A Mass Spectrometric Journey into Protein and Proteome Research

Ruedi Aebersold
Institute for Systems Biology, Seattle, Washington, USA

It is a frequently debated question whether technology drives biology or whether biology drives the development of new technologies. This issue is discussed in this manuscript as an account that covers approximately a decade during which mass spectrometry and protein biochemistry have intersected. It is shown that the capabilities of the mass spectrometric methods, initially developed to address the specific need to identify proteins reliably and at high sensitivity soon transcended the intended task. The rapid development of mass spectrometric technologies applied to protein research has catalyzed entirely new experimental approaches and opened up new types of biological questions to experimentation, culminating in the field of proteomics. Some conclusions from this case study relating to technological research and the environment in which it is carried out are also discussed. (J Am Soc Mass Spectrom 2003, 14, 685–695) © 2003 American Society for Mass Spectrometry

During the decade of the 1990s, changes in mass spectrometry (MS) instrumentation and techniques revolutionized protein chemistry and fundamentally changed the way protein analysis impacts biological research. These changes were catalyzed by two technical breakthroughs in the late 1980s—specifically, the development of the two ionization methods, electrospray ionization (ESI) [1, 2] and matrix-assisted laser desorption/ionization (MALDI) [3]. These methods solved, in an essentially general way, the difficult problem of generating ions of large, nonvolatile analytes, like proteins and peptides, to transfer them directly into the gas phase and into the MS for mass analysis, and to achieve all that without analyte fragmentation [4]. Due to the lack or minimal extent of analyte fragmentation during the ESI and MALDI processes, they are also referred to as "soft" ionization methods. ESI gained immediate popularity because of the ease with which it could be interfaced with popular chromatographic and electrophoretic liquid-phase separation techniques and it quickly supplanted fast atom bombardment [5] as the ionization method of choice for protein and peptide samples dissolved in a liquid phase. Furthermore, due to the propensity of ESI to produce multiply charged analytes, simple quadrupole instruments and other types of mass analyzer with limited *m/z* range could be used to detect analytes with masses exceeding the nominal *m/z* range of the instrument. For different—but no less compelling—reasons, MALDI also rapidly gained popularity. The time-of-flight (TOF) mass analyzer most commonly used with MALDI is robust, simple, sensitive and has a large mass range. Additionally, modern instruments have high mass accuracy and resolution. Furthermore, MALDI mass spectra are simple to interpret, due to the propensity of the method to generate predominantly singly charged ions, and the method is relatively resistant to interference with matrices commonly used in protein chemistry. In addition to the new ionization methods, rapid advances in mass spectrometers and computerized data processing capabilities also advanced MS technology in general and, specifically, its application to protein science.

The predominant goal of many applications of MS technology was, some ten years ago, the identification and characterization of selected, purified proteins at high sensitivity. Rapid technical advances and changes in biological experimentation cooperated in the emergence of a new field of research termed "proteomics." This account attempts to describe this exciting period in protein science and to illustrate how the interplay between technological and biological research resulted in quite unexpected consequences.

The story of the fruitful interaction between mass spectrometry and protein science is by no means complete. In contrast, over the last few years we have witnessed an accelerating pace of technological ad-

Published online June 2, 2003

Address reprint requests to Dr. R. Aebersold, Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103-8904, USA. E-mail: raebersold@systemsbiology.org

1044-0305/03/$30.00
doi:10.1016/S1044-0305(03)00289-7

Received March 3, 2003
Revised April 11, 2003
Accepted April 11, 2003

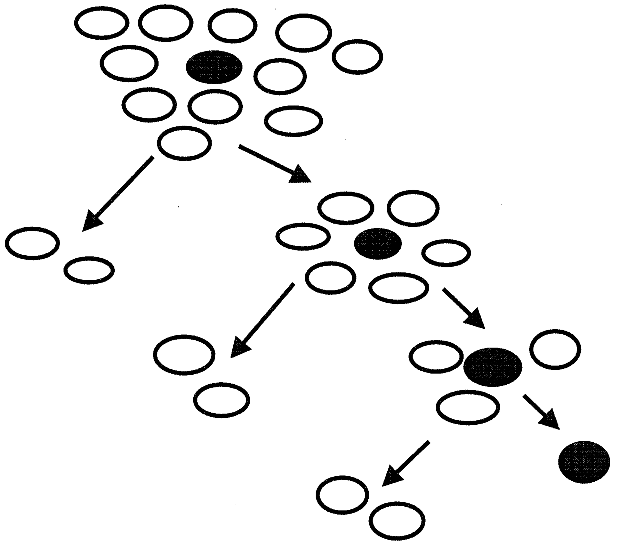

Figure 1. Schematic representation of activity-centered biology. An activity, typically catalyzed by a protein or a protein complex is purified to apparent homogeneity by a series of cycles of sample fractionation and assay for the activity in question. The target activity is indicated as a filled feature and proteins in the sample that are unrelated to the target activity are indicated as open features.

vances, breadth of questions asked and quality of results obtained, and these developments continue unabated. Within the space available for this article, it is not possible to do justice to the whole field, nor was it the intention to produce a comprehensive review. It is hoped, however, that the article can communicate some of the excitement of the field to its current and future practitioners.

Phase 1: Activity-Centered Biology

For several decades, the classical biochemical approach has been a mainstay to study biological activities; this strategy is schematically illustrated in Figure 1. Once an activity is identified, a suitable assay is developed to monitor the activity in biological samples. This assay is used, together with sequential protein purification methods such as salting out, ion exchange chromatography, or size exclusion chromatography to purify the protein catalyzing the particular activity to apparent homogeneity. Once the protein is purified, its different properties are studied. These include the 3-D structure, specific activity, dependence on co-factors and, most importantly for further experimentation, the amino acid sequence. With the advent of the powerful methods of molecular cloning some two decades ago, it became sufficient, in principle, to generate limited stretches of contiguous amino acid sequences as templates for degenerate oligonucleotide probes for the cloning the gene coding for the protein out of a gene library. Once the gene was isolated, it could be rapidly and completely sequenced and used for further experimentation into the structure and function of the protein in question. Knowledge of the gene sequence was particularly useful for the generation of larger amounts of the protein by overexpression of the gene in bacteria, yeast or other types of cells and to study the in vivo function of the protein using the powerful methods of site-directed mutagenesis, gene knock-out and gene replacement.

