be used for sequence comparisons. *AtpB* is the only thylakoid protein with a mass of about 53 kDa. In *Arabidopsis thaliana* a mass of 53,877.0 Da was observed, but the mass calculated for fMet-*AtpB* is 84.8 Da heavier (53,961.8 Da). Sequence alignment of *A. thaliana AtpB* with the few hundred *AtpB* orthologs revealed no obvious discrepancies in conserved residues (not shown). Changes were not observed in the *A. thaliana* sequence when compared with the consensus; however, the only other sequences from Brassicaceae (*Brassica napus* [AF267641] and *Raphanus sativus* [AJ277564]) both differ from the consensus at positions 11, 18, and 223 (Fig. 10A). The angiosperm consensus, $G^{11}K^{18}Q^{223}$, is similar to the *A. thaliana* sequence ($E^{11}K^{18}Q^{223}$). The Brassicaceae differ significantly at these residues ($A^{11}N^{18}L^{223}$). If the *A. thaliana* sequence is modified to match the other Brassicaceae, a mass of 53,875.8 Da is calculated, which is within 0.002% of the observed mass. Residue N^{9} in the *A. thaliana* sequence differs from the dicotyledon (including the Brassicaceae) consensus (D^{9}), but the 1-Da increase in mass introduced by this change is well within the experimental error and is presented only as an additional possible error.

Comparison of the Brassicaceae DNA sequences for residues 9, 11, and 223 reveals differences in the first or second position in the codon, and residue 18 differs at the wobble position (Fig. 10). It is likely that E^{11} (and probably N^{9}) in the *A. thaliana* sequence are genuine sequence errors, and that K^{18} and Q^{223} were assigned when a DNA sequence was compared with the existing ortholog sequences and reasonably called AAA in the case of K^{18} and CAA in the case of Q^{223}.

The previous examples from pea, spinach, and tobacco are from DNA sequences published in the mid-1980s. The *Arabidiopsis thaliana* chloroplast sequence was submitted to GenBank in 2000, yet *atpB* still appears to have a significant number of DNA sequence errors. Older DNA sequences had few orthologs available for comparison, and the sequence quality suffered from too little information. The *A. thaliana atpB* sequence probably suffered from too much information: 110 of 112 published

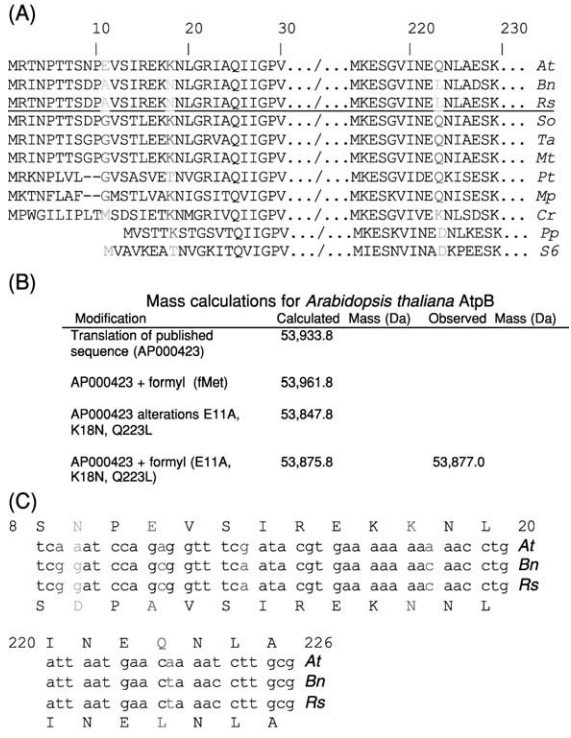


FIG. 10. (A) Alignment of specific regions of AtpB from a wide range of taxa. Taxa above the line are Brassicaceae. The three residues of interest are colored to indicate vascular plant consensus (red), Brassicaceae consensus (light blue), and others (green). (B) Mass calculations of possible *A. thaliana AtpB* masses compared with the observed mass. (C) Identification of nucleotide differences in the Brassicaceae. Blue, Conservative substitutions; green, possible mutant N^9; red, changes used to calculate revised mass. *At*, *Arabidopsis thaliana*, AP000423; *Bn*, *Brassica napus*, AF267641; *Rs*, *Raphanus sativus*, AJ277564; *So*, *Spinacia oleracea*, U23082; *Ta*, *Triticum aestivum*, M16843; *Mt*, *Magnolia tripetala*, AJ235526; *Pt*, *Pinus tunbergii*, D17510; *Mp*, *Marchantia polymorpha*, X04465; *Cr*, *Chlamydomonas reinhardtii*, M13704; *Pp*, *Porphyra purpurea*, U38804; *S6*, *Synechocystis* PCC 6803, X58129.

angiosperm sequences were conserved at residues 11, 18, and 223. It was unfortunate luck that the two differing sequences were from plants more closely related to *Arabidopsis thaliana* than the other 110 taxa. This example is used to demonstrate that care must be exercised when comparing orthologs as a means to check DNA sequence and that more recent DNA sequence data are not necessarily more error free.

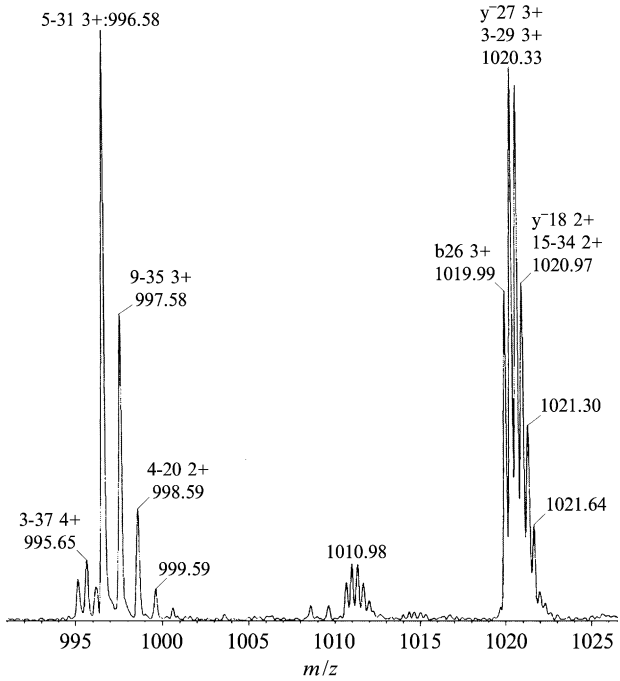


FIG. 11. MS/MS spectrum of a 4.4-kDa PsbF membrane protein recorded on a quadrupole time-of-flight mass spectrometer (QTOF, Micromass; courtesy of R. Martin). Isotopic resolution allows assignment of charge states based on isotopic spacing. Ion assignments were software generated on the basis of known sequence. A narrow section of the complete MS/MS spectrum is shown. The full spectrum provided confirmation of the PsbF sequence including a site altered as a result of an RNA-editing event. High-resolution MS/MS experiments will become increasingly important for characterization of membrane proteins, where proteolysis or chemical cleavage frequently generates large peptides.

5. Confirmation by MS/MS

Three of the four examples presented above have yet to be confirmed by resequencing. However, MS/MS has been used to reconcile sequence-altered mass calculations with IMTs for which nonconservative changes in sequence could not be identified by ortholog comparison. The IMT for pea PsbH (7697.3 1 Da) differs from the calculated mass (7727.0 Da) by 30 Da. There are no obvious mutations in the pea PsbH sequence that could explain this difference, and internal peptides were therefore subjected to MS/MS to pinpoint and identify the discrepancy. Residue

18 is translated as Ala in the published DNA sequence (Lehmbeck *et al.*, 1989), but MS/MS sequence showed the presence of Val instead. This change is insufficient to account for the difference between observed and calculated masses in pea PsbH and there are likely more conservative changes to be found. We have also sequenced the entire spinach PsbF protein by MS/MS (J. P. Whitelegge and L. B. Martin III, unpublished) (Fig. 11) and confirmed the observation that an RNA-editing site changes residue 26 from Ser to Phe (Bock *et al.*, 1993). This result is in complete agreement with other MS studies of pea PsbF (Sharma *et al.*, 1997a,b; Whitelegge *et al.*, 1998a).

IV. Conclusions

IMPs include many of the most important proteins found in cells and thus need to be fully covered in proteomics. Although 2D gels are undoubtedly a powerful tool in proteomics, it remains likely that at least some classes of IMPs are poorly represented in such analyses. Furthermore, the major premise of 2D gel technology is that expression changes can be visualized by changes in spot intensity and changes in isoelectric point and/or electrophoretic mobility. Alternative separation techniques incorporating mass spectra of intact proteins provide a means to track subtle covalent changes that would go unseen in a 2D analysis and thus a battery of techniques are under development. In actual fact the field is young and much intellectual space remains to be explored, especially with respect to separation of complex mixtures of membrane proteins and the miniaturization of these methods. Furthermore, the challenge of applying these techniques to the diverse heterogeneous IMPs of humans is immense. It thus seems likely that new approaches, perhaps involving capture of membrane proteins on antibody arrays combined with mass spectral analysis, will proliferate.

Acknowledgments

Supported in part by research grants from the NIH [DA 05010 (C. J. Evans, PI), K. F. F.; NS31271 (A. L. Fluharty, PI), K. F. F. and J. P. W.; AI 12601 (M. Lovett, PI), J. P. W.; and NS 07171 (A. D. Grinnell, PI), J. P. W.], the DOE [DE-FG03-01ER15251 (K. F. F. and J. P. W., PIs)], the Ford Foundation (S. M. G.), and equipment grants from the W. M. Keck Foundation and the Pasarow family. The authors gratefully acknowledge the advice from numerous collaborators and colleagues including Cameron Gunderson, Ernest Wright, H. Ronald Kaback, Johannes le Coutre, Eric Turk, John E. Walker, I. Fearnley, James Barber, Fred W. McLafferty, Richard L. Stevens, Arvan Fluharty, and Guido Zamphigi. Laboratory assistance from numerous UCLA undergraduate Student Research Project participants, including Kris Tjon, Rishi Desai, and Brandon Penn, is also acknowledged.

References

Anandan, S., Vainstein, A., and Thornber, J. P. (1989). *FEBS Lett.* **256,** 150–154.

Appleton, W. S. (2000). ''Prozac and the New Antidepressants.'' Dutton/Plume, New York.

Arkin, I. T., Brünger, A. T., and Engelman, D. M. (1997). *Proteins Struct. Funct. Genet.* **28,** 465–466.

Barber, M., Bordoli, R. S., Sedgwick, R. D., and Tyler, A. N. (1981). *Nature* **293,** 270–275.

Barber, M., Bordoli, R. S., Elliot, G. J., Sedwick, R. D., and Tyler, A. N. (1982). *Anal. Chem.* **54,** 654A–657A.

Barnidge, D. R., Dratz, E. A., Jesaitis, A. J., and Sunner, J. (1999). *Anal. Biochem.* **269,** 1–9.

Blanco, D. R., Whitelegge, J. P., Miller, J. N., and Lovett, M. A. (1999). *J. Bacteriol.* **181,** 5094–5098.

Bock, R., Hagemann, R., Kössel, H., and Kudla, J. (1993). *Mol. Gen. Genet.* **240,** 238–244.

Boyot, P., Trifilieff, E., Van Dorsselaer, A., and Luu, B. (1988). *Anal. Biochem.* **173,** 75–85.

Brett, M., and Findlay, J. B. C. (1983). *Biochem. J.* **211,** 661–670.

Cadene, M., and Chait, B. T. (2000). *Anal. Chem.* **72,** 5655–5658.

Curstedt, T., Johansson, J., Persson, P., Eklund, A., Robertson, B., Löwenadler, B., and Jörnvall, H. (1990). *Proc. Natl. Acad. Sci. USA* **87,** 2985–2989.

Demmers, J. A. A., Haverkamp, J., Heck, A. J. R., Koeppe, R. E.,II, and Killian, J. A. (2000). *Proc. Natl. Acad. Sci. USA* **97,** 3189–3194.

Demmers, J. A. A., Killian, J. A., and Heck, A. J. R. (2001a). *In* ''Proceedings of the 49th Conference on Mass Spectrometry and Allied Topics,'' Chicago, IL.

Demmers, J. A. A., van Duijn, E., Haverkamp, J., Greathouse, D. V., Koeppe, R. E., II, Heck, A. J. R., and Killian, J. A. (2001b). *J. Biol. Chem.* **276,** 34501–34508.

Eisenberg, D. (1984). *Annu. Rev. Biochem.* **53,** 595–623.

Farchaus, J., and Dilley, R. A. (1986). *Arch. Biochem. Biophys.* **244,** 94–101.

Fearnley, I. M., and Walker, J. E. (1996). *Biochem. Soc. Trans.* **24,** 912–917.

Feick, R. G., and Shiozawa, J. A. (1990). *Anal. Biochem.* **187,** 205–211.

Findlay, J. B. C. (1987). *In* ''Biological Membranes: A Practical Approach'' (J. B. C. Findlay and W. H. Evans, Eds.), pp. 179–217. IRL Press, Oxford.

Findlay, J. B. C., Brett, M., and Pappin, D. J. C. (1981). *Nature* **293,** 314–316.

Fridriksson, E. K., Baird, B., and McLafferty, F. W. (1999). *J. Am. Soc. Mass Spectrom.* **10,** 453–455.

Fuller, R. W., and Wong, D. T. (1990). *Ann. N.Y. Acad. Sci.* **600,** 68–80.

Gadsby, D. C., and Nairn, A. C. (1999). *Physiol. Rev.* **79,** S77–S107.

Gerber, G. E., Anderegg, R. J., Herlihy, W. C., Gray, C. P., Biemann, K., and Khorana, H. G. (1979). *Proc. Natl. Acad. Sci. USA* **76,** 227–231.

Gómez, S. M., Park, J. J., Zhu, J., Whitelegge, J. P., and Thornber, J. P. (1999). Photosynthesis: Mechanisms and effects. *In* ''Proceedings of the 11th International Congress on Photosynthesis'' (G. Garab, Ed.), Vol. I, pp. 353–356. Kluwer Academic, Dordrecht, The Netherlands.

Gómez, S. M., Nishio, J. N., Faull, K. F., and Whitelegge, J. P. (2002). *Mol. Cell. Proteomics* **1,** 46–59.

Gómez, S. M., Bil, K. Y., Aguilera, R., Nishio, J. N., Faull, K. F., and Whitelegge, J. P. (2003). In preparation.

Goodspeed, T. H., and Clausen, R. E. (1928). *Univ. Calif. Publ. Botany* **11,** 245–256.

Gray, J. C., Kung, S. D., Wildman, S. G., and Sheen, S. J. (1974). *Nature* **252,** 226–227.
Green, B. N., and Bordoli, R. S. (1986). *In* ''Mass Spectrometry in Biomedical Research'' (S. J. Gaskell, Ed.), pp. 235–250. John Wiley & Sons, New York.
Greer, K. L., and Golden, S. S. (1992). *Plant Mol. Biol.* **19,** 355–365.
Haltia, T., and Freire, R. (1995). *Biochim. Biophys. Acta* **1241,** 295–322.
Heukeshoven, J., and Dernick, R. (1982). *J. Chromatogr.* **252,** 241–254.
Hille, B. (1992). ''Ionic Channels of Excitable Membranes,'' 2nd ed., pp. 1–20. Sinauer Associates, Sunderland, MA.
Hoot, S. B., Magallon, S., and Crane, P. R. (1999). *Ann. Mo. Bot. Gard.* **86,** 1–32.
Hufnagel, P., Schweiger, U., Eckerskorn, C., and Oesterhelt, D. (1996). *Anal. Biochem.* **243,** 46–54.
Ikeda, T. M., Terachi, T., and Tsunewaki, K. (1992). *Jpn. J. Genet.* **67,** 111–123.
Ikeuchi, M., Takio, K., and Inoue, Y. (1989). *FEBS Lett.* **242,** 263–269.
Jensen, P. K., Pasa-Toli, L., Anderson, G. A., Horner, J. A., Lipton, M. S., Bruce, J. E., and Smith, R. D. (1999). *Anal. Chem.* **71,** 2076–2084.
Joppich-Kuhn, R., Corkill, J. A., and Giese, R. W. (1982). *Anal. Biochem.* **119,** 73–77.
Kelleher, N. L., Lin, H. Y., Valaskovic, G. A., Aaserud, D. J., Fridriksson, E. K., and McLafferty, F. W. (1999). *J. Am. Chem. Soc.* **121,** 806–812.
Kühlbrandt, W., Wang, D. N., and Fujiyoshi, Y. (1994). *Nature* **367,** 614–621.
Kung, S. D., Gray, J. C., Wildman, S. G., and Carlson, P. S. (1975). *Science* **187,** 353–355.
le Coutre, J., Whitelegge, J. P., Gross, A., Turk, E., Wright, E. M., Kaback, H. R., and Faull, K. F. (2000). *Biochemistry* **39,** 4237–4242.
Lehmbeck, J., Stummann, B. M., and Henningsen, K. W. (1989). *Physiol. Plant* **76,** 57–64.
le Maire, M., Deschamps, S., Mϕller, J. V., Le Caer, J. P., and Rossier, J. (1993). *Anal. Biochem.* **214,** 50–57.
Lin, D., Alpert, A. J., and Yates, J. R., III. (2001). *Am. Genomic/Proteomic Technol.* **1,** 38–46.
Loo, D. D. F., Zeuthen, T., Chandy, G., and Wright, E. M. (1996). *Proc. Natl. Acad. Sci. USA* **93,** 13367–13370.
Macfarlane, R. D. (1990). *Methods Enzymol.* **193,** 263–280.
Macfarlane, R. D., and Torgerson, D. F. (1976). *Science* **191,** 920–925.
Michel, H., Hunt, D. F., Shabanowitz, J., and Bennett, J. (1988). *J. Biol. Chem.* **263,** 1123–1130.
Mirza, U. A., and Chait, B. T. (1994). *Anal. Chem.* **66,** 2898–2904.
Molloy, M. P. (2000). *Anal. Biochem.* **280,** 1–10.
Obokata, J., Mikami, K., Hayashida, N., Nakamura, M., and Sugiura, M. (1993). *Plant Physiol.* **102,** 1259–1267.
Obokata, J., Mikami, K., Yamamoto, Y., and Hayashida, N. (1994). *Plant Cell Physiol.* **35,** 203–209.
Oesterhelt, D., Meentzen, M., and Schuhmann, L. (1973). *Eur. J. Biochem.* **40,** 453–463.
Pappin, D. J. C., and Findlay, J. B. C. (1984). *Biochem. J.* **217,** 605–613.
Picot, D., Loll, P. J., and Garavito, R. M. (1994). *Nature* **367,** 243–249.
Popot, J.-L., and Engelman, D. M. (2000). *Annu. Rev. Biochem.* **69,** 881–922.
Rosinke, B., Strupat, K., Hillenkamp, F., Rosenbusch, J., Dencher, N., Krüger, U., and Galla, H.-J. (1995). *J. Mass Spectrom.* **30,** 1462–1468.
Russell, W. K., Park, Z. Y., and Russell, D. H. (2001). *Anal. Chem.* **73,** 2682–2685.
Santoni, V., Molloy, M., and Rabilloud, T. (2000). *Electrophoresis* **21,** 1054–1070.
Savolainen, V., Chase, M. W., Hoot, S. B., Morton, C. M., Soltis, D. E., Bayer, C., Fay, M. F., de Bruijn, A. Y., Sullivan, S., and Qiu, Y. L. (2000). *Syst. Biol.* **49,** 306–362.

Schaller, J., Pellascio, B. C., and Schlunegger, U. P. (1997). *Rapid Commun. Mass Spectrom.* **11,** 418–426.
Schey, K. L., Papac, D. I., Knapp, D. R., and Crouch, R. K. (1992). *Biophys. J.* **63,** 1240–1243.
Schindler, P. A., Van Dorsselaer, A., and Falick, A. M. (1993). *Anal. Biochem.* **213,** 256–263.
Schmitz-Linneweber, C., Maier, R. M., Alcaraz, J. P., Cottet, A., Herrmann, R. G., and Mache, R. (2001). *Plant Mol. Biol.* **45,** 307–315.
Schroda, M., Vallon, O., Whitelegge, J. P., Beck, C. F., and Wollman, F. A. (2001). *Plant Cell* **13,** 2823–2839.
Sharma, J., Panico, M., Barber, J., and Morris, H. R. (1997a). *J. Biol. Chem.* **272,** 3935–3943.
Sharma, J., Panico, M., Barber, J., and Morris, H. R. (1997b). *J. Biol. Chem.* **272,** 33153–33157.
Sheppard, D. N., and Welsh, M. J. (1999). *Physiol. Rev.* **79,** S23–S45.
Shevchenko, A., Loboda, A., Shevchenko, A., Ens, W., and Standing, K. G. (2000). *Anal. Chem.* **72,** 2132–2141.
Smith, H. H. (1968). *In* ''Nicotiana: Procedures for Experimental Use'' (R. D. Durbin, Ed.), Technical Bulletin 1586, pp. 1–16. U.S. Department of Agriculture, Washington, D.C.
Stark, P., Fuller, R. W., and Wong, D. T. (1985). *J. Clin. Psychiatry* **46,** 7–13.
Stowell, M. H. B., and Rees, D. C. (1995). *Adv. Protein Chem.* **46,** 279–311.
Tarr, G. E., and Crabb, J. W. (1983). *Anal. Biochem.* **131,** 99–107.
Tjon, K., Faull, K. F., and Whitelegge, J. P. (2000). *UCLA Undergrad. Sci. J.* **13,** 122–126.
Turk, E., and Wright, E. M. (1997). *J. Membr. Biol.* **159,** 1–20.
Turk, E., Kim, O., le Coutre, J., Whitelegge, J. P., Eskandari, S., Lam, J. T., Kreman, M., Zampighi, G., Faull, K. F., and Wright, E. M. (2000). *J. Biol. Chem.* **275,** 25711–25716.
Vener, A. V., Harms, A., Sussman, M. R., and Vierstra, R. D. (2001). *J. Biol. Chem.* **276,** 6959–6966.
Victora, C. G., Bryce, J., Fontaine, O., and Monasch, R. (2000). *Bull. World Health Organ.* **78,** 1246–1255.
Villa, S., De Fazio, G., and Canosi, U. (1989). *Anal. Biochem.* **177,** 161–164.
von Heijne, G. (1999). *Q. Rev. Biophys.* **32,** 285–305.
Webber, A. N., Hird, S. M., Packman, L. C., Dyer, T. A., and Gray, J. C. (1989). *Plant. Mol. Biol.* **12,** 141–151.
Welsh, M., Tsui, L.-C., Boat, T. F., and Beudet, A. L. (1995). *In* ''The Metabolic and Molecular Bases of Inherited Disease'' (C. R. Scriver, A. L. Beaudet, W. S. Sly, and D. Balle, Eds.), 7th ed., pp. 3799–3876. McGraw-Hill, New York.
Wessel, D., and Flügge, U. I. (1984). *Anal. Biochem.* **138,** 141–143.
Westhoff, P., Farchaus, J. W., and Herrmann, R. G. (1986). *Curr. Genet.* **11,** 165–169.
White, S. H., and Wimley, W. C. (1999). *Annu. Rev. Biophys. Biomol. Struct.* **28,** 319–365.
Whitelegge, J. P. (1998). *In* ''The Handbook of Plant and Crop Stress'' (M. Pessarakli, Ed.), pp. 555–568. Marcel Dekker, New York.
Whitelegge, J. P., Jewess, P., Pickering, M. G., Gerrish, C., Camilleri, P., and Bowyer, J. R. (1992). *Eur. J. Biochem.* **207,** 1077–1084.
Whitelegge, J. P., Gundersen, C. B., and Faull, K. F. (1998a). *Protein Sci.* **7,** 1423–1430.
Whitelegge, J. P., Faull, K. F., and Fluharty, A. L. (1998b). *In* ''Proceedings of the 46th Conference on Mass Spectrometry and Allied Topics,'' Orlando, Florida, May 31–June 4, 1998.

Whitelegge, J. P., Faull, K. F., Gundersen, C., and Gómez, S. M. (1999a). *In* ''Photosynthesis: Mechanisms and Effects'' (G. Garab, Ed.), Vol. V, pp. 4381–4384. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Whitelegge, J. P., le Coutre, J., Lee, J. C., Engel, C. K., Privé, G. G., Faull, K. F., and Kaback, H. R. (1999b). *Proc. Natl. Acad. Sci. USA* **96,** 10695–10698.

Whitelegge, J. P., Penn, B., To, T., Johnson, J., Waring, A., Sherman, M., Stevens, R. L., Fluharty, C. B., Faull, K. F., and Fluharty, A. L. (2000a). *Protein Sci.* **9,** 1618–1630.

Whitelegge, J. P., Cerda, B., Horn, D., Ge, Y., Brueker, K., Young, R., Holowka, D., Baird, B., and McLafferty, F. W. (2000b). *In* ''Proceedings of the 48th Conference on Mass Spectrometry and Allied Topics,'' June 11–15, 2000, Long Beach, CA.

Whitelegge, J. P., Zhang, H., Aguilera, R., Taylor, R. M., and Cramer, W. A. (2002). *Mol. Cell. Proteomics* **1,** 816–827.

Wildner, G. F., Fiebig, C., Dedner, N., and Meyer, H. E. (1987). *Z. Naturforsch.* **42c,** 739–741.

Wong, D. T., Bymaster, F. P., and Engleman, E. A. (1995). *Life Sci.* **57,** 411–441.

Zhang, H., Whitelegge, J. P., and Cramer, W. A. (2001). *J. Biol. Chem.* **276,** 38159–38165.

Zheleva, D., Sharma, J., Panico, M., Morris, H. R., and Barber, J. (1998). *J. Biol. Chem.* **273,** 16122–16127.

PROTEOMICS IN DRUG DISCOVERY

By RODNEY M. HEWICK, ZHIJIAN LU, AND JACK H. WANG

Wyeth Research, Cambridge, Massachusetts 02140

I. Introduction

Wilkins defined the proteome as the entire protein complement expressed by a genome, a cell, or tissue type (Wasinger *et al.*, 1995; Wilkins *et al.*, 1996). Proteomics is the study of the proteome and involves the technology used to identify and quantify the various proteins, protein–protein and protein–nucleic acid interactions within the proteome, as well as the posttranslational modifications that affect protein activity. A detailed study of the human proteome and the proteomes of other organisms under various physiological conditions is essential for a complete understanding of the mechanism of disease and thus for the future discovery of new drugs. The power of proteomics technology will lead to new clinical markers of disease, new protein therapeutics, and new drug targets.

Because proteins are the primary effectors of disease, most current drugs are designed or chosen to act at the protein level and not at the nucleic acid level of cellular control. The current excitement regarding proteomics technology and its potential in drug discovery has been given an additional boost by the completion of the Human Genome Project. Proteomics technology and genomic/transcriptional studies are interlinked and the availability of complete and accurate nucleic acid databases will enhance

ADVANCES IN PROTEIN CHEMISTRY, Vol. 65
0065-3233/03 $35.00

the speed and efficiency of protein identifications and their functional assignments in proteomic studies.

