

# Proteomics: Theoretical and Experimental Considerations

Vassily Hatzimanikatis,<sup>‡</sup> Leila H. Choe,<sup>†</sup> and Kelvin H. Lee<sup>\*,†</sup>

School of Chemical Engineering, Cornell University, Ithaca, New York 14853-5201, and  
Biosciences Division, Wayzata, Minnesota 55391-2397

---

Cellular engineering relies on the ability to decipher the genetic basis of various phenotypes. Emerging technologies for analyzing the biological function of the information encoded in the genome of particular organisms and/or tissues focus on the monitoring of transcription (mRNA) and translation (protein) processes. Elementary theoretical considerations presented in this article strongly suggest that a combination of mRNA and protein expression patterns should be simultaneously considered to fully develop a conceptual understanding of the functional architecture of genomes and gene networks. We propose a framework of experimental and mathematical methods for acquiring and analyzing quantitative proteomic information and discuss recent developments in proteome analytical technology.

---

## Introduction

The field of proteomics (the simultaneous analysis of total gene expression at the protein level) represents one of the premiere strategies for studying biological systems and understanding the relationship between various expressed genes and gene products. One of the key issues in “functional genomics” is to relate linear sequence information to nonlinear cellular dynamics. Toward this end, significant scientific effort and resources are directed at mRNA expression monitoring methods and analysis. At the same time, while proteomics has become one of the important tools for functional genomics studies, there are relatively few examples of the effective application of proteomics technology to the solution of problems of biological relevance and even fewer research laboratories focused on improving the basic protein separation technologies which often limit proteome studies. It is in these two areas of proteomics technology development and application that biochemical engineers are well-suited to make an impact. Chemical engineers have a history in understanding separation processes at the chemical and biochemical levels to tune such technology for various applications. Moreover, biochemical engineers often take an integrated systems approach to engineering cells and tissues while tackling difficult problems of biological and medical relevance. At the heart of this perspective is the study of the system as a whole rather than the detailed study of individual components and their direct interactions.

The technology for performing proteome studies has been available for almost 25 years, even though the term proteomics is relatively new (1). At the core of proteomics is two-dimensional protein electrophoresis (2DE) (2). This technique is capable of resolving up to 11 200 proteins and peptides from a single, complex mixture in a single experiment (3). The tremendous resolving capability of 2DE has been one critical factor for its utility in functional genomics studies. Another important consideration is that the same technique can be scaled up and per-

formed with micropreparative amounts (5–10 mg) of protein and coupled to other analytical techniques such as mass spectrometry and amino acid sequencing. This ability to characterize the genetic basis for protein spot changes of interest provides a vital link between gene information at the DNA level (genomics) and observed phenotypic differences at the system level (functional genomics). Accordingly, proteomics promises to be an increasingly important field as more genomes are sequenced.

This paper is divided into two main sections: theoretical and experimental considerations of proteomics. In the first section, we consider the development of conceptual and mathematical models for biological phenomena which rely on genomic and proteomic information. The often employed Boolean modeling framework is described and contrasted with continuous mechanistic modeling approaches. The analysis presented suggests that both mRNA expression information and protein expression information are required to develop meaningful descriptions of gene function and regulation (4). Given a need to obtain experimental data on mRNA and protein expression, we consider in the second section the realm of experimental technology available to study proteomes, including standardized experimental methods, existing limitations, and future developments. For a discussion of the recent developments of microarray technology for mRNA expression, we refer the reader to other papers (5–7). In this second section we show that the expression of even a single reporter gene in *Escherichia coli* can have a tremendous impact on total gene expression at the protein level, and we also present a framework for obtaining and using experimentally derived information about gene expression.

## Theoretical Considerations

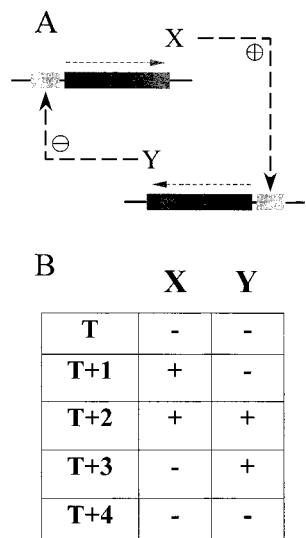
Powerful new technologies now permit the simultaneous identification and quantification of the expression of every gene in the genome or sets of gene by hybridization on a solid surface (5–7). An important motivation for the development of such mRNA-based technology is the hope that, by comparing complex quantitative gene expression patterns, one might be able to decipher the

---

\* Corresponding author.

† Cornell University.

‡ Biosciences Division.



