

# Functional Genomics: High-Throughput mRNA, Protein, and Metabolite Analyses

David J. Oliver,\* Basil Nikolau,<sup>†</sup> and Eve Syrkin Wurtele\*

\*Department of Botany and <sup>†</sup>Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, Iowa 50011

Received July 2, 2001; accepted September 27, 2001

A tremendous amount of DNA sequence information is now available to scientists and engineers. These DNA sequences provide the foundation for studying how the genome of an organism is functioning and they are particularly useful for metabolic engineers interested in manipulating plants for the production of chemicals and enzymes. Functional genomics relies on high-throughput techniques for measuring the mRNA (the transcriptome), protein (the proteome), and metabolite (the metabolome) components of plants as well as their organs and tissues. Microarray technologies, recent advances in protein mass spectrometry, and high-throughput metabolite analyses are beginning to provide detailed information on the total mRNA, protein, and metabolite components of plants. This knowledge will allow scientists to monitor changes in proteins and metabolites in plants. Ultimately, it may allow them to discover new metabolic pathways and to model metabolic and regulatory networks in plants. © 2002 Elsevier Science

## INTRODUCTION

During the last decade of the 20th century, basic biological research underwent a major revolution as life scientists switched their level of analyses from studying the expression of single genes and proteins to studying large numbers of genes and gene products simultaneously. During the first portion of the genomics era, researchers concentrated on accumulating DNA sequence information (both genomic and EST or expressed sequence tag data) from a range of economically important and model plants. While the most advanced work has been done on the small model dicot, *Arabidopsis thaliana*, public domain genomic efforts exist for many major crops, including rice, maize, soybeans, cotton, and sorghum (<http://ars-genome.cornell.edu> is an entry into this information). As genomics information has become available on a broad range of organisms, a new postgenomics era has arisen in which these data can be used as a resource base to characterize gene expression under different conditions. Because of the emphasis on gene function, this research is often referred to as functional genomics. The purpose of functional genomics is to use the

information made available upon sequencing a genome to quantitatively determine the spatial and temporal accumulation patterns of specific mRNAs, proteins, and important metabolites using high-throughput technologies.

## FUNCTIONAL GENOMICS FOR METABOLIC ENGINEERS

For metabolic engineers who are largely interested in using living organisms to produce proteins and metabolites for commercial purposes, functional genomics provides new tools and approaches for understanding, modeling, and ultimately manipulating plants. It provides a connection between the recent advances in genomics and attempts to use plants to accomplish a specific goal. In addition to producing plants with altered agricultural properties, these goals can include engineering plants or plant tissue culture systems that produce proteins, modified plant chemicals (particularly starches and lipids), or phytochemicals for pharmacological or industrial purposes. They can also contribute to producing plants to fill new functions such as phytoremediation of organic or heavy metal wastes.

Functional genomics relies heavily on three levels of high-throughput analyses: transcriptomics (or RNA profiling) for measuring levels of mRNA, proteomics for determining concentrations of individual proteins, and metabolomics (metabolite profiling) for determining the amounts of important metabolites. Tradeoffs exist among the analytical power of these systems in the amounts of data generated and the usefulness of the results obtained. These tradeoffs can readily be visualized in the construct of the central dogma of molecular biology, in which DNA can be transcribed into mRNAs, mRNAs translated into proteins, and proteins act to catalytically interconvert metabolites.

As you move through the central dogma from DNA to metabolites, the information becomes increasingly useful in terms of function. While DNA sequence indicates what genes are present in a plant, mRNA measurements show which of these genes are expressed. Similarly, protein measurements identify those specific mRNAs that are being

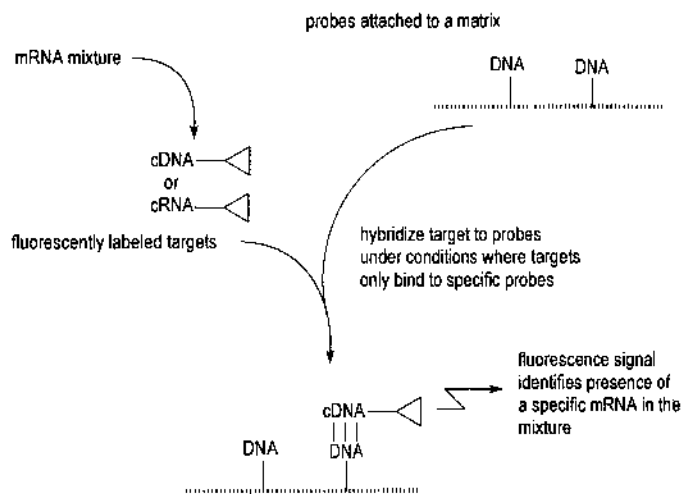
translated and the amount of the specific enzymes that are present. This may or may not be reflected in the mRNA level. Finally the amount of metabolite present (particularly if flux information can be deduced) may be more important than determining the potential for product formation as estimated by measuring enzyme levels. Unfortunately, while the information might become more valuable as we read through the central dogma, it becomes more difficult to obtain, often less quantitative, and certainly more fragmentary.

## RNA PROFILING

Analysis of the transcriptome in a given sample yields a measurement of the relative level of each mRNA. This level reflects the balance between the transcription rate of the gene and the rate at which that specific mRNA is being degraded. Several methods exist for determining the levels of cellular mRNAs; most utilize nucleic acid “probes” covalently bound to glass slides or “chips.” A major technology used is cDNA microarray, in which cDNAs (DNA copies of the mRNA population of a plant that are typically 200–2000 bases long) are used as probes (<http://www.cs.washington.edu/homes/jbuhler/research/array/>). Another predominant technology is Affymetrix chips (<http://www.affymetrix.com/>) on which cRNA oligonucleotides (25 bases long) are used as probes; the probes are designed specifically for each gene and chemically synthesized on the slide.