Proteomics has grown to encompass an increasing array of technologies that are being developed to analyze proteins and their interactions on a massive parallel scale (Fig. 1). These include indirect genetic procedures such as RNA profiling. This technology can be useful in determining possible changes in the proteome, although the transcript profile need not necessarily correlate well with the proteome (Anderson and Seilhamer, 1997; Gygi *et al.*, 1999b). Other genetic procedures such as yeast two-hybrid analysis (Fields and Song, 1989; Uetz *et al.*, 2000; Young, 1998; Colas and Brent, 1998) for identifying protein–protein interactions have had fair success but are indirect and do not easily measure multiple associations or

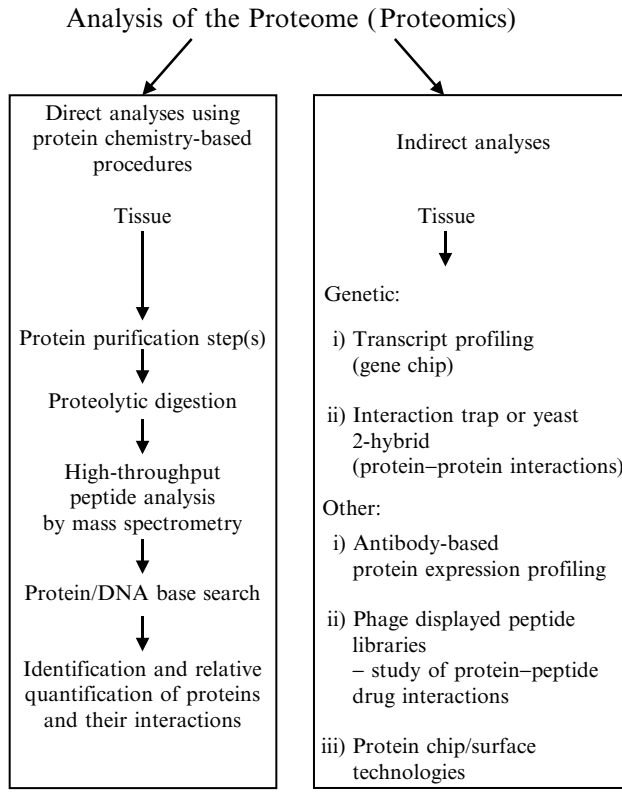


FIG. 1. Division of proteomics studies. *Left:* Direct analyses employing a series of procedures culminating in protein identification by mass spectrometry. *Right:* Indirect analyses using a variety of analytical technologies.

evaluate posttranslational modifications, which are important to defining drug targets. Protein and peptide libraries displayed by phage or by direct *in vitro* translation of RNA transcripts can be used to detect a broad range of protein–protein and drug–protein interactions (Winter *et al.*, 1994; Roberts and Szostak, 1997; Cho *et al.*, 2000). Also included within the field of proteomics is protein chip or protein surface technologies. Analogous to the approach used for DNA chips, this technology faces significant obstacles peculiar to proteins such as solubility, structural stability, specificity, and selectivity of binding as well as sensitivity, detection, and identification of the cognate binding ligand. Nevertheless, encouraging progress is being made (see Section VI).

A significant portion of proteomics is mass spectrometry (MS) based. As described in earlier articles in this volume, the final protein identifications are made by mass analysis comparisons of experimentally generated proteolytic fragments (or their derivatives) against a database containing the exact masses of proteolytic fragments (and their derivatives) from known protein sequences (Henzel *et al.*, 1993; James *et al.*, Mann *et al.*, 1993; Pappin *et al.*, 1993; Yates *et al.*, 1993; Hunt *et al.*, 1992; Eng *et al.*, 1994; Mann and Wilm, 1994), or from partial sequence information obtained by fragmenting individual peptides, using tandem MS.

In this article we describe proteomics technology and discuss the technical problems associated with proteome analysis, and how these are being overcome. We also include a representative selection of examples of the application of proteomic studies to the identification of markers and new drug targets.

II. The Complexity of Proteomic Analysis in Drug Discovery

The proteome is much more complex than the transcriptome because of posttranslational modifications. These alterations to the expressed protein include glycosylation, phosphorylation, sulfation, deamidation, oxidation, and proteolytic processing and degradation (Gooley and Packer, 1997). Thus, in the cell at any given time multiple isoforms and proteolytically cleaved versions of a single protein exist. To compare accurately the proteome under different physiological conditions all of these protein derivatives ideally need to be resolved and comparatively quantified. In addition, subtle changes in the proteome are likely to occur from moment to moment. Also, because proteins, unlike nucleic acids, have a wide range of solubilities, certain classes (e.g., membrane proteins) might be poorly represented in any given isolation procedure. Therefore, a single-step global analysis of the proteome is currently an unrealistic goal. Procedures

used to analyze and compare the complete proteome of a cell or tissue at any given time point must employ multiple purification steps to resolve adequately the rarest proteins from the more abundant. Consequently, provisions need to be made for chromatographic protein losses and for accurate relative quantification of protein samples. Sufficient sample or tissue also needs to be processed so that the rarest proteins can be identified in the final analytical procedure. This requirement may impose constraints on the design of the purification protocol and the apparatus used.

Even though it is an indirect form of proteomic analysis, the use of transcript profiling to predict differences in the relative expression of low-abundance proteins has certain practical advantages over direct protein analytical methods such as using two-dimensional (2D) gels and mass spectrometry (MS). Nucleic acids can be amplified and quantified such that rare and abundant mRNAs can be compared within and between samples. In addition, the uniform solubility of nucleic acids and the ability to isolate and detect any specific transcript by base-pair hybridization has made it possible to develop ''gene chips,'' which may eventually be able to compare and quantify the complete transcriptome (Lockhart *et al.*, 1996). Moreover, sample amount is less of a constraint for genomic studies because RNA transcripts can be amplified by ''quantitative'' polymerase chain reaction (PCR) and related procedures. There is no equivalent amplification step for proteomic analyses so investigative strategies are more dependent on sample scale-up. However, in both genomic and proteomic studies the quality and purity of the original samples to be compared are paramount in the generation of meaningful data.

Because a simple proteomics purification protocol is unlikely to detect rare proteins, and a more complicated multiple step procedure is likely to be too labor intensive and fraught with inaccurate quantification, current proteomic studies are usually general or specific in their approach.

III. Proteomic Strategies in Drug Discovery

The use of mass spectrometry-based proteomics in drug discovery can be divided broadly into two categories (Fig. 2): (1) general or global screens to identify proteins and to measure their relative abundances, and (2) focused inquiries into specific protein subsets, and studies of protein–protein interactions in cells and tissues of interest.

General screens look at a portion of the proteome that is revealed by the analytical purification procedures used. In their simplest form they may

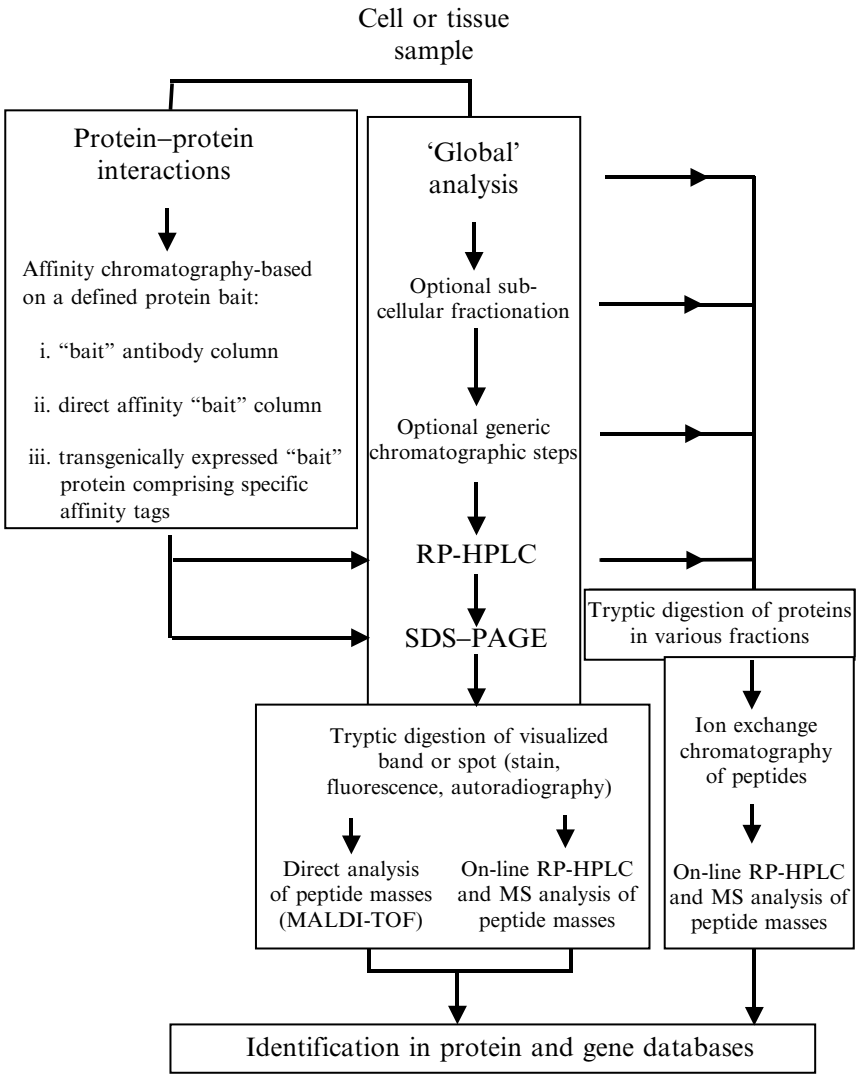


FIG. 2. Schematic flowchart showing strategies for proteomics studies using mass spectrometry. *Left:* Specific or affinity-based appoach. *Center:* General or global strategy. *Right:* Alternative strategy avoiding the use of 2D gels.

involve a single-step purification by 2D gel electrophoresis. In such analytical comparisons, the outputs of interest are protein expression patterns including their posttranslational modifications. Because disease mechanisms are generally pleiotropic, changes in expression patterns can

involve both rare and abundant proteins. Although rare proteins are often not detectable on stained gels, the changes seen in expression of the more abundant proteins can infer pathways associated with early disease etiology or may serve as useful markers of disease.

Inquiries of more focus take into account some prior knowledge, either from the scientific literature or from prior nonproteomic experimentation. In this approach, rare proteins in the proteome can be identified, quantified, and their interactions with other proteins and nucleic acids studied. Characteristic of this type of investigation is an initial subcellular fractionation to enrich for a specific subset of proteins, or an affinity purification using a previously characterized protein or nucleic acid as an affinity ''bait'' for identification of protein–protein and protein–nucleic acid complex formation. The gene for the bait protein can be engineered to encode specific affinity tags to enhance subsequent chromatographic purification of the bait–protein complex. The fusion protein specified by this gene can be expressed in large quantities in bacteria and then refolded and purified, or it can be expressed in mammalian cells with a greater certainty of having the correct conformation, albeit in lower yield. In either case, the protein comprises a specific amino acid sequence and conformation (e.g., FLAG sequence; see Section VI.B) that enables high-affinity noncovalent binding to a specific antibody that has itself been covalently attached to agarose beads (Brizzard *et al.*, 1994). In the case of cellular signal transduction studies, transfection of the appropriate mammalian target cells affords the expression of physiological levels of appropriately activated bait protein following stimulation by relevant extracellular factors. Using affinity chromatography, the isolated bait and associated proteins can be analyzed directly on sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) (see Section VI.B).

The 2D gel technology currently plays a critical role in traditional proteomics technology and as a single purification step has no parallel (Anderson and Anderson, 1996; Herbert *et al.*, 1997; Gorg *et al.*, 2000). A significant fraction of the complex mixture of proteins in the proteome can be resolved and arrayed by 2D gel technology. The first dimension of separation is usually based on the isoelectric point of proteins in the presence of various nonionic detergents and denaturing quantities of urea. Conventionally, the second dimension of separation is SDS–PAGE, which separates proteins on the basis of their apparent molecular weight. Detection of proteins as spots in the gel can be accomplished by several different staining methods or, if a radioisotope is used to label proteins metabolically or extrinsically, by autoradiography. Generally these detection methods are compatible with the main procedure used to identify the

protein spots, namely tryptic digestion followed by MS of the peptides generated.

As good as the separation is, 2D gel technology cannot display the whole proteome for several reasons: (1) protein spot congestion and distortion, which results in the overlap and masking of minor spots by other major protein components; (2) lack of solubility of certain hydrophobic proteins (e.g., membrane proteins) in the sample buffer and first dimension, even in the presence of nonionic detergents and in high concentrations of denaturants such as urea; (3) the poor representation of certain classes of proteins that fall outside of the normal separation boundaries. For example, very high or very low molecular weight proteins, and extremely acidic or basic proteins; and (4) finite protein-loading capacity (2–4 mg of total protein for standard 2D gels).

Proteins present at levels below detection by silver stain (less than 1 ng of protein) will be invisible. Such rare proteins often can be detected comfortably by autoradiography in the case of radioiodinated samples but may be of too low abundance to be routinely identified by current MS instrumentation.

Aside from the poor representation of certain classes of proteins, for most analyses of complex samples the greatest concerns with 2D gel technology are spot congestion and overlap, which affects visually based quantification, and loading capacity, because rare molecules of interest are not detectable. To deal with these two concerns other generic purification steps must precede 2D gel analysis (e.g., ion-exchange chromatography and reversed-phase high-performance liquid chromatography, i.e., RP-HPLC) to increase sample loading capacity and to provide additional dimensions of separation. Each additional purification step generates multiple fractions, which expands dramatically the number of 2D gel analyses. Furthermore, the extra chromatographic steps increase the potential for protein losses within the sample (Fig. 3). Thus quantitative differences observed between samples may be due to artifacts of the fractionation procedures.

IV. Quantification in Proteomic Studies

Because nearly all proteomics applications attempt to monitor differences in protein expression in response to disease, toxicity, or some physiological event, accurate relative quantification is critical in identifying true changes in the proteome. There is considerable activity in the commercial and academic sectors to develop diagnostic tests and therapeutics based on differences in protein abundances in diseased cells

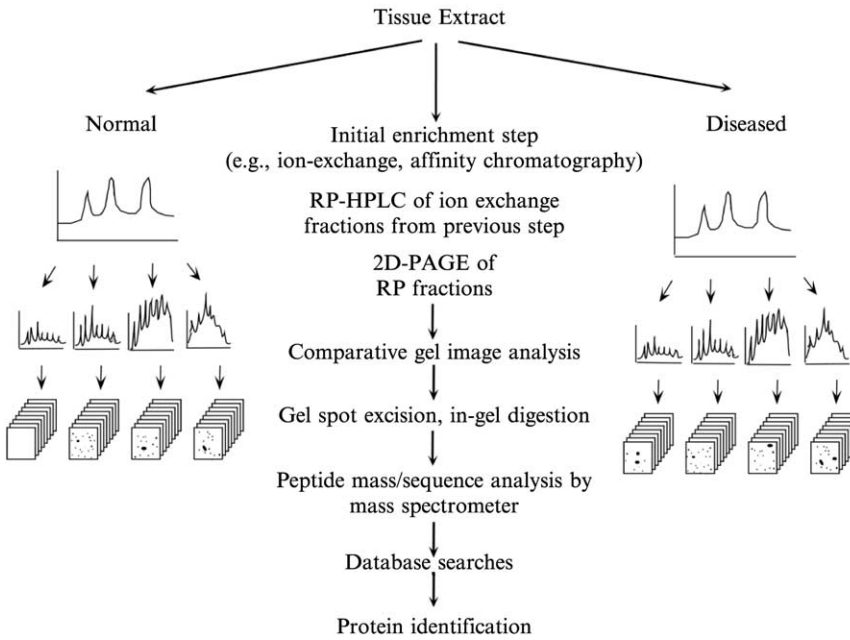


FIG. 3. Generic proteomics scheme. The number of two-dimensional (2D) gels that can be managed will determine the extent of fractionation at each step. The rapid generation of multiple 2D gels, as a result of added enrichment steps and associated fractionation, highlights the need for high-throughput automation of all aspects of gel spot difference analysis (i.e., gel handling, staining, excision, digestion, and MS analysis). [From *Drug Discovery Today*, Vol. 4, Jack H. Hang and Rodney H. Hewick, ''Protemics in Drug Discovery,'' pp. 129–133, Copyright (1999), with permission from Elsevier Science.]

and tissues. Until more recently, quantitative differences between tissue samples were almost exclusively determined by 2D gel spot densitometry after staining (dye or fluorescence) or autoradiography. These detection methods are compatible with the downstream sample-processing steps performed before analysis by MS, but can suffer from lack of a linear staining response. The stained 2D gel images then must be compared quantitatively after first correcting for artifactual differences caused by variations in sample loading and electrophoresis (Appel *et al.*, 1997).

One approach to dealing with the problem of relative quantification is the use of extrinsic labeling procedures in which two samples are differentially labeled, mixed, and processed together. For example, one of the samples is labeled with heavy isotopic reagent and the other with the equivalent light isotopic labeled reagent and then mixed and taken

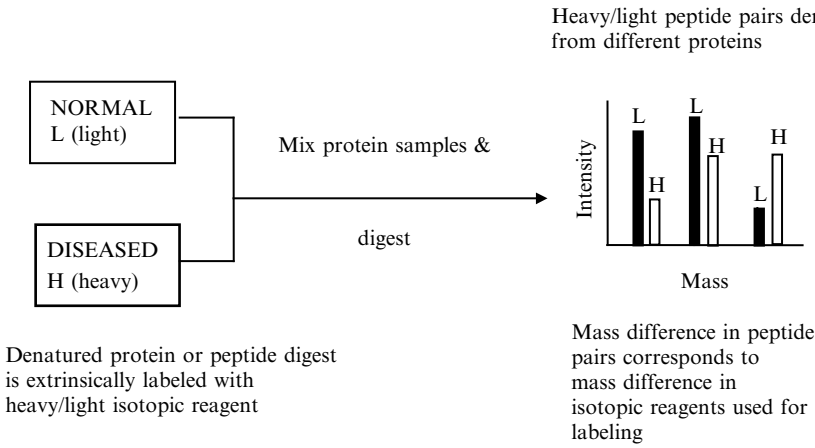


FIG. 4. The use of isotope-labeled reagents for quantitative proteomic analysis by mass spectrometry. Tagging a pair of proteomes with either a heavy or light isotopic version of the modification reagent affords comparison of the abundance of any given protein between the two proteomes.

through the proteomics analysis. Mass spectrometry is then used to quantify the relative abundance of similar proteins from the two samples by analyzing the relative abundance of the heavy and light forms of similar peptides generated via a digestion of the proteins as illustrated in Fig. 4. Although many isotopic labeling schemes have been developed, a few of the more promising techniques are discussed below.

A. ICAT Reagents

The introduction of isotope-coded affinity tags (ICATs) by Aebersold and co-workers for relative quantification of similar proteins in sample comparison promises to improve and simplify proteomic analyses (Gygi *et al.*, 1999a). The ICAT reagent, in either a heavy or light isotopic form, is designed to react quantitatively with cysteinyl residues within a protein. The reagent has also been designed to contain a biotin group to allow for selective purification of labeled peptides immobilized with avidin following trypsinization. With this approach, the heavy and light isotopically labeled proteins can be mixed, subjected to various protein purification techniques, and then proteolytically digested. After MS analysis, the relative abundance of heavy and light forms of similar peptides will reflect the ratio of the proteins from which the peptides were derived. Thus, true

relative expression differences for any given gene product can be measured. Sample-to-sample variability due to processing is minimized, because a single purification of the derivatized sample mixture is performed before MS analysis.

B. Other Protein Quantification Reagents Used in Proteomic Studies

Several other methods have been reported that use stable isotope labeling for quantification in proteomic analysis. Differential metabolic labeling was employed to quantitatively profile comparative systems in yeast (Oda *et al.*, 1999) and *Escherichia coli* (Jensen *et al.*, 1999). The cells were cultured in normal and isotopically enriched or depleted media and pooled. After extracting and processing the proteins, they were analyzed by MS and, similar to the ICAT example described above, pairs of peptides differing only by their isotopic content of heavy isotope were utilized for quantification. Another variation of the procedure used differential isotope labeling of peptides directly after proteolytic digestion followed by pooling of the two comparative samples for MS analysis and quantification (Munchbach *et al.*, 2000). In this example the N-terminal amino group was specifically labeled with either nicotinyl-*N*-hydroxysuccinimide or the deuterated form of this compound.

Our own group has designed a series of reagents that make use of an extrinsic heavy–light isotopic labeling strategy (Qiu *et al.*, 2002). These reagents can be broadly characterized as having a specific chemical reactivity, an isotopically coded linker, and a moiety that improves the yield and identification of the labeled peptides. One class of compounds, termed isotope-coded ionization-enhancing reagents (ICIERs), reacts quantitatively with specific amino acid residues (e.g., one reactivity could be for cysteine; another reactivity could be for lysine, etc.) within the protein molecule. These specific residues would become tagged with a functionality that enhances desorption of the labeled peptides by a laser beam. After proteolysis, ICIER-tagged peptides show greatly increased yield when analyzed by matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) MS compared with that of the unmodified peptides. This reagent will be particularly valuable for high-speed relative quantitative identifications of proteins on 2D gels, possibly eventually obviating the need for staining approaches to detect and identify protein expression differences in gels (see Section VI). This form of analysis will also reveal protein expression differences not seen by staining procedures where overlapping protein spots may totally obscure rare protein differences.

V. Applications

Proteomics has a broad range of applications in drug discovery. The fact that most disease processes and treatments are manifest at the protein level cannot be emphasized too strongly and argues for a prominent role for proteomics in the drug discovery process. At present proteomics is still perceived to be a more difficult technology to use than genomics. However, in many ways, it is an ideal complement to genomics and can be utilized in similar applications.

Because a significant amount of proteomics research is proprietary, being performed by pharmaceutical and proteomics companies, disclosure of the most informative data is understandably restricted. Attempts are made here to present a reasonably representative set of examples to demonstrate the application of the technology.

The use of proteomics in the drug industry can be categorized as follows: (1) drug target identification, (2) drug validation and toxicology, and (3) marker identification and pharmacoproteomics.

The examples given below suggest that proteomics will be successful in its application to various aspects of the drug discovery process.

A. Drug Target Identification

Proteomics technology when applied to drug target identification can be performed from either a global and comprehensive perspective or a focused and functional one. With the global approach, a proteome of interest is displayed, analyzed, and quantified. The process is performed invariably in a comparative mode with a control sample in order to identify proteins that have been differentially expressed and that may be relevant to the etiology or progression of the disease. Because there are no preconceived notions regarding which proteins may be relevant, the most comprehensive display of the protein complement for both the sample and control must be made for this comparison to be useful in identifying potential drug targets. Like genomic expression profiling, proteomic studies of this type will benefit from automation and high-throughput technology to enable massive parallel analyses, but will have the inherent technological disadvantage peculiar to proteins as outlined in Section II.

The common result of comprehensive proteomic analyses tends to be a list of proteins whose relevance needs to be determined. There is a greater likelihood that some proteins within this list identify downstream processes rather than events involved in the initiation of the disease. Therefore, these candidate proteins may be more relevant as markers for the disease.

An example of a global proteomic analysis is the study of rheumatoid arthritis by Oxford GlycoSciences (OGS, Oxfordshire, UK) (Ashton, 1998). In a comparative analysis of synovial fluid from multiple healthy and arthritic patients, 47 proteins that correlated with the disease were identified, out of more than 1500 that were screened. Within this short list, several proteins appeared novel with no known correlates elsewhere in the body. These candidates are being studied as potential therapeutic targets. The emphasis by OGS on a more global analysis using high-throughput, large-scale processing of multiple samples is one approach to improving the chance of success in identifying new drug targets.

Although a significant effort is being put into identifying drug targets in eukaryotic systems, much research is also being directed to studying infectious prokaryotes. Compared with higher organisms, prokaryotes have much smaller genomes and are relatively easy to grow. These two features make global proteomic strategies and analyses less complicated for prokaryotes than for mammalian cells or tissues. Protein expression of virulent strains of *Mycobacterium* associated with tuberculosis was compared with that of nonvirulent strains of the organism to facilitate the identification of potential therapeutics (Jungblut *et al.*, 1999a). When this group studied the proteomes of virulent and nonvirulent strains of the bacteria by 2D-PAGE, 263 mycobacterial proteins were revealed to differ either in quantity or position on the gels. The identities of these proteins were assigned following in-gel digestion and mass spectrometry. These proteins were classified further to assist in determining their correlation with virulence and thus their suitability as drug target candidates.

It appears from the limited published literature concerning the application of global proteomics to the discovery of new drug targets that more time will be needed to determine the success of this approach. The success will ultimately be determined by the validation of one or more of the proteins identified in a comprehensive proteomic analysis, as a viable drug target.

VI. Functional Proteomics Approach

One of the most important aspects of proteomics is its ability to study directly the dynamic interactions of proteins with other proteins, cofactors, substrates, membrane lipids, and nucleic acids. This feature differentiates functional proteomics from the previously described global approach. Because many biological functions are accomplished by large ordered complexes, or cascades of protein interactions of varying complexity, a functional proteomics approach provides a good opportunity for drug target identification.

A. *Protein Complexes*

Protein complexes perform all the essential biological functions such as signal transduction, cell cycle control, cell mobility and cytokinesis, DNA replication, RNA transcription, protein translation, posttranslational modification, translocation, and targeted degradation. In some of these examples, the protein levels of the entire proteome may remain relatively unchanged and yet, because of the formation or disassembly of certain protein complexes, biological processes can be initiated, modulated or terminated.

Protein folding and structural maintenance are helped greatly by interaction with molecular chaperones (Hartl, 1996). The important biological functions of chaperones make them potential targets for antibacterial drugs. This potential is illustrated by a mechanistic study, in which peptides derived from insects (previously shown to be antibacterial) were used as affinity-capture reagents. In an analysis of *E. coli* extracts, GroEL and DnaK were identified as the major interacting protein species (Otvos *et al.*, 2000). Subsequent experiments demonstrated that the antibacterial peptide action was species specific, as these peptides interacted only with the bacterial chaperone DnaK and not with Hsp70, the human equivalent protein.

Proteasomes are large protein complexes in the cytoplasm whose protease activity is regulated by multiple mechanisms (DeMartino and Slaughter, 1999). The biological function of the 26S proteasome (consisting of a 20S catalytic core and two 19S regulatory caps) is to degrade ubiquitinated proteins (Schwartz and Ciechanover, 1999). The 20S proteasome (the catalytic core) is thought to be involved in the elimination of oxidized cellular proteins resulting from oxidative stress (Ullrich *et al.*, 1999; Grune *et al.*, 1996). These proteolytic pathways play an important role in maintaining appropriate levels of different isoforms of regulatory proteins to ensure smooth and prompt occurrence of a variety of cellular processes (Yaron *et al.*, 1997; Zhu *et al.*, 1999; Lin *et al.*, 2000). Abnormalities in the ubiquitin–proteasome pathway have been implicated in many pathological processes including immune/inflammatory responses and muscle-wasting conditions (Schwartz and Ciechanover, 1999). In addition, viral proteins have been shown to interfere with proteasome activity, preventing proper antigen presentation (Levitskaya *et al.*, 1997; Dantuma *et al.*, 2000), or enhancing ubiquitination and proteasomal degradation of tumor suppressor p53 (Scheffner *et al.*, 1990, 1993), leading eventually to cellular transformation and tumorigenesis. In an effort to gain insight into proteasome assembly and proteolytic action, Verma and colleagues (2000) devised a one-step affinity

purification method to isolate intact 26S proteasomes, 19S regulatory caps, and 20S core particles, and then analyzed the purified samples by MS. Their results showed that there are more than 24 proteins interacting with proteasomes in an ATP-dependent manner, implying some regulatory roles of these interacting proteins in the proteolytic activities of the proteasome.

Degradation of target proteins by the 26S proteasome is specifically regulated by ubiquitination, a process that involves three steps catalyzed by E1, E2, and E3 enzymes (Hershko and Ciechanover, 1998). Proteomics technology has been used to determine the specific roles of enzymes within the E2 and E3 classes in attaching ubiquitin to regulatory proteins. Rapid proteolytic degradation of phosphorylated IκBa by ubiquitination and proteasome activity releases NF-κB for its nuclear translocation and subsequent transcription response to a variety of biological stimuli (May and Ghosh, 1998). In an effort to identify the specific E3 enzyme responsible for IκBa ubiquitination, proteins interacting with phosphorylated IκBa were isolated by immunoaffinity purification and analyzed by MS (Yaron *et al.*, 1998). The result shows that E3RSIκB is the core component of E3 activity in a large ubiquitination complex. In eukaryotic cells the proteolytic degradation of mitotic regulators such as securin and cyclin B is facilitated by a ubiquitin ligase E3 called anaphase-promoting complex (APC) or cyclosome (Grossberger *et al.*, 1999; Kramer *et al.*, 2000). Proteomic analysis of immunoaffinity purified human APC revealed two novel components, designated APC11 and Cdc26, which were previously identified only in yeast APC (Gmachl *et al.*, 2000). These researchers further demonstrated that APC11 is the core E3 activity in cyclosomes and that it can ligate ubiquitin to securin in the presence of relevant E1 and E2 enzymes (Gmachler *et al.*, 2000).