**Figure 1.** Genetic network consisting of two genes, X and Y. (A) The molecular model where the product of gene X induces the expression of gene Y, and the product of gene Y represses the expression of gene X. (B) State-transition table for the Boolean model of the genetic network.

regulatory wiring of the genetic networks responsible for various observed phenotypes, including developmental stages and disease (8). The enormous amount of data generated from these studies requires mathematical treatment and processing to extract relevant information and to draw meaningful conclusions. For example, statistical analysis of gene and protein expression information is currently used to cluster genes and proteins that share regulatory properties that help to determine phenotypic responses (9). However, a byproduct of the nonlinearity of many biological phenomena is that the use of these methods to identify such "coregulated" genes and proteins does not lead directly to conclusions about the regulatory wiring nor to appropriate genotype-phenotype mapping. Further, these approaches do not make full use of the laboriously derived information about the quantitative levels of mRNA and protein expression.

Early attempts to formulate mechanistic models of genetic networks employed Boolean models (10). Boolean modeling is based on logical rules where each gene is modeled as on or off, and this framework is consistent with human conceptual rationalization, which is also based on logic. The low computational complexity of Boolean models makes them attractive for describing large genetic networks. As a result, these concepts have been used to describe diverse biological phenomena such as circadian rhythms (11) and the lactose operon (12).

To illustrate the application of Boolean modeling, we will consider a simple genetic network consisting of two genes, X and Y. We will assume that the product of gene X directly induces the expression of gene Y, whose product directly represses the expression of gene X (Figure 1A). Boolean modeling of this system suggests that the unique steady state of the system will exhibit some oscillatory behavior (Figure 1B), which corresponds to biologically meaningful solutions (e.g., circadian rhythms). It is this simplicity and conceptual consistency in understanding that makes Boolean models appealing and justifies their wide application in the arena of functional genomics—particularly in the attempt to interpret mRNA expression information.

Unfortunately, the Boolean modeling approach suffers from a number of limitations. Information about gene

expression levels is not included when genes are modeled as only on or off. Thus, Boolean modeling fails to capture relevant physics. For example, this approach does not consider mRNA degradation, so phenomena such as RNA stability are not included in such models. Furthermore, the discrete, equally spaced time steps ignore differences in transcription rates which are expected from differing gene sequence lengths or differences in transcription efficiency. Certainly, the availability of a Boolean model for a genetic network does provide an approximate *in silico* system for hypothesis testing, redesigning genetic regulatory networks, knock-out studies, and steady-state analyses; however, the lack of quantification and time continuity does not permit the interpretation of and guidance for effective studies on dynamic systems, including exogenously controlled gene expression, drug dosage studies, and gene therapy approaches. Further, it has been shown that the application of Boolean modeling to reverse engineering algorithms (13) can lead to the identification of multiple ambiguous regulatory structures where one gene is identified as an inducer in one structure and as a repressor in another structure.

An alternate approach to overcome some of the limitations inherent in Boolean modeling is the formulation of continuous models. Using the same physical description as that presented above (Figure 1A), a continuous model can be constructed (4). Such a model accounts for three important aspects of the system: quantitative expression information, mRNA stability, and the real time evolution of the system. However, nonlinear stability analysis techniques can be used to prove that there are no oscillatory solutions to a continuous model of the two-component system depicted in Figure 1A. Similar analysis proves that there are, however, oscillatory solutions to continuous models of this genetic network if the network is expanded to include a description of translational processes (4). The appropriate physical description involves the product of gene X which is an mRNA, which results in the formation of a protein which induces gene Y, which results in the formation of an mRNA, which results in a protein which represses gene X. This simple example highlights the intellectual pitfalls associated with efforts toward modeling of complex multicomponent systems based solely on mRNA expression data. That is, mRNA and protein expression information must be collected together to develop a deep understanding of biological phenomena. This framework further suggests that current efforts to develop such an understanding of gene networks based on data from microarray experiments alone are incomplete (4). Indeed, the need for such a combination is highlighted by observations with yeast, in particular, where the quantitative expression level of many genes as measured by mRNA analysis is significantly different from that measured with a proteomics strategy (14).

