Proteomic research: potential opportunities for clinical and physiological investigators

K. Sreekumaran Nair, Abdul Jaleel, Yan W. Asmann, Kevin R. Short, and Sreekumar Raghavakaimal

Endocrinology Research Unit, Mayo Clinic School of Medicine, Rochester, Minnesota 55905

Proteomic research: potential opportunities for clinical and physiological investigators. Am J Physiol Endocrinol Metab 286: E863–E874, 2004; 10.1152/ajpendo.00370.2003.—Proteomics is the comprehensive and systematic study of proteins, which are functional molecules. Although proteins are products of gene expression, there are more proteins than genes due to the posttranslational modifications of proteins, making the study of proteins difficult. Protein expression is tissue specific, and its function is modulated by variety of factors, including other proteins, phosphates, sulfates, carbohydrates, and lipids, as well as other metabolites. Because of the dynamic nature of protein expression and posttranslational modifications, identification and quantification of proteins alone are not sufficient to understand functional changes. Emerging technologies will allow investigators to perform a combination of metabolic labeling and identification as well as quantification and measurement of the synthesis rates of a large number of proteins in a tissue. This offers the opportunity to better understand the regulation of tissue functions. Rapid advances in mass spectrometry, protein purification techniques, isotope labeling of proteins, and bioinformatics are likely to improve our understanding of physiological states and altered functions in diseased states. Such mechanistic information will improve the ability to perform early diagnosis of tumors and other diseases and develop prognostic indexes and novel therapies.

mass spectrometry; protein modification; protein quantification; metabolic labeling; protein synthesis

THE COMPLETION OF THE HUMAN GENOME SEQUENCE is perhaps the most important landmark in biological research thus far (84, 88). For physiologists and physicians interested in the regulation of bodily functions, an understanding of genes and their products is crucial to unraveling the underlying mechanisms of disease. Understanding the regulation of both normal physiology and pathology requires an investigation of genes, gene transcripts, proteins, and metabolites (Fig. 1), which have been termed the genome, transcriptome, proteome, and metabolome, respectively. We propose that perhaps the most important step in the expression of a gene occurs at the level of protein synthesis, since the protein product of a gene is what will ultimately be responsible for most biological functions. The term “proteome” refers to the collection of proteins within a cell, tissue, or entire organism and was first coined to describe large-scale protein identification and amino acid analysis (94). The study of the proteome, or proteomics, has moved beyond cataloguing of proteins in an organism and now includes the comprehensive and detailed analysis of proteins. The scope of proteomics is quickly expanding but presents many challenges. Proteomics is more complex than genomics, because proteins can undergo many posttranslational modifications that affect structure and function. Besides protein identification, proteomics also encompasses the regulation of protein synthesis at the translational level, factors regulating the folding of peptides, and interactions among proteins. The complexity of proteomics is further magnified by the fact that protein expression varies among different tissues as well as different physiological conditions, such as age, sex, fasting and feeding, changes in diet, physical activity, medications, pregnancy, disease status, and the like. Understanding how multitudes of proteins change under these conditions will be a great challenge to physiologists and clinicians.

There are several reviews that address the advancing technology available for proteomics studies (1, 32, 62, 85). In this developing field, no review can provide readers with a truly state-of-the-art description of proteomic techniques. The objective of the current review, therefore, is to specifically address potential applications of available proteomic methods for physiologists and physician scientists. We will briefly discuss some of the current techniques and their potential applications in examples of clinical research projects, future directions in the development of techniques relevant to physiologists and clinical researchers, and the state of bioinformatics, which is critical for interpretation of the vast amount of information emerging from proteomic research. The ultimate goal of proteomic research is to understand the regulation of body functions and how regulation is altered in disease states. To achieve these goals, there is need for greater integration of proteomic research with genomic, metabolic, and functional studies.
WHAT CAN THE CURRENT PROTEOMIC TECHNIQUES OFFER TO PHYSIOLOGISTS AND CLINICIANS?

Table 1 lists some of the emerging and existing proteomic techniques relevant for physiological studies. None of these techniques is without limitations; however, they still have many potential applications in physiological and clinical research. Moreover, the technology is rapidly advancing to improve the speed, accuracy, and precision of identifying and measuring proteins. Table 2 indicates some of the examples of potential applications of proteomics in human research. An example is the potential approach to investigate the underlying cause of muscle-wasting disease. The objective is to determine whether muscle wasting is related to changes in a single protein, protein complex, or cluster of proteins with a specific function. Muscle wasting may be regulated by systemic factors (present in circulation), local factors, or both. It is therefore prudent to study both serum/plasma and muscle. The purification and identification of individual proteins from a complex protein mixture are critical to achieving the specific aim of this study protocol. It is important to reduce the complexity of the protein mixtures by one or more initial strategies, as we describe. Several approaches are available (38), and the pres-