The above-described examples indicate the medical significance of the ubiquitination–proteasome pathway and demonstrate how proteomics can provide a direct and comprehensive analysis of protein–protein interactions potentially leading to new targets for therapeutic intervention.

B. Signal Transduction

One class of proteome subsets is multiprotein complexes formed in cells during signaling (Pandey and Mann, 2000). In response to biological stimuli, signaling proteins are altered by posttranslational modifications and changes in conformation, thus enabling them to interact with other proteins in mediating the signal transduction process (Wrana, 2000; Massague and Wotton, 2000; Wallach *et al.*, 1999; Takanashi *et al.*, 1999;

Karin and Ben-Neriah, 2000). Proteins that act as nodes within and between cell signaling pathways are prime targets for modulation by drugs. The dynamic interactions of signaling molecules have been studied by a variety of genetic methods including yeast two-hybrid (reviewed by Uetz and Hughes, 2000; Legrain and Selig, 2000) and protein/peptide display (Cabilly, 1998; Vaughan *et al.*, 1998; Cochran, 2000). Protein studies using MS provide distinct advantages, because they reveal the direct physical presence and identity of multiprotein complexes formed during signal tranduction (Andersen and Mann, 2000), instead of implicating pairwise interactions by genetic screens such as the yeast two-hybrid system (Fields and Song, 1989).

Proteomic studies of signal transduction require successful isolation of signaling protein complexes. Two broad strategies have been employed to isolate these complexes, both involving the use of affinity purification (Ford *et al.*, 1991; Sheibani, 1999). In one strategy a protein suspected to be involved in the signaling complex is expressed exogenously, using recombinant DNA technology, with affinity tags such as a glutathione *S*-transferase (GST), FLAG, hemagglutinin (HA), or a combination of these (Rigaut *et al.*, 1999; Sun *et al.*, 1999). Such recombinant proteins are then used as the purification bait to isolate interacting proteins from relevant biological sources, typically lysates of cells where signaling is occurring. The second strategy involves engineering the endogenous expression of an affinity-tagged signaling protein component in cell lines that contain the signaling pathway of interest (Stroschein *et al.*, 1999; Yang *et al.*, 2000). The lysates from these genetically engineered cell lines are used as the source for affinity purification of signaling complexes containing the tagged component.

Proteins that have a true biological role in signaling complexes interact with one another with varying degrees of strength. Therefore, isolation or washing procedures using buffers of high ionic strength or detergent concentration to remove nonspecifically bound proteins may be inappropriate in the study of weak protein–protein interactions. Thus in protein complex studies, washing conditions must be determined on the basis of the nature of the inquiry. It is also extremely important to compare isolated complexes with equivalent fractions from control samples where signaling has not been initiated. In this way, nonspecific protein interactions can be eliminated from the analysis. After proteomics identification, it is still important to devise some form of biological assay to establish the role and legitimacy of any newly identified proteins in the protein complex. The following section examines applications of proteomic strategies in signal transduction that may be critical in identifying drug targets.

The initial event in signal transduction is the ligand-induced receptor modification. For the T cell antigen receptor (TCR) this alteration includes subunit phosphorylation, recruitment of adaptor proteins, and interaction with intracellular signaling molecules (Grakoui *et al.*, 1999). In a proteomic study of TCR composition by Aebersold and co-workers, TCR complexes were immunoprecipitated with an antibody against CD3ϵ, a subunit of TCR. The protein complexes from detergent lysates of normal and activated cells from T cell line CD11.3 were analyzed by microcapillary liquid chromatography-tandem mass spectrometry (Heller *et al.*, 2000). Combining metabolic labeling and Western blotting with the use of phosphotyrosine-specific antibodies, this study revealed six known TCR subunits and ZAP70 (Cambier and Jensen, 1994), but also showed ligand-dependent phosphorylation of these proteins as well as activation-induced interaction of ZAP70 with TCR. In addition, they identified a novel set of proteins that appeared to interact with TCR. Our laboratory, using a similar strategy involving immunoprecipitation, is attempting to identify the receptors to several bone morphogenetic proteins (BMPs), members of the transforming growth factor-β (TGF-β) superfamily.

Immunoaffinity purification of transiently formed signaling complexes provides an opportunity to find novel factors involved in such complexes. The identification of Ski/Sno proteins as modulators of TGF-β signaling is an example. It is well established that TGF-β family members signal through a family of proteins called SMAD (Miyazono, 2000), and that abnormalities in this signaling network can lead to developmental disorders and cancer (Massague *et al.*, 2000). To study cofactors involved in SMAD signaling, Sun and colleagues (1999) first prepared a GST–SMAD3 fusion protein in which amino acid substitutions were introduced so that the protein mimicked the phosphorylated activated state of SMAD. This ''activated'' bait was used to capture interacting proteins from cell lysates and they were analyzed subsequently by MS. SnoN protein was identified as an 80-kDa species that bound to GST–SMAD3. Using the alternative strategy of endogenous expression discussed previously, Stroschein *et al.* (1999) engineered a 293T cell line so that it expressed a FLAG-tagged C-terminal domain of SMAD4 and then used anti-FLAG antibody to isolate the signaling complex, leading to the identification of SnoN and Ski proto-oncoproteins. After the identification of these protein factors by proteomics technology, both groups carried out biochemical and biological experiments and found that the role of Ski/SnoN proteins is to downmodulate TGF-β-induced, SMAD-mediated activation of transcription (Stroschein *et al.*, 1999; Sun *et al.*, 1999; Xu *et al.*, 2000). Our laboratory has long been interested in TGF-β family members (Wang *et al.*, 1988; Wozney *et al.*, 1988; Wolfman *et al.*, 1997; Lou *et al.*, 1999). The

biological signals of BMPs are mediated by SMAD1, SMAD5, and SMAD8 (Weinstein *et al.*, 2000; Massague *et al.*, 2000). In an effort to understand the signaling process and discover new protein cofactors involved, we are applying the approaches discussed above to cellular systems that respond to certain BMP molecules (our unpublished data). It is anticipated that a more detailed knowledge of these pathways will help develop effective drug treatments for diseases or disease syndromes associated with abnormal BMP function.

Some cytokine receptor signaling involves sequential phosphorylation of tyrosine residues on many signal-transducing molecules. Phosphotyrosine-specific antibodies can be useful tools in proteomics studies of these signaling pathways. Pandey and co-workers used antibodies of this specificity to purify tyrosine phosphorylated proteins from epidermal growth factor (EGF)-treated HeLa cells and then employed MS to identify the captured proteins. In addition to seven proteins previously known to be phosphorylated at tyrosine residues on receptor activation, two novel signal tranduction molecules, designated Vav-2 and STAM2 (Pandey *et al.*, 2000a,b), were identified. Further biological experiments established that both Vav-2 and STAM2 are phosphorylated at tyrosine residues on activation of epidermal growth factor receptor (EGF-R) or platelet-derived growth factor receptor (PDGF-R). It was also shown that Vav-2 is a direct substrate of the two receptors, and that STAM2 is phosphorylated by the Jac family of kinases. A similar approach was applied in searching for FGF-2 signaling pathway components in MCF-7 cancer cells (Vercoutter-Edouart *et al.*, 2000). Using immunoaffinity purification with anti-phosphotyrosine antibodies, combined with MS analysis, the researchers identified the phosphorylation of cyclin D2 as one link in the fibroblast growth factor 2 (FGF2) signaling pathway involving the mitogen-activater protein (MAP) kinase cascade.

C. Drug Target Validation/Toxicology

The utility of proteomics for drug validation and toxicology studies has found support in the pharmaceutical proteomics sectors. Proteomics technology can be used to profile proteins in largely acellular bodily fluids not suitable for analysis by genomic technologies. In the case of cellular studies, data obtained from proteomics can be used to complement protein expression data inferred from genomic or transcriptional studies. By comparing protein expression profiles of untreated controls, samples following exposure to standard drugs, and samples following exposure to novel drugs, determinations can be made as to drug efficacy and toxicity.

A major proponent for this use of proteomics is the company Large Scale Proteomics (LSP; Rockville, MD). LSP has developed a program to construct a database of protein expression profiles as a function of various drug treatments (Anderson *et al.*, 1996b). Their Molecular Effects of Drugs Database could prove valuable for clients who desire to compare their drug with known therapeutics with respect to efficacy or toxicity.

The effect of lovastatin, a lipid-lowering drug, on cholesterol biosynthesis and liver protein expression is a prime example of this application of proteomics by LSP. Initial studies validated the mechanism of action of lovastatin by demonstrating its effect on proteins involved with cholesterol metabolism (Anderson *et al.*, 1991). Subsequent mining and interpretation of the proteomics data have led to a more detailed appreciation of the various metabolic pathways in liver involved with lovastatin treatment (Steiner *et al.*, 2000). This result in turn has presented the investigators with insights into new drug targets for affecting cholesterol metabolism, as well as a better understanding of the potential toxicity and activity of lovastatin.

A similar large-scale effort to utilize proteomics for drug validation was initiated by the Developmental Therapeutics Program of the National Cancer Institute (Myers *et al.*, 1997). The aim in this study was to create a database of protein expression profiles from 60 different cancer cell lines in response to treatment with thousands of compounds from their chemical and natural products library. Specific correlations between protein expression patterns, cell growth inhibitory activity, and the pharmacology of these screened compounds were made by mining this database with a variety of bioinformatics analysis tools. It is hoped that the knowledge derived from analyzing these correlations can be incorporated into more rational drug design and individualized therapy.

Perhaps the prime exemplar for the use of proteomics in drug toxicology was the study of cyclosporin A toxicity in the kidney (Steiner *et al.*, 1996a), a common side effect for this immunosuppressive drug. Comparative protein expression profiling demonstrated a cyclosporin A-dependent decrease in expression of calbindin D. This protein is involved in calcium transport and its absence could explain the toxic side effect of kidney calcification observed with cyclosporin A. This finding has been confirmed by the discovery of a cyclosporin A-dependent nephrotoxicity and calbindin absence in other species including humans (Aicher *et al.*, 1998) and by an inverse correlation of immunosuppressive activity of cyclosporin A derivatives with calbindin expression (Aicher *et al.*, 1997).

Many examples exist of the use of proteomics for drug validation and toxicity as reviewed by Steiner and Witzmann (2000). Only a few of these examples are presented to conclude the discussion of this application.

A comparative study of the antihistamine methapyrilene, which was shown to be a rat hepatocarcinogen, and its noncarcinogenic derivative was performed to investigate possible mechanisms of carcinogenicity (Cunningham *et al.*, 1995). A serum protein expression profile from a rat inflammation model treated with indomethacin could potentially be used as a screen for drug efficacy and validation of new nonsteroidal antiinflammatory drugs (Eberini *et al.*, 1999). Proteomic analyses of liver extracts have been performed to study the mechanism of action or toxicity of etomoxir, a potential treatment for diabetes (Steiner *et al.*, 1996b); peroxisome proliferators (Anderson *et al.*, 1996a); oltipraz, an inhibitor of aflatoxin-induced liver cancer (Anderson *et al.*, 1995); and a hypoglycemic-inducing compound (Arce *et al.*, 1998).

Many of these proteomic studies, although sound in design and promising in their results, may require additional experimentation and further data interpretation before their success in contributing to the drug lead optimization process can be truly assessed.

D. Marker Identification

Programs initiated to identify potential drug targets for a disease may instead lead to the identification of putative markers for the disease. In many cases this is due to the limitations of the technology, which allows for the detection of only the most abundant proteins in a general screening approach, those that are more likely to be surrogate markers of the disease.

There are numerous examples of marker identifications in proteomic studies and only a partial listing is given here. Perhaps one of the best examples to date is the identification of psoriasin as a marker for bladder squamous cell carcinoma by Celis *et al.* (1996). This protein was identified from a readily available diagnostic sample source that does not require invasive procedures—the urine of patients. The specificity of this bladder squamous cell carcinoma marker was validated in subsequent experiments by the demonstration of its apparent absence in cells from the healthy urinary tract. However, further studies have demonstrated that healthy female subjects also exhibit the presence of psoriasin in their urine (Ostergaard *et al.*, 1999), making this a more appropriate bladder cancer marker for the male population. This complication emphasizes the need for closely matching samples and appropriate controls.

Another leading cause of death in humans, heart disease, is being actively studied by proteomic technology (Dunn, 2000). Extensive 2D-PAGE databases of cardiac proteins have been generated to lay the

foundation for a global analysis of markers associated with cardiac disease (Corbett *et al.*, 1994; Jungblut *et al.*, 1994). Preliminary protein identifications have been made from dilated cardiomyopathy studies (Jungblut *et al.*, 1999b) and from an animal model for cardiac hypertrophy (Arnott *et al.*, 1998). Further expansion of these studies will require larger data sets for statistical significance and disease models to monitor progression of the disease to demonstrate the relevance of the markers. The use of primary cell cultures as a more convenient and homogeneous surrogate tissue source for the study of disease states needs to be considered carefully, because cells can alter their phenotype *in vitro* over time (Celis *et al.*, 1999).

Disorders of the brain have been subjected to proteomic analysis, with some preliminary biomarkers having been identified (Edgar *et al.*, 1999, 2000; Johnston-Wilson *et al.*, 2000). Whether these proteins are related to the etiology of the disease or can serve as markers of the disease is still unclear and will depend on further research. Breast cancer is another major disease intensively studied by proteomics. In a collaboration between Oxford GlycoSciences (Oxfordshire, UK) and the Ludwig Institute for Cancer Research (London, UK) the obstacle of tissue heterogeneity, which often leads to confusing and erroneous differential comparisons between control and disease tissue samples, was addressed (Page *et al.*, 1999). This study identified differences between the proteome of breast lumenal and myoepithelial cells. Future comparative studies of breast tumor tissue, which is overwhelmingly of lumenal origin, will be inherently more meaningful because an exact comparison of the appropriate cell types can now be made.

Any comparative proteomic study will always be dependent on the quality of the samples to be analyzed. Direct comparisons require the samples to be as similar as possible for the results to be meaningful. Tissue samples, rather than cell lines, are preferable for analysis but by nature are heterogeneous at the cellular level. A relatively new technology, laser capture microdissection (LCM), has been applied to proteomic studies to address the issue of isolating homogeneous samples for comparison. LCM has the capability to collect from tissue sections specific cell populations, which are then subjected to standard proteomic analysis (Banks *et al.*, 1999; Emmert-Buck *et al.*, 2000). The potential of this technology for generating clean samples is clear. The major limitation is in generating sufficient quantities of sample for analysis. The effort needed to isolate enough sample for proteomic analysis of the less abundant cellular proteins may be prohibitive at this time. With the expected continued improvements in the sensitivity of proteomic analysis, this limitation may not be an issue in future.

The application of proteomics for marker identification in prokaryotes is appropriate for the same reasons outlined earlier for target identification. The same approach is taken whereby virulent and nonvirulent proteomes or subcellular fractions are compared to identify unique markers of disease such as tuberculosis (Hendrickson *et al.*, 2000), or to identify novel antigenic determinants that can be crucial for vaccine development (Chakravarti *et al.*, 2000). It is likely that proteomics will see its earliest success and make the most impact with drug and vaccine development for prokaryote-mediated diseases.

An extensive list of applications of proteomics to marker identification exists. A partial listing includes proteomic characterization of rat serum for generation of a reference database (Haynes *et al.*, 1998), rat islets of Langerhans for studying diabetes (Andersen *et al.*, 1997), rabbit kidney for lead nephrotoxicity (Kanitz *et al.*, 1999), and human nasal and bronchoalveolar lavage fluid for airway disorders such as asthma (Lindahl *et al.*, 1999; Wattiez *et al.*, 2000). A comparative study of colon mucosa was also performed for analysis of colorectal tumors (Stulik *et al.*, 1999) and cerebrospinal fluid for neurodegenerative diseases (Sickmann *et al.*, 2000).

Ultimately, as discussed previously, the quality of the samples to be compared will play a major role in determining the value of the data generated from any proteomic analysis.

VII. Future Trends in Proteomics

Proteomics technology will continue to evolve and improve. Some of the major issues that need to be considered include increased automation and sensitivity, miniaturization, and better bioinformatics. Successfully addressing these issues will result in a more comprehensive analysis of the proteome. Some novel technologies have been developed that begin to address some of these important issues. In an attempt to increase the throughput for analysis of proteins from polyacrylamide gels various groups have tried to automate the interface between the resolving gel and the mass spectrometric analysis. One approach is to directly analyze resolved proteins in a polyacrylamide gel by MALDI-TOF MS (Ogorzalek Loo *et al.*, 1997). A second approach expanded on this idea by transferring the proteins from a one-dimensional or two-dimensional polyacrylamide gel onto a polyvinylidene difluoride (PVDF) membrane while incorporating the digestion of the proteins as part of the blotting process (Bienvenut *et al.*, 1999; Binz *et al.*, 1999). The blotted peptides could then be scanned by MALDI-TOF mass spectrometry in an automated manner for protein identifications. These techniques are still in the process of

development and refinement and it remains to be determined whether they will be able to achieve their aims.

Two-dimensional gel electrophoresis technology is still the most proved means for resolving and displaying large numbers of proteins at one time. It is clear that, even in the case of relatively simple samples, many more protein species are present in 2D gels than can be detected by the usual methods of visualization. Two-dimensional gels comprising ^{125}I-labeled protein samples show a greater number of spot compared with gels that have been stained with silver reagents. Regions on 2D gels that appear blank after silver staining can often be shown to be populated with multiple trace proteins after autoradiography. By interfacing the technologies for post-2D gel sample manipulation with MS detection, losses from staining or blotting can be avoided. In this approach the whole gel is sectioned into small uniform pieces (pixels) for eventual automated in-gel digestion, processing, and analysis by MS. Our own instrument design using this strategy is in early development, but on the basis of this general concept, scientists at Hoffmann-LaRoche (Basel, Switzerland) have progressed further with such automated downstream processing (Lahm and Langen, 2000).

A growing trend in proteomic studies is to perform analyses without the use of 2D gels. As mentioned previously, the loading limitations and ultimate resolving power of 2D gels, though excellent, falls a long way short of being able to detect rare protein biological markers in complex bodily fluids such as plasma, cerebrospinal fluid, synovial fluid and urine. For this reason, multiple chromatographic steps need to precede 2D gel analyses to reduce sample complexity, and consequently hundreds of fractions are generated and hundreds of 2D gels need to be analyzed and accurately compared (Fig. 3). Non-2D gel proteomic studies rely primarily on the preferred chromatographic properties of peptides over proteins (e.g., peak shape, solubility, chromatographic yield) in the various purification steps and in the final quantification and identification of proteins in complex protein mixtures (see the right-hand pathway in Fig. 2). In this approach (Link *et al.*, 1999) compared samples or compared equivalent fractions from samples digested with trypsin or some other convenient protease. Ideally, at this point the two peptide populations are differentially labeled with heavy or light isotope (see Section IV) such that quantitative comparisons can be made at the end of the analysis. The approach relies heavily on an efficient and complete digestion so that peptides are generated from every protein species whether it is abundant or scarce. Because digestion is performed after protein denaturation, some soluble peptides may be generated from hydrophobic proteins and thus these proteins will be represented in the final analysis.

Simplification of the peptide mixtures is essential for detection of rare peptides (i.e., peptides efficiently generated from rare proteins). One approach to achieving this goal is to use serial, nonoverlapping chromatographic separations culminating in direct on-line chromatographies such as cation exchange and RP-HPLC (Fig. 5). In this way, multiple liquid fractions containing peptides are generated in a form suitable for direct MS analysis. This strategy is in contrast to the generation of multiple fractions containing proteins that would need to be resolved by 2D gel analysis before further processing could take place. Liquid fractions collectively representing the complete proteome in the form of peptides lend themselves more easily to massive automated analysis than do multiple 2D gels. In such a comprehensive comparison of important biological samples, the rate-limiting step is likely to be the on-line RP-HPLC analysis. This period of time multiplied by the total number of fractions generated from the procedure approximates the total time of a completely automated process to evaluate the proteome. Although this length of time may be considerable (several days or more of automated operation) the experiment is likely to be thorough and quantitatively meaningful. Clearly, in this approach more emphasis is placed on the ever-improving power of MS instrumentation to resolve, identify, and quantify relatively complex peptide mixtures, which in most cases display huge differences in abundance between the individual peptide species. The data obtained from such broad proteomic analyses could be used to establish which proteins change as a function of diseased state and thus which are good disease marker candidates for clinical screening using some form of multiplex ELISA or protein chip format.

If proteomics technology is to be automated for high-throughput analysis, the most efficient designs will eventually incorporate miniaturization and microfluidics to address the issues and limitations stated earlier. Ultimately protein separation technologies, sample handling, digestion, and interfacing with the mass spectrometers will be fabricated into microfluidic chips. This format easily lends itself to high-throughput automation and the miniaturized scale decreases sample requirements while optimizing recoveries for improved sensitivities. Initial applications of this technology for both protein and nucleic acid analysis have been reported (Figeys *et al.*, 1998; He *et al.*, 1998; Figeys, 1999; Zhang *et al.*, 1999; Wang *et al.*, 2000; Oleschuk and Harrison, 2000).

The ultimate goal in the proteomics field is the development of a protein chip that could be used to analyze changes in the proteome in a massively parallel way that is analogous to the use of DNA chips in monitoring transcriptional changes. Using this approach, diseases could be monitored in bodily fluids, such as plasma, as a multiprotein

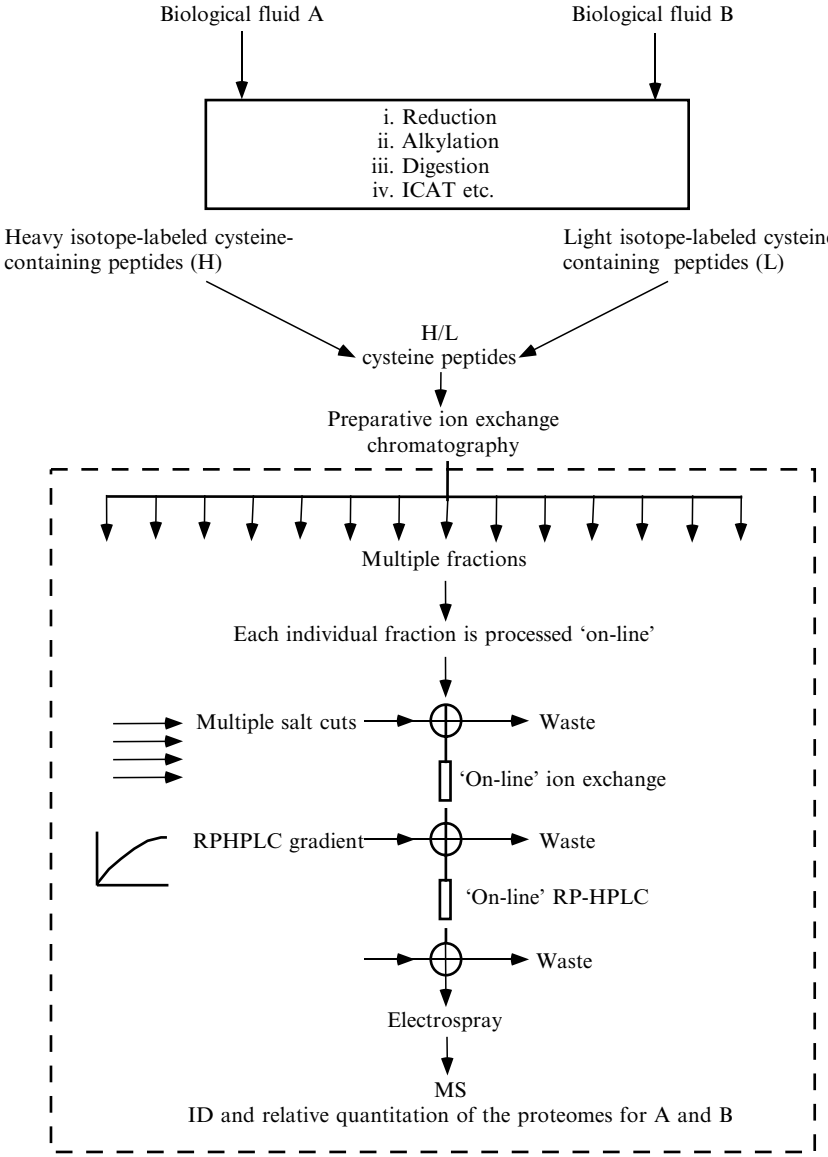


FIG. 5. Scheme for proteomics analysis in the absence of two-dimensional gels. In the scheme shown, samples to be compared are separately denatured, proteolytically digested, and then isotopically labeled. Cysteine-containing peptides are affinity purified with ICAT-type reagents. The two populations of differentially labeled cysteine-containing peptides are then mixed and subjected to serial chromatographies. The last two separation steps are directly on-line to the electrospray MS system. The boxed area depicts the part of the scheme that is readily automated.

''fingerprint.'' Characteristic patterns of protein intensities from such a chip could also be used in the case of cellular extracts to define new drug targets and to preview the potentially toxic side effects of new drug candidates by detecting subtle changes in protein expression.

Protein chips are difficult to design because protein selectivity is dependent on retention of the native conformational state in many affinity-based capture procedures. Initial work on protein chips has focused on the capturing agent being either specific antibodies generated in vast numbers from phage libraries (Winter *et al.*, 1994; Hoogenboom and Chames, 2000) or DNA aptamers generated by PCR technology (Brody *et al.*, 1999). These molecules can then be ''printed'' onto an appropriately treated surface, such as glass, as an ordered array, using robotic procedures (MacBeath and Schreiber, 2000; Haab *et al.*, 2001). Before chip fabrication in either case, the capturing agent must be selected for optimal binding to the desired target protein population. In the case of an all-encompassing proteomics chip, validation of selectivity for each aptamer or antibody with its cognate binding protein would require the generation of thousands of proteins of correct conformation. It thus follows that the first protein chips will be relatively small arrays of custom-selected antibodies or aptamers to a relatively small number of proteins of special interest. Another approach, which would not require the generation of huge numbers of correctly folded proteins, would be to use protein database information to synthesize target tryptic peptides (e.g., cysteine-containing peptides) from large numbers of proteins and to use these for aptamer or antibody selection. The sample monitored by this form of chip would be a thorough tryptic digest of the denatured protein substrate. With this sample treatment, selectivity of the aptamer or antibody to the synthesized target peptide would be more likely to match that of the experimentally generated peptide, because the peptide from the digest would have fewer conformational constraints than the equivalent domain in the native protein. Cysteine-containing peptides in the sample could be labeled with a fluorescent tag to quantify the relative abundance of similar peptides in the samples to be compared and thus the relative abundance of their parent proteins.

Protein chips are also being developed to explore protein–protein interactions. In this case, multiple protein ''baits'' are arrayed on the chip surface and allowed to interact with the tissue sample of interest, such as a cell lysate. After mild washing or after employing some form of chemical cross-linkage, followed by proteolysis, the appropriate protein–protein interactions can be identified by MS. Significant progress has been made in the chip surface chemistry used to enable this type of analysis by companies such as Zyomyx (Hayward, CA). Flat surfaces can be created

and treated to facilitate consistent binding and preservation of correctly folded protein. In addition, proteins can be packed in a dense and uniform configuration with the individual molecules sufficiently spaced and appropriately oriented to interact with other proteins in the sample milieu. The exposed surface of the chip between attached protein molecules has also been demonstrated to be nondenaturing and to have a low propensity for nonspecific protein binding.

In conclusion, there are numerous examples of the application of proteomics for drug discovery and reports of interesting leads, but to the best of our knowledge there has been no drug or diagnostic that has been commercialized as a direct result. The situation is not dissimilar for the more heralded genomics technology. In both cases further mining and validation of enormous quantities of data will be necessary for these technologies to realize their full potential. In the case of proteomics, it is a relatively new technology that is evolving at a tremendous pace. With developments in sample quantification and sample handling, including microfabrication and miniaturized on-line chromatography, and with the continuous improvements in the design of mass spectrometers, the diagnostic power of proteomics will likely rise by orders of magnitude. The result will be a more accurate understanding of changes in the proteome and a better knowledge of life processes. Improved drug discovery will be the natural outcome of this newly acquired knowledge.

References

Aicher, L., Meier, G., Norcross, A. J., Jakubowski, J., Varela, M. C., Corhdier, A., and Steiner, S. (1997). Decrease in kidney calbindin-D 28kDa as a possible mechanism mediating cyclosporine A- and FK-506-induced calciuria and tubular mineralization. *Biochem. Pharmacol.* **53,** 723–731.