## Experimental Considerations

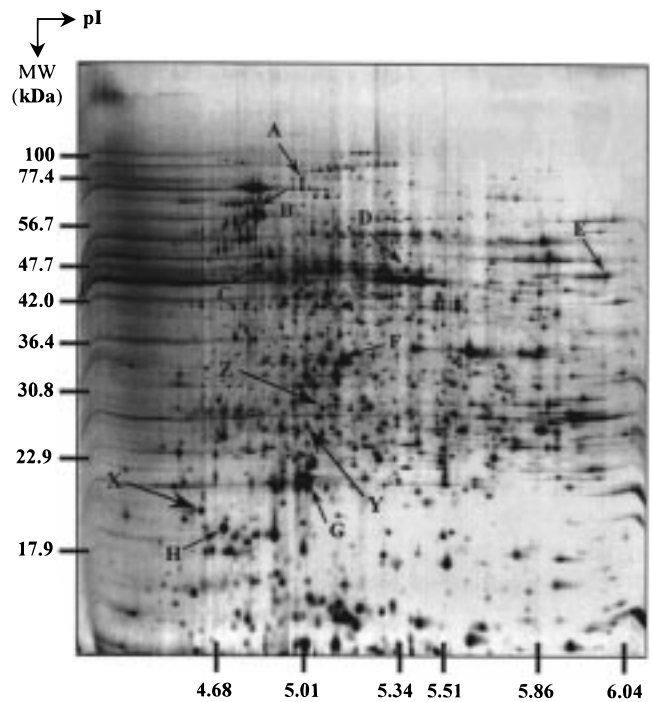
In the preceding section we demonstrated a need to obtain both nucleic acid-based gene expression information using microarray technology and protein-based gene expression information using proteomics technology to elucidate the genetic circuits that underlie dynamic biological phenomena. In this section we focus on experimental aspects of obtaining this information. In particular, the discussion centers on proteomics because nucleic acid-based studies (DNA sequencing, microarray technology, etc.) are well-described elsewhere (5–7). The generation of data about nucleic acids is aided by polymerase chain reaction technology, while no such amplifying

technology exists for proteins; thus, the available proteome analytical technology and experimental techniques limit the success of proteome experiments. Even so, it will be shown with a simple example that existing technology can provide detailed information on system responses in gene expression to simple perturbations. In particular, we will consider the use of reporter genes as a monitor of metabolism. Finally, we will consider impending developments in the field of proteomics.

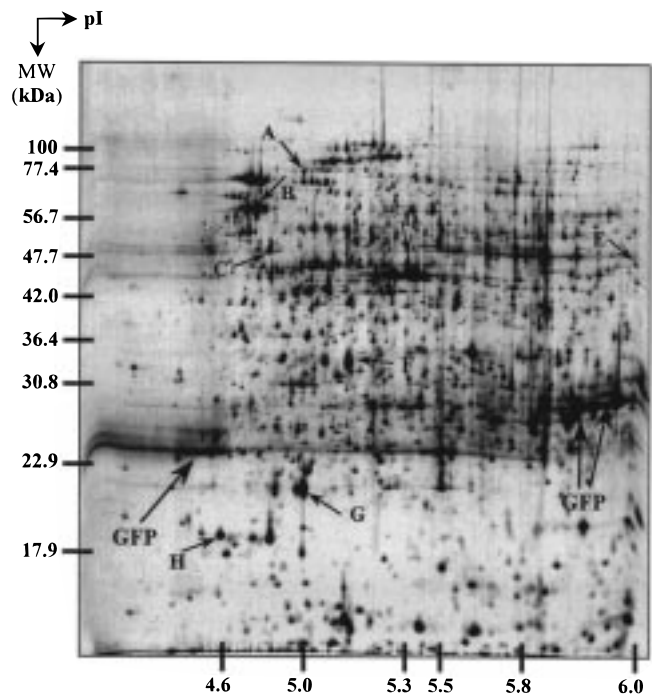
**Proteome Analysis of GFP-Expressing Cells.** One of the recently sequenced genomes is from *E. coli* (15). Because the genetics and metabolism of *E. coli* are relatively well understood, it is often genetically manipulated and used in bioprocess applications, and reporter genes are often used as a monitor for metabolism. Proteome analysis of the response of the host cell genetic network to the expression of a single reporter gene encoded on a plasmid should reveal changes in total gene expression caused by the expression of that particular protein as well as changes in total gene expression in response to any biological function that reporter protein may have in the host cell. Consider, for example, the expression of a single reporter gene, green fluorescent protein (GFP), in *E. coli* JM105 using pBluescript and grown to  $OD_{600} = 1.0$ . The simplest prediction of the effect of this gene network perturbation would include the same number of total genes being expressed as in untransformed control cells, plus the cloned gene product and other marker genes present on the expression vector. Taken further, one can predict that some small but significant number of peripherally related gene products will also be affected. It has been previously shown (16, 17) that the simple insertion of expression vectors into cells can affect numerous host cell proteins including enzymes involved in carbon metabolism, protein synthesis, and protein translation. Here we see that an even greater variety of host cell proteins are affected, suggesting that simple manipulations can have a dramatic effect on gene expression.

Figure 2 depicts part of the proteome of *E. coli* JM105 cells ("JM105 cells"), and Figure 3 depicts the same part of the proteome of GFP-expressing *E. coli* JM105 cells ("GFP cells"). The portion of the proteome depicted includes proteins and peptides from pH 4 to 6. The location of GFP, as determined by Western analysis, is shown in Figure 3, and GFP represents approximately 14.9% of total intracellular protein expression for these cells. The location of several landmark proteins (details in the figure captions) are indicated in Figures 2 and 3 and were identified using a combination of the approaches described in Protocols and Procedures, which is presented after this section.

