Proteomics: A powerful tool in the post-genomic era

Larysa Porubleva and Parag R Chitnis*
Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011, USA
Accepted 23 June 2000

Genomics is having a profound impact on biological research, including photosynthesis investigations. Genomes of many photosynthetic organisms have been sequenced. The information about ALL genes that govern and execute photoautotrophic metabolism provides many opportunities to understand genome function and details of known and uncharted pathways. Proteomics, analysis of the protein complement of the genome, is a powerful tool in understanding which proteins are present in a particular tissue under given conditions. Proteomics also allows us to estimate relative levels of proteins and to determine post-translational modifications of the gene products. In this minireview, we discuss the technology and its applications in plant sciences.

Genomics is one of the most influential fields in biology today. In an editorial in Science, Philip H. Abelson calls genomics the "third technological revolution". Recently, genomes of several bacteria, yeast and C. elegans have been sequenced completely. Genomes of several photosynthetic prokaryotes have been sequenced completely. Complete genome sequences of Arabidopsis and rice will be known shortly. Extensive databases of expressed sequence tags (EST) are now available for almost every major crop species. Thus plant biology has entered the genomics era. These huge amounts of data provide many challenges and opportunities. Much of the information in the sequenced genomes remains uncharted. For example, 34-68% of the open reading frames in the sequenced bacterial genomes have unknown functions. Systematic approaches in functional genomics are required to identify roles of these genes. Another opportunity in functional genomics is the ability to examine global changes in the genome expression in response to environmental or internal factors. Indeed, functional genomics is the next frontier.

Almost all genes function after they are expressed as proteins. Therefore, proteomics, analysis of the protein complement of the genome, is essential to understand the genome function. Proteome analysis involves identification of all proteins from an organism and determination of their abundance. Due to rapid developments in the technology, proteomics is destined to play an indispensable role in the post-genomic era. In this minireview, we discuss the proteomics technology and its applications, with an emphasis on photosynthetic organisms.

What is a proteome?

A proteome is the total protein complement of a genome. The science of proteomics, in its broadest sense, deals with the expression, structure and function of the genome at the protein level. The most active area of proteomics (and the major emphasis in this minireview) is characterization of the functional proteome, which is defined as gene-products expressed under specific environmental or intrinsic conditions. Other aspects of proteomics, such as genome-wide prediction of tertiary structures and high throughput analysis of protein structure and biochemical functions.

Analysis of a functional proteome provides a powerful tool for examining genome expression. Proteins are directly responsible for the function and phenotype of an organism. As Edmond H. Fischer, 1992 Nobel Laureate, points out, genome sequencing might enable us to predict the proteins that can potentially be generated, but not where, when or at what level. Nor can it take into account the enormous diversity generated by gene rearrangements, RNA splicing and editing, and post-translational modifications of proteins. DNA sequences alone tell us little about the

*Author for correspondence
Phone: 1-515-294-1657; Fax: 1-515-294-0453
Email: chitnis@iastate.edu

Abbreviations used: 2D, two dimensional; 2DGE, two-dimensional gel electrophoresis; MS, mass spectrometry; PMF, peptide mass fingerprinting.
physiological function of proteins. Proteome analysis provides a direct method for analyzing the complexity and diversity of proteins in a cell. Proteomics tells us what fraction of the genome is functional, and at what levels. Proteomics gives a global picture of metabolic and developmental changes in gene functions. Analysis of gene expression at the RNA level alone ignores translational and post-translational regulation. Recently, comparison of protein and transcript levels in yeast demonstrated that the correlation between mRNA and protein levels was insufficient to predict protein expression levels from quantitative mRNA data. Thus, proteomics is needed to truly understand genome expression.

How is a proteome studied?

Traditional biochemistry involved protein separation, characterization and linking proteins to their corresponding genes. Proteomics is an extension of these protein mapping and protein identification techniques that have been used routinely for small-scale analysis. The proteins are separated by two-dimensional gel electrophoresis (2DGE), estimated using image analysis and analyzed with mass spectrometry. Once proteins are identified, a reference map of 2D display can be used for comparative analysis. This approach complements the genome analysis, EST sequencing and analysis of RNA using microarray techniques. The recent emergence of proteomics as a powerful tool is due to many improvements in the technology.