Table 1. Proteomic techniques relevant to physiological studies

<table>
<thead>
<tr>
<th>Application</th>
<th>Methods</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein identification and quantification</td>
<td>1. 1D or 2D electrophoresis + MS</td>
<td>1. Most commonly used and sensitive.</td>
</tr>
<tr>
<td></td>
<td>2. MDLC/MS/MS</td>
<td>2. Multidimensional separation of complex protein mixtures can be performed</td>
</tr>
<tr>
<td></td>
<td></td>
<td>online, instead of electrophoresis, before introduction to MS. High sensitivity and high throughput and proteins can be labeled postextraction with ICAT-reagent or 18O/16O for quantification.</td>
</tr>
<tr>
<td></td>
<td>3. MALDI-TOF MS</td>
<td>3. Less sensitive but high throughput.</td>
</tr>
<tr>
<td>Protein synthesis</td>
<td>1. GC-MS</td>
<td>Highly sensitive and commonly used methods.</td>
</tr>
<tr>
<td></td>
<td>2. IRMS</td>
<td>Fractional synthesis rate of proteins in vivo can be measured using infusions of stable isotopes of amino acids and MS analysis.</td>
</tr>
<tr>
<td>Protein-protein interaction</td>
<td>1. Protein complex separation + MS</td>
<td>1. Affinity purification of protein complex followed by MS/MS</td>
</tr>
<tr>
<td></td>
<td>2. Yeast two-hybrid system</td>
<td>2. Mostly binary interactions, functional assignment to unknown protein</td>
</tr>
<tr>
<td></td>
<td>3. 1D or 2D electrophoresis + Western blot</td>
<td>1. Conventional and very reliable</td>
</tr>
<tr>
<td></td>
<td>4. Mass spectrometry</td>
<td>2. Examples are fluorescent stains for glycation and phosphorylation</td>
</tr>
<tr>
<td></td>
<td>5. Antibody array</td>
<td>3. Highly sensitive and very accurate</td>
</tr>
<tr>
<td>Protein function</td>
<td>1. Metabolic assays</td>
<td>4. Good for high-throughput data</td>
</tr>
<tr>
<td></td>
<td>2. High-throughput enzyme assays</td>
<td>Protein chips for functional protein arrays can be used for proteome-wide analyses of biochemical and enzyme activities and other molecular interactions.</td>
</tr>
<tr>
<td></td>
<td>3. Automated antigen-antibody assays</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4. Protein chips</td>
<td></td>
</tr>
<tr>
<td>Structural information</td>
<td>1. X-ray crystallography</td>
<td>Protein structure comparisons can help to identify the functions of newly discovered genes. Structural analysis can also show where drugs bind to proteins and where proteins interact with each other.</td>
</tr>
<tr>
<td></td>
<td>2. Nuclear magnetic resonance</td>
<td></td>
</tr>
<tr>
<td>Protein localization</td>
<td>1. Immunolocalization + fluorescence microscopy</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. Reverse transfection + indirect immuno-fluorescence</td>
<td>Defines subcellular compartment. Data help increase confidence of protein interaction and, in designing phenotypic screens, and in assays aimed at further elucidating their function.</td>
</tr>
</tbody>
</table>

1D or 2D electrophoresis, one- or two-dimensional electrophoresis; MS, mass spectrometry; MDLC, multidimensional liquid chromatography; MALDI-TOF, matrix-assisted laser desorption ionization time of flight; SELDI-TOF, surface-enhanced laser desorption ionization time of flight; GC-MS, gas chromatography-mass spectrometry; IRMS, isotope ratio mass spectrometry.
Table 2. Proteomics: objectives for potential application in human research

1) Identify proteins or protein involved in diseases
2) Identify modifications of proteins causing complications, e.g., non-enzymatic glycosylation of proteins in diabetes and structural modification (folding changes)
3) Identify molecular pathways involved in functional changes, e.g., signal transduction
4) Target drugs to specific proteins or pathway involved in disease

ently recommended approach for muscle tissue is subcellular fractionalization by ultracentrifugation (76). Different muscle protein subcellular fractions, mitochondrial, myofibrillar, and sarcoplasmic, can each be processed separately. Similarly, to achieve optimum purification and identification of individual proteins in plasma or serum samples, the complexity of the protein mixture often needs to be reduced. One of the most frequently used strategies to accomplish this goal has been an initial separation of proteins by two-dimensional gel electrophoresis (2DGE) followed by mass spectrometry (MS) (1). A more recently developed approach involves separation by multidimensional liquid chromatography (MDLC) followed by MS (91). Either of these approaches can be used to profile the proteins in both serum/plasma and subcellular fractions of muscle proteins. This type of study may provide important clues to the underlying mechanisms of the muscle-wasting condition. Once a protein or a cluster of proteins is identified, an analysis of the function(s) is important. Altered function may result from changes in protein concentration or posttranslational modification. Integration of proteomic data with gene (mRNA) array data will provide further insight into the molecular level at which the abnormality occurs. Such mechanistic information will form the basis of potential therapy for muscle-wasting conditions.

PROTEOMICS TOOLS

In this section we briefly discuss some of the methods and technology available for identification and quantification of proteins in biological samples. We also discuss some of the limitations of proteomic technologies and future directions needed to address these limitations.

Identification of Proteins and Comparative Analysis

Separation of complex mixtures of proteins. The number of proteins in any tissue is likely to be in the tens of thousands, and the expression levels of these proteins span at least six orders of magnitude at any one time. Besides the wide range in expression levels of proteins, wide variability in properties such as subcellular location, hydrophobicity, charge, size, shape, solubility, or affinity make it a daunting task for a protein chemist to achieve sufficient purification of individual proteins for their identification. The ability of MS to resolve complex mixtures of peptides and to identify proteins from peptide databases is rapidly improving (32). However, it is still necessary to reduce the complexity of biological samples before introduction to the mass spectrometer. There are several strategies that can be used for this initial step.