Aicher, L., Wahl, D., Arce, A., Grenet, O., and Steiner, S. (1998). New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis* **19,** 1998–2003.

Andersen, H. U., Fey, S. J., Larsen, P. M., Nawrocki, A., Hejnaes, K. R., Mandrup-Poulsen, T., and Nerup, J. (1997). Interleukin-1β induced changes in the protein expression of rat islets: A computerized database. *Electrophoresis* **18,** 2091–2103.

Andersen, J. S., and Mann, M. (2000). Functional genomics by mass spectrometry. *FEBS Lett.* **480,** 25–31.

Anderson, L., and Seilhamer, J. (1997). A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18,** 533–537.

Anderson, L., Steele, V. K., Kelloff, G. J., and Sharma, S. (1995). Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. *J. Cell Biochem. Suppl.* **22,** 108–116.

Anderson, N. G., and Anderson, N. L. (1996). Twenty years of two-dimensional electrophoresis: Past, present and future. *Electrophoresis* **17,** 443–453.

Anderson, N. L., Esquer-Blasco, R., Hofmann, J. P., and Anderson, N. G. (1991). A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. *Electrophoresis* **12,** 907–930.

Anderson, N. L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., and Eacho, P. (1996). The effects of peroxisome proliferators on protein abundances in mouse liver. *Toxicol. Appl. Pharmacol.* **137,** 75–89.

Anderson, N. L., Taylor, J., Hofmann, J. P., Esquer-Blasco, R., Swift, S., and Anderson, N. G. (1996). Simultaneous measurement of hundreds of liver proteins: Application in assessment of liver function. *Toxicol. Pathol.* **24,** 72–76.

Appel, R. D., Palagi, P. M., Walther, D., Vargas, J. R., Sanchez, J. C., Ravier, F., Pasquali, C., and Hochstrasser, D. F. (1997). Melanie II: A third-generation software package for analysis of two-dimensional electrophoresis images. I. Features and user interface. *Electrophoresis* **18,** 2724–2734.

Arce, A., Aicher, L., Wahl, D., Anderson, N. L., Meheus, L., Raymackers, J., Cordier, A., and Steiner, S. (1998). Changes in the liver protein pattern of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. *Life Sci.* **63,** 2243–2250.

Arnott, D., O'Connell, K. L., King, K. L., and Stults, J. T. (1998). An integrated approach to proteome analysis: Identification of proteins associated with cardiac hypertrophy. *Anal. Biochem.* **258,** 1–18.

Ashton, C. (1998). Proteomics—the new watchword for biotech. *Scrip Magazine* **65,** 32–35.

Banks, R. E., Dunn, M. J., Forbes, M. A., Stanley, A., Pappin, D., Naven, T., Gough, M., Harnden, P., and Selby, P. J. (1999). The potential use of laser capture microdissection to selectively obtain distinct populations of cells for proteomic analysis—preliminary findings. *Electrophoresis* **20,** 689–700.

Bienvenut, W. V., Sanchez, J. C., Karmime, A., Rouge, V., Rose, K., Binz, P. A., and Hochstrasser, D. F. (1999). Toward a clinical molecular scanner for proteome research: Parallel protein chemical processing before and during Western blot. *Anal. Chem.* **71,** 4800–4807.

Binz, P. A., Muller, M., Walther, D., Bienvenut, W. V., Gras, R., Hoogland, C., Bouchet, G., Gasteiger, E., Fabbretti, R., Gay, S., Palagi, P., Wilkins, M. R., Rouge, V., Tonella, L., Paesano, S., Rossellat, G., Karmime, A., Bairoch, A., Sanchez, J. C., Appel, R. D., and Hochstrasser, D. F. (1999). Toward a clinical molecular scanner for proteome research: Parallel protein chemical processing before and during Western blot. *Anal. Chem.* **71,** 4981–4988.

Brizzard, B. L., Chubet, R. G., and Vizard, D. L. (1994). Immunoaffinity purification of FLAG epitope-tagged bacterial alkaline phosphatase using a novel monoclonal antibody and peptide elution. *Biotechniques* **16,** 730–735.

Brody, E. N., Willis, M. C., Smith, J. D., Jayasena, S., Zichi, D., and Gold, L. (1999). The use of aptamers in large arrays for molecular diagnostics. *Mol. Diagn.* **4,** 381–388.

Cabilly, S., Ed. (1998). Combinatorial peptide library protocols. *Methods Mol. Biol.* **87**.

Cambier, J. C., and Jensen, W. A. (1994). The hetero-oligomeric antigen receptor complex and its coupling to cytoplasmic effectors. *Curr. Opin. Genet. Dev.* **4,** 55–63.

Celis, A., Rasmussen, H. H., Celis, P., Basse, B., Lauridsen, J. B., Ratz, G., Hein, B., Ostergaard, M., Wolf, H., Orntoft, T., and Celis, J. E. (1999). Short-term culturing of low-grade superficial bladder transitional cell carcinomas leads to changes in the expression levels of several proteins involved in key cellular activities. *Electrophoresis* **20,** 355–361.

Celis, J. E., Rasmussen, H. H., Vorum, H., Madsen, P., Honore, B., Wolf, H., and Orntoft, T. F. (1996). Bladder squamous cell carcinomas express psoriasin and externalize it to the urine. *J. Urol.* **155,** 2105–2112.

Chakravarti, D. N., Fiske, M. J., Fletcher, L. D., and Zagursky, R. J. (2000). Application of genomics and proteomics for identification of bacterial gene products as potential vaccine candidates. *Vaccine* **19,** 601–612.

Cho, G., Keefe, A. D., Liu, R., Wilson, D. S., and Szostak, J. W. (2000). Constructing high complexity synthetic libraries of long ORFs using in vitro selection. *J. Mol. Biol.* **297,** 309–319.

Cochran, A. G. (2000). Antagonists of protein–protein interactions. *Chem. Biol.* **7,** R85–R94.

Colas, P., and Brent, R. (1998). The impact of two-hybrid and related methods on biotechnology. *Trends Biotechnol.* **16,** 355–363.

Corbett, J. M., Wheeler, C. H., Baker, C. S., Yacoub, M. H., and Dunn, M. J. (1994). The human myocardial two-dimensional gel protein database: Update 1994. *Electrophoresis* **15,** 1459–1465.

Cunningham, M. L., Pippin, L. L., Anderson, N. L., and Wenk, M. L. (1995). The hepatocarcinogen methapyrilene but not the analog pyrilamine induces sustained hepatocellular replication and protein alterations in F344 rats in a 13-week feed study. *Toxicol. Appl. Pharmacol.* **131,** 216–223.

Dantuma, N. P., Heessen, S., Lindsten, K., Jellne, M., and Masucci, M. G. (2000). Inhibition of proteasomal degradation by the Gly-Ala repeat of Epstein-Barr virus is influenced by the length of the repeat and the strength of the degradation signal. *Proc. Natl. Acad. Sci. USA* **97,** 8381–8385.

DeMartino, G. N., and Slaughter, C. A. (1999). The proteasome, a novel protease regulated by multiple mechanisms. *J. Biol. Chem.* **274,** 22123–22126.

Dunn, M. J. (2000). Studying heart disease using the proteomic approach. *Drug Discov. Today* **5,** 76–84.

Eberini, I., Miller, I., Zancan, V., Bolego, C., Puglisi, L., Gemeiner, M., and Gianazza, E. (1999). Time-course of acute-phase protein expression and its modulation by indomethacine. *Electrophoresis* **20,** 846–853.

Edgar, P. F., Douglas, J. E., Knight, C., Cooper, G. J., Faull, R. L., and Kydd, R. (1999). Proteome map of the human hippocampus. *Hippocampus* **9,** 644–650.

Edgar, P. F., Douglas, J. E., Cooper, G. J., Dean, B., Kydd, R., and Faull, R. L. (2000). Comparative proteome analysis of the hippocampus implicates chromosome 6q in schizophrenia. *Mol. Psychiatry* **5,** 85–90.

Emmert-Buck, M. R., Gillespie, J. W., Paweletz, C. P., Ornstein, D. K., Basrur, V., Appella, E., Wang, Q. H., Huang, J., Hu, N., Taylor, P., and Petricoin, E. F., III (2000). An approach to proteomic analysis of human tumors. *Mol. Carcinog.* **27,** 158–165.

Eng, J. K., McCormack, A. L., and Yates, J. R. (1994). 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.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein–protein interactions. *Nature* **340,** 245–246.

Figeys, D. (1999). Array and lab on a chip technology for protein characterization. *Curr. Opin. Mol. Ther.* **1,** 685–694.

Figeys, D., Gygi, S. P., McKinnon, G., and Aebersold, R. (1998). An integrated microfluidics-tandem mass spectrometry system for automated protein analysis. *Anal. Chem.* **70,** 3728–3734.

Ford, C. F., Suominen, I., and Glatz, C. E. (1991). Fusion tails for the recovery and purification of recombinant proteins. *Protein Expr. Purif.* **2,** 95–107.

Gmachl, M., Gieffers, C., Podtelejnikov, A. V., Mann, M., and Peters, J.-M. (2000). The RING-H2 finger protein APC11 and the E2 enzyme UBC4 are sufficient to ubiquitinate substrates of the anaphase-promoting complex. *Proc. Natl. Acad. Sci. USA* **97,** 8973–8978.

Gooley, A. A., and Packer, N. H. (1997). The importance of co- and post-translational modifications in proteome projects. *In* ''Proteome Research: New Frontiers in Functional Genomics'' (M. R. Wilkins, K. L. Williams, R. D. Appel, and D. F. Hochstrasser, Eds.), pp. 65–91. Springer-Verlag, Berlin.

Gorg, A., Obermaier, C., Boguth, G., Harder, A., Scheibe, B., Wildgruber, R., and Weiss, W. (2000). The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **21,** 1037–1053.

Grakoui, A., Bromley, S. K., Sumen, C., Davis, M. M., Shaw, A. S., Allen, P. M., and Dustin, M. L. (1999). The immunological synapse: A molecular machine controlling T cell activation. *Science* **285,** 221–227.

Grossberger, R., Gieffers, C., Zachariae, W., Podtelejnikov, A. V., Schleiffer, A., Nasmyth, K., Mann, M., and Peters, J. M. (1999). Characterization of the DOC1/APC10 subunit of the yeast and the human anaphase-promoting complex. *J. Biol. Chem.* **274,** 14500–14507.

Grune, T., Reinheckel, T., and Davies, K. J. A. (1996). Degradation of oxidized proteins in K562 human hematopoietic cells by proteasome. *J. Biol. Chem.* **26,** 15504–15509.

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999a). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17,** 994–999.

Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999b). Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.* **19,** 1720–1730.

Haab, B. B., Dunham, M., and Brown, P. (2001). Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol.* **2,** RESEARCH 0004.

Hartl, F. U. (1996). Molecular chaperones in cellular protein folding. *Nature* **381,** 571–579.

Haynes, P., Miller, I., Aebersold, R., Gemeiner, M., Eberini, I., Lovati, M. R., Manzoni, C., Vignati, M., and Gianazza, E. (1998). Proteins of rat serum. I. Establishing a reference two-dimensional electrophoresis map by immunodetection and microbore high performance liquid chromatography-electrospray mass spectrometry. *Electrophoresis* **19,** 1484–1492.

He, B., Tait, N., and Regnier, F. (1998). Fabrication of nanocolumns for liquid chromatography. *Anal. Chem.* **70,** 3790–3797.

Heller, M., Goodlett, D. R., Watts, J. D., and Aebersold, R. (2000). A comprehensive characterization of the T-cell antigen receptor complex composition by microcapillary liquid chromatography-tandem mass spectrometry. *Electrophoresis* **21,** 2180–2195.

Hendrickson, R. C., Douglass, J. F., Reynolds, L. D., McNeill, P. D., Carter, D., Reed, S. G., and Houghton, R. L. (2000). Mass spectrometric identification of mtb81, a novel serological marker for tuberculosis. *J. Clin. Microbiol.* **38,** 2354–2361.

Henzel, W. J., Billeci, T. M., Stults, J. T., Wong, S. C., Grimley, C., and Watanabe, C. (1993). Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA* **90,** 5011–5015.

Herbert, B. R., Sanchez, J.-C., and Bini, L. (1997). Two-dimensional electrophoresis: The state of the art and future directions. *In* ''Proteome Research: New Frontiers in Functional Genomics'' (M. R. Wilkins, K. L. Williams, R. D. Appel, and D. F. Hochstrasser, Eds.), pp. 13–33. Springer-Verlag, Berlin.

Hershko, A., and Ciechanover, A. (1998). The ubiquitin system. *Annu. Rev. Biochem.* **67,** 425–479.

Hoogenboom, H. R., and Chames, P. (2000). Natural and designer binding sites made by phage display technology. *Immunol. Today* **21,** 371–378.

Hunt, D. F., Henderson, R. A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., Cox, A. L., Appella, E., and Engelhard, V. H. (1992). Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255,** 1261–1263.

James, P., Quadroni, M., Carafoli, E., and Gonnet, G. (1993). Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* **195,** 58–64.

Jensen, P. K., Pasa-Tolic, L., Peden, K. K., Martinovic, S., Lipton, M. S., Anderson, G. A., Tolic, N., Wong, K. K., and Smith, R. D. (2000). Mass spectrometric detection for capillary isoelectric focusing separations of complex protein mixtures. *Electrophoresis* **21,** 1372–1380.

Johnston-Wilson, N. L., Sims, C. D., Hofmann, J. P., Anderson, L., Shore, A. D., Torrey, E. F., and Yolken, R. H. (2000). Disease-specific alterations in frontal cortex brain proteins in schizophrenia, bipolar disorder, and major depressive disorder. *Mol. Psychiatry* **5,** 142–149.

Jungblut, P., Otto, A., Zeindl-Eberhart, E., Plessner, K. P., Knecht, M., Regitz-Zagrosek, V., Fleck, E., and Wittmann-Liebold, B. (1994). Protein composition of the human heart: The construction of a myocardial two-dimensional electrophoresis database. *Electrophoresis* **15,** 685–707.

Jungblut, P. R., Schaible, U. E., Mollenkopf, H. J., Zimny-Arndt, U., Raupach, B., Mattow, J., Halada, P., Lamer, S., Hagens, K., and Kaufmann, S. H. (1999a). Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: Towards functional genomics of microbial pathogens. *Mol. Microbiol.* **33,** 1103–1117.

Jungblut, P. R., Zimny-Arndt, U., Zeindl-Eberhart, E., Stulik, J., Koupilova, K., Pleissner, K. P., Otto, A., Muller, E. C., Sokolowska-Kohler, W., Grabher, G., and Stoffler, G. (1999b). Proteomics in human disease: Cancer, heart and infectious diseases. *Electrophoresis* **20,** 2100–2110.

Kanitz, M. H., Witzmann, F. A., Zhu, H., Fultz, C. D., Skaggs, S., Moorman, W. J., and Savage, R. E., Jr. (1999). Alterations in rabbit kidney protein expression following lead exposure as analyzed by two-dimensional gel electrophoresis. *Electrophoresis* **20,** 2977–2985.

Karin, M., and Ben-Neriah, Y. (2000). Phosphorylation meets ubiquitination: The control of NF-κB activity. *Annu. Rev. Immunol.* **18,** 621–663.

Kramer, E. R., Scheuringer, N., Podtelejnikov, A. V., Mann, M., and Peters, J.-M. (2000). Mitotic regulation of the APC activator proteins CDC20 and CDH1. *Mol. Biol. Cell* **11,** 1555–1569.

Lahm, H. W., and Langen, H. (2000). Mass spectrometry: A tool for the identification of proteins separated by gels. *Electrophoresis* **21,** 2105–2114.

Legrain, P., and Selig, L. (2000). Genome-wide protein interaction maps using two-hybrid systems. *FEBS Lett.* **480,** 32–36.

Levitskaya, J., Shapiro, A., Leonchiks, A., Ciechanover, A., and Masucci, M. G. (1997). Inhibition of ubiquitin/proteasome-dependent protein degradation by the Gly-Ala

repeat domain of the Epstein-Barr virus nuclear antigen 1. *Proc. Natl. Acad. Sci. USA* **94,** 12616–12621.

Lin, X., Liang, M., and Feng, X. H. (2000). Smurf2 is a ubiquitin E3 ligase mediating proteasome-dependent degration Smad2 in transforming growth factor-β signaling. *J. Biol. Chem.* **47,** 36818–36822.

Lindahl, M., Stahlbom, B., and Tagesson, C. (1999). Newly identified proteins in human nasal and bronchoalveolar lavage fluids: Potential biomedical and clinical applications. *Electrophoresis* **20,** 3670–3676.

Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R., III. (1999). Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676–682.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14,** 1675–1680.

Lou, J., Tu, Y., Ludwing, F. J., Zhang, J., and Manske, P. R. (1999). Effect of bone morphogenetic protein-12 gene transfer on mesenchymal progenitor cells. *Clin. Orthop.* **369,** 333–339.

MacBeath, G., and Schreiber, S. L. (2000). Printing proteins as microarrays for high-throughput function determination. *Science* **289,** 1760–1763.

Mann, M., and Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66,** 4390–4399.

Mann, M., Hojrup, P., and Roepstorff, P. (1993). Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **22,** 338–345.

Massague, J., and Wotton, D. (2000). Transcriptional control by the TGF-β/Smad signaling system. *EMBO J.* **19,** 1745–1754.

Massague, J., Blain, S. W., and Lo, R. S. (2000). TGF-β signaling in growth control, cancer, and heritable disorders. *Cell* **103,** 295–309.

May, J. J., and Ghosh, S. (1998). Signal transduction through NF-κB. *Immunol. Today* **19,** 80–88.

Miyazono, K. (2000). TGF-β signaling by Smad proteins. *Cytokine Growth Factor Rev.* **11,** 15–22.

Munchbach, M., Quadroni, M., Miotto, G., and James, P. (2000). 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.

Myers, T. G., Anderson, N. L., Waltham, M., Li, G., Buolamwini, J. K., Scudiero, D. A., Paull, K. D., Sausville, E. A., and Weinstein, J. N. (1997). A protein expression database for the molecular pharmacology of cancer. *Electrophoresis* **18,** 647–653.

Oda, Y., Huang, K., Cross, F. R., Cowburn, D., and Chait, B. T. (1999). Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96,** 6591–6596.

Ogorzalek Loo, R. R., Mitchell, C., Stevenson, T. I., Martin, S. A., Hines, W. M., Juhasz, P., Patterson, D. H., Peltier, J. M., Loo, J. A., and Andrews, P. C. (1997). Sensitivity and mass accuracy for proteins analyzed directly from polyacrylamide gels: Implications for proteome mapping. *Electrophoresis* **18,** 382–390.

Oleschuk, R. D., and Harrison, D. J. (2000). Analytical microdevices for mass spectrometry. *Trends Anal. Chem.* **19,** 379–388.

Ostergaard, M., Wolf, H., Orntoft, T. F., and Celis, J. E. (1999). Psoriasin (S100A7): A putative urinary marker for the follow-up of patients with bladder squamous cell carcinomas. *Electrophoresis* **20,** 349–354.

Otvos, L., Jr., Insug, O., Rogers, M. E., Consolvo, P. J., Condie, B. A., Lovas, S., Bulet, P., and Blaszczyk-Thurin, M. (2000). Interaction between heat shock proteins and antibacterial peptides. *Biochemistry* **39,** 14150–14159.

Page, M. J., Amess, B., Townsend, R. R., Parekh, R., Herath, A., Brusten, L., Zvelebil, M. J., Stein, R. C., Waterfield, M. D., Davies, S. C., and O'Hare, M. J. (1999). Proteomic definition of normal human luminal and myoepithelial breast cells purified from reduction mammoplasties. *Proc. Natl. Acad. Sci. USA* **96,** 12589–12594.

Pandey, A., and Mann, M. (2000). Proteomics to study genes and genomes. *Nature* **405,** 837–846.

Pandey, A., Podtelejnikov, A. V., Blagoev, B., Bustelo, X. R., Mann, M., and Lodish, H. F. (2000a). Analysis of receptor signaling pathways by mass spectrometry: Identification of Vav-2 as a substrate of the epidermal and platelet-derived growth factor receptors. *Proc. Natl. Acad. Sci. USA* **97,** 179–184.

Pandey, A., Fernandez, M. M., Steen, H., Blagoev, B., Nielsen, M. M., Roche, S., Mann, M., and Lodish, H. F. (2000b). Identification of a novel immunoreceptor tyrosine-based activation motif-containing molecule, STAM2, by mass spectrometry and its involvement in growth factor and cytokine receptor signaling pathways. *J. Biol. Chem.* **275,** 38633–38539.

Pappin, D., Hojrup, P., and Bleasby, A. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3,** 327–332.

Qiu, Y., Sousa, E. A., Hewick, R. M., and Wang, J. H. (2002). Acid-labile isotope-coded extractants: a class of reagents for quantitative mass spectrometric analysis of complex protein mixtures. *Anal. Chem.* **74,** 4969–4979.

Rigaut, G., Schevchenko, A., Rutz, A., Wilm, M., Mann, M., and Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17,** 1030–1032.

Roberts, R. W., and Szostak, J. W. (1997). RNA–peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci. USA* **94,** 12297–12302.

Scheffner, M., Werness, B. A., Huibregtse, J. M., Levine, A. J., and Howley, P. M. (1990). The E6 oncoprotein encoded by human papillomavirus types 16 and 18 promotes the degradation of p53. *Cell* **6,** 1129–1136.

Scheffner, M., Huibregtse, J. M., Vierstra, R. D., and Howley, P. M. (1993). The HPV-16 E6 and E6–AP complex functions as a ubiquitin-protein ligase in the ubiquitination of p53. *Cell* **3,** 495–505.

Schwartz, A. L., and Ciechanover, A. (1999). The ubiquitin-proteasome pathway and pathogenesis of human diseases. *Annu. Rev. Med.* **50,** 57–74.

Sheibani, N. (1999). Prokaryotic gene fusion expression systems and their use in structural and functional studies of proteins. *Prep. Biochem. Biotechnol.* **29,** 77–90.

Sickmann, A., Dormeyer, W., Wortelkamp, S., Woitalla, D., Kuhn, W., and Meyer, H. E. (2000). Identification of proteins from human cerebrospinal fluid, separated by two-dimensional polyacrylamide gel electrophoresis. *Electrophoresis* **21,** 2721–2728.

Steiner, S., and Witzmann, F. A. (2000). Proteomics: Applications and opportunities in preclinical drug development. *Electrophoresis* **21,** 2099–2104.

Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, N. L., and Cordier, A. (1996a). Cyclosporine A decreases the protein level of the calcium-binding protein calbindin-D 28kDa in rat kidney. *Biochem. Pharmacol.* **51,** 253–258.

Steiner, S., Wahl, D., Mangold, B. L., Robison, R., Raymackers, J., Meheus, L., Anderson, N. L., and Cordier, A. (1996b). Induction of the adipose differentiation-related protein in liver of etomoxir-treated rats. *Biochem. Biophys. Res. Commun.* **218,** 777–782.

Steiner, S., Gatlin, C. L., Lennon, J. J., McGrath, A. M., Aponte, A. M., Makusky, A. J., Rohrs, M. C., and Anderson, N. L. (2000). Proteomics to display lovastatin-induced protein and pathway regulation in rat liver. *Electrophoresis* **21,** 2129–2137.

Stroschein, S. L., Wang, W., Zhou, S., Zhou, Q., and Luo, K. (1999). Negative feedback regulation of TGF-β signaling by the SnoN oncoprotein. *Science* **286,** 771–774.

Stulik, J., Koupilova, K., Osterreicher, J., Knizek, J., Macela, A., Bures, J., Jandik, P., Langr, F., Dedic, K., and Jungblut, P. R. (1999). Protein abundance alterations in matched sets of macroscopically normal colon mucosa and colorectal carcinoma. *Electrophoresis* **20,** 3638–3646.

Sun, Y., Liu, X., Ng Eaton, E., Lane, W. S., Lodish, H. F., and Weinberg, R. A. (1999). Interaction of the Ski oncoprotein with Smad3 regulates TGF-β signaling. *Mol. Cell* **4,** 499–509.

Takahashi, N., Udagawa, N., and Suda, T. (1999). A new member of tumor necrosis factor ligand family, ODF/OPGL/TRANCE/RANKL, regulates osteoclast differentiation and function. *Biochem. Biophys. Res. Commun.* **256,** 449–455.

Uetz, P., and Hughes, R. E. (2000). Systematic and large-scale two-hybrid screens. *Curr. Opin. Microbiol.* **3,** 303–308.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403,** 623–627.

Ullrich, O., Reinheckel, T., Sitte, N., Hass, R., Grune, T., and Davies, K. J. (1999). Poly-ADP ribose polymerase activates nuclear proteasome to degrade oxidatively damaged histones. *Proc. Natl. Acad. Sci. USA* **11,** 6223–6228.

Vaughan, T. J., Osbourn, J. K., and Tempest, P. R. (1998). Human antibodies by design. *Nat. Biotechnol.* **16,** 535–539.

Vercoutter-Edouart, A.-S., Lemoine, J., Smart, C. E., Nurcombe, V., Boilly, B., Peyrat, J.-P., and Hondermarck, H. (2000). The mitogenic signaling pathways for fibroblast growth factor-2 involves the tyrosine phosphorylation of cyclin D2 in MCF-7 human breast cancer cells. *FEBS Lett.* **478,** 209–215.

Verma, R., Chen, S., Feldman, R., Schieltz, D., Yates, J., Dohmen, J., and Deshaies, R. J. (2000). Proteasomal proteomics: Identification of nucleotide-sensitive proteasome-interacting proteins by mass spectrometric analysis of affinity-purified proteasomes. *Mol. Biol. Cell* **10,** 3425–3439.

Wallach, D., Varfolomeev, E. E., Malinin, N. L., Goltsev, Y. V., Kovalenko, A. V., and Boldin, M. P. (1999). Tumor necrosis factor receptor and Fas signaling mechanisms. *Annu. Rev. Immunol.* **17,** 331–367.

Wang, C., Oleschuk, R., Ouchen, F., Li, J., Thibault, P., and Harrison, D. J. (2000). Integration of immobilized trypsin bead beds for protein digestion within a microfluidic chip incorporating capillary electrophoresis separations and an electrospray mass spectrometry interface. *Rapid Commun. Mass Spectrom.* **14,** 1377–1383.

Wang, E. A., Rosen, V., Cordes, P., Hewick, R. M., Kriz, M. J., Luxenberg, D. P., Sibley, B. S., and Wozney, J. M. (1988). Purification and characterization of other distinct bone-inducing factors. *Proc. Natl. Acad. Sci. USA* **85,** 9484–9488.

Wasinger, V. C., Cordwell, S. J., Cerpa-Poljak, A., Yan, J. X., Gooley, A. A., Wilkins, M. R., Duncan, M. W., Harris, R., Williams, K. L., and Humphery-Smith, I. (1995). Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis* **16,** 1090–1094.

Wattiez, R., Hermans, C., Cruyt, C., Bernard, A., and Falmagne, P. (2000). Human bronchoalveolar lavage fluid protein two-dimensional database: Study of interstitial lung diseases. *Electrophoresis* **21,** 2703–2712.

Weinstein, M., Yang, X., and Deng, C.-X. (2000). Functions of mammalian *Smad* genes as revealed by targeted gene disruption in mice. *Cytokine Growth Factor Rev.* **11,** 49–58.

Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J. C., Yan, J. X., Gooley, A. A., Hughes, G., Humphery-Smith, I., Williams, K. L., and Hochstrasser, D. F. (1996). From proteins to proteomes: Large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology* **14,** 61–65.

Winter, G., Griffiths, A. D., Hawkins, R. E., and Hoogenboom, H. R. (1994). Making antibodies by phage display technology. *Annu. Rev. Immunol.* **12,** 433–455.

Wolfman, N. M., Hattersley, G., Cox, K., Celeste, A. J., Nelson, R., Yamaji, N., Dube, J. L., Diblasio-Smith, E., Nove, J., Song, J. J., Wozney, J. M., and Rosen, V. (1997). Ectopic induction of tendon and ligament in rats by growth and differentiation factors 5, 6, and 7, members of the TGF-β gene family. *J. Clin. Invest.* **100,** 321–330.