Like a typical biochemical purification procedure, preparation of protein samples remains a crucial start in protein separation. The quality of 2D gels is strongly dependent on purity and an extent of protein solubilization. Any contamination such as ionic compounds, small charged molecules, DNA, lipids, pigments, etc. increase conductivity and viscosity of the sample, thus interfering with protein migration in an electric field. Recently new chaotropes and detergents have been introduced to overcome these problems. Solubilization of hydrophobic transmembrane proteins is still a very serious problem for proteomics.

The protein separation with 2DGE is a core technology in proteomics. In the first dimension, isoelectric focusing is performed using immobilized pH gradient (IPG) strips, in which buffering and titrant groups are covalently bound to the polyacrylamide matrix. The consistent performance of gel strips avoids instability, discontinuity and variations in the pH gradient. Computer standardized pre-dried IPG strips give run-to-run reproducibility. Thus the long awaited ideal of international rather than in-house reproducibility has become achievable. In the second dimension, proteins are separated according to their molecular mass. Such combination results in visualizing up to 10,000 proteins per a gel. An example of protein separation with 2DGE is shown in Fig. 1. Currently 2DGE remains to be the most powerful technique for separating a few thousand proteins simultaneously.

Another major advance in 2DGE is development of computer software for comparison and quantitative analysis. The genome remains constant in an organism and requires a single structural determination. Functional proteome varies and its analysis requires highly defined samples and highly reproducible experimentation. Researchers have used incorporation of radioactive isotopes for quantifying absolute amounts of proteins. Alternatively, chemically tagged peptides have been used for simultaneous quantification of proteins using mass spectrometry. However, quantitative analysis of patterns in the stained gels remains the most widely used technique for analyzing proteomes. For analysis of complex 2D patterns, significant advances have occurred in computerized image analysis of 2D gels, spot detection, background subtraction, spot matching, database construction, and numerical data analysis.

The most significant technological leap in proteomics comes from the application of mass spectrometry (MS) as a sensitive, inexpensive and rapid tool for protein identification. Despite the recent improvements in the sensitivity and automation of Edman protein microsequencing, this methodology has remained time and cost intensive to such an extent as to preclude large-scale proteome analysis.

Protein separation techniques yield several hundred to a few thousand resolved proteins. Hierarchical deployment of mass spectrometry is a promising solution in achieving high-throughput screening. The quickest and most inexpensive technology is used first, followed by increasingly time and cost-intensive procedures.

In a typical proteomic analysis, gel spots are excised and proteins in the gel are digested with enzymes, trypsin being the most commonly used protease. Peptide mass fingerprints (PMF) are collected with a matrix-assisted laser desorption ionization - time of flight (MALDI-TOF) mass spectrometer. Fig. 2
shows the typical mass spectrum obtained with MALDI-TOF MS. Many tools are available for protein identification and characterization using PMF data (see ExPaSy http://www.expasy.ch for a list of tools that are available on the web). Accurate masses of as few as four peptides with one or more orthogonal data sets (e.g. pl, organism, expression information) can be used to scan the existing protein sequence databases to identify a protein unambiguously\textsuperscript{27}. PMF data can also be used to predict post-translational modifications\textsuperscript{28}. When combined with robotics for gel excision, liquid-handling systems and automated data collection, one can obtain PMF on as many as 300 proteins in a day\textsuperscript{29}.

The next step would involve additional data such as fragmentation spectra, sequence tagging or a com-
bination of one or more additional protease digestion. In general, the larger the genome, the more rigorous the requirements for identifying a protein by mass spectrometry or sequence analysis. The use of peptide masses alone for identification can be compromised by the presence of extensive post-translational modifications, an incomplete genome database, errors in the DNA sequence, incomplete EST sequences, and other issues. In cases where the peptide masses are insufficient to identify a protein, it will be necessary to obtain direct sequence information via fragmentation analysis. While it is possible to obtain fragmentation information from MALDI mass spectrometry, this is not sufficiently reliable to be of routine use. An electrospray source on a tandem mass spectrometer or an ion trap is combined with HPLC or capillary electrophoresis and is used to obtain sequence information. This process can be automated for increasing throughput of analysis. Mass spectrometry has enabled researchers to characterize more than twice the number of yeast proteins in a few weeks than had previously been determined over a decade.