Traditional approaches to purify proteins by one-dimensional gel electrophoresis and antibody-based identification techniques do not have the scope to deal with such complexity. Purification of proteins by 2DGE (60), which separates proteins first by charge and then by molecular weight, has the resolving power to separate thousands of proteins (62) (Fig. 2). Protein spots are visualized in the gel after staining, and comparisons can be performed among gels representing the conditions of interest by use of densitometric analysis. Any number of protein spots can be selected for further analysis by excising them from the gel. The 2DGE approach has been used extensively for separation of proteins but has some limitations. Many proteins with low concentrations may not be visible for detection in the presence of other, high-abundance proteins, which take up a significant portion of the available analytical space. Furthermore, a comparative analysis of stained gels is not sensitive enough to detect small changes in protein concentrations. This is important, because many biological functions result from small changes in protein concentration that can escape detection with this method. Recent introduction of fluorescent-based stains such as SyproRuby has improved the sensitivity of detection and quantitative accuracy of gel-staining methods (65). However, variations in spot intensity and patterns have been observed even for identical spots on gels run in parallel, which makes it difficult to distinguish subtle quantitative differences.

To circumvent the limitations involved in the comparative analysis of different 2D gels, a method has been developed by Unlu (87) that permits two samples to be run under identical electrophoretic conditions (coelectrophoresis) in a type of differential display format. Here, the succinimydyl esters of propyl Cy3 and methyl Cy5 dyes have been used to prelabel two different protein samples before running them on the same 2D gel. Differences in the abundance of individual proteins between the two samples are detected by measuring the Cy3/Cy5 fluorescent signal ratio of the comigrating proteins. Although this approach is promising, there are some limitations resulting from the use of fluorescent dyes. Anomalous migration of proteins invariably does occur despite attempts to minimize differences in the electrophoretic behavior of derivitized and derivitized proteins on 2D gels (64). This renders spot excision difficult for identification. Also, it is not possible to directly excise protein spots from the gels, because the dye pair cannot be visualized effectively by placing a gel on a UV transilluminator. Moreover, unsulfonated Cy3/Cy5 dyes increase the hydrophobicity of proteins and thereby substantially reduce the solubility of proteins for doing 2DGE.

Another problem encountered with 2DGE is that more than one protein may be present in a single gel spot. Thus it becomes difficult to determine which of the proteins in the spot are up- or downregulated. A strategy to work around this problem is to perform additional sample separation to fractionate the sample even further before 2DGE. Various approaches, such as subcellular fractionalization, velocity gradient centrifugation, or anion exchange column separation, can help to reduce the complexity of protein mixture. Another way to achieve sample fractionation is to take advantage of the resolving power of immobilized pH gradients (IPGs) and fractionate according to protein pl by use of multiple narrow-range pH gradients. This has been one of the strategies for separating membrane proteins and other hydrophobic proteins, which are exceedingly difficult to analyze. However, profiling hydrophobic proteins using 2DGE has often been difficult, even with a
variety of new techniques and strategies, because of their chemical character and membrane compartmentalization, main contributing factors for their poor solubility and separation (49). Despite some of these limitations, the unique property of 2DGE, which is the ability for "visual selection" of specific protein(s) from a complex mixture, continues to make this technique a valuable tool for proteomics.

Proteins/peptides can be resolved on the basis of charge and molecular weight by using appropriate liquid chromatographic columns. An advantage of this method is that much of the separation process can be automated (91). A newer approach for separating and purifying complex protein/peptide (generated from trypsin digestion) mixtures is MDLC (46, 91). The amount of sample (a complex peptide mixture) can be effectively resolved and scaled up by exploiting the unique physical properties of charge and hydrophobicity. MDLC is most commonly accomplished by the combination of ion exchange chromatography (usually strong cation exchange) as a primary separation technique, because of its potential for increased loading capacity, followed by a reversed-phase (RP) chromatography as a secondary separation technique because of its ability to remove salts and its direct compatibility with MS through electron spray ionization. This method is useful for measuring the expression of low-abundance proteins. The protein mixture is digested using trypsin and is applied onto the MDLC system, where it is separated and introduced online onto the mass spectrometer for analysis as shown in Fig. 2C. Even though this method has the advantage of automation of the analysis to some extent, the time that is required to complete such analysis is relatively long.

Another recent development has been the use of protein chips (protein microarray) for high-throughput identification of proteins (97). There are two classes of protein microarrays. The first class comprises analytical microarrays, which are used to measure the presence and concentrations of proteins in a complex mixture. The most common form of these analytical microarrays is antibody arrays (21, 39). The principle of this approach is that protein antibodies are coated onto the chip and used to capture proteins of interest. Antibodies can be used that recognize specific proteins or a particular protein surface motif. These chips can be used for monitoring protein expression levels, protein profiling, and clinical diagnostics. A limitation of this strategy is that specific antibodies have to be generated against all proteins of interest, which can be challenging and labor intensive. Currently, attempts are being made by many groups to develop antibodies against a large number of known proteins so that protein chips can be used to perform large-scale protein identification and profiling. The second class comprises functional microarrays, in which proteins are prepared and arrayed for a wide range of biochemical activities. Functional protein chips are made by immobilizing large numbers of purified proteins or peptides onto a solid surface, and they can be used to analyze protein activities, binding properties, and posttranslational modifications. Again, the obstacle to overcome here is the purification of large numbers of proteins in a high-throughput manner. The potential applicability of protein chip technology in physiology has recently been described in detail in a special issue of *Proteomics* (Proteomics 3(11), 2003).
There have been intensive efforts in recent years to identify protein-protein interaction on a large scale, because it is essential to identify interacting partners of proteins to deduce the protein function. One approach is to use an endogenous protein as bait, and an antibody against that allows specific isolation of the protein with its bound partners. Because there is a lack of comprehensive specific antibody collection, a more generic strategy is to tag the protein of interest with a sequence readily recognized by an antibody specific for the tag. This biochemical copurification of complexes, coupled with protein identification using MS, can define the total spectrum of the complex (1). The genetic approach of using the yeast two-hybrid system (24) for the analysis of protein-protein interactions is also available. It is based on the modular nature of some eukaryotic transcriptional activators, which are characterized by two separable functional domains. One domain interacts with the DNA (binding domain) and must either be covalently linked to, or interact with, a second domain (the activation domain) to cause transcription. Thus, in practice, a protein of interest fused to the binding domain (bait) is screened against a library of activation domain hybrids (preys) to select interacting partners. Ordered arrays of strains expressing either DNA-binding domain or activation-domain fusion proteins are constructed for high-throughput proteomic analysis. Different strategies of two-hybrid systems have been used to analyze genome-wide protein interactions in yeast (26, 86). In addition to this, large-scale two-hybrid studies have been carried out in Helicobacter pylori (72) and in Caenorhabditis elegans (90).