Wozney, J. M., Rosen, V., Celeste, A. J., Mitsock, L. M., Whitters, M. J., Kriz, R. W., Hewick, R. M., and Wang, E. A. (1988). Novel regulators of bone formation: Molecular clones and activities. *Science* **242,** 1528–1534.

Wrana, J. L. (2000). Regulation of Smad activity. *Cell* **100,** 189–192.

Xu, W., Angelis, K., Danielpour, D., Haddad, M. M., Bischof, O., Campisi, J., Stavnezer, E., and Medrano, E. E. (2000). Ski acts as a co-repressor with Smad2 and Smad3 to regulate the response to type β transforming growth factor. *Proc. Natl. Acad. Sci. USA* **97,** 5924–5929.

Yang, X., Ji, X., Shi, X., and Xu, C. (2000). Smad1 domains interacting with Hoxc-8 induce osteoblast differentiation. *J. Biol. Chem.* **275,** 1065–1072.

Yaron, A., Gonen, H., Alkalay, I., Hatzubai, A., Jung, S., Beyth, S., Mercurio, F., Manning, A. M., Ciechanover, A., and Ben-Neriah, Y. (1997). Inhibition of NF-κB cellular function via specific targeting of the I-κB-ubiquitin ligase. *EMBO J.* **21,** 6486–6494.

Yaron, A., Hatzubai, A., Davis, M., Lavon, I., Amit, S., Manning, A. M., Andersen, J. S., Mann, M., Mercurio, F., and Ben-Neriah, Y. (1998). Identification of the receptor component of the IκBα-ubiquitin ligase. *Nature* **396,** 590–594.

Yates, J. R. D., Speicher, S., Griffin, P. R., and Hunkapiller, T. (1993). Peptide mass maps: A highly informative approach to protein identification. *Anal. Biochem.* **214,** 397–408.

Young, K. H. (1998). Yeast two-hybrid: So many interactions, (in) so little time. *Biol. Reprod.* **58,** 302–311.

Zhang, B., Liu, H., Karger, B. L., and Foret, F. (1999). Microfabricated devices for capillary electrophoresis-electrospray mass spectrometry. *Anal. Chem.* **71,** 3258–3264.

Zhu, H., Kavsak, P., Abollah, S., Wrana, J. L., and Thomsen, G. H. (1999). A SMAD ubiquitin ligase targets the BMP pathway and affects embryonic formation. *Nature* **400,** 687–693.

FROM CLONE TO CRYSTAL: MAXIMIZING THE AMOUNT OF PROTEIN SAMPLES FOR STRUCTURE DETERMINATION

By CHRIS M. KOTH AND ALED M. EDWARDS

Banting and Best Department of Medical Research and Department of Medical Genetics and Microbiology, C. H. Best Institute, University of Toronto, Toronto, Ontario, Canada, M5G 1L6

I. Introduction

Structural proteomics is a product of the postgenomic era. Broadly defined, its goal is to obtain the three-dimensional structures of all proteins, using either experimental or computational methods [1–4]. The role of structural biology in the post genomics era is being changed because of two observations. First, fewer than 60% of the open reading frames (ORFs) from even the most recently sequenced genomes share enough sequence identity with an assigned gene to be ascribed a function [5–7]. Second, the three-dimensional structure of a protein often yields clues to its function by uncovering structural homology to a protein of known function. This structural homology often occurs even in the absence of recognizable sequence conservation. Accordingly, structural biology is transitioning from a discipline that was used to explain function to one that can be used to predict function.

How soon will structure-based assignment of function make an impact? If we were to be availed immediately of structural information for all proteins, the impact would be immediate and significant. Already the majority of new structures show structural homology to other proteins [8–10], and this homology is associated with functional similarity [1, 11–13]. The promise of this use of structural biology is therefore great; the challenge is to overcome the technical hurdles that confront genome-wide structural biology efforts.

ADVANCES IN
PROTEIN CHEMISTRY, Vol. 65

0065-3233/03 $35.00

The first three-dimensional structure determined from a structural effort designed to reveal protein function vindicated the hypothesis that functional information could be derived from protein structure. In a pilot project using the thermophile *Methanococcus jannaschii* as a model organism, Zarembinski *et al.* determined the crystal structure of ORF MJ0577, whose function was previously unknown [14]. The protein structure contained a bound ATP, suggesting MJ0577 is an ATPase or an ATP-mediated molecular switch. Futhermore, the structure of MJ0577 revealed different ATP-binding motifs that are shared among many homologous hypothetical proteins in this family. This attributed function was subsequently confirmed biochemically. This result suggested that structure-based assignment of molecular function is a viable approach for the large-scale biochemical assignment of proteins and for discovering new motifs, a basic premise of structural genomics. Other efforts have also illustrated the feasibility of using three-dimensional structural information to uncover clues about the function of a protein [2, 15–19]. Of the first 10 structures determined by our own proteomics effort, 5 contained either a bound ligand or a ligand-binding site that could be inferred from structural homology. When coupled with biochemical assays, we were able to ascribe putative functions for many of the proteins [2].

With the feasibility of structural proteomics established, efforts must now focus on identifying and overcoming the numerous technical barriers encountered at each stage of the process. Although we have previously identified several key issues, those involving sample preparation and optimization are clearly the most difficult to overcome [3]. In this review, we briefly examine some of the key advances that have allowed large numbers of protein structures to be determined in relatively short periods of time. We follow with a detailed analysis of methods designed to increase the number of available samples for structural analysis via novel expression, preparation, and optimization systems.

II. Technology Drives the Process

We are in a position to propose structural proteomics efforts only because of significant advances and technological improvements in molecular, structural, and computational biology. Genome sequences provide the basis for identifying and obtaining new proteins. Recombinant DNA technology has made it relatively straightforward to obtain milligram amounts of highly purified sample. In crystallography, the advent of

high-flux synchrotron beamlines and new phasing methods has eliminated many of the potential problems in crystal structure determination. Electron density maps are usually of much higher quality than even 5 years ago and often initial three-dimensional protein models can be built in a matter of days or even hours [20]. Structure determination by nuclear magnetic resonance (NMR) spectroscopy is also making significant gains. These advances are derived from improved methodology, better software, and improved hardware. For example, the new cryogenically cooled probes (cryoProbes; Bruker BioSpin, Billerica, MA) dramatically increase the signal/noise ratio over conventional probes, and thus reduce data collection time. Taken together, these advances have dramatically reduced the amount of time required to go from protein sample to three-dimensional structure (Fig. 1; see Color Insert).

The benefits of the aforementioned improvements are driving advances in the field of structural proteomics. The current focus is now maximizing the number of samples suitable for structural analysis. Perhaps unexpectedly, this is not a problem of producing expression constructs for genes. Many structural proteomics efforts are initiating significant automation strategies for this step. However, it is likely that a small group of people could create tens of thousands of expression clones in a short period of time without robots. The critical step is to achieve adequate expression of soluble protein, and here the significant hurdles begin to emerge. For instance, an examination of current structural proteomics efforts indicates that at least 50% of successfully cloned genes do not ''survive'' the expression or purification processes, yielding insufficiently expressed or insoluble protein [3]. Therefore, as a first step in increasing the yield of suitable constructs, methods designed to optimize both expression and solubility are needed. Each construct could be expressed in a library of other bacterial or eukaryotic hosts or from alternative vectors. Equally promising strategies, and ones that are amenable to high-throughput approaches, would be to screen libraries of cofactors or buffer conditions with the aim of increasing solubility.

Expression and solubility, although prerequisite, do not guarantee a sample suitable for three-dimensional structure determination. In the case of X-ray crystallography, current screens are successful for only 20–40% of the proteins tested and those trials that do yield crystals invariably require optimization [2]. For proteins selected for NMR spectroscopy, as many as 60% are unsuitable for continued structural studies (Yee, Edwards, and Arrowsmith, unpublished data). The potential reasons for these difficulties are numerous, but one that attracts considerable attention is the notion that some proteins will prove challenging targets because they are composed of several individual domains. The conformational heterogeneity that results

from interdomain motion poses a severe problem for protein crystallization whereas the size of multidomain proteins is an impediment to NMR spectroscopy [3, 12]. An obvious solution is to focus on the protein domains, whose structure can be determined more readily by X-ray crystallography or NMR spectroscopy [21–24]. Given the success of this approach, by concentrating on such domains, determining the structure of a complete proteome is feasible. As with screens for solubility, many of the techniques designed to identify domain boundaries can be automated and are therefore amenable to high-throughput approaches.

III. From Sequence to Solubility: The First Step

Insolubility arises from either an intrinsic property of a protein (e.g., aggregation due to a hydrophobic patch on the surface) or because the protein is not susceptible to the folding mechanisms in the expression host; in which case there is an aggregation of folding intermediates. In our previous proteomics effort, insolubility and/or poor expression accounted for almost 60% of recalcitrant proteins [2, 3]. A preliminary examination of the primary sequence of these proteins provided two key observations. First, proteins that fulfilled the following criteria were likely to be insoluble: (1) had a hydrophobic stretch (>20 residues) with average hydrophobicity less than -0.85 kcal/mol (on the GES scale), (2) Gln composition $<4\%$, (3) Asp plus Gln composition $<17\%$, and (4) aromatic composition $>7.5\%$. Second, proteins that did not have a hydrophobic stretch and had less than 27% of their residues in low-complexity regions were likely to be soluble [2]. Although these ''rules'' were derived from a single proteomics effort within a single genome, they suggested that the primary sequence could be predictive of the solubility of a given protein. They did not, however, provide insight into the courses of action to be taken with insoluble proteins. Below, we examine various systems and strategies for increasing the pool of soluble proteins, using both currently available and rapidly emerging technologies.

IV. Alternative Expression Systems

The use of different expression systems such as insect, yeast, or mammalian cells often allows the expression of proteins that are insoluble when expressed in *Escherichia coli*. However, many of the advances that have paved the way for proteome-wide three-dimensional structure determination, including multiple-wavelength anomalous diffraction (MAD) phasing and metabolic labeling, were developed with *E. coli* as

an expression system. As such, alternative systems such as insect and human cell cultures, in their current forms, are expensive and time consuming. In addition, the development of metabolic labeling in these systems is in its infancy. An alternative approach, and one that is making a significant resurgence, is the use of cell-free expression systems with bacterial extracts [25]. By incorporating dialysis-based, semicontinuous flow methods to constantly regenerate substrates, the problems of efficiency and low protein yields that once plagued this process may have been overcome. Using this procedure, levels of protein production of up to 6 mg/ml per reaction mixture have been reported. Importantly, metabolic labeling is possible if the mixture is supplemented with labeled amino acids. These improvements in cell-free protein expression should serve to make the system more attractive for structural proteomics efforts.

V. Screening for the Most Soluble Ortholog

Despite considerable efforts from several groups, including our own, it is currently impossible to conclusively predict the degree of solubility of a protein from its gene sequence. It is known, however, that even subtle changes in amino acid sequence can dramatically affect protein solubility. Thus, for proteins that have many orthologs, a common strategy is to clone and express a selection in order to identify the ortholog with the best solubility properties. There are variations of this approach, many involving mutagenesis. One that is particularly attractive makes use of hybrids between related proteins. Termed SHIPREC (for sequence homology-independent recombination), the process generates libraries of sequence-independent cross-overs between two genes only at structurally related sites [26]. In combination with a powerful selection system, these are subsequently screened for expression and solubility. In this way, it may be possible to find hybrids with similar structures that are also more soluble than the parent proteins. In isolated instances, the above-described approaches have proved successful; however, we do not yet know how effective they will be on a proteome-wide scale.

VI. Cofactor Screens

Proteins may also be insoluble because they lack an obligate cofactor. With this in mind, structural proteomics efforts should include screens for small molecules that interact with newly synthesized proteins. These screens should incorporate known bioactive small molecules as well as a library of new chemical entities. Indeed, systematic identification of

interacting compounds may provide a means to purify more proteins, as well as a method to ascribe function to new proteins. This approach helps determine protein function by matching each member of a library of proteins with a small molecule chemical, perhaps a component of a combinatorial chemistry library.

VII. Solubility Screens

Methods designed to identify buffer conditions that increase the solubility of a given protein are analogous to the chemical proteomics methods described above. Several groups have designed sparse matrix screens, all of which attempt to effectively sample ''solubility space'' with any protein of interest [27–29]. In an effort to rapidly implement such a screen in our structural proteomics efforts, we have made successful use of readily available crystallization sparse matrix screens (Hampton Research, Laguna Niguel, CA) to identify buffer conditions that might prove amenable to NMR. As with other approaches, this samples a broad range of solubility space. Unlike other screens, however, it is simply an extension of crystallization methods and is therefore immediately adaptable to high-throughput analyses.

VIII. From Solubility to Structure: The Second Step

A soluble, pure, and concentrated protein does not always ensure a three-dimensional structure. However, we, and others, have observed an increased tendency for smaller proteins to crystallize or prove suitable for NMR spectroscopy [2, 21–24]. In fact, the vast majority of solved structures represent peptides < 30 kDa in size. One explanation for this is provided by the simple observation that large proteins are often composed of many individual domains. In the case of crystallography, the conformational heterogeneity that results from motion between such domains is a severe impediment to crystallization. For NMR analyses, intrinsic limitations on the size of molecules that can be studied often necessitate analysis of individual domains. Thus, the difficulties associated with determining the three-dimensional structures of conformationally heterogeneous proteins has prompted structural biologists to turn their attention to studying the protein domains. By focusing on these, determining the structure of a complete protein or group of proteins is feasible. Protein domains have additional features that render them attractive targets for structural analysis. First, individual domains are often much easier to express in recombinant form

in bacteria; protein domains express to higher levels and are often more soluble than multidomain proteins [21–23, 30]. Second, protein domains, which typically comprise fewer than 150 amino acids, are potentially amenable to structural analysis using either X-ray crystallography or NMR spectroscopy. The major challenge of studying the structure and function of protein domains is to identify their individual boundaries. In this section, we examine various strategies for identifying domain boundaries using methods suitable for both isolated and high-throughput approaches.

A. Domain Mapping by Sequence Analysis

One of the most common methods for estimating domain boundaries is via sequence homology. Here, the assumption is made that regions with the least amount of homology will be found in the regions between domains ([31], and references therein). This method is limited, however, because many structural and functional homologies escape detection by sequence-based approaches. Moreover, sequence conservation is often found in regions of proteins that do not adopt a stable tertiary structure (such as a protein interaction motif that folds only when bound to its partner).

B. Domain Mapping by Limited Proteolysis

As an alternative to sequence-based methods, we focus on the utility of limited proteolysis to identify domain boundaries [21–23, 30]. This method is predicated on the fact that regions between domains are typically more susceptible to protease digestion than are the domains themselves. Limited proteolysis has gained prominence because of significant advances in protein mass spectrometry (MS). Using MS, stable partial proteolytic products corresponding to individual domains can be identified rapidly and unambiguously in a matter of hours.

The aim of a partial proteolysis experiment (Fig. 2; see Color Insert) is to identify one or more proteolytically sensitive regions in a particular protein by ''tickling'' it with proteases. Such regions are sensitive to protease digestion because they are more accessible than those folded within a domain. For any given protein, we typically compare the digestion pattern produced by several proteases, including trypsin, chymotrypsin, pronase, and endoproteinase Glu-C. Digestions, monitored over several hours and with varying concentrations of protease, are resolved by Sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE). Subsequent

characterization of the products by MS indicates the approximate domain boundaries. Because of the specificity of individual proteases for specific residues, several mutants are cloned corresponding to sites near and including the identified boundary. The production of these clones is monitored in a suitable expression system and the most active and/or highest expressing and/or most stable constructs are chosen for structural analysis [30].

We modified the above-described procedure for high-throughput applications. We developed an immobilized protease platform that enables the partial proteolysis of several target proteins to be completed within a few hours. Briefly, different proteases are immobilized on plastic 96-well microtiter plates (Nuclon; Nunc, Roskilde, Denmark) resulting in an array of enzymes with varying concentrations. Equal amounts of a target protein are added to each well and the plates are incubated. The proteolytic products are stopped by the addition of SDS–PAGE sample buffer and analyzed by denaturing gel electrophoresis. Alternatively, the reaction is terminated by the addition of acetic acid and the products are purified by reversed-phase liquid chromatography. Mass spectrometry is then used to identify the purified proteolytic fragments. We have found that proteases bound to microtiter plates retain their activity after lyophilization. Thus, if several plates are prepared and stored beforehand, a partial proteolysis experiment could be performed on a relatively large number of target proteins in just a few hours.

IX. Conclusions

Considerable technological advances and numerous sequencing projects have paved the way for structural proteomics. Still, significant impediments remain. The dominant experimental barrier to a proteome-wide structural analysis will be the large number of proteins that are insoluble or unfolded when expressed or concentrated for structural efforts. Further studies and significant investment will be required to reveal whether high-throughput technologies intended to alter various parameters such as expression, solution conditions, or protein domain boundaries will increase the proportion of samples suitable for structure determination.

References

1. Kim, S. H. (1998). Shining a light on structural genomics. *Nat. Struct. Biol.* **5,** (Suppl.) 643–645.
2. Christendat, D. *et al.* (2000). Structural proteomics of an archaeon. *Nat. Struct. Biol.* **7,** 903–909.

3. Edwards, A. M. *et al.* (2000). Protein production: Feeding the crystallographers and NMR spectroscopists. *Nat. Struct. Biol.* **7,** (Suppl.) 970–972.
4. Christendat, D. *et al.* (2000). Structural proteomics: Prospects for high throughput sample preparation. *Prog. Biophys. Mol. Biol.* **73,** 339–345.
5. Venter, J. C. *et al.* (2001). The sequence of the human genome. *Science* **291,** 1304–1351.
6. Cho, Y., and Walbot, V. (2001). Computational methods for gene annotation: The *Arabidopsis* genome. *Curr. Opin. Biotechnol.* **12,** 126–130.
7. Adams, M. D. *et al.* (2000). The genome sequence of *Drosophila melanogaster. Science* **287,** 2185–2195.
8. Murzin, A. G. *et al.* (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540.
9. Gibrat, J. F., Madej, T., and Bryant, S. H. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6,** 377–385.
10. Orengo, C. A. *et al.* (1997). CATH—a hierarchic classification of protein domain structures. *Structure* **5,** 1093–1108.
11. Shapiro, L., and Lima, C. D. (1998). The Argonne Structural Genomics Workshop: Lamaze class for the birth of a new science. *Structure* **6,** 265–267.
12. Sali, A. (1998). 100,000 protein structures for the biologist. *Nat. Struct. Biol.* **5,** 1029–1032.
13. Montelione, G. T., and Anderson, S. (1999). Structural genomics: Keystone for a Human Proteome Project. *Nat. Struct. Biol.* **6,** 11–12.
14. Zarembinski, T. I. *et al.* (1998). Structure-based assignment of the biochemical function of a hypothetical protein: A test case of structural genomics. *Proc. Natl. Acad. Sci. USA* **95,** 15189–15193.
15. Elofsson, A., and Sonnhammer, E. L. (1999). A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics.* **15,** 480–500.
16. Lewis, H. A. *et al.* (2001). A structural genomics approach to the study of quorum sensing: Crystal structures of three luxs orthologs. *Structure (Camb)* **9,** 527–537.
17. Johnson, K. A. *et al.* (2001). Crystal structure and catalytic mechanism of the MJ0109 gene product: A bifunctional enzyme with inositol monophosphatase and fructose 1,6-bisphosphatase activities. *Biochemistry* **40,** 618–630.
18. Stec, B. *et al.* (2000). MJ0109 is an enzyme that is both an inositol monophosphatase and the ''missing'' archaeal fructose-1,6-bisphosphatase. *Nat. Struct. Biol.* **7,** 1046–1050.
19. Zhang, H. *et al.* (2000). Crystal structure of YbaK protein from *Haemophilus influenzae* (HI1434) at 1.8 Å resolution: Functional implications. *Proteins* **40,** 86–97.
20. Perrakis, A., Morris, R., and Lamzin, V. S. (1999). Automated protein model building combined with iterative structure refinement. *Nat. Struct. Biol.* **6,** 458–463.
21. Olmsted, V. K. *et al.* (1998). Yeast transcript elongation factor (TFIIS), structure and function. I. NMR structural analysis of the minimal transcriptionally active region. *J. Biol. Chem.* **273,** 22589–22594.
22. Pfuetzner, R. A. *et al.* (1997). Replication protein A: Characterization and crystallization of the DNA-binding domain. *J. Biol. Chem.* **272,** 430–434.
23. Barwell, J. A. *et al.* (1995). Overexpression, purification, and crystallization of the DNA-binding and dimerization domains of the Epstein-Barr virus nuclear antigen 1. *J. Biol. Chem.* **270,** 20556–20559.

24. Cohen, S. L. *et al.* (1995). Probing the solution structure of the DNA-binding protein Max by a combination of proteolysis and mass spectrometry. *Protein Sci.* **4,** 1088–1099.
25. Kigawa, T. *et al.* (1999). Cell-free production and stable-isotope labeling of milligram quantities of proteins. *FEBS Lett.* **442,** 15–19.
26. Sieber, V., Martinez, C. A., and Arnold, F. H. (2001). Libraries of hybrid proteins from distantly related sequences. *Nat. Biotechnol.* **19,** 456–460.
27. Bagby, S. *et al.* (1997). The button test: A small scale method using microdialysis cells for assessing protein solubility at concentrations suitable for NMR. *J. Biomol. NMR.* **10,** 279–282.
28. Lindwall, G. *et al.* (2000). A sparse matrix approach to the solubilization of overexpressed proteins. *Protein Eng.* **13,** 67–71.
29. Bagby, S., Tong, K. I., and Ikura, M. (2001). Optimization of protein solubility and stability for protein nuclear magnetic resonance. *Methods Enzymol.* **339,** 20–41.
30. Koth, C. M. *et al.* (2000). Elongin from *Saccharomyces cerevisiae*. *J. Biol. Chem.* **275,** 11174–11180.
31. Wilson, C. A., Kreychman, J., and Gerstein, M. (2000). Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297,** 233–249.

PROTEOMICS AND BIOINFORMATICS

By CAROL S. GIOMETTI

Argonne National Laboratory, *Argonne, Illinois 60439

I. Introduction

Wasinger and co-workers (1995) first used the term *proteome* to describe the proteins encoded within a genome, with the qualification that it is unlikely that all of the proteins encoded will be expressed at any given point in time. Thus, proteome analysis includes the characterization of the proteins that are expressed under a given set of conditions, including their relative abundance and their subcellular location. Proteome analysis also reveals posttranslational modifications of proteins and protein interactions with other biomolecules and ligands. To the fullest extent, proteome analysis can also be considered to cover the determination of protein three-dimensional structures and protein functions. Whereas a genome is a static collection of nucleotides in a specific sequence, a proteome is a dynamic population of polypeptides that changes in response to environmental triggers.

Bioinformatics has evolved as a field of research focused on the acquisition, analysis, and database management of nucleic acid and protein sequences (Butler, 1998). Once acquired, a sequenced genome is analyzed to decipher the codon triplets encoding the amino acid sequences of proteins. A number of algorithms are publicly available to ''call the ORFs'' (i.e., the protein-coding sequences referred to as *open-reading frames* or exons) in a genome (http://restools.sdsc.edu/biotools/

*Argonne National Laboratory, a U.S. Department of Energy Office of Science Laboratory, is operated by The University of Chicago under contract W-31-109-Eng-38.

ADVANCES IN
PROTEIN CHEMISTRY, Vol. 65
0065-3233/03 $35.00

biotools16.html). The ORF databases produced by such genome sequence analysis can then be used to predict which proteins are encoded, as well as their structure and function, on the basis of comparison with other completed genome sequences. Volumes of information are being produced as a result of these follow-up procedures to genome sequencing, and consequently numerous databases are being created for the storage of this information. In the context of proteomics, bioinformatics currently indicates which proteins could be produced by a given biological system, thus providing a starting point for proteome analysis. For example, the ORF databases are useful as guides to which cDNA sequences to include on microarrays for analysis of gene expression profiles and for the expression of proteins to be used as target molecules on protein microarrays. The ORF databases are also an essential component of proteome analyses focused on the identification of proteins expressed by biological systems.

Proteome methods such as two-dimensional gel electrophoresis coupled with peptide mass spectrometry (Patterson and Aebersold, 1995; Humphery-Smith *et al.*, 1997; Link *et al.*, 1997) or liquid chromatography coupled with mass spectrometry (Griffin and Aebersold, 2001; Mann *et al.*, 2001; Smith *et al.*, 2001; Washburn *et al.*, 2001) are used to detect the proteins actually expressed by a biological system at a specific point in time. Identifications of the proteins resolved by all of these methods are based on the comparison of the observed peptide masses of the separated proteins with the masses predicted by the DNA sequences in the appropriate ORF database. Thus, the deciphering of DNA sequences into ORFs and the annotations included in DNA sequence databases are essential to proteome analysis. As more proteomes are analyzed and the data are deposited in a growing number of databases, the need to integrate proteome databases with genome and protein sequence databases is growing. The proteome data, including such features as differential expression, posttranslational modifications, and subcellular location, add a new dimension to bioinformatics and provide new opportunities for development of strategies for data storage and the integration of genome and protein sequence data with protein expression data. Such integration will allow for the refinement of prediction algorithms and perhaps eventually lead to the bioinformatics tools needed to provide accurate computational modeling of biological functions. The discussion that follows is focused on the use of existing bioinformatics approaches in the analysis of global protein expression using two-dimensional electrophoresis and mass spectrometry and on the extension of bioinformatics to include the analysis and management of the proteome data being produced.

II. Bioinformatics Tools

A. *Sequence Databases*

The starting material for bioinformatics is composed of databases of DNA and protein sequence information. The first protein sequence database was created by Dayhoff and colleagues (1965) long before the advent of large-scale genome sequencing projects (i.e., pre-Human Genome Sequencing Project). This collection of protein sequences and structures eventually led to the creation of the Protein Information Resource (PIR) database (George *et al.*, 1997; http://www-nbrf.georgetown.edu/). Two other major protein databases now exist in addition to the PIR: the Swiss-Prot database of published protein sequences and annotations (http://ca.expasy.org/sprot/) and the Protein Data Bank (http://www.rcsb.org/pdb/) of three-dimensional protein structures. The major public repositories of published DNA sequences include the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/) maintained by the European Bioinformatics Institute, GenBank (http://www.ncbi.nlm.nih.gov) maintained by the National Center for Biotechnology Information, and the DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/).

Although the nucleic acid and protein sequence databases were originally created as separate entities and are currently still maintained as such, integration is beginning to appear. Whereas the trend has been to use genome sequence databases as an entry point to linked protein sequence databases through Web interfaces, comanagement of nucleotide and protein sequences is now appearing at the major public sites. With the EMBL nucleotide database, for example, TREMBL and Swiss-Prot can be accessed for protein sequence information. Within the GenBank genome sequence database, GenPep is provided for protein sequence information. The NCBI genome sequences are linked to the corresponding protein sequences through the search engine Entrez, facilitating the full search of available nucleic acid and protein sequences. It should be noted that, in many cases, the protein sequences provided are the translation of known nucleic acid sequences into the predicted amino acid sequences rather than protein sequence obtained by the physical sequencing of proteins themselves. These predicted protein sequences are delineated from actual protein sequences within the respective database annotations.

B. Sequence Analysis and Annotation

In addition to providing nucleic acid and protein sequences, bioinformatics includes the analysis of those sequences to deduce similarities and differences between the molecules expressed by different biological systems. A variety of algorithms are used to decipher coding sequences, that is, to detect the start and stop codons defining open reading frames and to determine the predicted protein sequence from the nucleic acid sequence within each open reading frame. Among the most widely used and publicly accessible algorithms are as follows:

GrailEXP: http://grail.lsd.ornl.gov/grailexp/index.html
GeneMark: http://opal.biology.gatech.edu/GeneMark/
GENSCAN: http://genes.mit.edu/GENSCAN.html

Each of these algorithms has a unique set of parameters and often more than one approach is applied to a genome sequence before the ORF list is considered by researchers to be optimized.