Even within this narrow pH window of total intracellular proteins, 44 protein spots are detected in JM105 cells but not in GFP cells. Similarly, 68 protein spots are present in the pH 4–7 proteome of GFP cells but not in JM105 cells to any significant extent. There are many quantitative changes as well. Among the proteins expressed in JM105 cells but not in GFP cells are sulfate starvation-induced protein 7, histidine-binding periplasmic protein, and glucose-permease IIA component (as labeled in Figure 2), and also ribosome releasing factor and phosphoglycerate kinase. Proteins expressed in GFP cells but not JM105 cells include 50S ribosomal protein L9, serine methylase, aerobic glycerol-3-phosphate dehydrogenase, and the amino acid ABC transporter binding protein in GltJ–CutE intergenic region. The ability to define with certainty the genetic basis for individual protein spots on 2DE gels is aided by recent



**Figure 2.** Silver-stained *E. coli* JM105 proteome separated on a pH 4–7 nonlinear gradient and 12%T SDS–PAGE system. Several reference proteins are noted and include (A) E2 component of pyruvate dehydrogenase complex, (B) GroEL protein, (C) ATP synthase  $\beta$  chain, (D) enolase, (E) serine methylase, (F) elongation factor TS, (G) inorganic pyrophosphatase, and (H) glucose-permease IIA component. Three of the many qualitative changes relative to GFP cells are also noted: (X) glucose-permease IIA component, (Y) sulfate starvation-induced protein 7, and (Z) histidine-binding periplasmic protein.



**Figure 3.** Silver-stained *E. coli* JM105 expressing GFP proteome separated on a pH 4–7 gradient and 12%T SDS–PAGE system. Several landmark proteins are noted using the same legend as Figure 2, and the location of GFP (determined by Western analysis) is also noted (GFP migrates as three bands at 23, 28, and 29 kDa by SDS–PAGE).

amino acid sequencing and mass spectrometry technologies as described below.

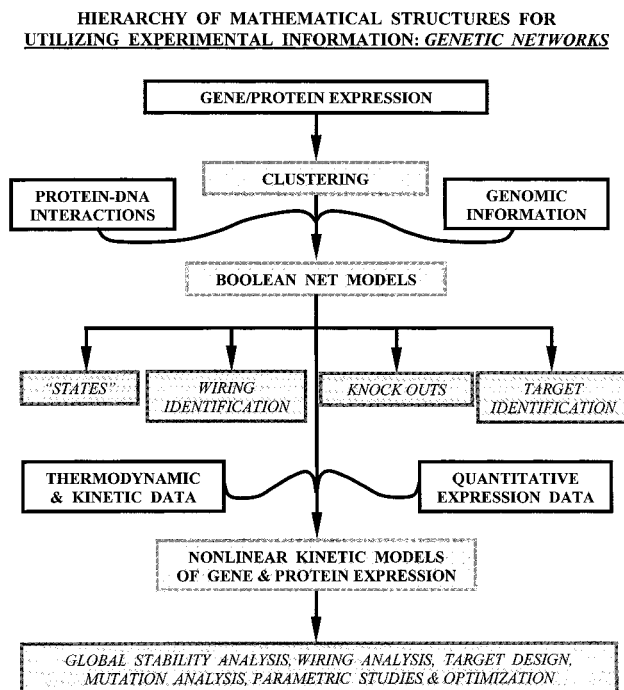
Based on the known function of some of these affected genes, it is apparent that many different and seemingly unrelated cellular processes are affected by the expression of GFP in these cells. The *E. coli* genome has 4269 open reading frames that encode proteins that range from approximately  $pI = 3.5$  to  $pI = 13$  (18). The above results represent only a portion (pH 4–6) of this entire proteome, suggesting that the effect of GFP on the entire intracellular and extracellular *E. coli* proteome is significant. This observation reinforces the notion that biological systems and phenomena are nonlinear and highly coupled and that reporter genes used to monitor metabolism can have a profound effect on various cellular processes by themselves. Such effects imply that there is no simple or obvious relationship between genotype and phenotype. Thus, one of the challenges for genetic engineering applications is the deconvolution of changes in gene expression caused by cloned gene expression from changes caused by other host cell responses.

There are many studies which do not involve the expression of heterologous message or protein in a bacterium. Such studies include the identification of protein markers for disease, the analysis of knock-out mice using 2D gels, and the comparison of expressed sequence tags from different organisms. The results presented above reinforce the concept that conclusions made on the basis of the study of an individual gene or state of an organism may have limited relevance as to how that gene and gene product function in the context of the whole cell, tissue, or organism.