Mass spectrometric methods to identify and quantify proteins. After the initial step(s) of sample fractionation, MS is used for protein identification and quantification. One of the most commonly used approaches is to perform 2DGE and select a number of protein spots of interest. Gel spots containing the proteins of interest are excised, and the proteins are fragmented by in-gel trypsin digestion. The extracted peptides are then spotted onto an MS sample plate (Fig. 2A). Molecular ions from the peptide samples are generated using a laser source and are introduced into the mass spectrometer by a method known as matrix-assisted laser desorption ionization (MALDI) coupled to a time-of-flight (TOF) mass spectrometer to obtain the corresponding molecular weights of the peptides. MALDI-TOF MS can rapidly determine peptide mass and estimate the probable protein source of the peptide by mass fingerprinting. The limitations of MALDI-TOF MS are that only ~60–70% of proteins in a mixture can be accurately identified, and quantification of differentially expressed proteins between healthy controls and patients is not accurate. These shortcomings of MALDI-TOF mass spectrometers were overcome to a certain extent by coupling a liquid chromatography (LC) system to a tandem mass spectrometer (LC-MS/MS) (Fig. 2B). The peptide fragments generated during trypsin digestion of protein spots from 2DGE can be separated further on an LC system and then directly introduced onto the mass spectrometer for analysis. The advantage of this system is that it can perform peptide sequencing, which increases the power for identifying proteins. However, sample throughput is slower than with MALDI-TOF MS, so the two methods can complement each other. The 2DGE separation of protein mixtures is labor intensive, and the amount of protein one can load onto 2DGE is limited. Hence, the identification and quantification of several low-abundant proteins are difficult using this approach.

The relatively new MDLC approach discussed earlier in this review can address some of these limitations. The amount of sample introduced onto the MDLC column can be scaled up by using appropriate chromatographic columns and conditions. This would allow investigators to examine the alterations in the expression of low-abundant proteins.

One of the newest developments in MS for proteomic research is two-dimensional liquid chromatography coupled to dual electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry, or 2D-LC-Dual ESI-FT-ICR-MS. This system features a protein separation step (2D-LC) coupled to a mass spectrometer with superior resolving power and dynamic mass range. This sensitive approach allows the detection and quantification of low-abundant and low-molecular-weight proteins (9). However, this technology is expensive and is not widely available.

Identification of proteins is an important step, but determination of differences in proteins is more critical for distinguishing between diseased and healthy states. A quantitative approach is needed to determine which proteins are up- or downregulated in a diseased state. One approach to quantify large numbers of proteins in a protein mixture is isotope-coded affinity tag (ICAT) labeling followed by MS (36) (Fig. 2). The ICAT reagent is internally labeled with either 8 deuterium atoms (D8) or regular hydrogen (D0) and will bind to cysteine residues. One sample to be compared (i.e., control sample) is labeled with D0, while the other sample (i.e., disease sample) is labeled with D8. The samples are then mixed together in equal proportion and run through the remaining procedures simultaneously. The results yield the relative abundance of each protein in one sample vs. that in the other sample (D8/D0 ratio). An advantage of this method is that comparative quantification of the two conditions of interest occurs in a single mixture of the two samples. This not only saves analysis time but also avoids some of the methodological error that arises from independent sample runs. There are, however, many limitations when using the ICAT approach. First, the quantification of peptides depends on the presence of cysteine residues, which may occur in low levels or not at all in some proteins of interest. Another limitation is that labeling is performed in vitro, and therefore the information on protein content does not represent the changes that may occur during an intervention period. In contrast, in vivo labeling provides quantitative or semi-quantitative information on the synthesis rate of specific proteins (5, 27), but these techniques are limited to measurements of synthesis rates of one or few proteins. Advances in protein purification techniques and MS may eventually make it possible to simultaneously perform large-scale analysis of synthesis rates of multiple proteins in a tissue by this in vivo approach. It remains to be determined whether synthesis rates and quantification of multiple proteins can be performed simultaneously by use of in vivo labeling. Labeling of proteins with isotopes of amino acids has also been utilized in cell lines and single cell organisms to quantify protein concentrations (61) and degradation rates (70). Isotope labeling in a single cell system has also been used for peptide mass fingerprinting (71). However, none of these substitute for the in vitro labeling and quantification of large number of proteins and tissue samples that are made possible by the ICAT approach. The D0- and D8-labeled peptides generated from a protein during tryptic digestion should run on a single peak of
the HPLC to get an accurate quantification on a mass spectrometer. In many cases, the same peptides labeled with D0 and D8 run as two partially resolved peaks on HPLC, which affects the quantification. To overcome these shortcomings, Applied Biosystems recently developed a new generation “cleavable” ICAT reagent in which the nine aliphatic carbon atoms on the backbone of the ICAT reagent were labeled with either 13C (heavy) or 12C (light). The new-generation ICAT reagents allow removal of the biotin affinity tag before MS and MS/MS analysis. This will improve MS/MS performance and significantly increase the number of proteins identified and quantified with higher confidence scores in a single experiment. Moreover, incorporation of 13C rather than deuterium into the ICAT heavy reagent promotes coelution of the heavy and light isotopes in RP chromatography, thereby increasing accuracy of quantification by MS/MS. Additional details regarding the merits and demerits of ICAT strategy and its comparison with 2DGE were presented in a recent review by Patton et al. (66).