Thousands of gene probes can be represented on a single chip. mRNA is isolated from tissue samples and used as a template to prepare a “target,” i.e., cDNA (or cRNA for Affymetrix chips) that is labeled, usually with fluorescent dye. The labeled target is hybridized to the probe on the microarray. Individual cDNAs/cRNAs from the target hybridize (bind) with the corresponding probe proportionally to their representation in a sample (Fig. 1). A specialized scanner detects the hybridized molecules by fluorescence.

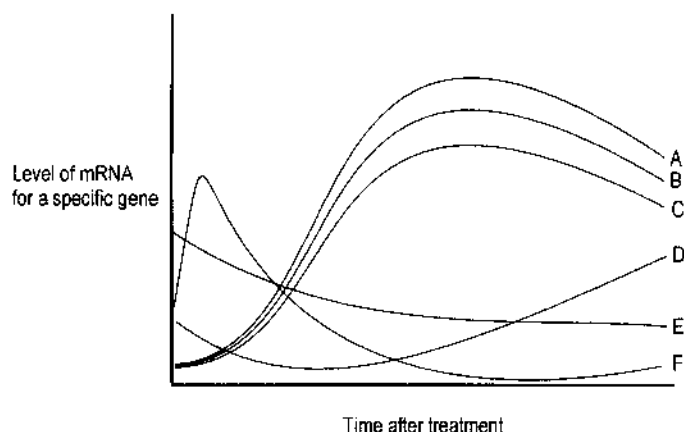
**Applications of this technology.** Microarray technology is expensive and time intensive both to set up and to use; however, the power of RNA profiling is immense. It can be used to analyze changes in gene expression during development, following chemical treatments or environmental stresses, or in different tissues. It can identify changes in gene activation in mutants compared to wild-type plants or between closely related cultivars. One particularly exciting application is gene discovery. Many metabolic pathways have not yet been fully elucidated, the number of the genes involved in these pathways are elusive, and their regulation



**FIG. 1.** Diagram of the mechanisms of mRNA microarrays. A mixture of mRNAs is extracted from a biological sample and copied into cDNA (or cRNA for Affymetrix chips) and labeled with a fluorescent tag. The DNA chip, a matrix (in this case a glass slide) spotted with specific DNA probes, is produced. The cDNA targets are incubated with the glass slide and the target cDNAs hybridize specifically to those probe DNAs that have exactly the complementary sequence. Fluorescence of a specific probe identifies the presence of a specific mRNA and the intensity of the fluorescence signal measures the amount of that mRNA.

is even less well understood. Genes whose expression patterns mirror one another over time and are associated with the same stimuli are candidates for being in the same pathway (Fig. 2). Likewise, regulatory genes might be expected to be up-regulated or down-regulated just prior to the onset of expression of genes encoding enzymes of that particular metabolic pathway (Fig. 2). Thus, microarrays provide a tool to identify or better understand the expression of candidate genes that function in metabolic pathways or as regulatory genes for that pathway (Devaux *et al.*, 2001).

As yet the power of transcriptome analysis to understand the function and regulation of plants is just beginning to be utilized. Such initial studies in plants characterize changes in mRNA accumulation during seed development (Girke *et al.*, 2000) and identify gene expression changes in a mutant (Helliwell *et al.*, 2001) and in response to pathogens (Dong, 2001) and oxidative stress (Desikan *et al.*, 2001). The technology is more mature in medical studies in which, in combination with other biological techniques, microarray technology is providing a new understanding of the basis of diseases, stresses, and genetic disorders, such as bipolar disorder (Niculescu and Kelsoe, 2001) and Alzheimer's disease (Hata *et al.*, 2001). Microarrays are being used clinically for drug discovery (Eyster and Lindahl, 2001) and disease diagnosis, including the staging of cancers (Li *et al.*, 2001). Identification of genes and even entire pathways that are involved in disease and medical processes like transplant



**FIG. 2.** Analysis of metabolic networks. How a small group of genes can be placed into a metabolic network by studying the time course of their expression using microarrays is illustrated. Genes A, B, and C show very similar expression patterns and this coregulation might indicate a common metabolic role. Genes D and E have expression patterns very different from those of genes A, B, and C and are much less likely to be part of the same metabolic system. Gene F has the type of expression pattern that might suggest it is a regulatory gene that controls the A, B, C cluster and must be activated before the genes it controls. While the data presented are in the form of a time course, the x axis could represent a group of mutant strains, different developmental states, or different drug treatments.

rejection (Xu *et al.*, 2001) and brain cell transplantation for Parkinson's disease (Beitner-Johnson *et al.*, 2001) have been undertaken with this methodology. Microarrays are used in conjunction with computational tools to identify potential metabolic or regulatory interconnections by visualizing expression patterns of an entire genome and determining associations between groups of genes. This approach has been pioneered in the model eukaryotic yeast, *Saccharomyces cerevisiae*, for which both the entire genome sequence and the total genome microarrays have been available for several years (Spellman *et al.*, 1998). In this system it is becoming possible to develop hypotheses as to regulatory and metabolic genes and networks that cannot be identified by single-gene methodologies (Devaux *et al.*, 2001; Natarajan *et al.*, 2001).

