Prospects and Challenges in Proteomics

Paul Bertone and Michael Snyder*

Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520–8103

Proteomics is one of the fastest growing areas in areas of research, largely because the global-scale analysis of proteins is expected to yield more direct understanding of function and regulation than analysis of genes. Although significant advances in the comprehensive profiling, functional analysis, and regulation of proteins has occurred in model organisms such as yeast (*Saccharomyces cerevisiae*) and in humans, proteomics research in plants has not advanced at the same pace. The availability of the complete Arabidopsis (*Arabidopsis thaliana*) genome, which is small compared to that of other plants, along with an increasingly comprehensive catalog of protein-coding information from large-scale cDNA sequencing (Seki et al., 2004) and transcript mapping (Yamada et al., 2003) experiments, set it apart as a complex but accessible model organism to study plant proteomics.

The application of proteomic approaches to plants entails three major challenges: (1) comprehensive identification of proteins, their isoforms, and their prevalence in each tissue; (2) characterizing the biochemical and cellular functions of each protein; and (3) the analysis of protein regulation and its relation to other regulatory networks.

GLOBAL ANALYSIS OF PROTEIN EXPRESSION: TOWARD CATALOGING THE PROTEOME

Unlike the genes of prokaryotes, which can be readily identified as long open reading frames, the genes of higher eukaryotes, including plants, typically contain introns that can be both large and numerous. The combinatorial exon usage evident in mRNA transcripts originating from complex gene structures often results in a multitude of splice variants, which in turn give rise to many different protein products from a given gene. Thus, the determination of the comprehensive pattern of expression of each protein isoform is expected to be challenging, particularly for those expressed at a low level, and will require new approaches.

A number of methods are currently available for profiling protein expression (Agrawal et al., 2005), including two-dimensional gel electrophoresis (2-DGE) or liquid chromatography followed by tandem mass spectrometry. The 2-DGE entails the separation of complex protein mixtures by molecular charge in the first dimension and by mass in the second dimension. Although recent advances in 2-DGE have improved resolution and reproducibility, the technique remains difficult to automate in a high-throughput setting. For this reason, alternative approaches that obviate the need for gel separation, such as multidimensional protein identification technology, have gained popularity for large-scale proteomics efforts and are able to generate a large catalog of proteins present in complex cell extracts. High-throughput protein analysis is expected to accelerate with the introduction of new robotic liquid chromatography systems and high-resolution analysis methods such as Fourier transform mass spectrometry, which should allow the detection of several thousand proteins.

Although current approaches allow the detection of thousands of proteins, even further advances will be required to be able to truly profile the tens to hundreds of thousands of different protein isoforms present in each cell. One method for detecting more proteins is subcellular fractionation. This approach can dramatically reduce the complexity of protein extracts, while rare proteins are enriched and thus more readily detected. An added benefit is the functional information gained from determining the subcellular compartments where individual proteins are localized. To this end, recent efforts have successfully characterized the nuclear, chloroplast, amyloplast, plasma membrane, peroxisome, endoplasmic reticulum, cell wall, and mitochondrial proteomes in Arabidopsis. Collectively, these projects are beginning to provide crucial genomic and proteomic data to the broader research community through the establishment of comprehensive databases that integrate sequence and functional information (Schneider et al., 2004).

In leveraging high-throughput technologies to characterize the proteome of Arabidopsis and other organisms, particular attention must be paid to classes of proteins that are recalcitrant to isolation and analysis, such as membrane and other hydrophobic proteins. Membrane proteins constitute 30% of the typical proteome, yet their propensity to aggregate and precipitate in solution confounds their analysis. Here, the focus on characterizing the subproteomes of discrete cellular compartments offers the opportunity to develop variations of extraction procedures that take into account the specific properties of membrane proteins.

* Corresponding author; e-mail michael.snyder@yale.edu; fax 203–432–3597.

www.plantphysiol.org/cgi/doi/10.1104/pp.104.900154.

Several groups have taken advantage of this approach to recover a higher percentage of membrane proteins from subcellular extracts using various nonionic and zwitterionic detergents or phase-partitioning methods. These efforts resulted in the successful determination of the protein complement of Arabidopsis plasma membrane (Nuhse et al., 2004), as well as the thylakoid and envelope membrane systems of the chloroplast (Froehlich et al., 2003).

LARGE-SCALE FUNCTIONAL ANALYSIS OF PROTEINS

One of the most important challenges in proteomics is determining the function of every isoform of each protein. More than one-third of plant genes identified by genome sequencing lack any obvious function, and our understanding of the cellular and biochemical roles of the majority of proteins is quite limited. As noted above, the determination of subcellular location of each gene product is expected to contribute information concerning the cellular function of many proteins. Another promising method includes the identification of proteins that interact with a protein of interest. The use of global two-hybrid interactions or affinity purification followed by mass spectroscopy in other organisms has provided considerable information about interacting partners. In other organisms such as yeast, these interaction maps often provide insight into the function of both new and previously characterized proteins (e.g. Uetz et al., 2000). The data can be even more robust when combined with other types of data, such as gene expression data (Xia et al., 2004). Such approaches applied to plants should be equally informative; they are most likely to be executed in organisms with relatively small genomes, such as Arabidopsis, in which such maps can be generated. Indeed, the large number of genes in many crop plants precludes a cost effective method for comprehensive analysis of protein-protein interactions.