As many proteins could only be purified with great difficulty and in small amounts, the major challenge for protein chemists at the time was the development of ever more sensitive methods to partially sequence proteins. In the 1980s, the vast majority of protein sequences were determined by the Edman degradation, a chemical process that removes one amino acid at a time from the *N*-terminal of a polypeptide; this process was first automatically implemented in 1967 [6]. The emerging mass spectrometric methods to sequence peptides pioneered by Biemann and co-workers, recounted in detail in Biemann, 2002 [7], while promising, were initially not of comparable sensitivity and generality to the Edman degradation. With the development of the gas-phase protein sequencer [8], a sequencing instrument was introduced that was able to sequence quantities of proteins (submicrogram amounts) that were far smaller than the milligram amounts usually purified by the traditional method of protein purification, column chromatography. The development of methods for the isolation, by gel electrophoresis, of proteins in a form compatible with gas-phase chemical sequencing, therefore, provided a significant increase in the overall sequencing sensitivity. Two such methods were particularly useful. In the first, proteins separated by gel electrophoresis were electroblotted on a solid support, to form a replica of the protein pattern in the gel. The protein spots, once detected by staining, could be excised and directly transferred into a sequencer [9, 10]. The second method, a variation on the theme, additionally involved tryptic digestion of the electroblotted proteins on the solid-support and the recovery and separation of the thus generated peptides (Figure 2). Selected peptides were then purified by reverse-phase HPLC and sequenced [11]. In effect, this method provided a peptide mixture from very small (microgram and submicrogram) amounts of proteins and was therefore a useful pre-cursor for the popular peptide mass mapping techniques developed a few years later. The method also provided a general solution to obtaining partial sequence information of all those proteins that contained *N*-terminal modifications that made them refractory to the Edman degradation.

A typical example illustrates the principle and the status of the technology in the mid 1980s. A heroic effort in protein purification by the group of Paul Patterson at Caltech had resulted in a sample that contained very low microgram amounts of a highly enriched activity defined as Cholinergic Differentiation Factor (CDF): A protein that was responsible for inducing a crucial differentiation step in the development of the mammalian nervous system. By applying the electroblotting, in

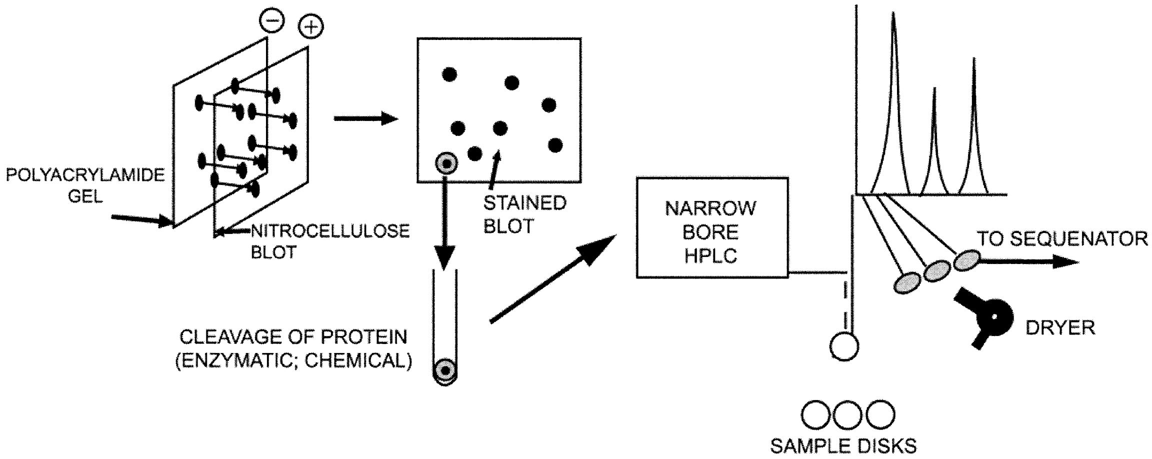

Figure 2. Schematic representation of the process for the generation of peptide fragments from small quantities of protein separated by 2-DE as described in [11]. A protein sample mixture is separated by 2-DE and the proteins are electroblotted onto a nitrocellulose membrane and stained. Specific protein spots are excised from the membrane and the proteins are digested on the membrane, typically with trypsin. The generated peptides are recovered and separated by reverse-phase HPLC and collected for sequencing. In later uses of the method the peptides were directly analyzed by MS or MS/MS for the purpose of identifying the protein(s) contained in the spot analyzed.

situ tryptic cleavage and peptide sequencing method, we were able to obtain partial sequence information from the purified activity [12]. Surprisingly, the obtained peptide sequences matched perfectly to a previously identified protein called leukemia inhibitory factor (LIF), a protein that later became important for maintaining embryonic stem cell lines used for the generation of transgenic mice in an undifferentiated state [13]. Further careful analyses of LIF and CDF—in particular, the comparison of the complete respective gene sequences—indeed conclusively indicated that the same protein carried out both activities. This example illustrates a number of characteristics of an activity-centered biology. First, the method critically depends on the availability of a suitable assay to follow the purification of the activity. In the case of CDF, this was a cell-based assay for which the read-out was obtained after several days. The generation and purification of the factor at an amount compatible with sequencing was a project of several months duration. Furthermore, proteins with no measurable activity or with an activity for which no assay could be developed, were outside the scope of the approach. Second, using gel electrophoresis as the final purification step was a major advance in the field, because the target protein no longer needed to be purified to homogeneity. It was sufficient to prepare a sample in which the target protein was highly enriched and corresponded to a defined electrophoretic band. Third, the protein sequence provided the link between the protein and the gene sequence and conclusively identified the activity under investigation and, as illustrated by the chosen example, it was not unusual to find proteins carrying out more than one defined activity. This multitasking of proteins was later appropriately termed "protein moonlighting" [14]. And fourth, the approach was reductionist; every activity was isolated and studied in isolation, and—a major goal of biochemical projects—the reconstruction of whole biological processes in vitro from the isolated components only rarely succeeded.

The challenges of the activity based biology for the analytical protein chemist were mainly related to obtaining stretches of uninterrupted amino acid sequence from ever decreasing amounts of protein. While of great practical utility, these methods did not represent a significant advance, conceptually. They were intended to obtain the sequence of purified proteins faster and more sensitively and therefore to support a traditional and well-developed approach to study biological systems.

Phase 2: From an Activity-Centered to a Sequence-Centered Biology

Using the electroblotting/microsequencing methods described above, in 1987 we succeeded in obtaining partial amino acid sequences of proteins separated by high-resolution two-dimensional gel electrophoresis (2-DE) [11, 15]. These results attracted a considerable amount of interest because they provided the key for the development of an approach that deviated substantially from the biochemical research method described above. 2-DE was developed in the 1970s independently by Klose [16] and O'Farrell [17] as a gel electrophoretic method for the separation of proteins at high resolution. In the most common implementation of the technique, proteins are separated by isoelectric focusing in a first dimension and then by SDS polyacrylamide gel electrophoresis (SDS-PAGE) according to their size in a second dimension. The separated proteins are then detected by staining and the staining intensity provides an estimate of the quantity of the protein present in each detected spot. It was soon recognized that highly reproducible protein patterns could be generated and that, therefore, the spot patterns from different samples could be overlaid and compared, providing a general method to perform subtractive analysis of the proteins contained in two or more samples. Using this method, proteins that were likely to be related to a particular biological process could be detected among hundreds of detected features.