**Limitations of this technology.** Microarrays for an entire eukaryotic genome are currently available only for yeast, although partial genome chips for human, rat, mouse, and *Arabidopsis* are being expanded rapidly. RNA profiling technology is expensive and not always accessible. Affymetrix chips for some *Arabidopsis* ESTs are expensive but are available commercially. An Affymetrix chip for the whole *Arabidopsis* genome is expected in Fall 2001. cDNA microarrays must be made by individual research groups for most plants. Additionally, access to instrumentation facilities for chip analyses is necessary. Technical problems abound, including contamination of DNA in array target

spots, uneven hybridizations, and spurious hybridizations that necessitate multiple replications of experiments and multiple on-slide controls. Because many genes are present in gene families, cross-hybridization of probes with targets from several related genes can confuse the results. This can be decreased by using either 3' or 5' untranslated sequences as probes. Cross-hybridization is less of a problem with Affymetrix arrays, which use oligonucleotide probes designed to detect unique gene sequences. Many transcribed genes have not yet been identified, particularly for complex eukaryotes like plants. Thus microarrays may not contain probes representing the entire collection of genes in an organism. Also, because of problems with sensitivity, low-abundance transcripts (less than 0.001% of total mRNA) cannot be detected. Thus, current microarray technology still does not yield information about changes in the levels of all cellular mRNAs.

Multicellular organisms contain complex tissues with different cell types and the same cell types at different stages of development, thus complicating analysis of mRNA profiles. The problem can be reduced by emerging *in situ* technologies, such as laser capture, which enables selection of specific cells, followed by isolation and amplification of mRNA from these cells. Another option is to focus on providing homogeneous cell samples. For example, cells can be cultured and thus exposed to a more uniform environment. A drawback to this approach for plants is that cultured cells are in an artificial setting; thus, gene expression and communications between cells are altered. Also, despite being grown in culture, cells may not be completely synchronous. Alternatively, sample heterogeneity problems can be alleviated by using microarrays in combination with single-gene analytical approaches such as *in situ* hybridization, which detects specific transcripts in individual cells.

## PROTEOMICS

Proteomics, the simultaneous large-scale analysis of the protein component of an organism (Pandey and Mann, 2000), was initiated in the 1970s with the availability of two-dimensional (2-D) gel electrophoresis. In the first dimension proteins are separated by a nondenaturing technique, usually isoelectric focusing (IEF). In a sequential second dimension the proteins are further resolved by denaturing sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE). The usefulness of this technology, however, was originally limited by difficulty in getting reproducible 2-D gels and the lack of high-throughput technology for identifying the proteins visualized in the stained gels. The reproducibility problem has been somewhat solved with the availability of commercial immobilized gradient IEF gels,

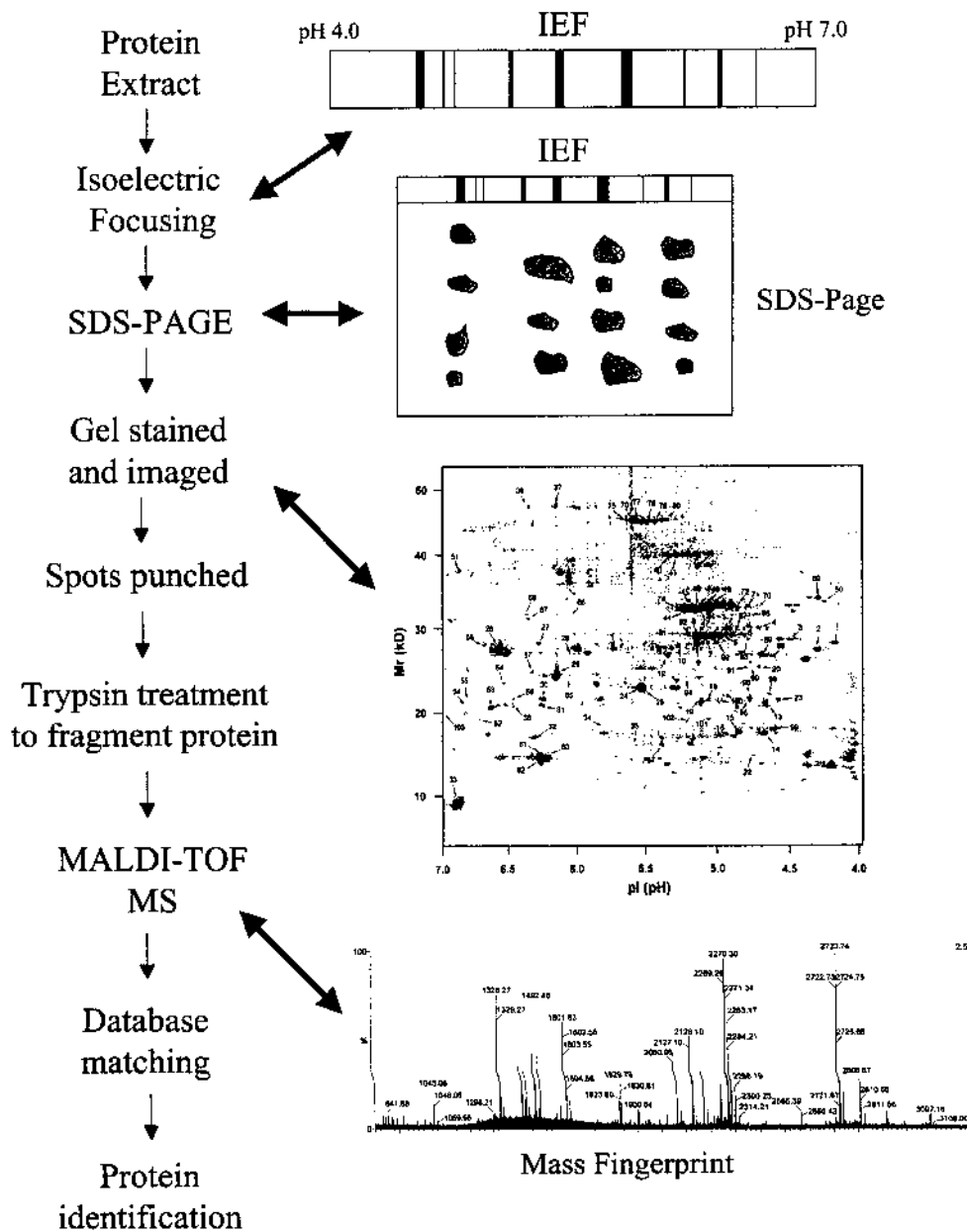


FIG. 3. Proteomic analysis using a combination of 2-D gel electrophoresis and MALDI-TOF mass spectrometry.

high-resolution gels that focus on small pH regions, and computer-controlled electrophoresis equipment.