**Hierarchy and Integration of Experimental Gene Expression Information: Current and Future Prospects.** Experimental and theoretical studies of biochemical reaction networks are advanced and mature within metabolic engineering (19, 20). The contributions of mathematical modeling frameworks and analysis have been recently reviewed, and a hierarchy of mathematical structures for effectively utilizing available experimental information has been proposed (21). In this same spirit, we propose a similar hierarchy for the interpretation of information generated within the field broadly defined as functional genomics (Figure 4).

Statistical analysis of gene (microarray) and protein (2DE) expression patterns permits the tentative clustering and identification of related genes and gene products as well as their association to phenotypes. DNA sequence information promotes further ordering of genes of interest according to similar upstream regulatory sites (22). Genomic information on upstream DNA protein-binding sites, based on a sequence similarity analysis, can also help to develop such information and to identify, within the statistically identified clusters, possible regulatory proteins. Still, even after a protein is identified as a regulator, one cannot confidently elucidate its particular role as an inducer or repressor without further information and analysis.

The formulation and analysis of Boolean networks, based on the previously derived information, helps in this regard. The examination of alternative Boolean models with respect to their *qualitative* consistency with observed expression patterns provides multiple, but a smaller subset of, possible conceptual models that fit the observations. The possible steady states can be qualitatively identified (on/off) along with the associated regulatory wiring of the genetic network of interest. As discussed earlier, Boolean models cannot describe differential expression of genes, but they can be applied to deletion studies (knock-out experiments). The deletion or insertion of individual genes (and therefore gene products) for



**Figure 4.** Hierarchy of mathematical structures for utilizing experimental information on gene networks. Primary experimental information is shown in bold, derived models are given by a dotted pattern, and results from model analyses are in italics on a hatch pattern. For more general information, see ref 21.

manipulation of undesirable phenotypes is a strategy commonly employed in medical applications (e.g., drug targets).

The Boolean models identified above as possible candidates for the gene network of interest can be further refined using thermodynamic and kinetic data along with the *quantitative* information from the expression experiments to construct nonlinear mechanistic kinetic models of gene and protein networks (23, 24). Such models can permit a wide array of studies from stability analysis and parametric optimization to targeted mutagenesis analysis and target design.

Consider the proposed methodology for integrating mathematical tools for the analysis of experimental data from genetic networks in the context of a two-gene system. A statistical analysis of the data for the temporal patterns of mRNA and protein levels of the two genes could help to identify the oscillatory nature of the system. Next, assisted by genomic analysis, one could formulate a set of Boolean models that would suggest protein–DNA interactions. In particular, it can be shown that there is only one regulatory network that can result in oscillatory behavior (depicted in Figure 1A). One could proceed to construct possible mechanistic models for gene and protein expression for this system. The data from microarrays and proteome analysis would be used to estimate key parameters of these mechanistic models. Simulation studies of the nonlinear mechanistic models could suggest the experiments required to discriminate between the mechanistic models and to identify the best candidate model.

One of the critical determinants of the successful application of the framework presented in Figure 4 is in the quality of experimentally derived information on gene expression. In proteomics, a key issue is the ability to separate, detect, and characterize all of the proteins present in a given biological sample in a rapid and

reproducible manner. Often, technological achievements in the separations procedure address only one of these concerns at the expense of progress in the other areas of interest. However, there are several emerging technologies which hold promise for alleviating existing bottlenecks in proteome analysis.

No amplifying technology exists for proteins, which focuses much interest on improving detection technologies. Current technology for ammoniacal silver stains can detect proteins in 1–10-fmol quantities; however, new technology from Biotraces Inc. (Fairfax, VA), based on multiphoton detection with radiolabeled iodine, improves on this sensitivity by more than 3 orders of magnitude, permitting the identification of proteins present at 1–10 amol per sample. Applied to amino acid sequencing technology, these developments may enable automated amino acid sequencers to generate reliable sequence information from only 25 zmol of material.

For the microchemical characterization of small quantities of protein, mass spectrometry (MS) is currently the preferred technology, because up to 500 MS-based protein identifications can be made per day (25). Various MS techniques are available which use either liquid feed for on-line analysis or a solid matrix support. For details, the reader is referred to recent reviews (26, 27). In general, ion trap MS offers an optimal combination of relatively low cost and high sensitivity (femtomoles) in concert with the ability to perform data-dependent MS/MS fragmentation to generate *de novo* sequence information. Sequence data from these tandem MS experiments are used to probe on-line protein sequence databanks (e.g., Swiss-Prot) as well as translations of expressed sequence tag (EST) databases. This technology permits routine identification of the genetic basis for unknown spots from 2DE gels. A further advantage of MS techniques is the ability to generate *de novo* sequence data on multiple protein species present in a single spot, which begins to address limitations inherent in parallelizing microchemical spot characterization. Also of note in this regard is the Molecular Scanner developed by Denis Hochstrasser and colleagues (28). This new technology takes 2DE gels and combines protease digestion and electroblotting to a membrane into a single step, followed by the generation of MS fingerprints (by matrix-assisted laser desorption/ionization time-of-flight MS) of individual regions on the resulting 2DE membrane. As a result, a standard 16-cm × 16-cm gel provides 160 GB of mass spectra data on the gene identity of *all* spots present in femtomole or greater quantity. Such developments herald quantum leaps in the ability to do functional genomics studies and in understanding the complex nonlinear genotype–phenotype relationship for biological systems.