Based on such 2-DE protein profiles, ideas were developed in the 1970s and 1980s, to build protein databases (e.g., the human protein index) [18] and to apply strategies based on subtractive pattern analysis [19–21] akin to today's popular strategies for the analysis of data obtained from gene expression array experiments. In fact, at that time many of the principles now commonly used for global, quantitative analysis of gene expression patterns, such as the use of clustering algorithms and multivariate statistics, were developed in the context of 2-DE [22, 23]. At that time, however, these ideas were not substantially implemented, mainly because 2-DE by itself was an essentially descriptive technique that did not indicate the identity of the separated proteins. The ability to identify proteins from 2-D gels provided the required link between the observed protein pattern and the sequence of the proteins constituting interesting patterns. Therefore, subtractive 2-D gel pattern analysis could now serve as the assay to identify proteins that seemed to be part of a biological process and these proteins could be sequenced before their function or activity was known.

The following example illustrates the approach and highlights some of its limitations. In collaboration with John Leavitt, then at the Linus Pauling Institute, we performed comparative analysis of 2-D gel protein patterns of two human fibroblast cell lines. The control cell line was a normal fibroblast cell line and the sample cell line had been chemically transformed. It was expected that subtractive analysis of the protein patterns extracted from the two cell lines would identify protein differences and that these differences were related to the chemical transformation. Figure 3 shows a section of the silver stained 2-D gels from the control and transformed cell line, respectively, and indicates that the expected pattern differences indeed were observed. The protein spot marked lpl in Figure 3 was clearly present in the transformed fibroblasts but absent in the control cells. The protein was identified by sequencing as plastin, a protein now known to be involved in the organization of the actin network [24]. The attraction of this approach lies in the fact that complex biological processes such as cell transformation could be studied at the protein level without the need for an assay to probe for a specific activity. In fact, no hypotheses as to which proteins might be involved in the process were required. The limitation of the sequence first approach was the difficulty of identifying the molecular and cellular function of the proteins that were identified. A further limitation of the approach as described was also the limited dynamic range of protein abundance that could be explored by the 2-D gel method [25] and the slow and relatively insensitive protein identification by chemical sequencing.

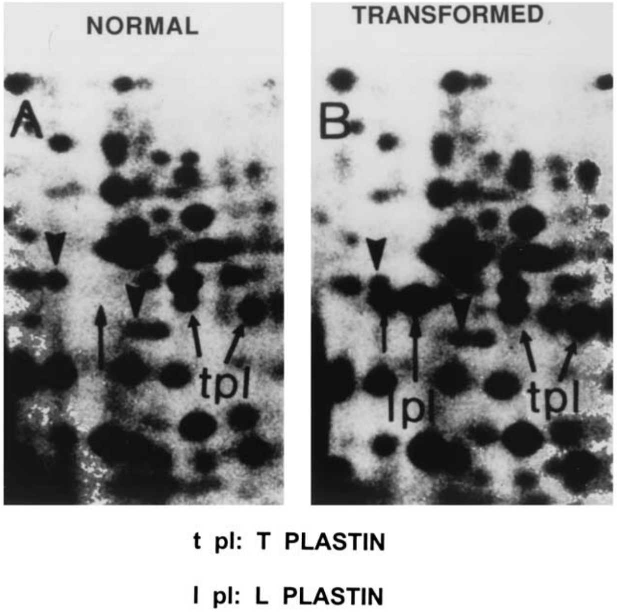

Figure 3. Identification of plastin by subtractive 2-DE [15]. A section of silver stained 2-DE gel patterns of proteins extracted from normal fibroblasts (A) or from transformed fibroblasts (B) is shown. Within a pattern of spots common to both samples a doublet of spots marked with lpl is clearly up regulated in the pattern of the transformed cell. The protein was identified as L Plastin by microsequencing. A related protein T Plastin tpl appeared unchanged under the conditions examined. This example illustrates the sequence first approach to identify proteins involved in complex physiological functions without the need for a specific biochemical assay.

For the analytical chemist, the challenges of the activity-based and sequence-based biology remained similar and were mainly related to the task of conclusively identifying proteins with increasing sensitivity and speed. For the biologist, the main challenge posed by the sequence-centered approach was that s/he was frequently confronted with a large list of seemingly important observations, e.g., proteins, the abundance of which changed under specific experimental conditions; such data are difficult to interpret in a biological context. Conceptually, the process described above marked a significant departure from the traditional protein biochemical approach because no specific hypothesis and the development of no specific assays were needed to study a particular question. This development illustrates that the application of an established technique in a different manner can have a significant impact on the experimental approach used and on the questions asked. These developments also pioneered the idea that the large-scale analysis of gene expression patterns contained a wealth of useful information, an idea that was later very successfully emulated by gene expression array technologies and used with great success [26]. Before the systematic, large-scale analysis of the proteins present in a complex sample could be routinely applied, much improved analytical technologies were needed. It was in this context that in 1994, at the first 2-DE meeting in Siena, Italy, the term "proteome" was coined [27]. The term was defined as the *prote*in complement of the gen*ome*, and the process of studying the proteome was soon thereafter called "proteomics" and mass spectrometry was its most important component.

From First to Second Generation Proteomic Technologies

In the early 1990s, a bewildering array of methods were developed that had the common goal of identifying proteins separated by gel electrophoresis rapidly and sensitively [28]. These methods were the result of a convergence of rapidly improving MS technologies, most notably the development of the ESI and MALDI ionization methods that were capable of routinely ionizing even large polypeptides, the availability of sequence databases with rapidly increasing contents, and the development of computer search algorithms that were capable of correlating mass spectrometric information obtained from single peptides or from the collective peptide mixture generated by the digestion of a protein with sequence databases, thus identifying proteins without the need for de novo sequencing. While the methods differed in detail, two main procedures could be distinguished. The first is known as peptide mapping, peptide mass mapping, or peptide mass fingerprinting. In this method, the match of a list of experimental peptide masses with the calculated list of all peptide masses of each entry in a database (e.g., a comprehensive protein database), identifies the protein. Five groups almost coincidentally developed computer tools for the identification of proteins by peptide mapping [29–33]. Since mass mapping requires an essentially purified target protein, the technique has been commonly used in conjunction with prior protein fractionation by 2-DE. This method, which is mainly carried out with MALDI-TOF mass spectrometers is still widely used today. In the second procedure, protein identification was based on sequence information generated from selected peptides in a tandem mass spectrometer. Since the information contained in collision-induced dissociation spectra (CID) is not readily convertible into a full, unambiguous peptide sequence, the CID spectra are scanned against comprehensive protein sequence databases using one of a number of different algorithms of which Sequest is the prototypical tool [34]. It rapidly became apparent that a strictly 2-DE-based proteomics technology platform was technically complex, labor- and therefore cost-intensive and fundamentally limited as shown below. The increased use of MALDI-MS and ESI-MS/MS for the identification of 2-DE separated proteins also led to the realization that the incidence of co-migration of proteins even in this, the highest resolving protein separation method known, was more prevalent than first thought [25, 35]. Since quantification in 2-DE relies upon the assumption that one protein is present in each spot, co-migration compromises such analyses. It was also observed that with conventional protein staining methods, only a relatively small subset of a cellular proteome is apparent if unfractionated cell lysates are separated [25, 36]. Therefore, in spite of its maturity and unmatched performance for separating intricate patterns of differentially modified and processed proteins [37], and in spite of continued evolution of 2-DE separation and detection technology, alternative methods for large-scale protein expression analysis began to be more vigorously investigated.