The biggest technical advance in increasing the usability of proteomics has been in the use of mass spectrometry (MS) combined with genomic information to determine the identity of the proteins resolved by the 2-D gels. Excellent recent reviews on mass spectrometry are available (Blackstock and Mann, 2000; Mann *et al.*, 2001). The most accessible type of mass spectrometry for this work is matrix-assisted laser desorption ionization (MALDI-MS). Most MALDI mass spectrometers are time of flight

(MALDI-TOF) instruments. The peptides to be analyzed are cocrystallized on a metal substrate with a matrix, such as  $\alpha$ -cyano-4-hydroxycinnamic acid, and are subsequently ionized by short pulses with a nitrogen laser. The mass spectrometer then determines the molecular mass of each peptide fragment suspended in the matrix. To identify one of the proteins revealed by staining the 2-D gel with Coomassie blue or silver staining, the gel containing the protein spot is removed and incubated with a protease, usually trypsin (Fig. 3). After the digestion is complete, the protein fragments are analyzed by MALDI-MS. The

resulting peptide “mass fingerprint” is unique and diagnostic for the protein analyzed. The specific identity of an unknown protein spot can often be determined by searching protein or EST databases for sequences that upon *in silico* digestion with virtual trypsin would generate the same mass fingerprint. Because of the need to identify proteins by correlation of multiple fragments, MALDI-MS works best with single proteins or simple mixtures of a few proteins. It usually takes about 100 fmol of a peptide to ensure a good mass fingerprint by MALDI-MS.

**Applications of this technology.** The 2-D gel technology is fairly straightforward and can readily visualize, in a semiquantitative manner, about 1000 proteins from a sample. For bacteria or a virus, this means that nearly all soluble proteins can be visualized. It is also readily possible to visualize all of the major proteins present in the culture medium of a plant tissue culture. For plant tissues or organs, however, this may represent only the most abundant 10 or 20% of the water-soluble proteins. It is normally possible to visualize and record changes in any protein that represents 0.1% of the total protein present. Uses in plants include following changes in protein expression after environmental stresses such as anoxia (Chang *et al.*, 2000) or during normal development such as seed germination (Gallardo *et al.*, 2001). This technology has also been used to identify proteins in subcellular compartments like chloroplasts (vanWijk, 2000) and the plasma membrane (Santoni *et al.*, 1999) or in organs like maize leaves (Porubleva *et al.*, 2001).

One use that is very promising in metabolic engineering is to follow the accumulation of transgenic proteins in whole plants, individual organs, or tissue culture. Two-dimensional gels can be very useful for optimizing production of a protein. This technology also is excellent for designing a purification scheme for a protein because of the amount of information (i.e., size, relative amount, isoelectric point, presence in protein complexes) gathered for the protein of interest as well as the contaminants present.

Mass spectrometry has the potential for providing the most readily useable detailed molecular analysis of transgenic proteins. Comparison of the peptide mass fingerprint of the transgenic protein to that of the protein predicted by the transgene sequence can be used to address the following questions (Burlingame *et al.*, 1998):

1. Is the correct protein being produced or has some undetected mutation occurred? Contemporary MALDI-MS instruments can resolve the molecular mass of peptides to within a few parts per million and can readily detect most single amino acid substitutions.

2. Is some unexpected protease removing amino acids at the N- or C-termini? If the protein is targeted to a specific

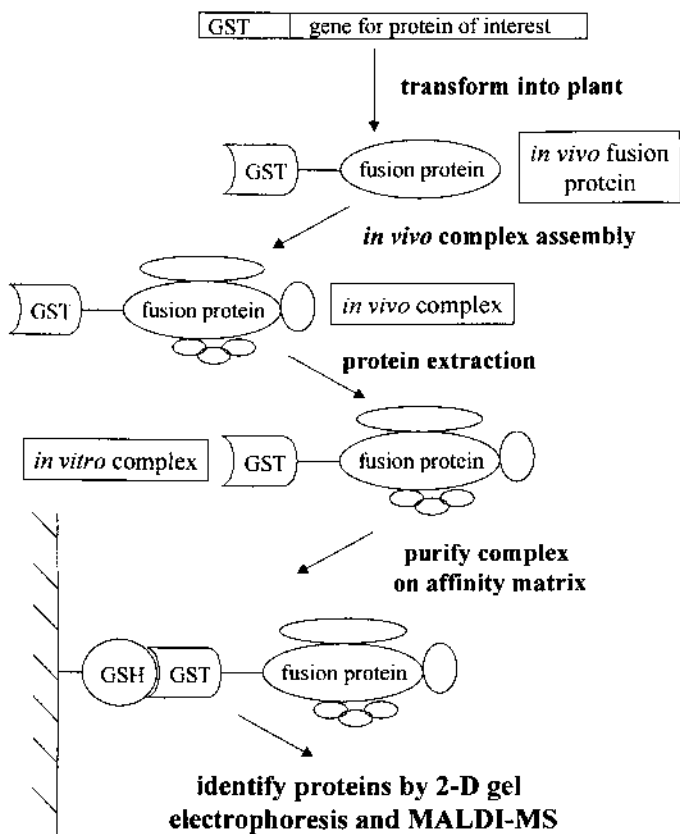
subcellular organelle, is the targeting sequence being processed properly?