### Protocols and Procedures

We provide here some standard protocols and guidelines for proteome analysis. 2DE is a technically complex and demanding procedure, and there are several opportunities to introduce technical errors into experiments or to make interlaboratory comparisons difficult. Thus, we exercise care in the selection of grade and source for reagents. Furthermore, deionized water (dI H<sub>2</sub>O) of the highest purity is necessary to obtain quantitative stains with low background, and we use a Barnstead Nanopure system with ultrafilter. Finally, care must be taken to minimize contamination by keratins and other proteins which can complicate microchemical analysis of spots. All steps should be performed with powder-free gloves and while wearing labcoats.

**Preparation of Samples.** Green fluorescent protein (GFP) is expressed in *E. coli* JM105 using pBluescript (Stratagene). Cells are grown to OD<sub>600</sub> = 1.0 in LB media and harvested. The culture is pelleted at 4000 rpm and 15 °C for 10 min and washed four times in a solution containing 3.0 mM KCl, 1.5 mM KH<sub>2</sub>PO<sub>4</sub>, 68 mM NaCl, and 9.0 mM NaH<sub>2</sub>PO<sub>4</sub>. The washed pellet is resuspended in a solution containing 10 mM Tris-HCl pH 8.0, 1.5 mM MgCl<sub>2</sub>, 10 mM KCl, 0.5 mM dithiothreitol (DTT), 0.5 mM Pefabloc SC protease inhibitors (Boehringer), and 0.1% sodium dodecyl sulfate (SDS). This solution is sonicated at full power on ice for 30 s in a Fisher F550 cup horn sonifier and stored at –75 °C until use. One hundred micrograms of protein (corresponding to 40 μL of frozen sample) is mixed with 20 μL of 8 M urea, 4% (w/v) 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS), 65 mM DTT, 40 mM Tris pH 8.0, and a trace of bromophenol blue. This sample mixture is loaded by in-gel rehydration (29) by adding 340 μL of reswelling solution containing 8 M urea, 2% CHAPS, 0.3% DTT, 1.33% BioLyte 3-10 (Bio-Rad Laboratories), 0.67% BioLyte 5-7 (Bio-Rad Laboratories), and a trace of bromophenol blue.

**Isoelectric Focusing.** Isoelectric focusing (IEF) is performed in 18-cm, pH 3–10 nonlinear Immobiline gels (Amersham-Pharmacia-Hoefer). These Immobiline gel strips, which have an immobilized pH gradient (IPG) to improve reproducibility, are rehydrated overnight in the Reswelling cassette (Amersham-Pharmacia-Hoefer) with the combination of reswelling solution and sample mixture, complete with 100 μg of protein, as mentioned above. Alternatively, IEF can be performed in 18-cm pH 4–7 linear Immobiline gels. These gels are reswelled with the same sample mixture and in the same manner as above with a corresponding pH 4–7 reswelling solution containing 8 M urea, 2% CHAPS, 0.3% DTT, 1.0% BioLyte 3-10, 1.0% BioLyte 4-7, and a trace of bromophenol blue.

The rehydrated gels are placed on an alignment card, which is then placed in an Immobiline Strip Tray (Amersham-Pharmacia-Hoefer). Kerosene oil (J.T. Baker) is used as heat-transfer fluid between the cooling plate of the Multiphor II and the Immobiline Strip Tray, and between the tray and the alignment card. Electrode strips (Amersham-Pharmacia-Hoefer) are presoaked in 0.5 mM NaOH and 6 mM H<sub>3</sub>PO<sub>4</sub> and placed over the tips of the gels at the cathode and anode, respectively. The cathode and anode electrodes are then positioned over the electrode strips, and paraffin oil (Fisher Scientific) is used to overlay the entire assembly within the tray.

For pH 3–10 nonlinear strips, isoelectric focusing is performed for a total of 71 750 V-h at 20 °C. The voltage is ramped linearly from 500 to 3500 V during the first 5 h and is maintained at 3500 V for a subsequent 15.5 h using an EPS 3500 XL power supply (Amersham-Pharmacia-Hoefer). For pH 4–7 strips, isoelectric focusing is performed for a total of 84 000 V-h at 20 °C. The voltage is ramped linearly from 500 to 3500 V during the first 5 h and is maintained at 3500 V for a subsequent 19 h.