The elegant proteome studies in animal systems, E. coli and yeast have provided extensive information about changes in proteomes in response to disease conditions, stress conditions, or drugs. Although proteomes of many bacteria, yeast and animal systems have been investigated in detail, plants have been studied with proteomic techniques only recently. Similarly, proteome analyses of the cyanobacterium Synechocystis sp. PCC 6803 have commenced only in the last few years. One difficulty in analysis of plant proteomes is that separation of plant proteins with 2DGE is complicated by relatively low content of proteins in tissues and presence of phenolic compounds. Table 1 lists some recent 2DGE studies describing plant proteomes. In most studies so far, the plant proteomes were identified by amino-acid microsequencing. Consequently, very few proteins have been identified in the plant proteomes. Additionally, since most plant genomes have not been sequenced completely, a significant part of the analyzed proteins could not be identified using current databases (Table 1).

The technological challenges in the future research

Proteomics is a science in its infancy. The success of this endeavor will be determined by the extent to which this technology-driven science can overcome its shortcomings. Here we discuss the challenges for proteomics. Note that mRNA profiling with microarrays also has similar weaknesses with normalized representation, variation, proper controls and reproducibility.

(i) Variation: The genome remains constant in an organism and requires a single structural determination. Functional proteome varies and its analysis requires highly defined samples and highly reproducible experimentation. Proteome analysis must depend upon rigorous experimental design and the need to examine several physiological states both in vitro and in vivo.

(ii) Sensitivity: A weakness of the current proteome technology is the poor detection threshold of 2DGE. Protein detection lacks the advantage of amplification that has revolutionized nucleic acid analysis. From current estimates, silver staining of a 2D gel can be used to detect proteins that are 1000 copies per an eukaryotic cell. Thus, only 40%-70% of all genome activity within a given system can be followed simultaneously and quantitatively. However, there are several ways to increase sensitivity. If antibodies are available, immunodetection with enhanced chemiluminescence can be used to increase sensitivity by 10-100 fold over silver staining. Such level of sensitivity enhancement (as little as 125 picograms of single proteins) can also be reached with radioactive iodine and fluorescent dyes. At present these dyes of various colors allow to analyze up to three different samples at the time. Sample fractionation based upon different cellular localization of the proteins, their relative solubilities, charges, sizes, affinity to ligands and use of narrow pH ranges in the first dimension increase detection sensitivity by up to 100 fold. After analyzing multiple subproteomes, results can be pooled to generate information about the complete proteome. Even when all proteins cannot be detected and identified in proteomic analysis, by placing points of reference on 2DGE maps, our understanding of biological processes is destined to become more holistic.