Analyzing peptides extracted from MHC class I and class II proteins, Hunt and colleagues laid the groundwork for a gel-independent approach to proteomics by demonstrating the ability of LC-MS/MS systems to handle extremely complex peptide mixtures [38], and it is this method that is today at the core of mass spectrometry-based proteomics. However, before LC-MS/MS could be used for both the identification of protein mixtures and for quantitative proteomic experiments, a number of technical issues had to be addressed, the main one being the inherently poor correlation between the quantity of an analyte present in a sample and the signal intensity generated for that analyte. To add a quantitative dimension to peptide LC-MS/MS experiments, we applied the proven technique of stable isotope dilution [39] to proteome analysis. Stable isotope dilution makes use of the fact that pairs of chemically identical analytes of different stable isotope composition can be differentiated in a mass spectrometer, due to their mass difference, and that the ratio of signal intensities for such analyte pairs accurately indicates the abundance ratio for the two analytes.

To generate pairs of labeled peptides, we synthesized a class of reagents termed "isotope coded affinity tags" (ICAT) reagents and a mass spectrometric method for gel-independent quantitative proteome profiling [40]. The structure of the reagents and the method are schematically illustrated in Figure 4. ICAT reagents consist of three functional elements: a thiol reactive group for the selective labeling of reduced Cys residues, an isotopically coded linker in an isotopically normal (d0) or heavy (d8) form and a biotin affinity tag to allow for selective isolation of labeled peptides. A typical experiment is schematically illustrated in Figure 5. The disulfide bridges of the proteins contained in the sample are reduced under denaturing conditions, and the free sulfhydryl groups of the proteins from the two related samples to be compared are labeled with the isotopically light and heavy forms of the reagent, respectively. The samples are then combined, proteolyzed

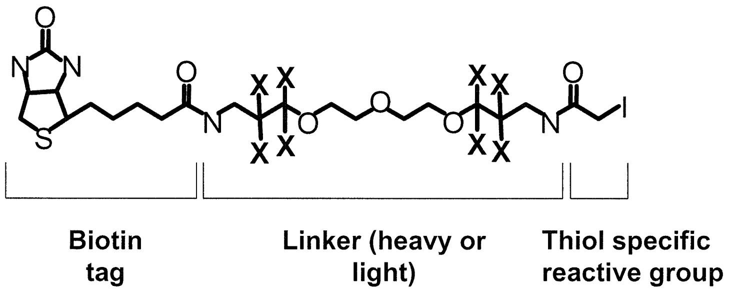

Figure 4. Structure of isotope coded affinity tag (ICAT) reagents. The groups designated with X in the linker indicate the location at which stable heavy isotopes are incorporated into the reagent. In the initial reagents [40] the heavy isotope was deuterium. In second generation reagents the H/D isotope pair was replaced with $^{12}C/^{13}C$.

with trypsin, and the resulting peptides can be separated by any number of optional fractionation steps, including the removal of untagged peptides (i.e., not containing a Cys residue) via avidin affinity chromatography. Peptide/protein identifications are finally made by MS/MS analyses of the individual fractions followed by sequence database searching the observed MS/MS spectra. The observed ratio of the signal intensities for the unfragmented, isotopically light and heavy forms of the same peptide finally yields the relative abundance of that peptide, and hence the protein from which it was derived, in the original samples.

Results from a series of applications of the method have illustrated its versatility, documented current technical limitations and showed new uses for quantitative proteomic analyses. The applications can be broadly grouped into three classes. In the first, quantitative proteomics was used to ask whether the analysis of perturbation-induced changes measured at the mRNA and protein levels provide redundant or complementary information [41, 42]. In the second class, quantitative proteomics was used to gain new insights into specific cellular mechanisms [43–45] and in the third class, applications that use quantitative proteomics for purposes that go beyond simple protein profiling were explored. These include the application of the technique for the analysis of macromolecular complexes [46] and for the systematic analysis of protein phosphorylation [47, 48]. Two of the accompanying papers in this volume [49, 50] illustrate typical applications of the ICAT reagent-based quantitative profiling method.

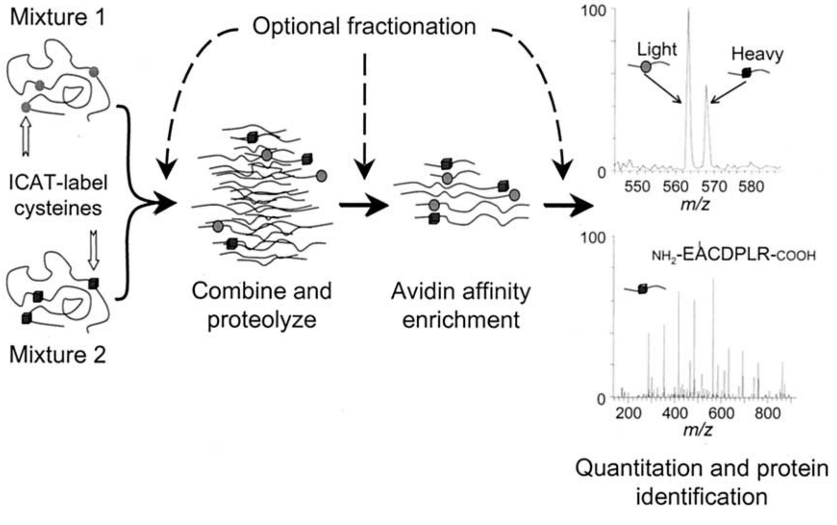

Figure 5. Schematic representation of typical ICAT experiment for quantitative protein profiling. Two protein mixtures are treated with the isotopically light and heavy ICAT reagents, respectively. The labeled protein mixtures are then combined and proteolyzed, tagged peptides are selectively isolated by avidin affinity chromatography and analyzed by MS and MS/MS. The relative abundance is determined by the ratio of signal intensities of the tagged peptide pairs. Every other scan in the mass spectrometer is devoted to fragmenting a peptide. The CID spectra are recorded and searched against large protein sequence databases to identify the protein. Therefore, in a single operation, the relative abundance and sequence of a peptide are determined. The peptide samples analyzed can include cell lysates or fractions thereof. Specifically, proteins in subcellular fractions including microsomes [45], secreted proteins, proteins in protein complexes [46], and proteins in body fluids such as serum have been analyzed by the method.