**Equilibration.** After isoelectric focusing, gels are incubated for 15 min in a solution containing 6 M urea, 2% DTT, 30% glycerol, 2% SDS, and 0.024 M Tris pH 6.8 and subsequently for 5 min in a solution containing 6 M urea, 2.5% iodoacetamide, 30% glycerol, 2% SDS, and 0.024 M Tris pH 6.8. Both equilibration steps are performed on a rocking platform and in a custom-built equilibration tray modeled after the Pharmacia Reswelling cassette.

**SDS-Polyacrylamide Gel Electrophoresis (PAGE).** SDS-PAGE is performed using the Bio-Rad Protean II xi Multicell, which is a vertical second dimension system. Second dimension 12%T, 2.6%C gels (18 cm × 16 cm × 1.5 mm) are prepared by mixing 132.16 mL of acrylamide/piperazine diacrylamide (PDA) solution (30% Bio-Rad acrylamide, 0.8% Bio-Rad PDA), 88 mL of 1.5 M Tris pH 8.6, and 110.24 mL of dI H<sub>2</sub>O (makes four gels). This solution is degassed for 10 min. To polymerize, 335 μL of freshly made 10% ammonium persulfate (Bio-Rad) is added, followed by 140 μL of *N,N,N,N*-tetramethylethylenediamine (Bio-Rad). The solution is immediately poured to a height of 18 cm and overlaid with 0.8 mL of water-saturated *sec*-butanol. After gel polymerization (approximately 2 h) on a vibration-free table, *sec*-butanol is replaced with dI H<sub>2</sub>O, and the gels are covered with plastic wrap and stored at room temperature overnight.

A solution of 0.5% agarose, 25 mM Tris (pH 8.3), 198 mM glycine, and 0.1% SDS is prepared prior to the transfer of strips to the second dimension gels. The agarose solution is heated in a 900-W microwave for 90 s at high power, and the temperature is maintained at 75 °C on a heated stirplate. The IPG gel strips are transferred to the previously prepared 12%T gels and trimmed approximately 3 mm at the acidic end and 10 mm at the basic end to fit onto the 16-cm-wide 12%T second dimension gels. The agarose solution is poured over the strips to a height of 2 cm. Electrophoresis is carried out at 40 mA per gel until the bromophenol blue dye front migrates to the end of the gel.

**Detection.** Proteins are detected using an ammoniacal silver stain (30), and all steps are performed on an orbital shaker in 250 mL of liquid per staining tray with two gels per tray. After SDS-PAGE, the gels are fixed for 1 h in 50% dI H<sub>2</sub>O, 40% ethanol, and 10% acetic acid, and the gels are rehydrated overnight in 90% dI H<sub>2</sub>O, 5% ethanol, and 5% acetic acid. The following day, proteins are cross-linked in 10% glutaraldehyde for 30 min. The glutaraldehyde is subsequently washed out by four 5-min dI H<sub>2</sub>O washes and three 30-min dI H<sub>2</sub>O washes.

Ammoniacal silver solution is made fresh, about 30 min prior to staining. For 500 mL of ammoniacal silver solution, 4 g of silver nitrate dissolved in 20 mL of dI H<sub>2</sub>O is slowly pipetted into 105.9 mL of dI H<sub>2</sub>O, 768 μL of 50%NaOH, and 6.66 mL of ammonium hydroxide. dI H<sub>2</sub>O is then added to a final volume of 500 mL. After the last water wash, the gels are stained in the ammoniacal silver solution for 20 min. This is followed by three 5-min dI H<sub>2</sub>O washes. The gels are then developed in a 0.01% (w/v) citric acid, 0.037% (v/v) formaldehyde solution until spots are optimally visible against the background. Development is stopped by incubation in a 30% (w/v) citric acid solution for 5 min before laser scanning. Gels can be stored in a 5% acetic acid solution for longer periods of time without further image development. However, some bleaching of spots can occur under these conditions.

**Gel Analysis.** Computer-assisted gel analysis (Melanie II) is performed on images captured with a Molecular Dynamics Personal Densitometer at 50-μm resolution with 12-bit dynamic range. Melanie II default parameters are used for feature detection and matching and corrected by visual inspection. An estimate of relative quantitative changes is made on the basis of the change in percent volume among silver-stained gels. Data from spots that are very faint or very strongly stained are not included in quantitative comparisons, as the staining intensity either is not linear or is saturated for these cases. Spot changes of interest are tested on multiple gels for reproducibility.