(iii) Speed: Despite the recent improvements in the sensitivity and automation of Edman protein microsequencing, this methodology has remained time and cost intensive to such an extent as to preclude large-scale proteome analysis. For organisms with completely sequenced genomes as well as for those lacking significant DNA sequence information, mass spectrometry has an essential role to play in achieving
<table>
<thead>
<tr>
<th>Plant species</th>
<th>Subproteome</th>
<th>Method of 2-D electrophoresis*</th>
<th>Number of protein resolved</th>
<th>Number of spots analysed</th>
<th>Number of protein identified</th>
<th>Proteins with no assigned functions</th>
<th>Method of identification**</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barley (<em>Hordeum vulgare</em> L.)</td>
<td>Seeds, endosperm</td>
<td>CA/U/N</td>
<td>no report</td>
<td>26</td>
<td>12</td>
<td>0</td>
<td>AAS</td>
<td>[71]</td>
</tr>
<tr>
<td>Spinach (<em>Spinacea oleracea</em> L.)</td>
<td>Thylakoid membrane</td>
<td>CA/U/N</td>
<td>about 100 spots</td>
<td>15</td>
<td>11</td>
<td>0</td>
<td>IB</td>
<td>[72]</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>Total proteins from five tissues</td>
<td>IPG/U/T</td>
<td>4763 spots</td>
<td>125</td>
<td>48</td>
<td>27</td>
<td>AAS</td>
<td>[73, 74]</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>Total proteins from nine tissues</td>
<td>IPG/U/T</td>
<td>4892 spots</td>
<td>144</td>
<td>36</td>
<td>24</td>
<td>AAS</td>
<td>[74, 75]</td>
</tr>
<tr>
<td>Maize (<em>Zea mays</em> L.)</td>
<td>Total coleoptile proteins ***</td>
<td>IPG/U/CHAPS</td>
<td>no report</td>
<td>56</td>
<td>12</td>
<td>2</td>
<td>AAS</td>
<td>[76]</td>
</tr>
<tr>
<td>Maize (<em>Zea mays</em> L.)</td>
<td>Total leaf proteins***</td>
<td>IPG/U/CHAPS</td>
<td>no report</td>
<td>18</td>
<td>13</td>
<td>1</td>
<td>AAS</td>
<td>[76]</td>
</tr>
<tr>
<td>Tobacco (<em>Nicotiana tabacum</em>)</td>
<td>Plasma membrane</td>
<td>IPG/U/CHAPS, CA/U/N</td>
<td>about 600 spots</td>
<td>16</td>
<td>0</td>
<td>6</td>
<td>AAS</td>
<td>[77]</td>
</tr>
<tr>
<td><em>Pine (Pinus pinaster Ait.</em>)</td>
<td>Total needle proteins</td>
<td>CA/U/T</td>
<td>about 900 spots</td>
<td>28</td>
<td>19</td>
<td>3</td>
<td>AAS</td>
<td>[78]</td>
</tr>
<tr>
<td><em>Pine (Pinus pinaster Ait.</em>)</td>
<td>Total xylem proteins</td>
<td>CA/U/T</td>
<td>about 600 spots</td>
<td>35</td>
<td>16</td>
<td>3</td>
<td>AAS</td>
<td>[78]</td>
</tr>
<tr>
<td><em>Rice (Oryza sativa</em> L.)</td>
<td>Total proteins from green and etiolated shoots</td>
<td>IPG/U/N</td>
<td>about 300 spots</td>
<td>85</td>
<td>30</td>
<td>16</td>
<td>AAS</td>
<td>[79]</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>Plasma membrane ***;</td>
<td>IPG/U/CHAPS, SB IPG/U/TU/CHAPS</td>
<td>total proteins specific in plasma membrane fraction</td>
<td>about 700 spots</td>
<td>104</td>
<td>52</td>
<td>31</td>
<td>AAS</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td></td>
<td></td>
<td>about 350 spots</td>
<td>22</td>
<td>8</td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Abbreviations: IPG, immobilized pH gradient; CA, carrier ampholites IEF; U, urea; N, NP-4; T, Triton X-100; TU, thiourea; SB, sulphobetaine SB 3-10.
** AAS, amino-acid sequencing; IB, immunoblotting; MS, mass spectrometry.
*** 2-D Database were created
high-throughput screening. Sensitive high-throughput screening of proteins can employ mass spectrometry in a hierarchical manner. Initial protein characterization can proceed rapidly, cost-effectively and in a manner that lends itself to automation and robotics. Accurate mass of a few peptides with one or more orthogonal data sets (e.g. pl, organism, expression information) can be used to scan the existing protein sequence databases to identify a protein unambiguously. The next step could involve additional data such as fragmentation spectra, sequence tagging or a combination of one or more additional protease digestions. Such strategies have enabled researchers to characterize more than twice the number of yeast proteins in a few weeks than had previously been determined over a decade. High throughput screening has made large-scale proteome analysis a reality.