3. Is the protein undergoing covalent posttranslational modifications? For example, it is possible to determine if one or more phosphorylations occur on any trypsin fragment and the percentage of the molecules that are phosphorylated. Glycosylation, myristylation, sulfonation, acetylation, etc., can also be followed by changes in the resulting mass of the peptides.

4. Have the correct disulfide bonds reformed? MALDI-MS analysis of peptides derived from proteins that have been partially reduced and before cysteine modification makes it possible to reconstruct the disulfide bond pattern of the protein. This also provides some indication that the transgenic protein is acquiring the correct three-dimensional structure.

5. Is this protein part of a multienzyme complex? Many proteins exist as part of more or less stable enzyme complexes. Two-dimensional gels and MALDI-MS often make it possible to separate and identify the proteins that interact with a transgenic enzyme after the whole complex has been removed from solution by immunoprecipitation or by means of a polyhistidine or GST tag on the transgenic protein (Fig. 4).

**Limitations of this technology.** Two-dimensional gel electrophoresis resolves many components from a single sample, but is optimized for processing only one or a few samples at a time and is difficult to convert to a fully automated or high-throughput technology. In addition, while it is readily possible to resolve over 1000 proteins via 2-D gels there is an upper limit to the number of spots that can be visualized and to the sensitivity of the staining techniques available. Two-dimensional gel techniques require proteins to be solubilized in a detergent that does not alter the isoelectric charge of the proteins, limiting use of the technique for integral membrane proteins. Two-dimensional gels have a limited dynamic detection range and a moderate sensitivity. MALDI, on the other hand, is not quantitative and, for reasons that are not well understood, identical amounts of peptides give very different sized peaks. Finally, while the mass fingerprints from MALDI-MS are highly diagnostic, they are dependent on extensive protein, EST, and genomic sequences for identification (van Wijk, 2001). In a recent study that we completed on the soluble proteins of maize leaves (an organism with a lot of public domain genomic information available) we were able to identify only about half of the proteins with good mass fingerprint data from the public databases (Porubleva *et al.*, 2001). Fortunately, recent advances in mass spectroscopy outlined below are providing answers to these problems.



**FIG. 4.** Use of proteomics to identify the components of an enzyme complex. A gene for a protein in a complex is fused to the gene for glutathione *S*-transferase (GST) and this fusion construct is transformed into plants. The GST-fusion protein is expressed in the plant and the complex assembles spontaneously. A crude protein extract is then made under conditions under which the enzyme complex is stable. This complex can be purified by using the affinity of GST for immobilized GSH. The protein can then be separated by 2-D gel electrophoresis and identified by MALDI-MS.

Electrospray ionization (ESI) technology can produce a fine mist of a protein solution that can be introduced into a mass spectrometer for mass determination of the proteins and peptides it contains. ESI is usually used on quadrupole mass spectrometers. By linking a liquid chromatograph to an electrospray MS it is possible to devise a system in which proteins are separated and analyzed completely in a liquid phase. Thus the problems of solubilizing proteins while maintaining a native conformation before IEF analyses are avoided. The resolution of the system is thereby limited by the resolution of the chromatography system instead of that of the 2-D gels. Similarly, the sensitivity of this system is limited by the sensitivity of the MS and not that of the gel stains. Finally, since no gel systems are involved, this system is more applicable to high-throughput and automation.

Emerging MS techniques provide amino acid sequence information for the identification of proteins that are not

available in the databases. In these tandem mass spectrometers, the first MS separates the peptides, which are then further fragmented and analyzed in the second MS. Deconvolution of these MS/MS spectra can allow reconstruction of the amino acid sequence of the original peptide. Since individual peptides are sequenced it is not necessary to link multiple peptides with a single protein as in MALDI-MS. It is therefore possible using a combination of liquid chromatography, electrospray ionization, and tandem mass spectrometry to analyze complex protein mixtures in a single pass. The peptides are separated by liquid chromatography and introduced into the MS by ESI. Select peptides from the first MS are then fragmented and sequenced by the second MS. Even with existing technology, it is possible to identify hundreds of proteins from a complex mixture in a single run. As recent technical advances continue in the application of MS techniques to high-throughput protein analyses, more MS facilities become available, and accessibility of this instrumentation increases for nonspecialists, these techniques will play an increasing role for academic and industrial biotechnology researchers.

## METABOLITE PROFILING

The ultimate level at which genomic information is expressed is in the form of metabolites. For metabolic engineers who seek to optimize the production of valuable biochemicals, this is often the level that is the target for manipulation. Indeed, most genetically, developmentally, or environmentally induced alterations are ultimately manifested as changes in the concentration of metabolites, either in quasi-steady-state levels of intermediates or in the final accumulation levels of terminal metabolites. Hence, one of the elements of metabolomics is the science of determining the concentrations of metabolites in the tissue of interest at a given time (i.e., metabolite profiling). Metabolite profiling provides a snapshot of the chemical composition of that tissue. By comparing metabolite profiles between two tissue states, separated either in time or in space or by genetic variation, differences in genome functionality can be assessed.

Another element of metabolomics is identification of the biochemical and/or genetic mechanisms that regulate the flux through metabolic pathways. This issue is critical to metabolic engineers who have the goal of optimizing the concentration of a specific metabolite or end-product of metabolism (for example, optimizing the accumulation of biomedicines in plant tissue culture cells). Thus, knowing how the flux through a metabolic pathway is regulated provides metabolic engineers with insights as to how to modulate that flux in order to optimize the concentration of the metabolite that are derived from that pathway.