By comparing the resulting ORF databases with preexisting databases from other genomes, similarities and differences can be determined. Annotation of new genomes is dependent on these comparisons, since the annotations are carried over to new genome sequences on the basis of sequence similarity. The first algorithm widely used to search biological sequence databases for sequence similarities was FASTA, originally designed to search protein sequences (Lipman and Pearson, 1985; Pearson and Lipman, 1988) and then applied to nucleic acid sequence databases (Pearson, 1994). More recently, the Basic Local Alignment Search Tool was developed as a starting point to search for genome sequence similarities (BLAST; Altschul *et al.*, 1990). Using BLAST (available through the NCBI server at http://www.ncbi.nlm.nih.gov/BLAST/), new genome sequences, generally in FASTA format, are compared with existing sequences for similarities and the results of the comparison are ranked. The amount of similarity, that is, homology, between sequences is based on these rankings, and predictions of protein identity, three-dimensional structure, and function for new genome sequences are often based on such similarity searches. PSI-BLAST (Altschul *et al.*, 1997), that is, position-specific iterated BLAST, provides a more refined search based on several iterative database search runs.

Both FASTA and BLAST are restricted to pairwise comparisons. Comparative analysis of multiple genomes, however, especially across different branches of the evolutionary tree, is a valuable approach in bioinformatics for the accurate annotation of genomes as well as the prediction of protein structure and function and of metabolic pathways

within a given organism. Therefore, numerous approaches to comparative genome analysis and annotation have been developed to provide cross-species data use in structure and function prediction, and for computational metabolic pathway reconstructions (Thompson *et al.*, 1994; Gaasterland *et al.*, 2000; Overbeek *et al.*, 2000).

The analysis of genome sequences and their corresponding protein sequences for clues to protein structure and function includes detection of specific diagnostic sequence patterns. Algorithms have been developed, for example, to predict which encoded proteins are involved in signal transduction (SENTRA; D'Souza *et al.*, 2000). Protein phosphorylation sites (NetPhos; Blom *et al.*, 1999; http://www.cbs.dtu.dk/services/NetPhos/) and membrane association (SignalP, Nielsen *et al.*, 1997, http://www.cbs.dtu.dk/services/SignalP/; TSEG, Kihara *et al.*, 1998, http://www.genome.ad.jp/SIT/tseg.html) can also be predicted from putative amino acid sequences. Bader and Hogue (2000) have produced the Biomolecular Interaction Network Database (BIND; http://bioinfo.mshri.on.ca/) that provides information about protein interactions with other proteins, DNA, or ligands. BIND can be used to compare new DNA sequences with existing database entries to search for indications of possible interaction capabilities. Metabolic pathway databases for specific organisms, such as EcoCyc for the *Escherichia coli* genome (Karp *et al.*, 2000; http://www.biocyc.org/ecocyc/), or for all available completed genomes, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto, 2000; http://www.genome.ad.jp/kegg/kegg2.html), provide the capability to use new DNA sequences to search for metabolic pathway associations. This is a small sample of the computational approaches currently available to researchers who have access to new DNA sequences and are interested in querying existing databases to find similarities indicative of specific protein functions.

C. *Bioinformatics and Proteomics*

Fully, and even partially, annotated genome databases have multiple uses in the broad context of proteomics, including protein structure and function prediction as well as protein expression analysis. In the context of protein expression, the ORF sequences are used to select specific nucleotide sequences to include on the cDNA microarray chips used to quantify changes in specific mRNA abundance. The ORF sequences are also used as the starting material for protein arrays, providing the nucleic acid sequence that is incorporated into host cells to produce over expression of the target proteins for the arrays. Obviously, the accuracy of

the annotation in these ORF databases is essential to the successful selection of targets for such proteome analyses. For global proteomics studies that seek to identify and quantify each protein component actually synthesized by a biological system at a specific point in time, the ORF sequences and their annotation are the source of the protein identifications. Whether the proteins are separated in a gel matrix by either one- or two-dimensional gel electrophoresis or by liquid separation methods such as liquid chromatography, the use of protein or peptide masses to identify the separated proteins is completely dependent on the accuracy of the ORF databases available. With the rapid proliferation of proteome data in addition to genome information, expansion of the field of bioinformatics to include the collection and organization of proteome data as well as the integration of proteome data with genome data will produce an accurate summary of cellular functions at the molecular level.

III. Proteomics Tools

A. *ORF Databases and Proteomics*

The proteins actually expressed by a biological system can be detected and quantified by a variety of techniques. The best established of these proteomics tools is the combination of two-dimensional gel electrophoresis (2DE) to separate and detect the proteins in a complex mixture with peptide mass spectrometry to aid in the identification of the proteins detected. Methods to obtain the protein or peptide mass data independent of gel separation are also under development (Griffin and Aebersold, 2001; Mann *et al.*, 2001; Smith *et al.*, 2001; Washburn *et al.*, 2001). The identification of proteins is dependent on the comparison of actual peptide masses (after proteolytic digestion) with predicted peptide masses based on the hypothetical digestion of all the predicted proteins encoded in the corresponding ORF database (Link *et al.*, 1997). The accuracy of the ORF database for the identification of proteins is of obvious importance in such proteome studies, because the identity of the expressed proteins is directly linked to the ORF annotation.

Numerous examples demonstrate the identification of expressed proteins based on the availability of complete genome sequence data. Using 2DE coupled with peptide mass analysis in conjunction with complete genome sequence information, 150 proteins from *Saccharomyces cerevisiae* were identified by Shevchenko *et al.* (1996). More recently, the number of *S. cerevisiae* proteins identified was reported to be 502 after numerous subfractions of the protein samples were analyzed by 2DE (Langen *et al.*,

2000). Also using 2DE with peptide mass spectrometry, Link *et al.* (1997) identified 235 proteins from *Haemophilus influenzae* and Wasinger *et al.* (2000) identified 158 proteins from *Mycoplasma genitalium*. Other articles have reported the identification of proteins in the context of biological and biochemical studies rather than simply identifying all detectable proteins. Using comparative 2DE analysis, Giometti *et al.* (2001) identified *Methanococcus jannaschii* proteins that were differentially expressed under conditions of hydrogen depletion, including the characterization of structural alterations in the major flagellin proteins. Jungblut *et al.* (2001) reported the comparison of protein expression in *Mycobacterium tuberculosis* strain H37Rv and a clinical isolate of *M. tuberculosis* strain CDC1551, using the completed genome sequence of H37Rv and a partial sequence of CDC1551 for protein identifications. In this latter report, six proteins that were actually missed in the genome annotation were detected, demonstrating the importance of cross-referencing between genome and proteome databases so that such gaps in genome annotation can be corrected.

Demonstrating the high throughput of liquid chromatography combined with tandem mass spectrometry for protein identifications in proteome studies, Washburn *et al.* (2001) identified 1484 proteins from *S. cerevisiae*. Li *et al.* (2001) described the combination of capillary liquid chromatography with Fourier transform ion cyclotron resonance mass spectrometry to both separate and identify *Deinococcus radiodurans* proteins on the basis of their peptide masses. As the liquid separation methods interfaced with a variety of mass spectrometry methods become more prevalent, the number of proteins identified on the basis of peptide mass identity with ORF predictions will increase exponentially because of the high-throughput capability of the separation methods. The accuracy of the identifications, however, will continue to be totally dependent on the accuracy of the corresponding ORF databases used for identity searches.

To date, as the preceding references show, the primary means for the communication of proteome study results has been as lists of ORFs in publications. Such presentations oversimplify the volume and complexity of the data that underlie the correlation of protein expression with ORFs. For each protein associated with a specific ORF there are one or more mass spectra. For each mass spectrum, there is a list of possible identifications with correlation scores from the ORF database queries. In addition, data describing the actual and predicted molecular weights and isoelectric points of each detected protein, posttranslational modifications, and, in some cases, the relative abundance of each detected protein are produced as part of the proteome analysis, but not reflected in most published reports. Therefore, a new field of proteome bioinformatics is evolving in parallel with the existing bioinformatics of

nucleic acid and protein sequences. In addition to the acquisition and analysis of data pertaining to protein expression in biological systems, proteome bioinformatics includes the development of database architectures for managing and interfacing proteome data in forms as accessible to the research community as DNA and protein sequence databases.

B. Proteome (2DE) Databases

In the context of this discussion, ''proteome databases'' refer to those collections of data that reflect the protein expression within a given biological system under a specific set of conditions gathered through the use of 2DE and mass spectrometry. In the near future, databases will also be available that include protein expression information collected by liquid separation methods coupled with mass spectrometry (Washburn *et al.*, 2001; Smith *et al.*, 2001) and by a myriad of protein microarray methods (e.g., surface enhanced laser desorption ionization, antibody chips, ligand chips) now under development. Many of the issues currently being addressed by the curators of 2DE databases will also confront the scientists producing data by these alternative proteome analysis methods. Thus, this discussion is pertinent to researchers utilizing any combination of established and evolving methods for analysis of protein expression.

Swiss-2DPAGE (http://www.expasy.ch/ch2d/ch2d-top.html) was the first to support a Web-based database of proteome data in the context of 2DE patterns with protein identifications and links to related genome and protein sequence databases. By establishing a set of presentation criteria in order to be defined as ''federated,'' Swiss-2DE provides links to numerous Web sites containing 2DE patterns with varying amounts of annotation (Table I). These databases, although they provide protein identifications as well as links to nucleic acid and protein sequence databases, do not, however, reveal the complexity of data available from proteome analysis and do not include query tools that provide the users with the capability to explore the data in the context of genome information. Perhaps most important, they do not provide information about the relative abundance of each protein expressed or about the possible posttranslational modifications. Such information is important to understanding the function of the proteins in the whole cell.

The complete, annotated genome sequence of *Methanococcus jannaschii* (www.tigr.org) provides the opportunity to explore the complete proteome of a free-living biological system with only 1736 predicted ORFs. This proteome has been used to develop a prototype Web-based database (http://proteomeweb.anl.gov) that incorporates not only links

TABLE I
A Sampling of Two-Dimensional Gel Electrophoresis Web Sites.[a]

Site	URL	Status
SIENA-2DPAGE	http://www.bio-mol.unisi.it/2d/2d.html	F
The Danish Centre for Human Genome Research	http://biobase.dk/cgi-bin/celis	F
Harefield Heart Science Centre	http://www.harefield.nthames.nhs.uk/nhli/protein/	F
Aberdeen Proteome Facility, University of Aberdeen, Scotland	http://www.abdn.ac.uk/~mmb023/2dhome.htm	PF
Max Planck Institute for Infection Biology	http://www.mpiib-berlin.mpg.de/home.html	F

[a]The sites listed represent a subset of the ''federated'' (**F**) or ''partially federated'' (**PF**) sites listed by Swiss- 2D-PAGE. The rules for becoming a federated site (Appel *et al.*, 1996) can be found at http://ca.expasy.org/ch2d/fed-rules.html.

to existing genome and protein sequence databases, but also tools for querying the proteome data in the context of genome information. The use of hyperlinked documents in a Web interface offers the opportunity to organize the complex volumes of data produced by proteome analyses. In this manner, all of the information associated with each protein spot in a 2DE pattern can be nested within a series of linked pages. Users have the opportunity to look at the associated genome sequence and annotation as well as Web sites that provide metabolic pathway information, results of experiments designed to modulate the proteins expressed, and tools for querying the 2DE results. This sort of Web-based data management is one approach to the integration and management of proteome and genome databases and demonstrates how the integration of genome and proteome information enhances the value of both data types.

Analysis of the *M. jannaschii* proteome by 2DE coupled with peptide mass spectrometry has, thus far, produced a total of 170 protein identifications for 166 proteins analyzed (Giometti *et al.*, 2002). The ORFs associated with each *M. jannaschii* protein cut from 2DE gels and then analyzed by peptide mass spectrometry are presented as hyperlinks in a master 2DE gel pattern (Fig. 1; see color insert). Deviation from a one-to-one correlation between the number of proteins analyzed (isolated as distinct entities from 2DE patterns) and the number of ORFs identified has been observed in a variety of biological systems, including *Helicobacter pylori* (Lock *et al.*, 2001), *Saccharomyces cerevisiae* (Santucci *et al.*, 2000), *Mycobacterium tuberculosis* (Jungblut *et al.*, 2001), and human keratinocytes (Celis *et al.*, 1998), as well as

M. jannaschii and has several explanations. The identification of multiple ORFs in what appears to be a single protein spot in a 2DE gel patterns is an indication that multiple proteins comigrate or migrate closely to one another. If the genome sequence predicts isoelectric point and molecular weights for the proteins that are similar, the assumption can be made that the proteins comigrate under the separation conditions used and additional experiments to improve the separation can be done. In some cases, however, there is a significant deviation between the predicted and observed isoelectric points and molecular weights (Fig. 2; see Color Insert, and Table II), suggesting actual physical associations were retained through sample preparation and separation. These data are important to understanding the total proteome of the organisms, providing function as well as expression information, and, therefore, are retained as part of the interactive database for retrieval and analysis.

Another reason for the lack of correlation between the number of proteins analyzed and the number of ORFs represented is the occurrence of posttranslational modifications that produce multiple protein components arising from the same gene product. Figure 3 (see Color Insert), for example, shows six different *M. jannaschii* proteins that contain tryptic

TABLE II
Methanococcus jannaschii Open Reading Frames Identified within Protein Spot 4 in Whole Lysate Two-Dimensional Gel Electrophoresis Patterns Based on Peptide Mass Data[a]

ORF	Annotation	Predicted pI	Predicted MW
Mj0842	Methyl coenzyme M reductase I, subunit β	5.6	47,789
Mj0845	Methyl coenzyme M reductase I, subunit γ	5.3	30,174
Mj0846	Methyl coenzyme M reductase I, subunit α	5.1	61,266

[a]An apparently single protein spot (master spot 4) was cut from replicate 2DE gels after separation of proteins from *M. jannaschii* lysates. The mass spectra of the tryptic digest from spot 4 yielded three reproducible matches (a fourth, MJ0083, was detected in only one of two analyses) with high levels of confidence. The predicted molecular weights for these three proteins are significantly different, suggesting that they would not comigrate in the sodium dodecyl sulfate environment of the second-dimension separation unless they were associated as a complex. The sum of the predicted molecular weights for these three proteins approximates the observed molecular weight for spot 4 (see Fig. 2).

peptides similar to those predicted for MJ ORF0324, annotated as translation elongation factor EF-1, subunit α (http://www.tigr.org). Comparison of the isoelectric point and molecular weight predicted from the genome sequence with the actual values observed on calibrated 2DE patterns indicates that one of the proteins within this set of proteins is the unmodified gene product (Fig. 4; see Color Insert) whereas the other protein species deviate from the predicted value in either molecular weight or isoelectric point. The molecular weight deviations suggest peptide removal through a proteolysis process whereas the isoelectric point variants indicate posttranslational modification such as deamidation or phosphorylation. Although these modifications remain to be characterized, information about the heterogeneity in expression of this gene product is preserved in the proteome database and the appropriate annotations will be added as data relevant to the modifications are obtained. Access to this and other information related to the proteins encoded by a genome should also be provided through the genome sequence databases in order to provide complete annotation.

Changes in the pattern of protein expression in response to specific circumstances are the major component of proteome studies because such changes are central to the biological response of an organism. Quantitative changes in protein expression can be assessed from 2DE patterns, using a variety of commercial [Progenesis (Nonlinear USA, Durham, NC); PDQuest (Bio-Rad, Hercules, CA); Melanie III (GeneBio, Geneva, Switzerland)] and proprietary software [e.g., Tycho (Anderson *et al.*, 1981; Giometti *et al.*, 1991)]. Integration of these quantitative changes with protein and genome annotations provides valuable information with respect to gene regulation and the regulation of protein function. For example, when *M. jannaschii*, a methanogen, is grown with lower than normal hydrogen pressure, numerous proteins are expressed in altered abundance relative to cells grown under control conditions (Fig. 5; see Color Insert). By comparing tryptic peptide masses with the ORF database, some of these proteins were identified as the enzymes involved in methanogenesis (Table III). Overlaying these quantitative data from the proteome analysis on the linear arrangement of *M. jannaschii* genes (Fig. 6; see color insert) indicates that these proteins are encoded by genes located within a cluster and are therefore possibly subject to coregulation. Thus, the integration of the quantitative data from the proteome analysis with the genome sequence provides the opportunity to interpret changes in protein expression in the context of the physical arrangement of the genome. A database that provides the capability to accumulate and assimilate such quantitative protein expression data provides a new dimension to the genome information.

TABLE III
Proteins Significantly[a] Decreased in Abundance in *Methanococcus jannaschii* Cells Grown with Decreased Hydrogen Pressure[b]

2DE protein spot number	CT/LOH	Mj ORF	Annotation
8	1.7	MJ0217	H^+-transporting ATP synthase, subunit A
51	1.6	MJ0219	H^+- transporting ATP synthase, subunit C
57	1.6	MJ0845	Methyl CoM reductase I, γ
64	1.5	MJ0851	N^5-Methyl tetrahydromethanopterin CoM methyltransferase subunit A
65	1.5	MJ0854	N^5-Methyl tetrahydromethanopterin CoM methyltransferase subunit H
106	2.1	MJ1338	H_2-dependent methylenetetrahydro-methopterin DH
115	2.5	MJ1338	H_2-dependent methylenetetrahydro-methopterin DH

[a] $p < 0.05$.
[b] These protein spots were found to be decreased in abundance in the two-dimensional gel electrophoresis patterns of *M. jannaschii* whole cell extract proteins when cells were grown with reduced hydrogen pressure relative to control cultures, that is, 5 kPa compared with 100 kPa. The relative abundance of each protein spot in control compared with experimental patterns is expressed as the average integrated density of the control divided by the average integrated density of the experimental (CT/LOH) (Giometti *et al.*, 2001). Averages were calculated from three replicate patterns, each, of one control and one experimental sample. The Mj ORF identifications and annotations are taken from the *M. jannaschii* genome database at the Institute for Genome Research (TIGR) Web site (http://www.tigr.org).

Links to metabolic pathway databases provide an additional level of integration between proteome and genome databases. Quantitative data from the proteome analysis can be superimposed on metabolic pathway diagrams to visualize the location of the enzymes that are altered in expression within the relevant metabolic pathways. Using the example of *M. jannaschii* protein expression under conditions of depleted hydrogen again, the enzymes found to be significantly altered in abundance (Fig. 5 and Table III) can be located in the KEGG metabolic pathway database according to their ORF assignments or EC numbers (Fig. 7; see Color Insert). By using colors that correlate with the observed change in abundance, the coordinated response of the organism to downregulate methanogenesis is quite obvious. No significant change was observed in many of the other enzymes in the pathway, suggesting those proteins are encoded by genes regulated through an independent mechanism. The

interpretation of the changes in protein expression obviously gains a greater depth when placed in the framework of metabolic pathways.

IV. Database Integration

A critical issue that obviously needs to be addressed under the topic of proteomics and bioinformatics is the need to integrate nucleic acid and protein sequence databases with proteome data and databases (Fig. 8). Whereas some standardization of genome and protein sequence database architecture is in place, no standards currently exist for proteomics data. More expression data, both from the mRNA and the protein levels, as well as protein identifications, need to be made public in some type of standard format. To further complicate the situation, the integration of genome and proteome databases should also include, for the sake of comprehensive

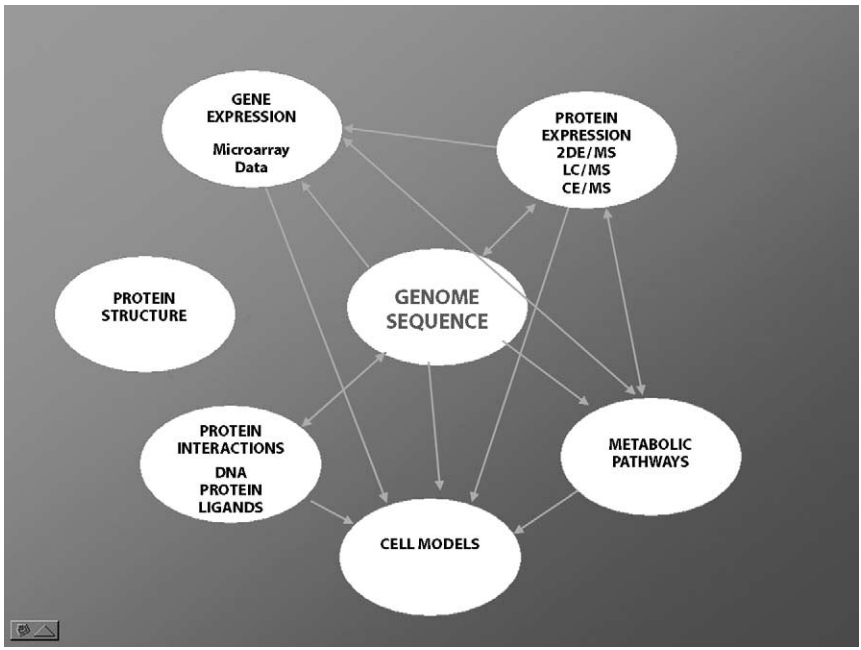


FIG. 8. Genome and proteome database integration. Bioinformatics must evolve to incorporate the data output from the diverse fields of proteomics, as well as functional genomics. Integration of these diverse databases is essential to the full utilization of both genome and proteome information. This diagram depicts an idealized view of how the feedback between databases should evolve in order to fully capitalize on the information available within the individual databases.

coverage, gene expression databases (i.e., those databases containing the results of microarray analyses), metabolic pathway databases, and protein interaction databases. The example of the *M. jannaschii* Web database discussed above illustrates how connections can be made from proteome data to genome and protein sequence databases using current software capabilities. There are, at the present time, however, no obvious connections between existing genome and protein sequence databases and proteome and gene expression databases. Because, for a majority of researchers, the first portal of entry to the biological databases is still most often through the genome sequence, links need to be provided from the genome to the proteome databases. In addition, methods must be developed for the revision of genome sequence annotation when proteome analyses show inconsistencies, omissions, or provide new information.

As shown by Jungblut *et al.* (2001), the algorithms currently used to define ORFs sometimes miss genes, and the information produced by analysis of proteomes should be incorporated into the corresponding genome annotations to provide more comprehensive information. Genes annotated as ''hypothetical'' and ''conserved hypothetical'' in the genome sequence databases are another example of how feedback between genome and proteome databases would provide added value to the genome sequence. As more proteome data are assimilated, the assignment of function to such hypothetical proteins based on patterns of protein expression in response to varied stimuli should be possible.

Proteome analysis can also provide evidence of posttranslational modifications that could be included in the genome sequence annotation or cross-referenced from the genome sequence database to the appropriate proteome database. For example, *M. jannaschii* has been observed to produce flagellin proteins with altered isoelectric point and molecular weight under a variety of growth conditions, including reduced hydrogen concentration (Giometti *et al.*, 2001). An additional annotation in the *M. jannaschii* genome sequence database indicating this post translational modification would be invaluable to researchers searching gene sequence databases for similar structural changes in other organisms.

V. Conclusions

As this work is written, in the minds of most scientists bioinformatics relates to the acquisition, analysis, and management of genome sequence data. Although there is some effort in the area of integrating genome and protein sequence databases, the integration of proteome data, defined as information pertaining to the abundance, subcellular location, and

conditions under which a protein is expressed, with genome databases is an issue that is as yet unaddressed. In the future, however, bioinformatics will of necessity come to include the acquisition, analysis, and management of such protein expression data as well as the integration of those data with genome and protein sequence databases. Once in place, the integration of genome and proteome databases will no doubt reveal the regulatory mechanisms and metabolic interactions that control biological processes. Comparative analysis of the metabolic processes of whole cells will be facilitated through the comparison of protein expression and genome sequence data from diverse cell types and the dream of computational cell modeling will become a reality.

Acknowledgments

This work was supported by the U.S. Department of Energy, Office of Biological and Environmental Research, under Contract W-31-109-ENG-38. The editorial comments as well as the scientific contributions of Sandra L. Tollaksen, Tripti Khare, and Gyorgy Babnigg in the preparation of this manuscript are gratefully acknowledged.

References

Anderson, N. L., Taylor, J., Scandora, A. E., Coulter, B. P., and Anderson, N. G. (1981). The TYCHO system for computer analysis of two-dimensional gel electrophoresis patterns. *Clin. Chem.* **27,** 1807–1820.

Appel, R. D., Bairoch, A., Sanchez, J. C., Vargas, J. R., Golaz, O., Pasquali, C., and Hochstrasser, D. F. (1996). Federated 2-DE database: A simple means of publishing 2-DE data. *Electrophoresis* **17,** 540–546.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402.

Bader, G. D., and Hogue, C. W. (2000). BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **16,** 465–477.

Blom, N., Gammeltoft, S., and Brunak, S. (1999). Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294,** 1351–1362.

Butler, B. A. (1998). Sequence analysis using GCG. *In* ''Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins'' (A. D. Baxevanis, and B. F. F. Ouellette, Eds.), pp. 74–97. John Wiley & Sons, New York.

Celis, J. E., Østergaard, M., Jensen, N. A., Gromova, I., Rasmussen, H. H., and Gromov, P. (1998). Human and mouse proteomic databases: Novel resources in the protein universe. *FEBS Lett.* **430,** 64–72.

Dayhoff, M. O., Eck, R. V., Chang, M. A., and Sochard, M. R. (1965). *"Atlas of Protein Sequence and Structure,"* Vol. 1. National Biomedical Research Foundation, Silver Spring, MD.

D'Souza, M., Romine, M. F., and Maltsev, N. (2000). SENTRA, a database of signal transduction proteins. *Nucleic Acids Res.* **28,** 335–336.

Gaasterland, T., Sczyrba, A., Thomas, E., Aytekin-Kurban, G., Gordon, P., and Sensen, C. W. (2000). MAGPIE/EGRET annotation of the 2.9-Mb *Drosophila melanogaster* Adh region. *Genome Res.* **10,** 502–510.

George, D. G., Dodson, R. J., Garavelli, J. S., Haft, D. H., Hunt, L. T., Maarzec, C. R., Orcutt, B. C., Sidman, K. E., Srinivasarao, G. Y., Yeh, L. S. L., Arminski, L. M., Ledley, R. S., Tsugita, A., and Barker, W. C. (1997). The Protein Information Resource (PIR) and the PIR-International Protein Sequence Database. *Nucleic Acids Res.* **25,** 24–28.

Giometti, C. S., and Taylor, J. (1991). The application of two-dimensional electrophoresis to mutation studies. *In* ''Advances in Electrophoresis'' (M. J. Dunn, Ed.), pp. 359–389. Walter de Gruyter, New York.

Giometti, C. S., Tollaksen, S. L., Babnigg, G., Reich, C. I., Olsen, G. J., Lim, H., and Yates, J. R.,III. (2001). Structural modifications of *Methanococcus jannaschii* flagellin proteins revealed by proteome analysis. *Eur. J. Mass Spectrom.* **7,** 195–205.

Giometti, C. S., Reich, C., Tollacksen, S., Babnigg, G., Lim, H., Zhu, W., Yates III, J., and Olsen, G. (2002). Global analysis of a ''simple'' proteome: *Methanococcus jannaschu. J. Chromatog. B.* **782,** 227–243.

Griffin, T. J., and Aebersold, R. (2001). Advances in proteome analysis by mass spectrometry. *J. Biol. Chem.* **276,** 45497–45500.

Humphery-Smith, I., Cordell, S. J., and Blackstock, W. P. (1997). Proteome research: Complementarity and limitations with respect to the RNA and DNA worlds. *Electrophoresis* **18,** 1217–1242.

Jungblut, P. R., Muller, E.-C., Mattow, J., and Kaufmann, S. H. E. (2001). Proteomics reveals open reading frames in *Mycobacterium tuberculosis* H37Rv not predicted by genomics. *Infect. Immun.* **69,** 5905–5907.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28,** 27–30.

Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Paley, S. M., and Pellegini-Toole, A. (2000). The EcoCyc and MetaCyc databases. *Nucleic Acids Res.* **28,** 56–59.

Kihara, D., Shimizu, T., and Kanehisa, M. (1998). Prediction of membrane proteins based on classification of transmembrane segments. *Protein Eng.* **11,** 961–970.

Langen, H., Takacs, S., Evers, S., Berndt, P., Lahm, H.-W., Wipf, B., Gray, C., and Fountoulakis, M. (2000). Two-dimensional map of the proteome of *Haemophilus influenzae. Electrophoresis* **21,** 411–429.