Caprioli et al. (15) recently introduced MALDI-MS-based imaging MS. This technology utilizes MALDI-TOF-MS to generate profiles and two-dimensional ion density maps of peptide and protein signals directly from the surface of thin tissue sections. This molecular imaging tool has tremendous potential, since specific information can be obtained on the local proteomic composition, relative abundance, and spatial distribution of these components. This has potential application in identification of tumors during surgery and localizing invasion (82). However, improvements in sample preparation and handling, instrumentation, imaging acquisition, image resolution, and data mining need to be developed for this technology to be useful in clinical research (16).

Posttranslational Modifications

Direct analyses of protein modifications are important, since they cannot be predicted from genomic data. Protein modification studies often center on signal transduction pathways, since signals are most often transmitted by protein modifications such as phosphorylation. There are several types of experiments required for a proteomic approach to study protein modifications. Functional changes of proteins in cells occur because of modification by the attachment of groups such as phosphates, sulfates, carbohydrates, and lipids. There are more than 100 different types of posttranslational modifications that can occur to proteins, of which two of the most important are phosphorylation and glycosylation.

Protein phosphorylation. Reversible protein phosphorylation is one of the most common protein modifications. Analysis by protein labeling with 32P, or [γ-32P]ATP is the classic approach to study protein phosphorylation. Integration of this method with 2DGE and autoradiography can be used to visualize phosphoproteins in cells after in vivo labeling (3). Phosphoproteins are identified by MS analysis of gel spots (35). Direct detection of phosphoproteins from cell lysates can be achieved with use of a specific antibody directed against phosphorylated amino acid residues, with methods such as Western blot (41) or antibody array. Because phosphoproteins comprise only a small fraction of total cell-lysate proteins, methods to obtain a fraction enriched in phosphoproteins have been developed to facilitate isolation and identification. Purification of phosphopeptides with immobilized metal affinity chromatography followed by identification with MS is one approach (23). Another way is to derivatize phosphorylated serine, threonine, and tyrosine to moieties that can help in purifying and enriching the phosphopeptides by chromatography (59, 96). Tandem MS can be used to identify the specific sites of protein phosphorylation, which may be important for understanding the effect of the phosphorylation event.

Protein glycosylation. Glycosylation, the most common posttranslational modification of proteins, plays a major role in a variety of cellular events. Glycoproteins consist of proteins covalently linked to carbohydrates through either O-glycosidic or N-glycosidic bonds. The predominant carbohydrate attachments in glycoproteins of mammalian cells are via N-glycosidic linkage. The roles for glycosylation have been shown in protein folding in endoplasmic reticulum, transport and secretion, immune recognition, anchoring of proteins, and protease protection. Alteration of the expression of carbohydrate structures is strongly associated with many different types of cancer, as well as other diseases such as rheumatoid arthritis (20). Traditionally, gel electrophoresis-based methods are used for the detection of glycoproteins (44). The most commonly used method involves chromogenic detection by periodic acid staining using acid fuschin dye. Fluorescent stains have been developed (81) on the basis of the same principle and increase the sensitivity and quantitative accuracy of this approach, thereby enhancing the applicability of the 2DGE technique. The integration of these 2DGE approaches with MS systems and protein arrays will allow detection and characterization of glycosylation in a global way.

Carbohydrate structure analysis has become an important prerequisite for the functional studies on glycoproteins, because alterations in the glycosylation status have been found to accompany many diseased states. Structural investigations can be carried out at different levels, including identification of attached carbohydrates by staining procedures, characterization of glycan types by use of glycosidases, lectins, or antibodies, release of sugar chains by chemical or enzymatic treatment, labeling, chromatographic profiling, structural analysis of glycan pools or individual species by MS or NMR spectroscopy, and finally, localization of glycoproteins by immunohistochemical or cytochemical methods (30).

Hyperglycemia-induced nonenzymatic glycation of proteins is an important posttranslational modification and is believed to be responsible for the micro- and macrovascular complications of diabetes (13). It involves the condensation reaction of the carbonyl group of sugar aldehydes with the free amino groups or NH2 terminus of proteins, resulting in the formation of a Schiff base. This condensation product undergoes rearrangements through reversible acid-based catalysis to early or intermediate glycation products called amadori adducts. Advanced glycation end products (AGE) are the complex end product of the irreversible chemical reactions of these amadori adducts. Antibody-based techniques are available to establish the presence and role of amadori adducts as well as AGEs in the tissues and biological fluids of diabetic subjects (77). 1-Deoxyfructosyl lysine and carboxymethyl lysine are the major antigens of amadori adducts and AGEs, respectively (78). Antibodies directed against these antigens can be used in combination with 2DGE and MS for profiling of nonenzymatic glycated proteins in the tissues and biological fluids of diabetic subjects. This is
vital for the discovery of better indexes for glycemic control and in elucidating the mechanisms of complications of diabetes.