(iv) Membrane proteins: Membrane proteins constitute a major component of a proteome. Approximately 30% of the bacterial genes code for integral membrane proteins. Membrane proteins drive and mediate many essential cellular processes such as signal transduction, cell adhesion, hormone perception and transport of proteins, metabolites, nutrients and ions across cell membranes. Unfortunately the integral membrane proteins are difficult to study using current proteomic technology. Only 66 integral membrane proteins from bacteria, yeast, animals and plants have been identified in contrast with thousands soluble and peripheral membrane proteins that have been identified on 2-D gels. Recently a consortium of European researchers has undertaken meticulous analysis of the plant plasma membrane proteins (PPMdb) (ref. 51). It contains comprehensive two-dimensional polyacrylamide gel electrophoresis map, partial amino acid sequences and expression data.

Applications in the Functional Genomics

Genomics is moving now from a focus on genome structure and comparisons of nucleotide sequences to a focus on gene function. Functional genomics encompasses the development and application of global (genome-wide or system-wide) experimental approaches to assess gene function by making use of the information and reagents provided by structural genomics. Proteomics provides a powerful problem-solving tool in functional genomics.

Proteome analysis can aid in genomic annotation

The ability of proteomics to confirm the existence of gene-products predicted from DNA sequence is a major contribution to genomic science. Proteomic analysis of very small proteins (<5 kDa) will enable better annotation of genome sequences. Small open reading frames (ORFs) are generally not taken into account during genomic annotations. Proteomics will allow us to identify which of these small ORFs are functional genes. When DNA sequences are available, identification of proteins in a proteomic project could link a protein to a gene with known function or could show that a protein is a product of an unidentified open reading frame. In the latter case, proteomics demonstrates that a putative ORF is indeed a functional gene. For example, of 129 abundant proteins identified in a study on thylakoids of the cyanobacterium Synechocystis sp. PCC 6803, 37% proteins were derived from unidentified open reading frames.

Regulation of genes can be understood by comparative proteomics

Proteomics tells us what fraction of the genome is functional, and at what levels during development and under stress conditions. Such analyses must not be limited to changes in the RNA levels alone. Frequently, plant gene expression is regulated at translational and post-translational levels. It is imperative to examine changes in the proteome if we want to obtain a true understanding of how a phenotype is altered by changes in the genome expression. Cluster analysis of the expression patterns can reveal roles of the novel proteins in cellular physiology and metabolism. Cellular proteins can be labeled in vivo and then used for protein analysis. Thus, proteome analysis can be used to understand dynamics of genome expression. Proteomics also provides a way to characterize mutants rapidly. The altered expression pattern reveals possible interconnections in gene expression cascades and allows analysis of global effects due to mutation in a specific gene. Subcellular fractionation prior to proteome analysis demonstrates intracellular location of a gene-product and indicates possible function. Identification of subcellular location of a protein is a unique application of proteomics. Genomic sequences or mRNA profiling cannot demonstrate intracellular localization. Thus, proteome analysis has several implications in functional genomics.

The resolving power of 2-D electrophoresis has been used to study genetic variability in plants and to
### Table 2—The changes in the protein composition in response to environmental stresses studied with 2-D electrophoresis

<table>
<thead>
<tr>
<th>Stress</th>
<th>Plant, organ</th>
<th>Number of proteins changed</th>
<th>Method of identification</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phosphate deprivation</td>
<td>Tomato roots</td>
<td>2 NI</td>
<td>NI</td>
<td>[82]</td>
</tr>
<tr>
<td>Chilling acclimation</td>
<td>Soybean</td>
<td>2 AAS</td>
<td>AAS</td>
<td>[83]</td>
</tr>
<tr>
<td>Salt excess</td>
<td>Barley root</td>
<td>2*</td>
<td>NI</td>
<td>[84]</td>
</tr>
<tr>
<td>Drought</td>
<td>Maize</td>
<td>78 (38)**</td>
<td>11</td>
<td>AAS</td>
</tr>
<tr>
<td>Jasmonic acid</td>
<td>Rice</td>
<td></td>
<td>11</td>
<td>AAS</td>
</tr>
<tr>
<td>UV Irradiation</td>
<td>Rice</td>
<td></td>
<td>4</td>
<td>AAS</td>
</tr>
<tr>
<td>Copper chloride</td>
<td>Rice</td>
<td></td>
<td>5</td>
<td>AAS</td>
</tr>
<tr>
<td>Iron deficiency</td>
<td>Maize root plasmalemma</td>
<td>4***</td>
<td>NI</td>
<td>[87]</td>
</tr>
<tr>
<td>Ozone</td>
<td>Arabidopsis leaves</td>
<td>2</td>
<td>1</td>
<td>NI</td>
</tr>
<tr>
<td>Heat shock</td>
<td>Maize</td>
<td>6</td>
<td>35</td>
<td>MS</td>
</tr>
<tr>
<td>Anoxia</td>
<td>Maize root</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The number of proteins increased most strikingly.
** The number of proteins with different expression in two genotypes.
*** The changes in the mutant against the wild type.
**** Only 45 kDa heat shock protein were studied.