One promising method to obtain biochemical functional information is to systematically express proteins in recombinant form and subject them to biochemical analyses. Recently, numerous advances in high-throughput protein production and microarray surface technologies have enabled the development of innovative formats for proteins ordered at high spatial density (Zhu et al., 2001). The microarray format provides a robust and convenient platform for the simultaneous analysis of thousands of individual protein samples, facilitating the design of sophisticated and reproducible biochemical experiments under highly specific conditions.

The principal challenges in protein array development are 3-fold: (1) creation of a comprehensive expression clone library; (2) high-throughput protein production, including expression, isolation, and purification; and (3) adaptation of DNA microarray technology to accommodate protein substrates. By far the greatest obstacle in developing functional protein microarrays is the construction of a comprehensive expression clone library from which a large number of distinct protein samples can be produced. Additionally, the substrate materials associated with conventional protein assays are not often compatible with robotic arrayers, cannot provide the sensitivity or dynamic range expected from microarray experiments, or contribute high fluorescence background, resulting in low signal-to-noise ratios. However, the practice of printing directly onto chemically treated glass surfaces is now in wider use for protein microarrays (Jona and Snyder, 2003).

The advent of protein-based microarrays allows the global observation of biochemical activities on an unprecedented scale, where hundreds or thousands of proteins can be simultaneously screened for protein-protein, protein-nucleic acid, and small-molecule interactions, as well as posttranslational modifications. In many cases, proteins are likely to have multiple cellular roles—the ability to systematically analyze them in a high-throughput fashion should reveal biochemical properties for proteins not previously appreciated. For example, in yeast a metabolic enzyme was also found to have a role in the regulation of gene expression (Hall et al., 2004). It is possible that many proteins will have multiple cellular roles.

INTEGRATION OF PROTEOMIC DATA FOR CONSTRUCTING COMPREHENSIVE REGULATORY NETWORKS

Protein interaction information can provide information as to the functions of proteins. Moreover, they can be used to build complex networks. In this form, individual proteins appear as nodes in a graph, and edges represent the physical relationships between proteins. Additionally, these data can be combined with other genetic information, using existing functional classification of proteins to reconstruct intricate biochemical pathways. These associations can be instrumental in mapping the relationship between individual proteins as well as their shared participation in subnetworks. For example, it has been shown that proteins that interact with many partners, and subsequently influence many biological processes, tend to be essential. When applying these techniques to an organism whose genetic and proteomic constituents are ill defined, it is useful to infer functionality through evolutionary conservation of gene sequences and protein domains between species.

In addition to assembling protein-protein interaction networks, many other types of information need to be obtained on a global scale. Global analysis of gene expression, transcription factor binding, and posttranslational modifications are active areas of interest in other eukaryotes. Acquisition of these data and combining them into a fully integrated network should provide a comprehensive analysis of

regulatory circuits. Such information is essential for improving plants for human benefit.

Arabidopsis has proven to be a valuable framework for plant biology, providing a bridge between different botanical research organisms. The adaptation of various analytical methodologies to the Arabidopsis model has facilitated the generalization and integration of findings from diverse research areas (Bevan and Walsh, 2004). As such, it is an excellent candidate for the development of new proteomic techniques and as a model for their eventual extension to important crop species. Through the combined application of emerging technologies to isolate, classify, and interrogate the gene products of model organism genomes, proteomics can be expected to have a wide-ranging impact on the future of plant research.

LITERATURE CITED

Agrawal GK, Yonekura M, Iwahashi Y, Iwahashi H, Rakwal R (2005) System, trends and perspectives of proteomics in dicot plants Part I: technologies in proteome establishment. J Chromatogr B **815:** 109–123

Bevan M, Walsh S (2004) Positioning Arabidopsis in plant biology. A key step toward unification of plant research. Plant Physiol **135:** 602–606

Froehlich JE, Wilkerson CG, Ray WK, McAndrew RS, Osteryoung KW, Gage DA, Phinney BS (2003) Proteomic study of the *Arabidopsis thaliana* chloroplastic envelope membrane utilizing alternatives to traditional two-dimensional electrophoresis. J Proteome Res **2:** 413–425

Hall DA, Zhu H, Royce T, Gerstein M, Snyder M (2004) Regulation of gene expression by a metabolic enzyme. Science **306:** 482–484

Jona G, Snyder M (2003) Recent developments in analytical and functional protein microarrays. Curr Opin Mol Ther **5:** 271–277

Nuhse TS, Stensballe A, Jensen ON, Peck SC (2004) Phosphoproteomics of the Arabidopsis plasma membrane and a new phosphorylation site database. Plant Cell **16:** 2394–2405

Schneider M, Tognolli M, Bairoch A (2004) The Swiss-Prot protein knowledgebase and ExPASy: providing the plant community with high quality proteomic data and tools. Plant Physiol Biochem **42:** 1013–1021

Seki M, Satou M, Sakurai T, Akiyama K, Iida K, Ishida J, Nakajima M, Enju A, Narusaka M, Fujita M, et al (2004) RIKEN *Arabidopsis* full-length (RAFL) cDNA and its applications for expression profiling under abiotic stress conditions. J Exp Bot **55:** 213–223

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, et al (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature **403:** 623–627

Xia Y, Yu H, Jansen R, Seringhaus M, Baxter S, Greenbaum D, Zhao H, Gerstein M (2004) Analyzing cellular biochemistry in terms of molecular networks. Annu Rev Biochem **73:** 1051–1087

Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, et al (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. Science **302:** 842–846

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, et al (2001) Global analysis of protein activities using proteome chips. Science **293:** 2101–2105