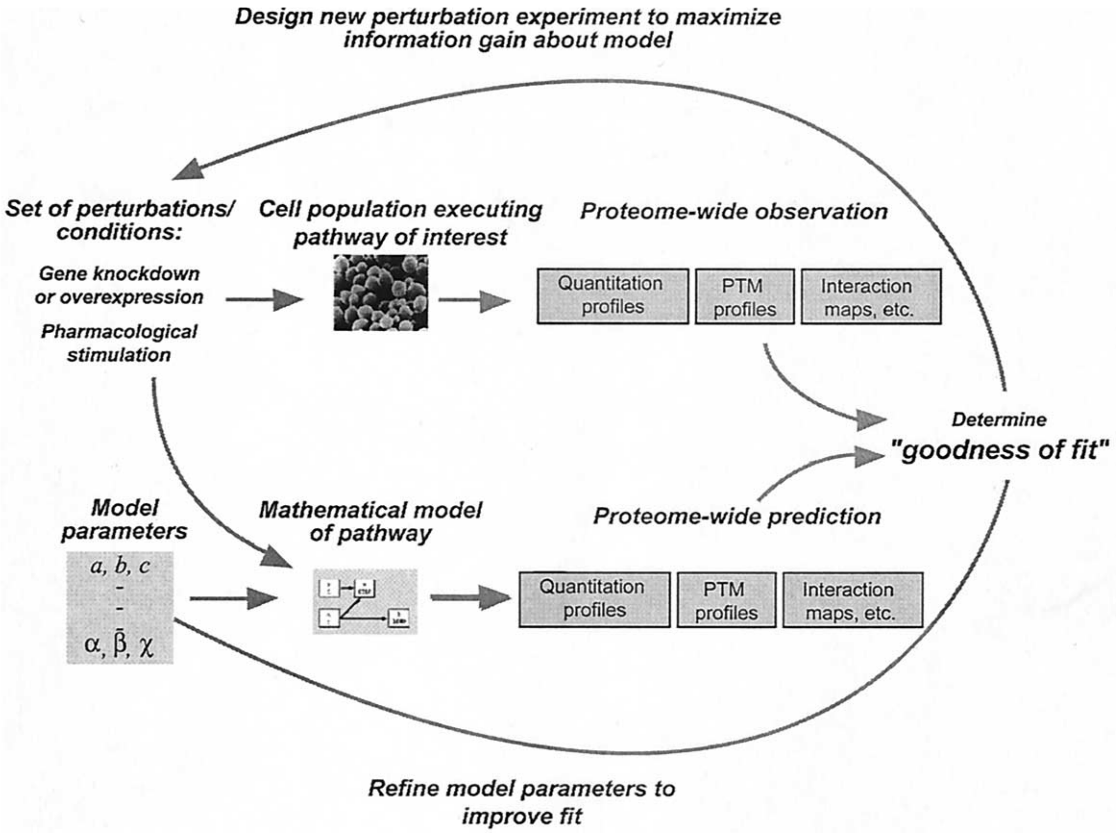
Design new perturbation experiment to maximize information gain about model
Set of perturbations/ conditions:
Gene knockdown or overexpression
Pharmacological stimulation
Cell population executing pathway of interest
Proteome-wide observation
Quantitation profiles
PTM profiles
Interaction maps, etc.
Determine
"goodness of fit"
Model parameters
a, b, c
-
-
α, β̄, χ
Mathematical model of pathway
Proteome-wide prediction
Quantitation profiles
PTM profiles
Interaction maps, etc.
Refine model parameters to improve fit

Figure 6. Schematic illustration of method for the analysis of complex biological systems using large data sets obtained by systematic genomic and proteomic measurements (adapted according to [41]. Cells are subjected to targeted perturbation of selected elements of the system studied. The perturbed sample is subjected to quantitative, systematic measurements and the data obtained are explained within the framework of a model of the process studied. The model is used to make predictions about the expected effects of each perturbation. The predicted and measured effects are reconciled and the perturbation/analysis cycle is continued until the measured and predicted patterns converge.

From a Sequence-Centered Biology to Systems Biology

Above, we identified as one of the major challenges of a sequence-based biology the common situation that a biologist is faced with a long list of seemingly interesting observations, i.e., proteins that change in their abundance under specific experimental conditions and that are difficult to explain in biological terms. Currently, the investigator is forced to select, based on experience, references in the literature, or arbitrarily, one or a few of these observations to carry out validation and follow-up experiments to eventually publish a report on the findings related to that protein or a selected group of proteins. This is process is illustrated by two accompanying papers in this volume [49, 50]. While this approach is successful, it suffers from the obvious disadvantage that all those observations that are not followed up are essentially lost, even though they might also contain very useful information. It has, therefore, become apparent, that new approaches are required to fully exploit the emerging capacity to collect large sets of quantitative data using the proteomic technologies discussed above. A recent study carried out in the yeast *S. cerevisiae* points towards a possible solution of the data interpretation problem [41]. The principle of the approach is illustrated in Figure 6. In this study, both genomic and proteomic data were collected from yeast cells in which all the known components of the galactose induction pathway had been systematically perturbed or eliminated by targeted gene knock out, or by metabolic stimulation. The different types of data, systematically collected, were integrated into a mathematical model consistent with the available information. This model was then used to predict new, previously unknown interactions within the pathway and between the galactose induction pathway and other cellular processes. Some of these predictions were subsequently verified experimentally [41]. The study revealed the striking observation that perturbations in this seemingly simple and relatively confined process that consists of nine core elements resulted in changes in the expression of close to 1000 yeast genes, indicating that this systematic approach has the potential to indicate connections between cellular processes that are difficult to determine by the traditional reductionist research methods. The systematic approach developed in this study is based on the seemingly paradoxical position that it is easier to interpret the information contained in multiple large datasets than in a single one.

For the researcher who plans to apply this iterative approach to systematically study biological processes, several challenges are immediately apparent. First, the capacity to generate large proteomic data sets needs to be available. Currently, every large-scale quantitative proteomic profiling experiment is a major effort requiring several days of mass spectrometer time and countless hours of data analysis. Second, the data from different experiments can only be meaningfully compared if the quality of the data is known, consistent and verifiable. Currently, the identification of proteins in most published proteomic data sets involve at least partial manual interpretation and are thus expected to be inconsistently scored. To make data interpretation portable and transparent, it will minimally be necessary to develop algorithms that assign a score to each observation, such as peptide or protein identification or quantitative ratio, that estimates the probability that the observation is correct, independent of experimental variables such as the type of instrument used to generate the data, the database search tool used, or the quality of the sample. The recently developed computer tools PeptideProphet [51] and ProteinProphet [52] (see also http://www.systemsbiology.org/Default.aspx?pagename=proteomicssoftware) address these needs and will be useful to rigorously test whether large proteomic data sets can be consistently and transparently analyzed. Third, the data need to be organized in relational databases and software tools need to be developed for higher order analyses such as hierarchical clustering and multivariate statistics; and fourth, the capabilities of generating high quality quantitative proteomic analyses need to become readily accessible to broadly impact biology and medicine.

Ordered Peptide Arrays: A Path to the Future?

The sample throughput of the current LC-MS/MS based proteomics technology is mainly limited by the need to de novo identify every peptide/protein in each experiment. This is currently accomplished by peptide mass fingerprinting and increasingly, by tandem mass spectrometry and sequence database searching. Proteomics, therefore, currently operates in a perpetual "discovery mode" in which the observations made prior to a current experiment are disregarded. The genome projects have taught us that the universe of observable biological events in a species, e.g., the number of different proteins produced, is large but finite. Therefore, once all the possible proteins within a species have been discovered and described, proteomics will be transformed from a discovery mode of identifying and describing proteins, to a 'browsing' mode, in which the universe of possible events is searched for constellations that correlate with a particular state or function. Genomics-style biology, including proteomics, can be separated into two distinct phases, a discovery phase to characterize the universe, and a browsing phase, in which system-wide biological assays search the universe. To reach its potential as a high impact, high throughput technology, proteomics needs to advance from a discovery mode to a browsing mode of operation.