**Protein Turnover**

Most of the recent developments in proteomics have focused on improving the technology for protein identification and quantification. Another aspect of protein regulation that must be considered and incorporated into a comprehensive proteomic analysis is protein turnover, the combination of protein synthesis and breakdown. The balance between synthesis and breakdown determines the protein concentration in the cell or tissue. Quantification of proteins in the absence of turnover information may overlook some proteins that are affected by a particular biological condition. For example, the concentration of a protein may not change much, but the rate of turnover can be altered by a condition of interest. In such a situation, the function of the protein may change as older, damaged copies are replaced with newer proteins. A promising approach to solve this problem is to measure the synthesis rate by using in vivo metabolic labeling of proteins with isotope-labeled amino acids and measuring the increment of the protein-bound isotopic enrichment during a study period (50, 53, 56, 61). The calculation of synthetic rate of the protein also requires the isotopic enrichment in the precursor pool (40). The technology for large-scale measurement of synthetic rates of individual proteins remains to be established, although some individual protein synthetic rates can be measured in tissue samples (5, 6, 75, 76).

Protein breakdown is also essential to maintain the quality of proteins and their functional integrity. Proteins within cells are continually being degraded to amino acids and replaced by newly synthesized proteins (34). This is a highly regulated process that prevents accumulation of nonfunctional and potentially toxic proteins. A detailed review of this process is available elsewhere (33). The importance of regulation of protein breakdown has been demonstrated in the wasting conditions associated with human immunodeficiency virus infection and cancer cachexia (2, 3, 20, 31).

A simple and accurate method to study protein breakdown on a protein-by-protein basis has yet to be developed. Protein degradation can be measured across a tissue bed or at whole body level (8, 17, 54, 55). In vivo measurement of degradation rates of individual proteins is fraught with many problems. Attempts have been made to achieve this goal in yeast culture maintained at steady state with mL-[U-14C]leucine (70). It is important to determine the rate of breakdown of individual protein with a high degree of accuracy and precision to understand the selectivity of the proteolytic process whereby different proteins are committed to breakdown at significantly different rates. Although protein synthesis and breakdown are coordinately regulated in the physiological state, their mechanisms are independent (92). This difference in regulation explains the marked disparity that is sometimes seen between transcriptome and proteome data. For example, changes in mRNA levels can affect protein synthesis, which may or may not result in a change in protein concentration, depending on how protein breakdown is affected.

**Bioinformatics**

To complete the picture of protein regulation requires merging of the information from proteomic analysis, which we have described, with data from gene expression profiling (i.e., transcript microarrays) in the same tissues and conditions (79–81, 89, 90). The main challenge is to integrate the vast amount of information derived from these two (gene array and proteomic) levels of analysis. The biostatistical, biomodeling, and bioinformatic tools necessary to analyze the large data sets that are derived from these approaches are still not sufficiently powerful to fully interpret the results. Thus a major new challenge facing clinical proteomics is bioinformatics.

Bioinformatics is crucial in data management, mining, and interpretation (11, 12). In clinical research, data from muscle, plasma, and serum proteomics have been of particular interest because of the potential diagnostic and therapeutic values (37, 38). Proteomics has evolved from studying expression proteomics (mapping proteins on 2DGE) to functional proteomics (protein-protein interactions and activations with functional context). The bioinformatics techniques have evolved accordingly.

**Data management.** Current protein sequence databases were designed to collect nonredundant protein sequences, with annotation such as the molecular and cellular features (48). The organization of the data in these databases makes it ideal for sequence-related searches, such as finding a homologous family member, but not for function-related searches, which are often the ultimate goal in proteomics studies. Clinical proteomics requires the addition of physiological and clinical information about the proteins into the databases. Information about protein phosphorylation and glycosylation should also be included in the protein annotation because of the essential roles posttranslational modification plays in protein functions. There are pathway and ontological databases (e.g., KEGG database, Ingenuity, and GenMAPP) available for functional information. Appropriate bioinformatics tools are needed to access and integrate these information sets with proteomics data. Other databases include a compilation of 2D protein gel patterns that can be used to study protein expression profiles (47). A key challenge is to establish positions of corresponding protein spots across all gel images so that this information will be widely useful.

**Data mining.** The most frequently used data mining tool in proteomics when MS is used for protein identification and quantification is peptide-mass fingerprinting. The MS-generated protein profiles [mass-to-charge (m/z) ratio vs. signal intensity] are compared with the “virtually” predicted spectra of peptide m/z values from a protein database. The accuracy of this method greatly depends on the accuracy and resolution of the MS output. With the advancement of mass spectrometric techniques, such as tandem MS and dual electrospray ionization Fourier transform ion cyclotron resonance mass spectrometers, one could not only generate the peptide-mass fingerprinting for protein identification/quantification but also determine the sequence of those peptides. Searching proteomic databases such as Mascot, Sequest, or pro-ICAT with the molecular weight of the peptide and its sequence information usually provides the identification of the protein from which this peptide is generated. In addition, in differential proteomics, one can use these tools to identify and quantify the proteins.
that are differentially expressed between the healthy control subjects and patients.