Although metabolite profiling cannot directly determine flux through a pathway, it may provide indirect evidence of changes in flux regulation among different tissue states. For example, metabolite profiling could reveal changes in flux through a pathway if the change in flux alters the quasi-steady-state concentrations of all or a subset of the intermediates of that pathway.

The recent developments in functional genomics have propelled the need for global profiling of gene expression at the level of metabolites. In addition, recent technological advances have enhanced the sensitivity of analytical capabilities, which make metabolite profiling a powerful tool in functional genomics. These advances have led to the development of technologies that allow for the unbiased determination of concentrations of metabolites. The analytical technologies that are being utilized combine chromatographic procedures for separating metabolites, based upon their physical and chemical properties, coupled with mass spectral-based identification of each metabolite. A variety of chromatographic methods can be used for separation purposes, including gas-liquid chromatography, liquid chromatography (LC), or capillary electrophoresis (CE). Each of these chromatographic methods provides unique capabilities to separate different chemical classes of metabolites. The mass spectrometric-based identification of metabolites has expanded with the development of different ionization capabilities. In particular, the advent of electrospray ionization and matrix-assisted laser desorption ionization are the two biggest success stories in this regard, making possible the coupling of LC and CE separation methods to mass spectrometry. Another analytical tool that is also being utilized in the analysis of plant metabolism is NMR. Although this technology is not as sensitive as the chromatographic-based methods for profiling metabolites, NMR holds the promise of elucidating the regulation of metabolism and possibly imaging of metabolites in intact tissues. For a recent review of the status of NMR-based studies of plant metabolism, see Ratchiffe and Shachar-Hill (2001).

The major limitation to applying these technologies to functional genomics is the need for high-throughput analysis of large numbers of samples. The absence of high-throughput technologies for metabolite profiling is primarily due to the excessive cost associated with conducting mass spectral analyses in parallel. This limitation may be overcome in the near future by developments in the use of parallel CE separation methodologies. Recent advances made in response to the need for high-throughput DNA sequencing for the human genome project have led to the development of CE instrumentation that can separate, in parallel, up to 96 samples. If this technology can be coupled with either pre- or postseparation chemistries that can fluorescently tag metabolites, such instrumentation will provide for high-throughput and highly sensitive methodologies for metabolite profiling of large numbers of samples.

Metabolite profiling is a relatively new branch of functional genomics and has been used primarily in the biomedical area (Duez *et al.*, 1996; Gopaul *et al.*, 2000; Nicholson *et al.*, 2000; Flurer, 1999; Beaudry *et al.*, 1999; Hempel *et al.*, 1999; Lim *et al.*, 1999). However, metabolite profiling technologies are now being applied to plant systems (e.g., Fiehn *et al.*, 2000; Roessner *et al.*, 2001). For example, comparing the metabolite profiles of wild-type potato tubers to those of transgenic tubers that are altered in the expression of a single sugar-metabolizing gene has provided valuable insights into the regulation of the complexity of plant metabolism (Roessner *et al.*, 2001). Our own work has focused on metabolite profiling of the cuticular waxes of a large collection of plants (Schnable *et al.*, 1994) that carry single mutations that affect the normal accumulation of cuticular waxes on the surface of seedlings. By comparing the composition of the cuticular waxes associated with each mutant plant to the cuticular waxes of wild-type siblings, we are obtaining insights as to the metabolic function that is associated with each of the mutant genes.

## METABOLIC MODELING

High-throughput mRNA, protein, and metabolite technologies generate large amounts of data. A major and exciting challenge of the future is how to extract the useful information from these data sets and combine it together with what we already know about pathways and their regulation, to achieve a better understanding of how metabolism is regulated. This effort has many aspects. Computational schemes must be developed for capturing and integrating RNA, protein, and metabolite data (Dickerson *et al.*, 2001; Stoeckert *et al.*, 2001; Tomita, 2001). Hierarchical text searching methods are needed to mine the literature. Models need to be built that integrate metabolic and regulatory reactions and the factors that control them. A common ontology needs to be developed (Gene Ontology Consortium, 2000; [http://www.cbil.upenn.edu/Ontology/MGED\\_ontology.html#ontology](http://www.cbil.upenn.edu/Ontology/MGED_ontology.html#ontology)) to facilitate universal organization and understanding of information in databases; a plant ontology is still in its infancy. Methods need to be worked out for clustering global profiling data and for visualizing the complex information that results. All these are current areas of research.

Modeling approaches that have been used for genetic networks, include Boolean networks (Liang *et al.*, 1998; Akutsu *et al.*, 1999), linear weighting networks (Weaver *et al.*, 1999), differential equations (Akutsu *et al.*, 2000), and Petri nets (Alla and David, 1998). Modeling techniques for enzymatic reactions include differential equations and metabolic flux analysis methods (Edwards *et al.*, 2001;

Morgan and Rhodes, this volume). The integration of RNA, protein, and metabolite data requires yet other approaches. In plant and other eukaryotic cells, the sub-cellular compartmentation of metabolism must be considered in the model. Plants regulate metabolism at many levels and maps can be constructed of pathways for various levels of regulation (Dickerson *et al.*, 2001). Several algorithms can be used to represent different types of relationships, for example, enzymatic reactions, in which substrates are consumed as products are made, compared to regulatory reactions, in which the regulatory molecule is not consumed but can act repeatedly or continuously. Fuzzy methods for modeling metabolic and regulatory networks can be superimposed on combined metabolic and regulatory maps (Dickerson *et al.*, 2001).