analyze mutant phenotypes. For example, analysis of the 2-D gels of total proteins extracted from developmental mutants of *Arabidopsis thaliana* (L.) Heynh. and from wild-type plants grown in the presence of various hormones allowed to find mutants overaccumulating cytokinins. The possibility of 2DGE to visualize hundreds of proteins at the time have been successfully applied for studying the protein changes induced by diseases. Comparative proteomics has also been used for identifying differentially expressed proteins under abiotic stress conditions (Table 2).

### Epilogue

The current microarray and proteome analyses focus on the absolute quantity of a gene-product (RNA and/or protein). This emphasis will shift to more relevant parameters, such as, molecular half-life; synthesis rate; the influence of the environment, cell cycle, stress and disease on gene-products; and the collective roles of multigenic and epigenetic phenomena. Cellular proteins can be labeled *in vivo* and then used for protein analysis. Thus, proteome analysis can be used to understand dynamics of genome expression. Another emphasis needs to be on protein function and interactions. Several new developments hold promise in this area. The traditional biochemical techniques for studying protein-protein interactions such as co-immunoprecipitation, cross-linking, and cofractionation are labor- and time-extensive. New approaches have been developed recently to make these investigations more amenable to a high throughput analysis. To define both indirect and direct protein interactions, protein complexes can be purified with a novel tandem affinity purification (TAP) tag from a relatively small number of cells without prior knowledge of complex composition, activity or function. The component of isolated complex then can be identified with mass spectrometry. Biosensors that are based on surface plasmon-resonance biosensors are ideal for characterization and identification of protein-protein interactions in purified samples, as well as in complex mixtures. Combination of SPR-biomolecular interaction analysis and MALDI-TOF provides a powerful approach in proteome analysis. Detection limits for both SPR-BIA and MALDI-TOF are at low-femtomole to subfemtomole level. Applying these tools in proteomics could lead to ligand and protein-complex identification and to estimation of the kinetic parameters of these interactions.

Since the pioneering work of O’Farrell some twenty five years ago, protein analysis by 2DGE has come a long way. In the post-genomic era, several key improvements are driving proteomics: reproducible 2D gel technology, staining and scanning technology, mass spectrometry for identification, and bioinfor-
matics for database construction and searching. In the future, many new developments in the technology and major investments from genomics industry are bound to propel proteomics to a scale that will be comparable to the high throughput genomics projects. Microfabrication at different steps will aid protein display and analysis. Some methods bypass the 2D electrophoretic separation. For a recently proposed method identifies intact proteins from genomic databases using a combination of accurate mass spectrometry with a magnetic sector mass spectrometer holds a promise for a high throughput analysis of phosphorylated proteins. Thus, many innovative approaches are being examined to improve the proteomic technology. Genomic sequences when complemented with the information derived from microarray hybridization assays and proteome analysis may herald a new era in holistic plant biology.

Acknowledgement

The authors’ research on proteomics was supported in part by grants from NSF-NATO and Iowa Corn Promotion Board. Journal Paper No. J-XXXX of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. 3416 and supported in part by Hatch Act and State of Iowa funds.

References

1 Abelson P H (1998) Science 279, 2019
5 Duboule D (1997) Science 278, 555
38 Suzuki T & Ohara O (1997) Electrophoresis 18, 1252-8