Protein-protein interaction model for drug design. Bioinformatics and functional proteomic methods can be used to predict and validate protein complex formation (4). These methods take advantage of the known protein structures recorded in the Protein Data Bank database and use information from protein homology, protein functional domains, pathway profiling, and the like, to model the interaction conformations between two or more proteins. This approach has been widely used in the computer-aided drug design process (25, 29). The challenge in this field is the limited number of proteins with known structure because of the difficulties in obtaining enough proteins with crystallographic purity.

Pattern recognition as a diagnostic tool. For some poor prognostic malignancies, such as pancreatic and ovarian cancers, early diagnosis and surgery are the best therapeutic approaches. There are no specific and highly sensitive biomarkers available for these diseases. A self-trained pattern recognition algorithm has been proven capable of identifying proteomic patterns in MS spectra to completely segregate cancer from normal (67), although no specific proteins were identified. These pattern recognition algorithms involve complicated neural networking technologies. This pattern recognition strategy has been applied to 2D gel patterns by “translating” the 2D gel densitogram into a unique zigzag “fingerprint” pattern. The gel spots were ordered to represent proteins in a 2D gel map according to their relative intensities (73).

APPLICATIONS OF PROTEOMICS IN CLINICAL RESEARCH

In this section, we provide examples of how proteomics methods have been used to understand specific diseases or conditions. There are several examples in the literature of proteomic approaches used to address problems in yeast or cell cultures. The examples provided below illustrate how proteomics has been used in human studies of heart disease and ovarian cancer. We are not aware of any published studies to date that have applied the types of tools described here to investigate questions in endocrinology or metabolism in humans. There are many research groups interested in these areas, however, so it is expected that results will begin to appear in the near future.

Heart Disease

Heart diseases resulting in heart failure are the leading cause of morbidity and mortality in developed nations. The pathophysiology of various heart diseases is still unknown, but it is likely that significant changes in cardiac gene and protein expression underlie the disease processes and determine their progression and outcome. With this purpose in mind, several heart proteome databases were developed on the basis of 2DGE analysis of heart proteins (45a, 52, 68). These databases contain information on several hundred cardiac proteins that have been identified by various approaches.

Proteomic investigations of human heart disease have so far concentrated on dilated cardiomyopathy (DCM), a severe cardiac dysfunction characterized by impaired systolic function resulting in heart failure. The combined results from several studies identified about 100 cardiac proteins that are markedly altered in DCM and are classified into primarily three classes, 1) cytoskeletal and myofibrillar proteins, 2) mitochondrial proteins, and 3) stress response proteins (18, 43). Such a global approach should provide new insights into the cellular mechanisms involved in cardiac dysfunction, and the major challenge will now be to investigate the contribution of these changes to altered cellular functions resulting in heart failure.

Proteomics has also been used to identify cardiac-specific antigens that produce an antibody response in heart disease and after heart transplantation (45, 69, 93). These data were obtained using a 2DGE separation of cardiac proteins followed by Western blot analysis using monoclonal antibodies against these proteins. With this strategy, several cardiac proteins were identified in DCM and myocarditis (63). Studies in heart transplant patients identified several cardiac antigens that mediate acute and chronic allograft rejection (63).

Ovarian Cancer

One of the difficulties in treating ovarian cancer is that it presents at a late clinical stage in ~80% of patients, and the average 5-year survival period is only 35–40%. Thus it would be extremely valuable to be able to identify the disease at an earlier stage of development, when treatments could be more effective. Proteomic strategies are being used to identify important biomarkers at an early stage of ovarian cancer. Petricoin et al. (67) identified a protein pattern in the serum proteome obtained from 50 each of control (noncancer) and ovarian cancer patients by use of surface-enhanced laser desorption and ionization time-of-flight (SELDI-TOF) MS. These authors claim that the sensitivity is 100% and that specificity is 95%, with a positive predictive value of 94% in this analysis. SELDI-TOF technology utilizes solid supports or “chips” made of aluminum or stainless steel engineered with bait surface 1 to 2 mm in diameter. The bait surfaces can be hydrophobic, cationic, or anionic chromatographic supports or can be based on “affinity supports,” comprised of biochemical molecules such as antibodies, purified proteins such as receptors or ligands, or DNA oligonucleotides. Small amounts of solubilized tissue or body fluid are directly applied to the bait surface. After washing to remove unbound proteins, the proteins specifically interacting with the bait surface are analyzed by MS-TOF, similar to standard MALDI-TOF analysis. The low-molecular-weight ionized proteins and peptides are recorded as mass signature peaks as they strike a detector plate based on their differential time of flight, and the data are displayed as a standard chromatograph (95). SELDI analysis software allows data to be viewed as a standard mass chromatograph or as a gel-like density graph. It is very promising in terms of identifying biomarkers from a simple blood analysis, and there may be potential applications of this approach to other diseases. However, an important limitation is that the interaction between the modified surface of the protein chips used to capture the protein from a mixture and various proteins is nonspecific, and washing with different solvents could remove the proteins in a nonuniform way. This can increase the variability in the measurements between samples and patients.

It should be noted that, because MS analysis is very time and labor intensive, it might not be practical to perform this type of work on every patient sample. Rather, these detailed proteomic techniques will oftentimes be used mainly in the discovery process to identify early biomarkers of disease. Subsequently, methods with rapid throughput, such as an ELISA or other
antibody-based assays, can be developed for more widespread use in clinical settings. Together, it will be possible to find new biomarkers to diagnose a disease and follow the treatment regimen.