To mine global data, clustering tools are needed to identify relationships and groupings within and among gene products. While a variety of clustering programs have been developed for microarray data (e.g., <http://genome-www4.stanford.edu/MicroArray/SMD/restech.html>), research is required to determine which methods and algorithms are most appropriate to gain maximal information from the expression data under particular conditions (Lukashin and Fuchs, 2001). These tools have not yet begun to address the concept of clustering RNA, protein, and metabolite data together.

Finally, in order to maximize their usefulness to scientists and engineers, clusters and metabolic and regulatory maps need to be visualized. This entails viewing multidimensional data on a two-dimensional computer screen. One interesting approach is data touring (<http://www.ggobi.org/>). Multiple levels of data made up of information from the literature and global expression data could be visualized by mapping the gene expression and clustering data onto metabolic and regulatory network maps. These maps can then be linked to literature searches (Dickerson *et al.*, 2001; Ding *et al.*, 2002). It is critical that these systems combine power and flexibility for the scientist to view, manipulate, and model different areas.

Since the 1930s, biochemists have identified many of the basic plant metabolic processes. In the past 15 years, molecular biology has provided ever more powerful tools for discovering and isolating genes and for manipulating plant metabolism. Genomic technologies have now revolutionized the paradigms by which we study metabolism. These technologies offer the promise for us to understand and alter metabolism to engineer plants to produce useful proteins and chemicals. Visualizing and modeling global expression data in new ways may provide new insights into the function of metabolism and ultimately expand our ability to more precisely direct metabolism through metabolic engineering.

## REFERENCES

- Alla, H., and David, R. (1998). Continuous and hybrid Petri nets. *J. Circuits Syst. Comput.* 8, 159–188.
- Akutsu, T., Miyano, S., and Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Presented at Pacific Symposium on Biocomputing 4, Hawaii, 1999.
- Akutsu, T., Miyano, S., and Kuhara, S. (2000). Algorithms for inferring qualitative models of biological networks. Presented at Pacific Symposium on Biocomputing 5, Hawaii, 2000.
- Beaudry, F., Yves Le Blanc, J. C., Coutu, M., Ramier, I., Morcau, J. P., and Brown, N. K. (1999). Metabolite profiling study of propranolol in rat using LC/MS/MS analysis. *Biomed. Chromatogr.* 13, 363–369.
- Beitner-Johnson, D., Seta, K., Yuan, Y., Kim, H., Rust, R. T., Conrad, P. W., Kobayashi, S., and Millhorn, D. E. (2001). Identification of hypoxia-responsive genes in a dopaminergic cell line by subtractive cDNA libraries and microarray analysis. *Parkinsonism Relat. Disord.* 7, 273–281.
- Blackstock, W., and Mann, M., Eds. (2000). "Proteomics: A Trends Guide," Elsevier, Amsterdam.
- Burlingame, A. L., Boyd, R. K., and Gaskell, S. J. (1998). Mass spectrometry. *Anal. Chem.* 70, R647–R716.
- Chang, W. W., Huang, L., Shen, M., Webster, C., Burlingame, A. L., and Roberts, J. K. (2000). Patterns of protein synthesis and tolerance of anoxia in root tips of maize seedlings acclimated to a low-oxygen environment, and identification of proteins by mass spectrometry. *Plant Physiol.* 122, 295–318.
- Desikan, R., Mackerness, S. A., Hancock, J. T., and Neill, S. J. (2001). Regulation of the Arabidopsis transcriptome by oxidative stress. *Plant Physiol.* 127, 159–172.
- Devaux, F., Marc, P., and Jacq, C. (2001). Transcriptomes, transcription activators and microarrays. *FEBS Lett.* 498, 140–144.
- Dickerson, J. A., Bericant, D., Cox, Z., Qi, W., and Wurtele, E. (2001). Creating metabolic network models using text mining and expert knowledge. Presented at the Atlantic Symposium on Molecular Biology and Genome Information Systems and Technology (CBGIST 2001), Durham, North Carolina, 2001.
- Ding, J., Berleant, D., Nettleton, D., and Wurtele, E. (2002). Mining Medline: Abstracts, sentences, or phrases? In "Pacific Symposium on Biocomputing," pp. 1–12, in press.
- Dong, X. (2001). Genetic dissection of systemic acquired resistance. *Curr. Opin. Plant Biol.* 4, 309–314.
- Duez, P., Kumps, A., and Mardens, Y. (1996). GC-MS profiling of urinary organic acids evaluated as a quantitative method. *Clin Chem.* 42, 1609–1615.
- Edwards, J. S., Ibarra, R. I., and Palsson, B. O. (2001). In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* 19, 125–130.
- Eyster, K. M., and Lindahl, R. (2001). Molecular medicine: A primer for clinicians. Part XII. DNA microarrays and their application to clinical medicine. *S. D. J. Med.* 54, 57–61.
- Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R. N., and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* 18, 1157–1161.
- Flurer, C. L. (1999). Analysis of antibiotics by capillary electrophoresis. *Electrophoresis* 20, 3269–3279.
- Gallardo, K., Job, C., Groot, S. P. C., Puype, M., Demol, H., Vandekerckhove, J., and Job, D. (2001). Proteomic analysis of Arabidopsis seed germination and priming. *Plant Physiol.* 126, 835–848.