Li, L., Masselon, C. D., Anderson, G. A., Pasa-Tolic, L., Lee, S.-W., Shen, Y., Zhao, R., Lipton, M. S., Conrads, T. P., Tolic, N., and Smith, R. D. (2001). High-throughput peptide identification from protein digests using data-dependent multiplexed tandem FTICR mass spectrometry coupled with capillary liquid chromatography. *Anal. Chem.* **37,** 3312–3322.

Link, A. J., Hays, L. G., Carmack, E. B., and Yates, J. R.,III. (1997). Identifying the major proteome components of *Haemophilus influenzae* type-strain NCTC 8143. *Electrophoresis* **18,** 1314–1334.

Lipman, D. J., and Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* **227,** 1435–1441.

Lock, R. A., Cordwell, S. J., Coombs, G. W., Walsh, B. J., and Forbes, G. M. (2001). Proteome analysis of *Helicobacter pylori*: Major proteins of type strain NCTC 11637. *Pathology* **33,** 365–374.

Mann, M., Hendrickson, R. C., and Pandy, A. (2001). Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70,** 437–473.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heihne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10,** 1–6.

Overbeek, R., Larsen, N., Pusch, G., D'Souza, M., Selkov, E., Jr., Kyrpides, N., Fonstein, M., Maltsev, N., and Selkov, E. (2000). WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28,** 123–125.

Patterson, S. D., and Aebersold, R. (1995). Mass spectrometric approaches for the identification of gel-separated proteins. *Electrophoresis* **16,** 1791–1814.

Pearson, W. R. (1994). Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol.* **24,** 307–331.

Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.

Santucci, A., Trabalzini, L., Bovalini, L., Ferro, E., Neri, P., and Martelli, P. (2000). Differences between predicted and observed sequences in *Saccharomyces cerevisiae*. *Electrophoresis* **21,** 3717–3723.

Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Shevchenko, A., Boucherie, H., and Mann, M. (1996). Linking genome to proteome by mass spectrometry: Large-scale identification of yeast proteins from two-dimensional gels. *Proc. Natl. Acad. Sci. USA* **93,** 14440–14445.

Smith, R. D., Pasa-Tolic, L., Lipton, M. S., Jensen, P. K., Anderson, G. A., Shen, Y., Conrads, T. P., Udseth, H. R., Harkewicz, R., Belov, M. E., Masselon, C., and Veenstra, T. D. (2001). Rapid quantitative measurements of proteomes by Fourier transform ion cyclotron resonance mass spectrometry. *Electrophoresis* **22,** 1652–1668.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680.

Washburn, M. P., Wolters, D., and Yates, J. R., III. (2001). Large-scale analysis of yeast proteome by multidimensional protein identification technology. *Nat. Biotech.* **19,** 242–247.

Wasinger, V. C., Cordell, S. J., Cerpa-Poljak, A., Yan, J. X., Gooley, A. A., Wilkins, M. R., Duncan, M. W., Harris, R., Williams, K. L., and Humphery-Smith, I. (1995). Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*. *Electrophoresis* **16,** 1090–1094.

Wasinger, V. C., Pollack, J. D., and Humphery-Smith, I. (2000). The proteome of *Mycoplasma genitalium*: CHAPS-soluble component. *Eur. J. Biochem.* **267,** 1571–1582.

AUTHOR INDEX

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z

SUBJECT INDEX

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z

(A)

Taxa	Translation of published DNA sequence
1	*S6* MAQR----------TRLGDILRPLNSEYGKVVPGWGTTPVMGVFMALFLVFLLIILQIYNSSLILEGFSVDWAG
	Ph MGQK----------TALSNFLKPFNSNAGKVVPGWGTTPLMGLFMGLLFVFLLIILQIYNSTIVLDAFSVNVGG
2	*Gt* MALR----------TRLGELLRPLNSEYGKVVPGWGTTPAMGFVMLLFFLFLLIILQIYNSSLILENVDVDWASLGN
3	*Cp* MPQR----------TALGNILRPLNSEYGKVAPGWGTTPLMAVFMLLFFVFLLIIIQIYNSSLLLENVQVSWTAATA
4	*Pp* MALR----------TRLGEILRPLNSEYGKVAPGWGTTPIMGIFMLLFFLFLLIILQIYNSSLVLENVDVDWATLGS
	Cc MALK----------TRLGELLRPLNSQYGKVAPGWGTTPIMGIFMALFLLFLIIILQIYNSSLILENLDISWTTLGI
5	*Osi* MALR----------TRLGEILRPLNAEYGKVAPGWGTTPIMGVVMLAFLVFLLIILQIYNSSLIIENVDVDWSNGIV
6	*Mv* MADTSQGKR----------TVVGNFLKPLNSEYGKVAPGWGTTVLMGVFMALFAVFLVIILQIYNASVLLDGIPANWSSLSKY
	No MATGST----SKAKADEST--SKITPLGTALKPLNSEYGKVAPGWGTTPVMGLFMALFAVFLLIILEIYNSSLILDDVGVSWYSLGK
	Cr MATGTSKAKPSKVN-SDFQEPGLVTPLGTLLRPLNSEAGKVLPGWGTTVLMAVFILLFAAFLLIILEIYNSSLILDDVSMSWETLAKVS
	Cv MATGTT----SKVKVSDT---GVSTPLGTLLKPLNSEYGKVAPGWGTTVLMGIFMALFAVFLVIILEIYNSSVLLDDVTMSWESLSK
7	*Eg* MTT-ISKNKTSNSKGKT-------TTLGTILKPLNSKYGKVLPGWGTAGIMLIFMTLFAIFLTIILEIYNSSVILDIK
8	*Mp* MATQIIDD-TPKTKGKK-------SGIGDILKPLNSEYGKVAPGWGTTPLMGIMMALFAVFLVVILELYNSSVLLDGVSVSW
9	*Zf* MATQTIDD-TFRSGPRR-------TVVGNLLKPLNSEYGKVAPG#
10	*Gb* MATQTIDD-TSGSGPRR-------TVIGNLLKPLNSEYGKVAP#
11	*Pt* MATQTIDD-TSKTTPKE-------TLVGTTLKPLNSEYGKVAPGWGTTPLMGFAMALFAVFLSIILEIYNSSVLLDGIPVSWG
12	*Gg* MATQTVND-TSRPRPKK-------TGVGSYLKPLNSEYGKVAPG#
13	*Lt* MATQTVEG-SSRSGPRR-------TITGDLLKPLNSEYGKVAPG#
14	*Cca* MATQIIED-SSRSAPRR-------TLVGDLLKPLNSEYGKVAPG#
	Nod MATQTIEG-SSRSGPRR-------TIVGDLLKPLNSEYGKVAPG#
15	*Lf* MATQTVED-SSRSGPRR-------TLVGDLLKPLNSEYGKVAPG#
	Sc MATQTVED-SSRSGPRR-------TIIGDLLKPLNSEYGK#
16	*Dw* MATQTVEG-SSRSRPRR-------TTTGNLLKPLNSEYGKVAPG#
17	*Cd* MATQTIEG-SSRSGPRR-------TVVGDLLKPLNSEYGKVAPG#
18	*Ac* MATQTVEG-SSRSGPRR-------TIVGDLLKPLNSEYGKVAPG#
	Zm MATQTVED-SSRPKPKR-------TGAGSLLKPLNSEYGKVAPGWGTTPFMGVAMALFAIFLSIILEIYNSSVLLDGILTN
	Os MATQTVED-SSRPGPRQ-------TRVGNLLKPLNSEYGKVAPGWGTTPFMGVAMALFAVFLSIILEIYNSSVLLDGILMN
	Ta MATQTVED-SSKPRPKR-------TGAGSLLKPLNSEYGKVAPGWGTTPFMGVAMALFAIFLSIILEIYNSSVLLDGILTN
	Hv MATQTVED-SSKPRPKR-------TGAGSLLKPLNSEYGKVAPGWGTTPFMGVAMALFAIFLSIILEIYNSSILLDGILTN
	Sce MATQTVED-SSKPRPKR-------TGAGSLLKPLNSEYGKVAPGWGTTPFMGVAMALFAIFLSIILEIYNSSVLLDGILTN
19	*Cj* MATQTVEG-SSRSGPRR-------TIVGDLLKPLNSEYGK#
	Tar MVTQTVEG-SSRSGPRR-------TITGDLLKPLNSE#
	Nt MATQTVEN-SSRSGPRR-------TAVGDLLKPLNSEYGKVAPGWGTTPLMGVAMALFAVFLSIILEIYNSSVLLDGISMN
	Oa MATQTAEE-SSRARPKK-------TGLGGLLKPLNSEYGKVAPGWGTTPLMGLAMALFAVFLSIILEIYNSSVLLDGISMN
	Oe MATQTAEE-SSRARPKK-------TGLGGLLKPLNSEYGKVAPGWGTTPLMGLAMALFAVFLSIILEIYNSSVLLDGISMN
	At MATQTVED-SSRSGPRS-------TTVGKLLKPLNSEYGKVAPGWGTTPLMGVAMALFAVFLSIILEIYNSSVLLDGISVN
	Pd MATQSVEG-SSRSGPRR-------TIVGDLLKPLNSEYGKVAPGWGTTPLMGVAMALFAVFLSIILEIYNSSVLLDGISMN
	Lj MATQTVED-SSRARPRQ-------TSVGSLLKPLNSEYGKVAPGWGTTPLMGIAMALFAIFLSIILEIYNSSILLDGISMN
	Ps MATQTVEN-SSRSGPRR-------TAVGDLLKPLNSEYGKVAPGWGTTPLMGIAMALFAVFLSIILEIYNSSLLDQISMN
	So MATQTVES-SSRSRPKP-------TTVGALLKPLNSKYGKVAPRWGTTPLMGVAMALFAVFLSIILEIYNSSVLLDGISMN

(B)

Confirmation of predicted errors in Spinach PsbH sequence by Spinach chloroplast genome sequence (Observed mass 7599.5 ±1.2 Da)

		Calc. mass (Da)
1986	MATQTVESSSRSRPKPTTVGALLKPLNSKYGKVAPRWGTTPLMGVAMALFAVFLSIILEIYNSSVLLDGISMN	7697.1
1998	MATQTVESSSRSRPKPTTVGALLKPLNSEYGKVAPGWGTTPLMGVAMALFAVFLSIILEIYNSSVLLDGISMN	7598.9
2000	MNIIRFMATQTVESSSRSRPKPTTVGALLKPLNSEYGKVAPGWGTTPLMGVAMALFAVFLSIILEIYNSSVLLDGISMN	8505.0

(C)

```
27 N   S   K   Y   G   K   V   A   P   R   W   G
   aat tcg aaa tat ggt aaa gta gct cct agg tgg gga  1986
   aat tcg gaa tat ggt aaa gta gct cct ggg tgg gga  2000
27 N   S   E   Y   G   K   V   A   P   G   W   G
```

WHITELEGGE *ET AL.*, FIG. 8. (A) PsbH sequences aligned to show the probable errors in the spinach sequence at residues 29 and 36 (red), as well as the conserved residues (blue) and other changes from the consensus (purple). PsbH sequences that provided no information in the region of interest were not included. #, Gene sequence information is truncated in a partial clone. (B) Confirmation of spinach sequence errors predicted by this methodology. 1986, Translation of first published DNA sequence (Westhoff *et al.*, 1986; X07106); 1998, predicted amino acid sequence based on LC-MS data and PsbH consensus (changes in red) (S. M. Gómez and J. P. Whitelegge, unpublished); 2000, Translation of the revised DNA sequence submitted in the complete spinach chloroplast genome [Schmitz-Linneweber (2001), direct submission to GenBank March 6, 2000; AJ400848). Residues not present in the mature protein are underlined. (C) Comparison of the DNA sequence of spinach *psbH* from the 1986 (Westhoff *et al.*, 1986; X07106) sequence and the 2000 (AJ400848) sequence. The two G-to-A transitions creating the E29K, G36R ''mutant'' are indicated. Red, ''mutant''; blue, ''wild type.'' Taxa abbreviations: 1, Cyanobacteria; 2, Cryptomonadacea; 3, Glaucocystophyta; 4, Rhodophyta (red algae); 5, Bacillariophyta (diatoms); 6, Chlorophyta (green algae); 7, Euglenida; 8, Marchantiophyta (liverworts); 9, Cycadophyta; 10, Ginkgophyta; 11, Coniferophyta; 12, Gnetophyta; 13, Magnoliales; 14, stem Magnoliophyta incertae sedis; 15, Piperales; 16, Winterales; 17, Ceratophyales; 18, Liliopsida (monocots); 19, eudicotyledons (dicots); 13–19, Magnoliophyta (flowering plants). *S6*, *Synechocystis* PCC6803, X58532; *Ph*, *Prochlorothrix hollandica*, X60314; *Gt*, *Guillardia theta*, AF041468; *Cp*, *Cyanophora paradoxa*, U30821; *Pp*, *Porphyra purpurea*, U38804; *Cc*, *Cyanidium caldarium*, AF022186; *Osi*, *Odontellasinensis*, Z67753; *Mv*, *Mesostigma viride*, AF166114; *No*, *Nephroselmis olivacea*, AF137379; *Cr*, *Chlamydomonas reinhardtii*, Z15133; *Cv*, *Chlorella vulgaris*, AB001684; *Eg*, *Euglena gracilis*, X70810; *Mp*, *Marchantia polymorpha*, X04465; *Zf*, *Zamia furfuracea*, AF188846; *Gb*, *Ginkgo biloba*, AF123851; *Pt*, *Pinus thunbergii*, D17510; *Gg*, *Gnetum gnemon*, AF123852; *Lt*, *Liriodendron tulipifera*, AF123855; *Cca*, *Cabomba caroliniana*, AF123845; *Nod*, *Nymphaea odorata*, AF188851; *Lf*, *Lactoris fernandeziana*, AF123854; *Sc*, *Saururus cernuus*, AF123856; *Dw*, *Drimys winteri*, AF123850; *Cd*, *Ceratophyllum demersum*, AF123847; *Ac*, *Acorus calamus*, AF123843; *Zm*, *Zea mays*, X86563; *Os*, *Oryza sativa*, X12695; *Ta*, *Triticum aestivum*, X04710; *Hv*, *Hordeum vulgare*, X14107; *Sce*, *Secale cerale*, X07672, *Cj*, *Cercidiphyllum japonicum*, AF123848; *Tar*, *Trochodendron aralioides*, AF123857; *Nt*, *Nicotiana tabacum*, Z00044; *Oa*, *Oenothera argilicolla*, X55899; *Oe*, *Oenothera elata* subsp. *hookeri*, AJ271079; *At*, *Arabidopsis thaliana*, AP000423; *Pd*, *Populus deltoides*, Y13328; *Lj*, *Lotus japonicus*, AP002983; *Ps*, *Pisum sativum*, AF153442; *So*, *Spinacia oleracea*, X07106.

(A)

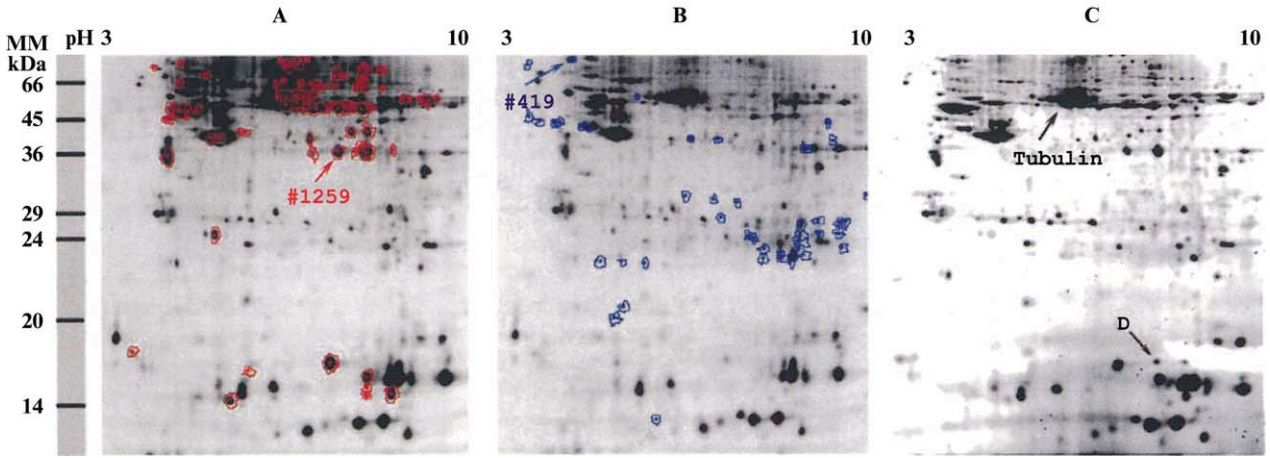
	Taxa	Translation of Published DNA Sequence
1	*S6*	MESVYAILVLTMALAVLFFAIAFREPPRIEK
2	*Gt*	METLVYTFLLIGTLAVLFAAVFFRDPPRIAKK
3	*Cp*	MEALVYTFLLVTTLGILFFSIIFRDPPRINQ
4	*Pp*	MEALVYVFLLTGTLMVIFFAIFFREPPRIAK
	Cc	MEALVYVFLLIGTLVVIFFAIFFRDPPRIAKK
5	*Osi*	MEALVYTFLLIGTLMVIFFAVFFRETPRILRK
6	*Mv*	MEALVYTFLLVGTLGIIFFAIFFREPPRI
	No	MEALVYTFLLISTLGIIFFGIFFREPPRI
	Cr	MEALVYTFLLVGTLGIIFFSIFFRDPPRMIK
	Cv	MEALVYTFLLVGTLGIIFFAIFFREPPRIVK
7	*Eg*	MEALVYTFLLIGTLGVIFFAIFFRESPRIN
8	*Mp*	MEALVYTFLLVGTLGIIFFAIFFREPPKVPSKGKK
9	*Zf*	MEALVYTFLLVSTLGIIFFAIFFREPPKVPDRGSK
10	*Gb*	MEALVYTFLLVSTLGIIFFAIFFRDPPKVPNEGRK
11	*Pt*	MEALVYTFLLVSTLGIIFFAIFFREPPKLPTKGGK
12	*Gg*	MEALVYTFLLVSTLGILFFAIFFREPPRVTPKGGK
13	*Lt*	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKKMK
14	*Cca*	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKKAK
	Nod	MEALVYTFLLVSTLGIIFFAIFFREPPKVPNKKMK
	Atr	MEALVYTFLLVSTLGIIFFAIFFREPPKVPNKKIK
15	*Lf*	MEALVYTFLLVSTLGIIFFAIFFREPPRISTKKMK
	Sc	MEALVYTFLLVSTLGIIFFAIFFREPPRISTKKMK
	Aca	MEALVYTFLLVSTLGIIFFAIFFREPP?ILTKKTK
16	*Dw*	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKR
17	*Cd*	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKK
18	*Cf*	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKKMK
19	*Ip*	MEALVYTFLLVSTLGIIFFAIFFREPPKVPNKKMK
20	*Ac*	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKKMK
	Db	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKKMK
	Zm	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKK
	Os	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKKVK
	Ta	MEALVYTFLLVSTLGIIFFAIFFREPPKVPPTPTKRIK
	Hv	MEALVYTFLLVSTLGIIFFAIFFREPPKVPPTPTKRIK
	Sce	MEALVYTFLLVSTLGIIFFAIFFREPPKVPPTPTKRIK
21	*Cj*	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKKMK
	Tar	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKKMK
	Nt	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKKN
	Oa	MEALVYTFLLVSTLGIIFFAIFFREPPKIQTKRRNDF
	Oe	MEALVYTFLLVSTLGIIFFAIFFREPPKIQTKRRNDF
	So	MEALVYTFLLVSTLGIIFFAIFFREPPKISTKK
	At	MEALVYTFLLVSTLGIIFFAIFFREPPKISTKK
	Pd	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKKVK
	Lj	MEALVYTFLLVSTLGIIFFAIFFREPPKVPTKK
	Ps	MEALVYTFLLVSK-ELVFFAIFFREPPKVPTKNTK

(B)

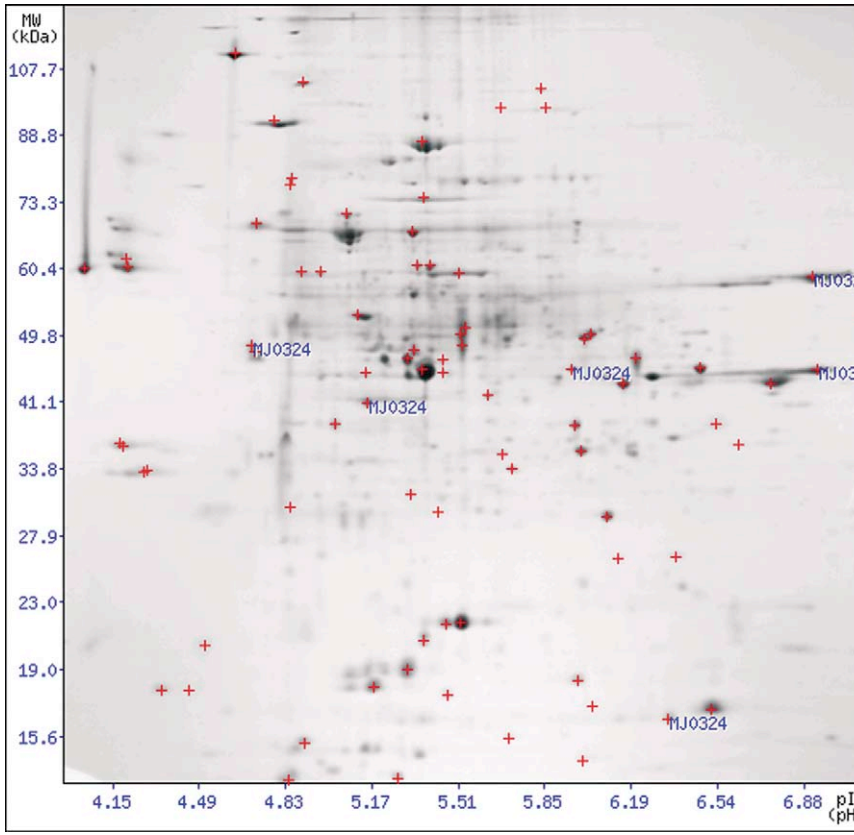
Pea *psbT* (Observed peptide mass-4060.1 Da) compared to related *psbT* sequences showing an insertion and a deletion in the pea gene.

```
10 L   V   S   K       E   L   V   F   F       Published - 4032.8 Da
   tta gtg tcg a^^ ^^a gga att agtt ttt ttc Ps
   tta gtc tcg act tta gga atc a tt ttt ttc Lj
   tta gta tcc act tta ggg ata a tt ttt ttc At
   tta gtc tcg act cta ggg ata a tt ttt ttc Nt
10 L   V   S   T   L   G   I   I   F   F       Corrected - 4060.9 Da
```

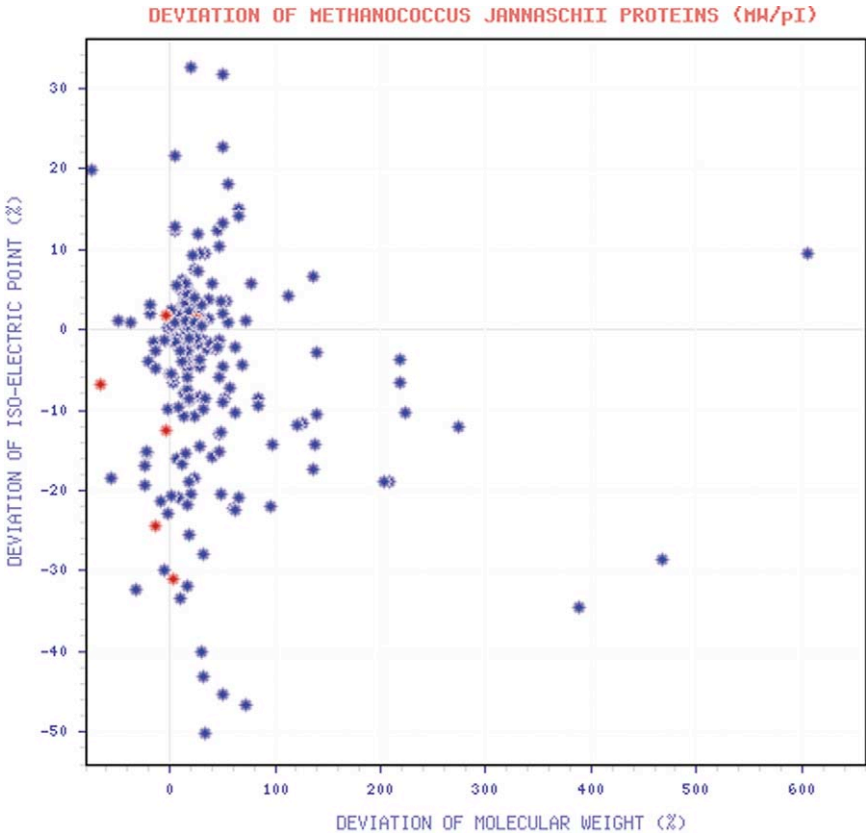
WHITELEGGE *ET AL.*, FIG. 9. (A) PsbT sequences aligned to show the probable errors in the pea sequence from residues 13 to 16 (red), as well as conserved residues (light blue) and other changes from the consensus (purple) at residues 13 to 17. (B) Identification of the ''mutation'' in pea *psbT*. The ''mutation'' is shown in red and the ''wild type'' in blue. Taxa abbreviations: 1, Cyanobacteria; 2, Cryptomonadacea; 3, Glaucocystophyta; 4, Rhodophyta (red algae); 5, Bacillariophyta (diatoms); 6, Chlorophyta (green algae); 7, Euglenida; 8, Marchantiophyta (liverworts); 9, Cycadophyta; 10, Ginkgophyta; 11, Coniferophyta; 12, Gnetophyta; 13, Magnoliales; 14, stem Magnoliophyta incertae sedis; 15, Piperales; 16, Winterales; 17, Ceratophyales; 18, Laurales; 19, Illiciales; 20, Liliopsida (monocots); 21, eudicotyledons (dicots); 13–21, Magnoliophyta (flowering plants). *S6*, *Synechocystis* PCC6803, D64000; *Gt*, *Guillardia theta*, AF041468; *Cp*, *Cyanophora paradoxa*, U30821; *Pp*, *Porphyra purpurea*, U38804; *Cc*, *Cyanidium caldarium*, AF022186; *Osi*, *Odontellasinensis*, Z67753; *Mv*, *Mesostigma viride*, AF166114; *No*, *Nephroselmis olivacea*, AF137379; *Cr*, *Chlamydomonas reinhardtii*, X64066; *Cv*, *Chlorella vulgaris*, AB001684; *Eg*, *Euglena gracilis*, X70810; *Mp*, *Marchantia polymorpha*, X04465; *Zf*, *Zamia furfuracea*, AF188846; *Gb*, *Ginkgo biloba*, AF123851; *Pt*, *Pinus thunbergii*, D17510; *Gg*, *Gnetum gnemon*, AF123852; *Lt*, *Liriodendron tulipifera*, AF123855; *Cca*, *Cabomba caroliniana*, AF123845; *Nod*, *Nymphaea odorata*, AF188851; *Atr*, *Amborella trichopoda*, AF235042; *Lf*, *Lactoris fernandeziana*, AF123854; *Sc*, *Saururus cernuus*, AF123856; *Aca*, *Asarum canadense*, AF123844; *Dw*, *Drimys winteri*, AF123850; *Cd*, *Ceratophyllum demersum*, AF123847; *Cf*, *Calycanthus floridus*, AF123846; *Ip*, *Illicium parviflorum*, AF123853; *Ac*, *Acorus calamus*, AF123843; *Db*, *Dioscorea bulbifera*, AF123849; *Zm*, *Zea mays*, X04422; *Os*, *Oryza sativa*, X15901; *Ta*, *Triticum aestivum*, X54947; *Hv*, *Hordeum vulgare*, X14107; *Sce*, *Secale cerale*, X07672; *Cj*, *Cercidiphyllum japonicum*, AF123848; *Tar*, *Trochodendron aralioides*, AF123857; *Nt*, *Nicotiana tabacum*, Z00044; *Oa*, *Oenothera argilicolla*, X55899; *Oe*, *Oenuthera elata* subsp. *hookeri*, X55900; *So*, *Spinacia oleracea*, X02945; *At*, *Arabidopsis thaliana*, AP000423; *Pd*, *Populus deltoides*, Y13328; *Lj*, *Lotus japonicus*, AP002983; *Ps*, *Pisum sativum*, AF153442.