There are at least two clinical trials being conducted by the National Cancer Institute (NCI) for the ovarian cancer field with proteomic approaches (58). The first of these trials is an observational study investigating whether protein patterns can be used to predict recurrence of ovarian cancer. The patients being studied have either stage III or IV primary peritoneal, fallopian tube, or ovarian epithelial cancer, or stage IIC ovarian clear cell cystadenocarcinoma in first clinical remission. Serum and tumor samples will be analyzed to identify protein patterns that emerge in the event of a relapse. Approximately 40 patients are in this pilot study (www.cancer.gov/newscenter-clinical). Patients are seen at baseline, 1 mo, 3 mo, and thereafter every 3 mo for history, physical examination, and various other kinds of laboratory testing. Patient evaluation continues until biopsy-proven relapse.

The second clinical trial will monitor tumor cell protein response to the cancer drug imatinib mesylate (Gleevec). The specific aim of this NCI-based study is to understand more about the interaction of this drug with various tumor cell-signaling pathways. Gleevec is a molecularly targeted therapy, designed to interfere with specific tyrosine kinases that promote cancerous growth. This drug is already approved for use in cancer treatment, but this study will focus on ovarian cancer and how the drug regulates protein expression in ovarian tumors by use of proteomic methods like those described here.

LIMITATIONS AND FUTURE DEVELOPMENTS

There are already efforts underway to catalog all of the proteins in tissues and serum, much like the large-scale sequencing efforts used on the genome projects for humans and other species. However, the approaches described above are unlikely to provide the complete profile of proteins in either muscle or serum. First, it is hard to achieve a high-throughput screening of complex protein mixtures in muscle or serum while being able to identify low abundant proteins. Many proteins with major roles, especially during short intervention studies, are enzymes (e.g., kinases, proteases, ligases, phosphatases, etc.) present in low concentrations, and they are difficult to separate and identify from a complex mixture that includes many high-concentration proteins. In addition, during short intervention studies, changes in concentration of proteins with slow turnover (e.g., myosin heavy chain isoforms) (5) are too small to detect by any available techniques. Moreover, changes in concentration of proteins with fast turnover (rapid rates of synthesis and degradation) are unlikely to be detected by the available proteomic techniques. Isotope tracer approaches in combination with MS to identify the changes in synthesis rates of these proteins are likely to be sensitive enough to detect changes occurring during short intervention studies. Although isotope tracer techniques are available to measure fractional synthesis rates of mixtures of several proteins (10, 53, 57) and individual proteins (5, 19, 28), these techniques are not currently applicable to large-scale analysis of synthesis rates of individual proteins.

It is highly likely that alterations in one protein may occur in isolation, causing a disease state, or during a physiological perturbation. For example, a primary change in concentration of a protein like insulin initiates a cascade of events resulting in alterations of gene transcription (79), translation (42), and increase or decrease in production of several other proteins (55). These may also result in modification of proteins (e.g., glycosylation, phosphorylation, etc.), all of which will determine the ultimate functional alterations manifested in a disease state (13, 40, 51). It is now well known that type I diabetes results from insulin deficiency. Insulin deficiency results in a cascade of events resulting in alterations in expression of many genes (79), proteins (55), and metabolites (74). A mass scale proteomic analysis will demonstrate changes in many proteins, and these changes will vary from tissue to tissue. A fundamental understanding of the basic mechanism that causes these changes would have eluded investigation for many years if only a simple quantitative proteomic approach had been used, even on a large scale. Thus we should be mindful that the classical experimental approach, which led to the discovery of insulin, will still form the basis of future research in many diseases. However, the advances in technology available for identifying proteomes in combination with the classical experimental approach, as was adopted by Banting and Best (7), should accelerate the process of discovery of the causes of many diseases. For example, the current proteomic approaches will be invaluable for identifying all of the proteins produced in the pancreas. This can be done in different pancreatic components, such as islet cells and exocrine cells. A physician will also be interested to identify the changes in the pattern of proteins secreted from the pancreas in response to various stimuli, such as substrates (e.g., glucose, amino acids, fatty acids, etc.) or drugs. In addition, it is possible to identify the changes in proteins (concentration, modification, etc.) in various tissues (e.g., muscle, fat, liver, etc.) occurring in response to changes in insulin concentration (with varying levels of substrates that are altered by insulin). The current and
developing proteomic approaches have the potential to perform these measurements.

SUMMARY AND CONCLUSION

Proteins are functional molecules, and the study of proteins is of major interest to physiologists and physician scientists. The new technologies and the sequencing of the genomes of humans and other species offer opportunities to study genes and proteins on a larger scale than ever before. Because one protein may cause cascade effects and change the synthesis and functions of other proteins on a larger scale than ever before. Because one gene is of major interest to physiologists and physician scientists. Ford, Jane Kahl, and Larry Ward.

ACKNOWLEDGMENTS

We acknowledge the skillful technical support of Mai Persson, Charles Ford, Jane Kahl, and Larry Ward.

GRANTS

The work of the authors is supported by the National Institutes of Health Grants R01 DK-41973 and R01 AG-09531 (to K. S. Nair), General Clinical Research Center Grant M01 RR-00585, the Mayo Foundation, and the Murdock-Dole Professorship (to K. S. Nair).

REFERENCES