- Gene Ontology Consortium (2000). Gene ontology: Tool for the unification of biology. *Nat. Genet.* 25, 25–29.
- Girke, T., Todd, J., Ruuska, S., White, J., Benning, C., and Ohlrogge, J. (2000). Microarray analysis of developing Arabidopsis seeds. *Plant Physiol.* 124, 1570–1581.
- Gopaul, S. V., Farrell, K., and Abbott, F. S. (2000). Gas chromatography/negative ion chemical ionization mass spectrometry and liquid chromatography/electrospray ionization tandem mass spectrometry quantitative profiling of *N*-acetylcysteine conjugates of valproic acid in urine. *J. Mass Spectrom.* 35, 698–704.
- Hata, R., Masumura, M., Akatsu, H., Li, F., Fujita, H., Nagai, Y., Yamamoto, T., Okada, H., Kosaka, K., Sakanaka, M., and Sawada, T. (2001). Up-regulation of *Calcineurin abeta* mRNA in the Alzheimer's disease brain: Assessment by cDNA microarray. *Biochem. Biophys. Res. Commun.* 284, 310–316.
- Helliwell, C. A., Chin-Atkins, A. N., Wilson, I. W., Chapple, R., Dennis, E. S., and Chaudhury, A. (2001). The Arabidopsis *ampl* gene encodes a putative glutamate carboxypeptidase. *Plant Cell* 13, 2115–2125.
- Hempel, R., Schupke, H., McNeilly, P. J., Heinecke, K., Kronbach, C., Grunwald, C., Zimmermann, G., Griesinger, C., Engel, J., and Kronbach, T. (1999). Metabolism of retigabine (D-23129), a novel anticonvulsant. *Drug Metab. Dispos.* 27, 613–622.
- Li, S., Ross, D. T., Kadin, M. E., Brown, P. O., and Wasik, M. A. (2001). Comparative genome-scale analysis of gene expression profiles in T cell lymphoma cells during malignant progression using a complementary DNA microarray. *Am. J. Pathol.* 58, 1231–1237.
- Liang, S., Fuhrman, S., and Somogyi, R. (1998). REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Presented at Pacific Symposium on Biocomputing 3, Hawaii, 1998.
- Lim, H. K., Stellingweif, S., Sisenwine, S., and Chan, K. W. (1999). Rapid drug metabolite profiling using fast liquid chromatography, automated multiple-stage mass spectrometry and receptor-binding. *J. Chromatogr. A* 831, 227–241.
- Lukashin, A. V., and Fuchs, R. (2001). Analysis of temporal gene expression profiles: Clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics* 17, 405–414.
- Mann, M., Hendrickson, R. C., and Pandey, A. (2001). Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* 70, 437–473.
- Natarajan, K., Meyer, M. R., Jackson, B. M., Slade, D., Roberts, C., Hinnebusch, A. G., and Marton, M. J. (2001). Transcriptional profiling shows that *Gen4p* is a master regulator of gene expression during amino acid starvation in yeast. *Mol. Cell. Biol.* 13, 4347–4368.
- Nicholson, J. K., Lindon, J. C., Scarfe, G., Wilson, I. D., Abou-Shakra, F., Castro-Perez, J., Eaton, A., and Preece, S. (2000). High-performance liquid chromatography and inductively coupled plasma mass spectrometry (HPLC-ICP-MS) for the analysis of xenobiotic metabolites in rat urine: Application to the metabolites of 4-bromoaniline. *Analyst* 125, 235–236.
- Niculescu, A. B., 3rd, and Kelsoc, J. R. (2001). Convergent functional genomics: Application to bipolar disorder. *Ann. Med.* 33, 263–271.
- Pandey, A., and Mann, M. (2000). Proteomics to study genes and genomes. *Nature* 405, 837–846.
- Porubleva, L., Vander Velden, K., Kothari, S., Oliver, D. J., and Chitnis, P. R. (2001). The proteome of maize leaves: Use of gene sequences and EST data for identification of proteins with peptide mass fingerprints. *Electrophoresis* 22, 1724–1738.
- Ratcliffe, R. G., and Shachar-Hill, Y. (2001). Probing plant metabolism with NMR. *Annu. Rev. Plant Physiol.* 52, 499–526.
- Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L., and Fernie, A. (2001). Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* 13, 11–29.
- Santoni, V., Rabilloud, T., Doumas, P., Rouquie, D., Mansion, M., Kieffer, S., Garin, J., and Rossignol, M. (1999). Towards the recovery of hydrophobic proteins on two-dimensional electrophoresis gels. *Electrophoresis* 20, 705–711.
- Schnable, P. S., Stinard, P. S., Wen, T.-J., Heinen, S., Weber, D., Zhang, L., Hansen, J. D., and Nikolau, B. J. (1994). The genetics of cuticular wax biosynthesis. *Maydica* 39, 279–287.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9, 3273–3297.
- Stoeckert, C., Pizarro, A., Manduchi, E., Gibson, M., Brunk, B., Crabtree, J., Schug, J., Shen-Orr, S., and Overton, G. C. (2001). A relational schema for both array-based and SAGE gene expression experiments. *Bioinformatics* 17, 300–308.
- Tomita, M. (2001). Whole-cell simulation: A grand challenge of the 21st century. *Biotechnology* 19, 205–210.
- van Wijk, K. J. (2000). Proteomics of the chloroplast: Experimentation and prediction. *Trends Plant Sci.* 5, 420–425.
- van Wijk, K. J. (2001). Challenges and prospects of plant proteomics. *Plant Physiol.* 126, 501–508.
- Weaver, D. C., Workman, C. T., and Stormo, G. D. (1999). Modeling regulatory networks with weight matrices. Presented at Pacific Symposium on Biocomputing 4, Hawaii, 1999.
- Xu, B., Sakkas, L. I., Slachta, C. A., Goldman, B. I., Jeevanandam, V., Oleszak, E. L., and Platsoucas, C. D. (2001). Apoptosis in chronic rejection of human cardiac allografts. *Transplantation* 71, 1137–1146.