Fortunately, it is possible to suggest a browsing technology for proteomics. The following proposed scheme is conceptually simple and schematically illustrated in Figure 7. For each protein, protein isoform or specifically modified form of a protein, a peptide sequence that is idiotypic (or uniquely identifies) for that polypeptide is selected, chemically synthesized and labeled with tags of a heavy stable isotope. These peptides are therefore definitive markers for the proteins to be studied. Precisely measured amounts of these reference peptides are then added to a sample in which the proteins or peptides have been labeled with tags of a light stable isotope. The combined peptide sample can be separated reproducibly, and fractions deposited on the sample plate of a mass spectrometer, effectively generating an ordered peptide array. Each array element can then be interrogated by a mass spectrometer and will generate two types of signals: one representing the signals of the peptides for which no reference peptide has been added - appearing as single peaks, and the other representing the signals for those peptides for which a reference peptide was added—appearing as paired signals with a mass difference that precisely corresponds to the mass differential encoded in the stable isotope tag. In this method, a protein is identified by correlating the position and the accurately measured mass of each isotope-peptide pair in the array. Proteins are quantified by determining the ratio of the size of the signal of a peptide derived from the protein mixture with the signal of the corresponding reference peptide.

There are several advantages of this proteome browsing method. First, one peptide is sufficient for the unambiguous identification and quantification of each protein. Therefore, the number of peptides that need to be analysed to identify and quantify the product of every gene approaches the number of genes in a genome. Second, data analysis becomes trivial because each protein is identified and quantified by correlating the acquired data with a look-up table, rather than by de novo sequencing. Third, the method is easily standardized between laboratories. Fourth, the absolute quantity of each protein is determined, thus making data sets easily comparable. Fifth, any subset of proteins, for example, proteins contained in organelles, sub-cellular fractions or differentiated cells, can be selectively interrogated. Sixth, splice isoforms, differentially modified or processed proteins, can similarly be absolutely quantified, provided that appropriate refer-

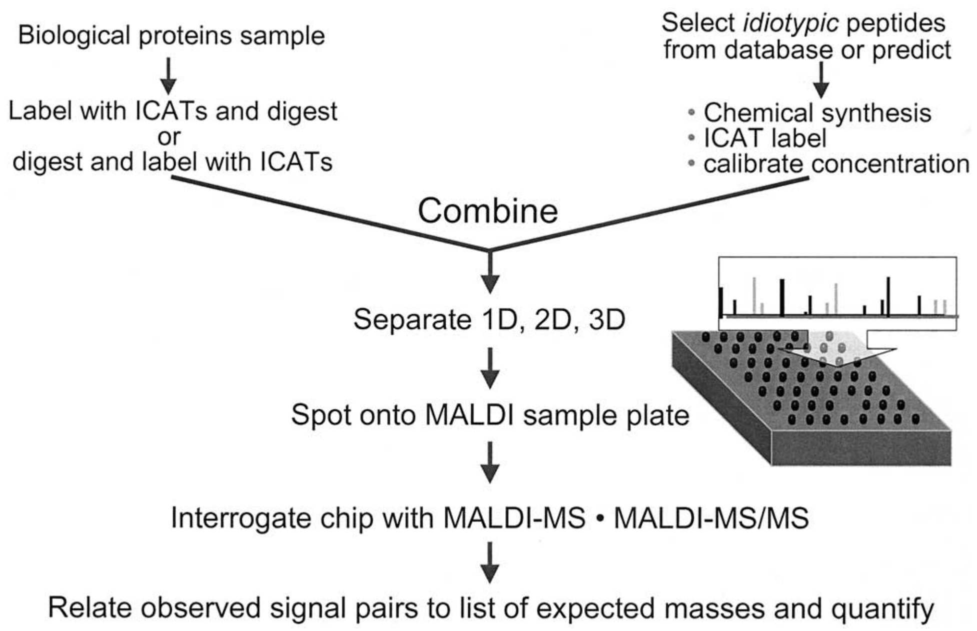
Biological proteins sample
Label with ICATs and digest
or
digest and label with ICATs
Select *idiotypic* peptides
from database or predict
• Chemical synthesis
• ICAT label
• calibrate concentration
Combine
Separate 1D, 2D, 3D
Spot onto MALDI sample plate
Interrogate chip with MALDI-MS • MALDI-MS/MS
Relate observed signal pairs to list of expected masses and quantify

Figure 7. Schematic representation of ordered peptide array technology for quantitative proteomics. A protein sample is labeled with stable isotope tagging reagents (e.g., light reagent) and processed to generate a sample of isotopically tagged peptides (left branch). Concurrently, a reference peptide sample consisting of calibrated, isotopically tagged (e.g., heavy reagent) synthetic peptides is prepared and added to the peptide sample derived from the protein sample (right branch). The reference peptides are selected such that each target protein is represented by one or more idiotypic peptides. The combined peptide sample is fractionated by 1-D, 2-D, or 3-D chromatography and spotted onto the sample plate of a MALDI mass spectrometer. MS analysis of the resulting ordered array (insert) indicates that each sample spot contains multiple peptide signals of two types. The first type of signal is a singlet and represents a peptide that is either a protein-derived peptide for which no reference sample was added (usual case) or a reference peptide for which no protein-derived peptide was detected (rare case). The second type of signal is a doublet of which one peak represents the calibrated reference peptide and the other peak represents the protein-derived peptide. Focusing the analysis on the doublets rapidly and conclusively identifies and absolutely quantifies proteins contained in the sample mixture.

ence peptides can be synthesized. Finally, the method is relatively cheap, as only minuscule (nanogram to subnanogram) amounts of the peptide standards are used per assay. However, similar to other genomics technologies, the proposed proteomics technology requires a sizable initial cost and labour investment that will pay dividends by the wide dissemination of a rapid, robust and simple quantitative technology. The initial investment is required for the synthesis and calibration of thousands of isotopically labeled peptides. A project of this scope not only exceeds the scope of a typical laboratory, but mandates a collaborative approach, since if different research groups were to generate their own sets of reference peptides, correlation of data between studies would be difficult. It can be expected that such a robust, simple and quantitative proteomics technology will have the sample throughput required to carry out measurements on differentially perturbed cell for carrying out clinical studies, and therefore to realize the potential of proteomics.