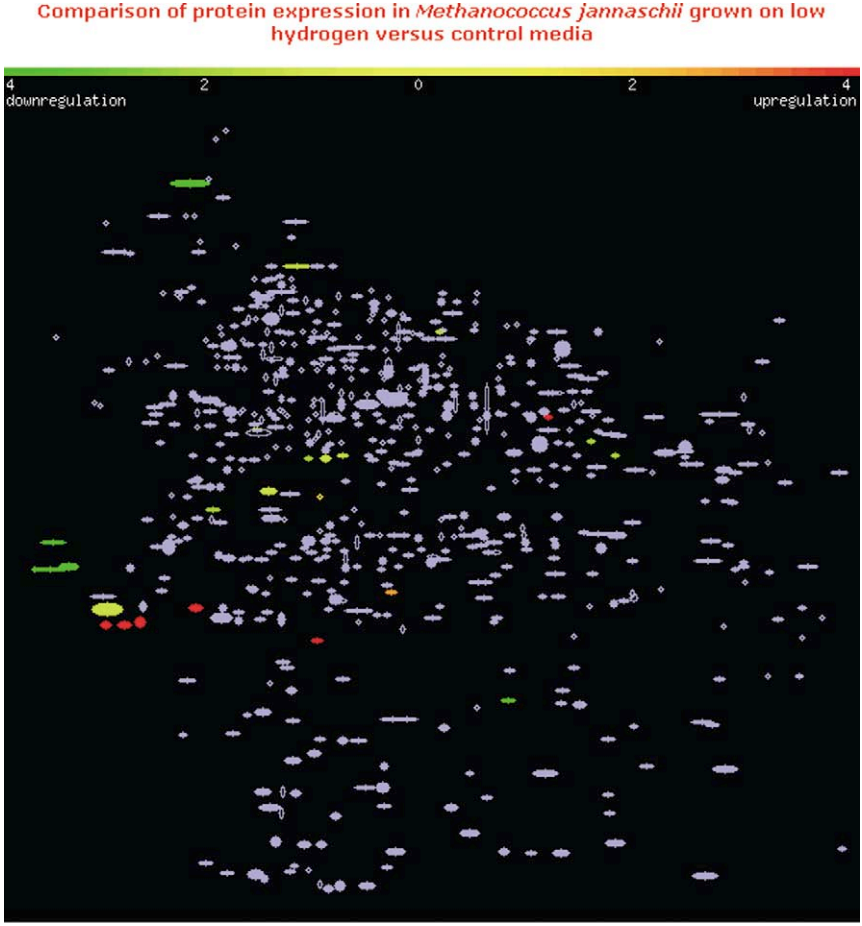
ZHOU AND YU, FIG. 10. Differential gel electrophoresis (DIGE) gel images of normal and esophageal cancer cells. Images of the Cy3- and Cy5-labeled proteins separated by 2D-PAGE are shown in (A) and (B), respectively. A SYPRO Ruby-stained gel image, which gives a combined protein profile of both cell types, is shown in (C). The subtle difference in protein profiles between the Cy3 and Cy5 results and the SYPRO Ruby-stained gel results from the change in molecular weight of the proteins that are covalently labeled with the fluorescent dyes (D). Spots identified as annexin I (#1259), tumor rejection antigen gp96 (#419), and tubulin are shown on the gels.



GIOMETTI, FIG. 3. Multiple *M. jannaschii* proteins associated with the same ORF. MJ0324, the *M. jannaschii* ORF annotated as translation elongation factor EF-1 α subunit, has been identified as a component of more than one protein spot in the 2DE patterns of *M. jannaschii* proteins. Although the biochemical analysis to characterize these different proteins is in progress, the hypothesis is that the molecular weight differences could be the result of protein processing through proteolysis, while the isoelectric point heterogeneity could indicate posttranslational modification such as deamidation or phosphorylation.

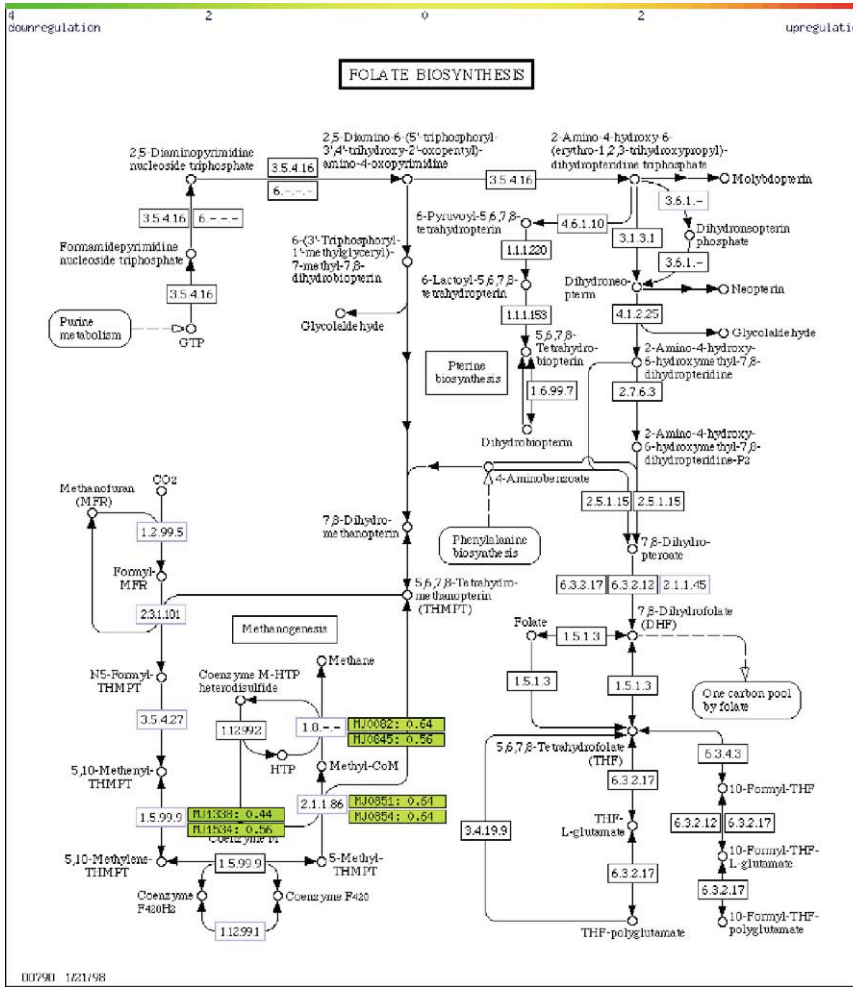


GIOMETTI, FIG. 4. Comparison of the predicted and actual 2DE migration position of the protein product of MJ0324. The red spots indicate the actual measured isoelectric point and molecular weight of the *M. jannaschii* protein spots (see Fig. 3) producing tryptic peptides with masses comparable to those predicted for the product of ORF MJ0324. The red spot closest to the zero intercept for both the *x* and *y* axes indicates the protein that most closely matches the isoelectric point and molecular weight predicted by the ORF sequence. Deviations greater than zero in molecular weight suggest incompletely dissociated protein complexes, while deviations less than zero indicate proteolysis. Deviations in the isoelectric point suggest posttranslational modifications.

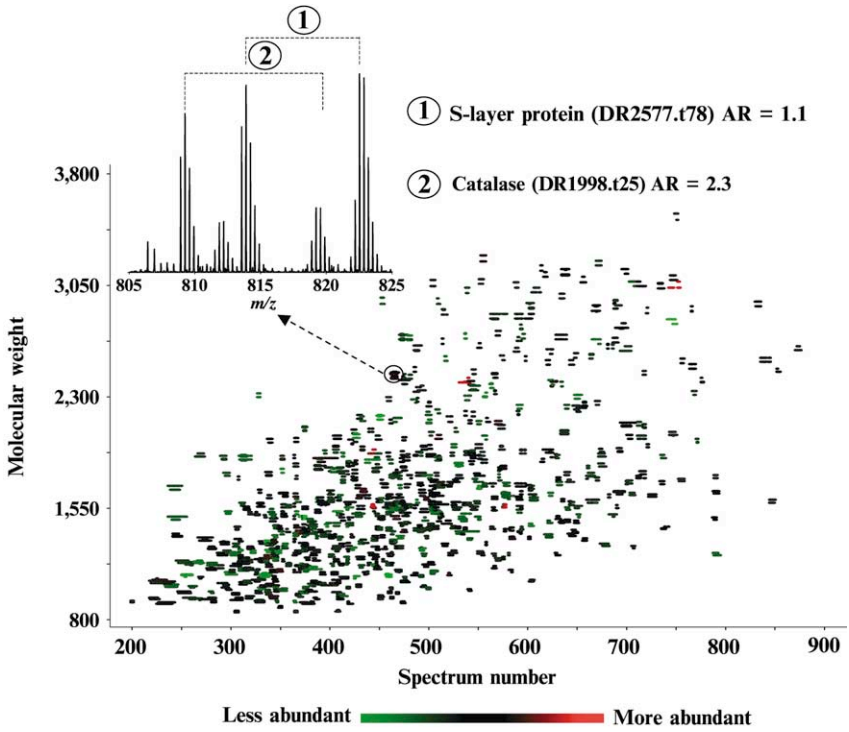


GIOMETTI, FIG. 5. Quantitative differences in *M. jannaschii* protein expression when cells are grown with lower than control levels of hydrogen. This display is an example of how quantitative data from 2DE analyses can be interfaced with protein identifications and genome sequence. By using color tables to indicate quantitative differences, users can see which proteins are modulated by growth conditions. Hyperlinks to genome (see Fig. 6) and protein identification information are provided. In this example, *M. jannaschii* cells were grown with either control (100 kPa) or decreased (5 kPa) hydrogen pressure and the expressed proteins were analyzed for changes in abundance after separation by 2DE. Quantitative analysis of the silver-stained 2DE patterns was done using the TYCHO suite of algorithms (Giometti and Taylor, 1991)

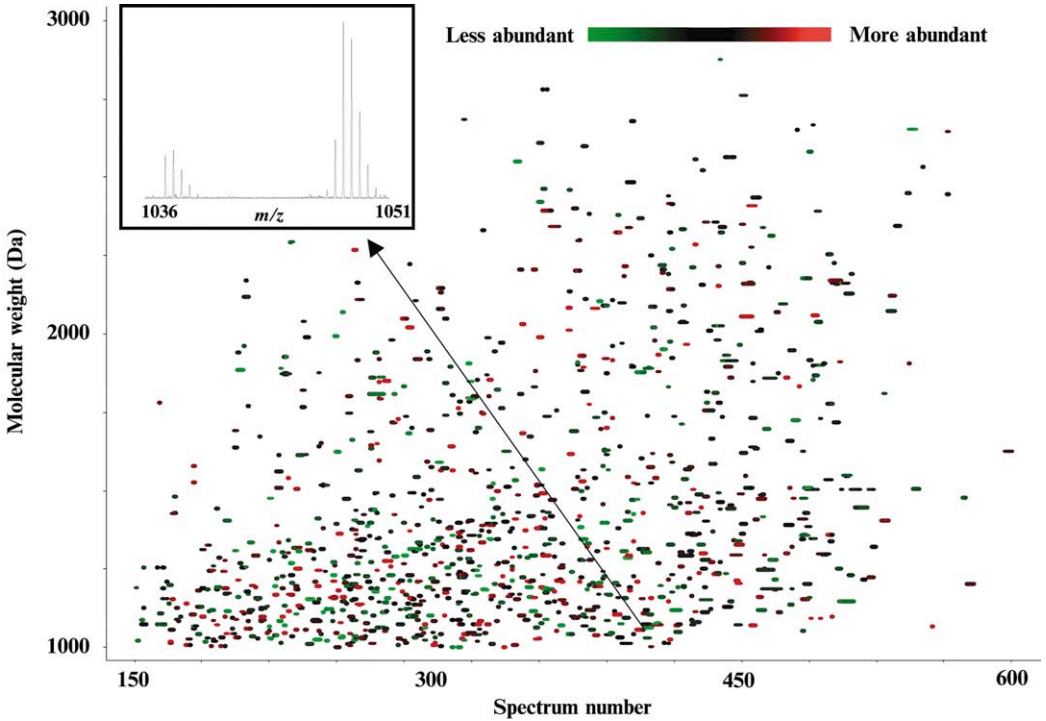
GIOMETTI, FIG. 6. Protein expression data overlaid on DNA sequence information provides indication of coregulated genes. Open reading frames can be clustered into predicted regulatory groups according to specified DNA sequence parameters. When the ORFs associated with expressed proteins are identified, significant changes in the abundance of proteins resulting from changes in growth conditions can be overlaid on the gene sequence and physical associations within the genome can be deduced. This is one example of interfacing protein expression data with genome sequence information.



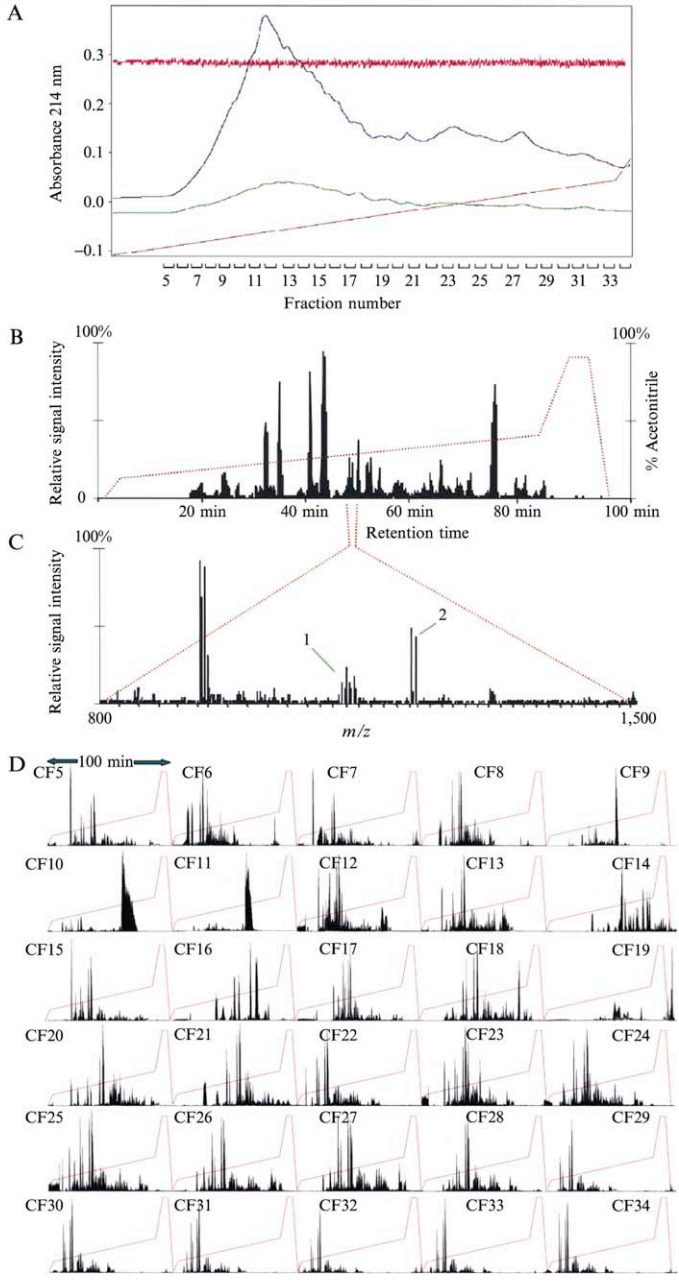
GIOMETTI, FIG. 7. Protein expression in the context of metabolic pathways. The metabolic effects of fluxes in the abundance of subsets of proteins within a cell are crucial to cell survival. Placing observed changes in the abundance of proteins within the context of metabolic pathways provides an overview of the total cellular response to changing environments. In this diagram, the *M. jannaschii* proteins observed to change in abundance in response to growth under lower than normal hydrogen pressure are shown within a copy of the KEGG metabolic pathway for folate biosynthesis. The colors used are the same as shown in Fig. 5 and indicate the relative change in abundance for these proteins compared with cells grown with control hydrogen pressure. Using such visual representations of the quantitative data collected from proteome analyses, changes in the abundance of proteins can be directly related to metabolic changes.



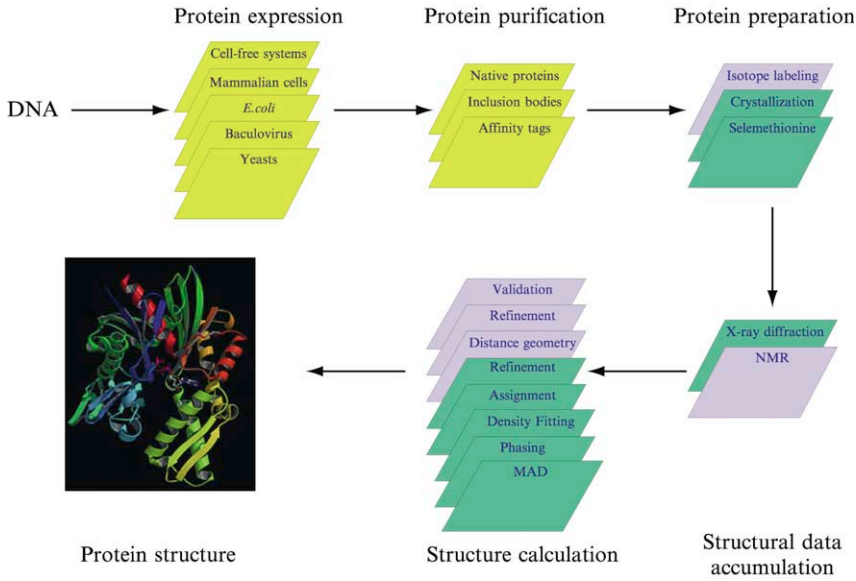
SMITH *ET AL.*, FIG. 11. Two-dimensional display comparing relative protein abundances of control and H_2O_2-treated *D. radiodurans* cells. The colored spots provide a representation of the relative expression level of the peptides (green represents a decrease in abundance, black unchanged, and red an increase). *Inset*: Results for two selected AMTs corresponding to S-layer protein and catalase observed in the control and H_2O_2-treated cells, along with their calculated abundance ratios (ARs). The ARs are determined by the summation of the set of peak intensities (all isotopic peaks) for the contiguous set of spectra during which each peptide elutes.



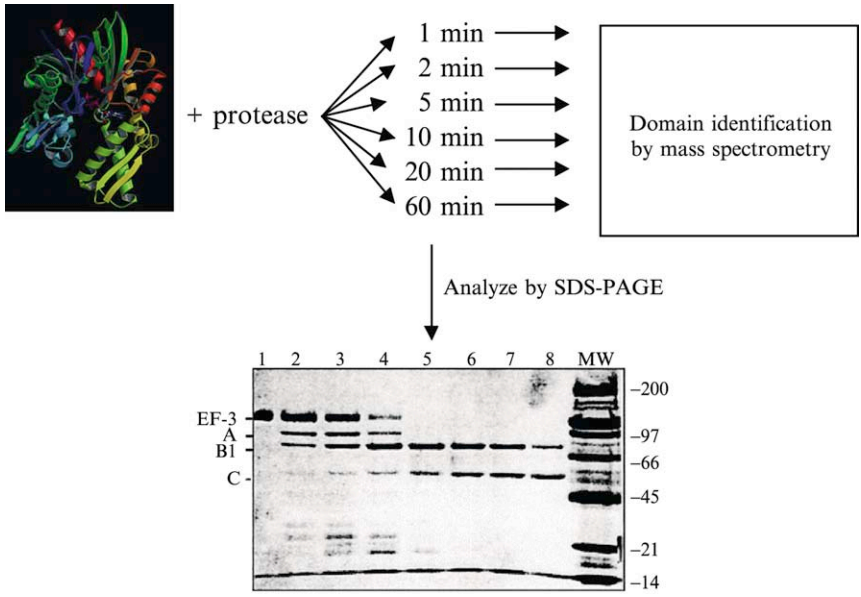
SMITH *ET AL.*, FIG. 12. Two-dimensional (2D) comparative display of a capillary LC-FTICR analysis of *D. radiodurans* after exposure to 17.5 kGy [65] of ionizing radiation. Cells were harvested 0, 3, 7, 9, and 12 h after exposure. The 2D comparative display corresponds to 3 h after exposure for unlabeled cells compared with the ^{15}N-labeled reference proteome (control, nonirradiated cells). The colored spots provide a representation of the relative expression level of the peptide (green represents a decrease in abundance, black unchanged, and red an increase) compared with the reference. For example, each spot corresponds to two peptides having well-defined mass differences (*inset*) whose abundances are used to calculate an abundance ratio (AR).


A
Absorbance 214 nm
0.3
0.2
0.1
0.0
−0.1
5 7 9 11 13 15 17 19 21 23 25 27 29 31 33
Fraction number
B
100%
100%
Relative signal intensity
% Acetonitrile
0
20 min
40 min
60 min
80 min
100 min
Retention time
C
100%
Relative signal intensity
1
2
800
1,500
m/z
D
100 min
CF5
CF6
CF7
CF8
CF9
CF10
CF11
CF12
CF13
CF14
CF15
CF16
CF17
CF18
CF19
CF20
CF21
CF22
CF23
CF24
CF25
CF26
CF27
CF28
CF29
CF30
CF31
CF32
CF33
CF34

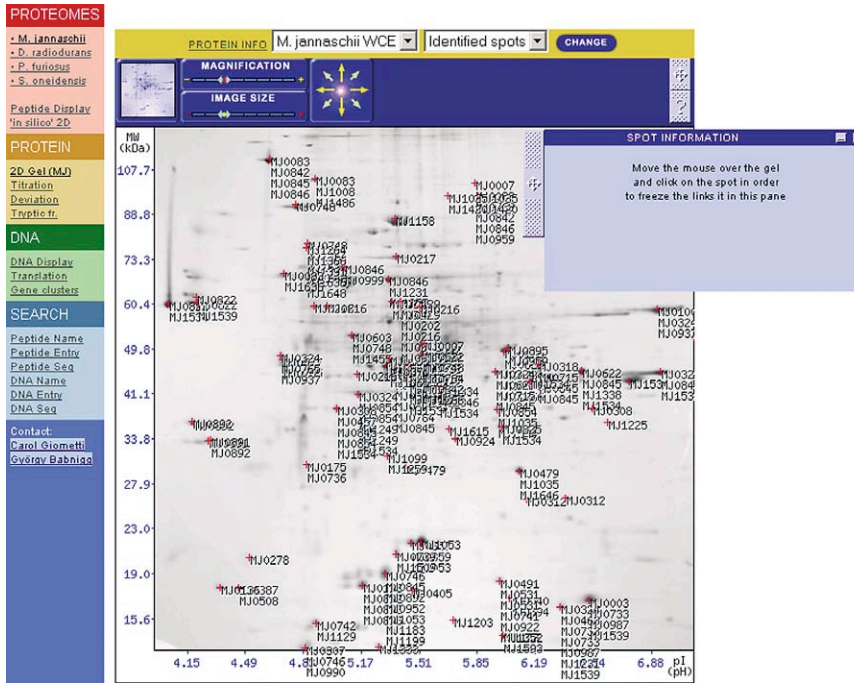
ISSAQ, FIG. 5. Multidimensional liquid chromatography tandem mass spectrometric analysis of ICAT-labeled peptides. (A) ICAT-labeled peptides contained in trypsin-digested HL-60 microsomal fraction are initially separated by strong cation-exchange chromatography into different fractions. (B) The biotinylated, cysteine-containing peptides contained in cation-exchange fractions are then analyzed by LC-ESI-MS/MS, as shown for cation-exchange fraction 18. (C) The MS spectrum of peptides detected in thre 30-s time window indicated in (B) displays several pairs of ICAT-labeled peptides. (D) Base peak ion chromatogram of all of the cation-exchange fractions collected, indicating ICAT peptide distribution. The solvent gradient is indicated by the red lines.



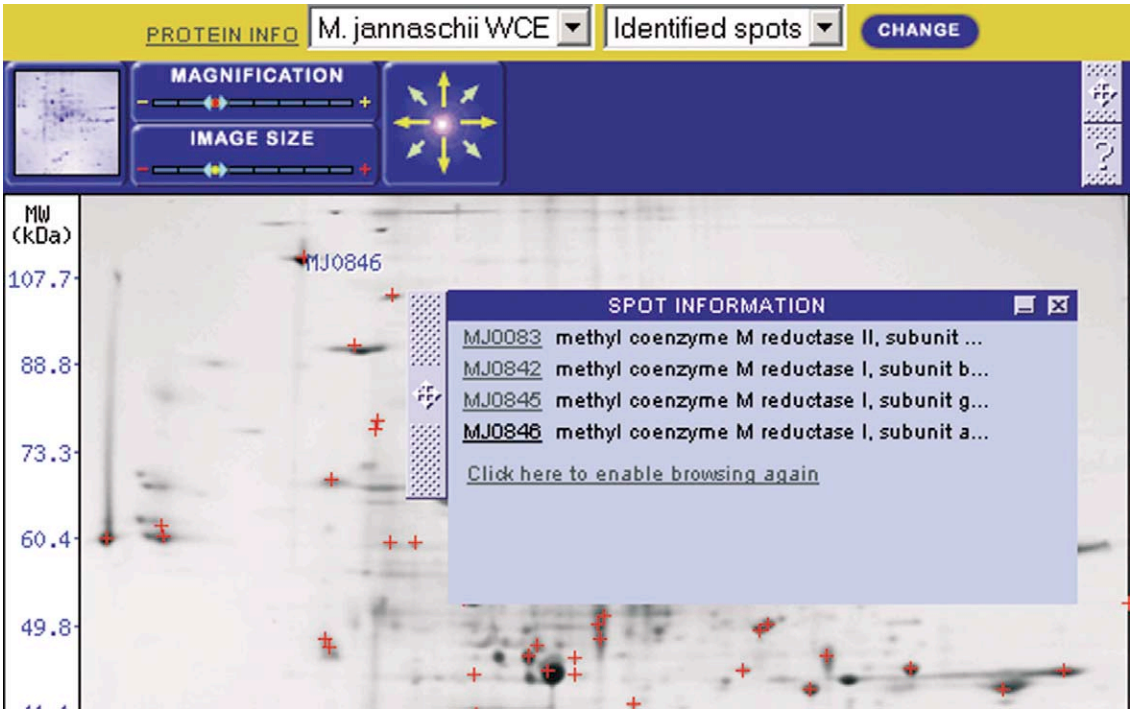
KOTH AND EDWARDS, FIG. 1. The sequence of preparative and analytical steps proceeding from the coding DNA to the three-dimensional protein structure. At each step from DNA expression to structure calculation, procedures or options common to X-ray crystallography and NMR spectroscopy are shown in yellow, X-ray-specific steps are shown in green, and NMR-specific steps are shown in blue.



KOTH AND EDWARDS, FIG. 2. Schematic showing the steps carried out during a limited proteolysis experiment designed to identify a functional domain(s) within a multidomain protein.



GIOMETTI, FIG. 1. Identified proteins in the *Methanococcus jannaschii* proteome display on http://proteomeweb.anl.gov. This display demonstrates the complexity of information provided by proteome analyses using 2DE. All the ORFs associated with *M. jannaschii* proteins cut from 2DE gels and analyzed by peptide mass spectrometry after trypsin digestion are shown. The red crosses indicate hyperlinks to text windows that contain links to genome and protein sequence databases as well as to tools for querying those databases.



GIOMETTI, FIG. 2. Multiple *M. jannaschii* proteins in a single protein spot on a 2DE pattern. This display shows the ORFs identified on the basis of the masses of tryptic peptides from the protein spot indicated as MJ0846 (master spot 4; see Table II). All these proteins are involved in the terminal step of methanogenesis and are hypothesized to be present in the intact cells as a complex that is incompletely disrupted by the conditions that were used for 2DE sample preparation (9 *M* urea, 2% α-mercaptoethanol, 4% Nonidet P-40, and 2% ampholytes, pH 8–10).