Conclusions

"Does technology development drive biology or does biology drive the development of new technologies" was the question posed at the outset of this manuscript. Using the development of advanced mass spectrometric techniques and their application to protein based biological research as an example, it was illustrated that there is no clear answer since the interplay of biology and technology development is complex. Protein identification via MS was initially used to better support the biochemical, activity-based research method than had been possible with the traditional protein sequencing methods. By combining these increasingly sensitive protein identification tools with high resolution 2-DE, a new research strategy was created that no longer relied on the availability of assays for specific activities. In fact, the protein pattern displayed in the 2-D gels, if subjected to comparative pattern analysis, could by itself be used as an assay to probe complex biological processes, and the strategy of systematically studying the proteins expressed in a sample was termed "proteomics." Limitations in the 2-D gel based approach to quantitative proteomics led to the development of second generation proteomic techniques, notably the ICAT reagent technique in conjunction with tandem mass spectrometry. Furthermore, the challenge of interpreting the large amounts of data generated by proteomic studies in terms of biological function catalyzed the development of a new strategy that is based on the comparative analysis of systematically collected data sets from differentially perturbed cells. As this strategy requires the repeated analysis of substantially identical samples, it seemed inefficient to de novo discover each protein in each experiment, and we propose to enter into a "browsing" mode of proteomics with the potential to generate accurately quantitative data at very high throughput.

This interrelatedness of technology and biology research also has important implications for the working environment in which successful work is being carried out. It follows that technologies with an early and large impact on biology are best developed in a culture in which important problems posed by biology are apparent and understood. It also follows that the research should be carried out in an environment in which biologists are being made aware of emerging technical capabilities to advance their research. In short, it appears that an integrated, multidisciplinary research environment provides a fertile ground for advancing high impact technologies as well as for pioneering advances in biology. How such environments are being created is, of course, another question altogether.

Acknowledgments

The author gratefully acknowledges the contributions of all group members, past and present, who have made this proteomics journey memorable, interesting, and productive. The author specifically acknowledges the excellent editorial help of Laurel Feltz.

References

1. Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* **1989,** *246,* 64–71.
2. Cole, R. B. *Electrospray Ionization Mass Spectrometry: Fundamentals, Instrumentation, and Applications.* Wiley: New York, 1997; 19th ed.; p 557.
3. Karas, M.; Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Mass Exceeding 10,000 Daltons. *Anal. Chem.* **1988,** *60,* 2299–2301.
4. Chait, B. T.; Kent, S. B. H. Weighing Naked Proteins: Practical High-Accuracy Mass Measurement of Peptides and Proteins. *Science* **1992,** *257,* 1885–1894.
5. Barber, M.; Bordoli, R. S.; Sedgwick, R. D.; Tyler, A. N. J. *Chem. Soc. Commun.* **1981**, 7, 325–327.
6. Edman, P.; Begg, G. A Protein Sequenator. *Eur. J. Biochem.* **1967,** *1,* 80–91.
7. Biemann, K. Four Decades of Structure Determination by Mass Spectrometry: From Alkaloids to Heparin. *J. Am. Soc. Mass Spectrom.* **2002,** *13,* 1254–1272.
8. Hewick, R. M.; Hunkapiller, M. W.; Hood, L. E.; Dreyer, W. J. A Gas-Liquid Solid Phase Peptide and Protein Sequenator. *J. Biol. Chem.* **1981,** *256,* 7990–7997.
9. Aebersold, R. H.; Teplow, D. B.; Hood, L. E.; Kent, S. B. Electroblotting onto Activated Glass. High Efficiency Preparation of Proteins from Analytical Sodium Dodecyl Sulfate-Polyacrylamide Gels for Direct Sequence Analysis. *J. Biol. Chem.* **1986,** *261,* 4229–4238.
10. Matsudaira, P. J. Sequence from Picomole Quantities of Proteins Electroblotted onto Polyvinylidene Difluoride Membranes. *Biol. Chem.* **1987,** *262,* 10035–10038.
11. Aebersold, R. H.; Leavitt, J.; Saavedra, R. A.; Hood, L. E.; Kent, S. B. Internal Amino Acid Sequence Analysis of Proteins Separated by One- or Two-Dimensional Gel Electrophoresis After in Situ Protease Digestion on Nitrocellulose. *Proc. Natl. Acad. Sci. U.S.A.* **1987,** *84,* 6970–6974.
12. Yamamori, T.; Fukada, K.; Aebersold, R.; Korsching, S.; Fann, M. J.; Patterson, P. H. The Cholinergic Neuronal Differentiation Factor from Heart Cells is Identical to Leukemia Inhibitory Factor. *Science* **1989,** *246,* 1412–1415.
13. Hilton, D. J. LIF: lots of interesting functions. *Trends Biochem. Sci.* **1992,** *17,* 72–76.
14. Jeffery, C. J. Moonlighting Proteins. *Trends Biochem. Sci.* **1999,** *24,* 8–11.
15. Lin, C. S.; Aebersold, R. H.; Kent, S. B.; Varma, M.; Leavitt, J. Molecular Cloning and Characterization of Plastin, a Human Leukocyte Protein Expressed in Transformed Human Fibroblasts. *Mol. Cell Biol.* **1988,** *8,* 4659–4668.
16. Klose, J. Protein Mapping by Combined Isoelectric Focusing and Electrophoresis of Mouse Tissues. A Novel Approach to Testing for Induced Point Mutations in Mammals. *Humangenetik* **1975,** *26,* 231–243.
17. O'Farrell, P. H. High Resolution Two-Dimensional Electrophoresis of Proteins. *J. Biol. Chem.* **1975,** *250,* 4007–4021.
18. Anderson, N. G.; Anderson, L. The Human Protein Index. *Clin. Chem.* **1982,** *28,* 739–748.
19. Garrels, J. I. The QUEST System for Quantitative Analysis of Two-Dimensional Gels. *J. Biol. Chem.* **1989,** *264,* 5269–5282.
20. Garrels, J. I.; Franza, B. R., Jr. Transformation-Sensitive and Growth-Related Changes of Protein Synthesis in REF52 Cells. A Two-Dimensional Gel Analysis of SV40-, Adenovirus-, and Kirsten Murine Sarcoma Virus-Transformed Rat Cells Using the REF52 Protein Database. *J. Biol. Chem.* **1989,** *264,* 5299–5312.
21. Garrels, J. I.; Franza, B. R., Jr. The REF52 Protein Database. Methods of Database Construction and Analysis Using the QUEST System and Characterizations of Protein Patterns from Proliferating and Quiescent REF52 Cells. *J. Biol. Chem.* **1989,** *264,* 5283–5298.
22. Anderson, N. L.; Hofmann, J. P.; Gemmell, A.; Taylor, J. Global Approaches to Quantitative Analysis of Gene-Expression Patterns Observed by Use of Two-Dimensional Gel Electrophoresis. *Clin. Chem.* **1984,** *30,* 2031–2036.
23. Tarroux, P.; Vincens, P.; Rabilloud, T. HERMeS: A Second Generation Approach to the Automatic Analysis of Two-Dimensional Electrophoresis Gels. Part V: Data Analysis. *Electrophoresis* **1987,** *8,* 187–199.
24. de Arruda, M. V.; Watson, S.; Lin, C. S.; Leavitt, J.; Matsudaira, P. Fimbrin is a Homologue of the Cytoplasmic Phosphoprotein Plastin and Has Domains Homologous with Calmodulin and Actin Gelation Proteins. *J. Cell Biol.* **1990,** *111,* 1069–1079.
25. Gygi, S. P.; Corthals, G. L.; Zhang, Y.; Rochon, Y.; Aebersold, R. Evaluation of Two-Dimensional Gel Electrophoresis-Based Proteome Analysis Technology. *Proc. Natl. Acad. Sci. U.S.A.* **2000,** *97,* 9390–9395.
26. Duggan, D. J.; Bittner, M.; Chen, Y.; Meltzer, P.; Trent, J. M. Expression Profiling Using cDNA Microarrays. *Nat. Genet.* **1999,** *21(Suppl 1),* 10–14.
27. Wilkins, M. R.; Sanchez, J. C.; Gooley, A. A.; Appel, R. O.; Humphrey-Smith, I.; Hochstrasser, D. F.; Williams, K. L. Progress with Proteome Projects: Why All Proteins Expressed by a Genome Should be Identified and How to Do It. *Biotech. Gen. Eng. Rev.* **1995,** *13,* 19–50.
28. Aebersold, R.; Goodlett, D. R. Mass Spectrometry in Proteomics. *Chem. Rev.* **2001,** *101,* 269–295.

29. Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. Identifying Proteins from Two-Dimensional Gels by Molecular Mass Searching of Peptide Fragments in Protein Sequence Databases. *Proc. Natl. Acad. Sci. U.S.A.* **1993,** *90,* 5011–5015.

30. Yates, J. R., III; Speicher, S.; Griffin, P. R.; Hunkapiller, T. Peptide Mass Maps: A Highly Informative Approach to Protein Identification. *Anal. Biochem.* **1993,** *214,* 397–408.

31. James, P.; Quadroni, M.; Carafoli, E.; Gonnet, G. Protein Identification in DNA Databases by Peptide Mass Fingerprinting. *Protein Sci.* **1994,** *3,* 1347–1350.

32. Pappin, D. J. C.; Hojrup, P.; Bleasby, A. J. Rapid Identification of Proteins by Peptide-Mass Fingerprinting. *Curr. Biol.* **1993,** *3,* 327–332.

33. Mann, M.; Hojrup, P.; Roepstorff, P. Use of Mass Spectrometric Molecular Weight Information to Identify Proteins in Sequence Databases. *Biol. Mass Spectrom.* **1993,** *22,* 338–345.

34. Eng, J. K.; McCormack, A. L.; Yates, J. R., III. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994,** *5,* 976–989.

35. Smolka, M.; Zhou, H.; Aebersold, R. Quantitative Protein Profiling Using Two-Dimensional Gel Electrophoresis, Isotope-Coded Affinity Tag Labeling, and Mass Spectrometry. *Mol. Cell Proteomics* **2002,** *1,* 19–29.

36. Corthals, G. L.; Wasinger, V. C.; Hochstrasser, D. F.; Sanchez, J. C. The Dynamic Range of Protein Expression: A Challenge for Proteomic Research. *Electrophoresis* **2000,** *21,* 1104–15.

37. Rabilloud, T. Two-Dimensional Gel Electrophoresis in Proteomics: Old, Old Fashioned, but It Still Climbs up the Mountains. *Proteomics* **2002,** *2,* 3–10.

38. Hunt, D. F.; Henderson, R. A.; Shabanowitz, J.; Sakaguchi, K.; Michel, H.; Sevilir, N.; Cox, A. L.; Appella, E.; Engelhard, V. H. Characterization of Peptides Bound to the Class I MHC Molecule HLA-A2.1 by Mass Spectrometry. *Science* **1992,** *255,* 1261–1263.

39. De Leenheer, A. P.; Thienpont, L. M. Application of Isotope Dilution-Mass Spectrometry in Clinical Chemistry, Pharmacokinetics, and Toxicology. *Mass Spectrom. Rev.* **1992,** *11,* 249–307.

40. Gygi, S. P.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. Quantitative Analysis of Complex Protein Mixtures Using Isotope-Coded Affinity Tags. *Nat. Biotechnol.* **1999,** *17,* 994–999.

41. Ideker, T.; Thorsson, V.; Ranish, J. A.; Christmas, R.; Buhler, J.; Eng, J. K.; Bumgarner, R.; Goodlett, D. R.; Aebersold, R.; Hood, L. Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science* **2001,** *292,* 929–34.

42. Griffin, T. J.; Gygi, S. P.; Ideker, T.; Rist, B.; Eng, J.; Hood, L.; Aebersold, R. Complementary Profiling of Gene Expression at the Transcriptome and Proteome Levels in Saccharomyces cerevisiae. *Mol. Cell Proteomics* **2002,** *1,* 323–333.

43. Shiio, Y.; Donohoe, S.; Yi, E. C.; Goodlett, D. R.; Aebersold, R.; Eisenman, R. N. Quantitative Proteomic Analysis of Myc Oncoprotein Function. *EMBO J.* **2002,** *21,* 5088–5096.

44. Guina, T.; Purvine, S. O.; Yi, E. C.; Eng, J.; Goodlett, D. R.; Aebersold, R.; Miller, S. I. Quantitative Proteomic Analysis Indicates Increased Synthesis of a Quinolone by *Pseudomonas aeruginosa* Isolates from Cystic Fibrosis Airways. *Proc. Natl. Acad. Sci. U.S.A.* **2003,** *100,* 2771–2776.

45. Han, D. K.; Eng, J.; Zhou, H.; Aebersold, R. Quantitative Profiling of Differentiation-Induced Microsomal Proteins Using Isotope-Coded Affinity Tags and Mass Spectrometry. *Nat. Biotechnol.* **2001,** *19,* 946–951.

46. Ranish, J. A.; Yi, E. C.; Leslie, D. M.; Purvine, S. O.; Goodlett, D. R.; Eng, J.; Aebersold, R. The Study of Macromolecular Complexes by Quantitative Proteomics. *Nat. Genet.*, **2003**, *33*, 349–355.

47. Zhou, H.; Watts, J. D.; Aebersold, R. A Systematic Approach to the Analysis of Protein Phosphorylation. *Nat. Biotechnol.* **2001,** *19,* 358–375.

48. Oda, Y.; Huang, K.; Cross, F. R.; Cowburn, D.; Chait, B. T. Accurate Quantitation of Protein Expression and Site-Specific Phosphorylation. *Proc. Natl. Acad. Sci. U.S.A.* **1999,** *96,* 6591–6596.

49. Shiio, Y.; Yi, E.C.; Donohoe, S.; Goodlett, D.R.; Aebersold, R.; Eisenman, R.N. Quantitative Proteomic Analysis of Chromatin-Associated Factors. *J. Am. Soc. Mass Spectrom.*, **2003**, *14*, 696–703.

50. Guina, T.; Wu, M.; Purvine, S. O.; Yi, E. C.; Lee, K. A.; Eng, J.; Goodlett, D. R.; Aebersold, R.; Miller, S. I. Proteomic Analysis of Pseudomonas aeruginosa Grown Under Magnesium Stress. *J. Am. Soc. Mass Spectrom.*, **2003**, *14*, 742–751.

51. Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **2002,** *74,* 5383–5392.

52. Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry, unpublished.