**Microchemical Characterization of Spots.** The microchemical characterization of spot changes can be made on the basis of a comparison to generally available proteome databases such as those available at Swiss 2D-Page (www.expasy.ch). N-Terminal sequence tagging (the generation of 3–4 residues of sequence information by Edman sequencing in an automated sequenator) can be performed from gels electroblotted to poly(vinylidene difluoride) (PVDF) and stained with Coomassie blue or amido black. Internal sequence information can be obtained from spots digested with trypsin (or other protease) and collected by HPLC before Edman sequencing. Mass spectra (for peptide fingerprinting) and tandem MS data (for *de novo* sequencing) from proteins present in femtomole amounts can be obtained from HPLC-purified peptides run on a triple-quadrupole mass spectrometer with a nanospray needle or an ion trap mass spectrometer. Tandem MS experiments for amino acid sequencing of peptides permit the characterization of spots of interest, even in the face of contaminating proteins. The selection of appropriate microchemical characterization technique and the ensuing sample preparation can be spot dependent, so the reader is referred to recent reviews (26, 27).

### Acknowledgment

The authors thank Wilfred Chen for relevant strains and genes and Mike Shuler for useful comments and suggestions. K.H.L. is generously supported by Merck Inc. and by grants from Intel Corp. (98-238), the Lederman Family Foundation, and the New York State Science and Technology Foundation and Industrial Partners.

### References and Notes

- (1) Wasinger, V. C.; Cordwell, S. J.; Cerpa-Polijak, A.; Yan, J. X.; Gooley, A. A.; Wilkins, M. R.; Duncan, M. W.; Harris, R.; Williams, K. L.; Humphery-Smith, I. Progress with gene-product mapping of the Mollicutes. *Mycoplasma genitalium*. *Electrophoresis* **1995**, *16*, 1090–1094.
- (2) O'Farrell, P. H. High-resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **1975**, *250*, 4007–4021.
- (3) Klose, J. Genotypes and phenotypes. Presented at From Genome to Proteome: 3rd Siena 2D Electrophoresis Meeting, Siena, Italy, Aug 31–Sept 3, 1998.
- (4) Hatzimanikatis, V.; Lee, K. H. Submitted.
- (5) Winzeler, E. A.; Richards, D. R.; Conway, A. R.; Goldstein, A. L.; Kalman, S.; McCullough, M. J.; McCusker, J. H.; Stevens, D. A.; Wodicka, L.; Lockhart, D. J.; Davis, R. W. Direct allelic variation scanning of the yeast genome. *Science* **1998**, *281*, 1194–1197.
- (6) Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **1995**, *270*, 467–470.
- (7) Lockhart, D. J.; Dong, H.; Byrne, M. C.; Follettie, M. T.; Gallo, M. V.; Chee, M. S.; Mittmann, M.; Wang, C.; Kobayashi, M.; Horton, H.; Brown, E. L. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **1996**, *14*, 1675–1680.
- (8) Loomis, W. F.; Sternberg, P. W. Genetic networks. *Science* **1995**, *269*, 649.
- (9) Michaels, G. S.; Carr, D. B.; Askenazi, M.; Fuhrman, S.; Wen, X.; Somogyi R. Cluster analysis and data visualization of large-scale gene expression data. *Pac. Symp. Biocomput.* **1998**, *3*, 42–53.
- (10) Kauffman, S. A. *The Origins of Order*; Oxford University Press: New York, 1993.
- (11) Crosthwaite, S. K.; Dunlap, J. C.; Loros, J. J. *Neurospora wc-1* and *wc-2*: transcription, photoresponses, and the origins of circadian rhythmicity. *Science* **1997**, *276*, 763–769.
- (12) Thomas, R. In *Kinetic Logic: A Boolean Approach to the Analysis of Complex Regulatory Systems*; Thomas, R., Ed.; Springer-Verlag: New York, 1979; pp 352–402.

- (13) Liang, S.; Fuhrman, S.; Somogyi, R. REVEAL: a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.* **1998**, *3*, 18–29.
- (14) Haynes, P. A.; Gygi, S. P.; Figeys, D.; Aebersold, R. Proteome analysis: biological assay or data archive? *Electrophoresis* **1998**, *19*, 1862–1871.
- (15) Blattner, F. R.; Plunkett, G.; Bloch, C. A.; et al. The complete genome sequence of *Escherichia coli* K-12. *Science* **1997**, *277*, 1453–1474.
- (16) Lee, S. B.; Bailey, J. E. A mathematical model for  $\lambda$ dv plasmid replication: analysis of wild-type plasmid. *Plasmid* **1984**, *11*, 151–165.
- (17) Lee, S. B.; Bailey, J. E. A mathematical model for  $\lambda$ dv plasmid replication: analysis of copy number mutants. *Plasmid* **1984**, *11*, 166–177.
- (18) Link, A. J.; Robison, K.; Church, G. M. Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. *Electrophoresis* **1997**, *18*, 1259–1313.
- (19) Bailey, J. E. Toward a science of metabolic engineering. *Science* **1991**, *252*, 1668–1675.
- (20) Cameron, D. C.; Tong, I. T. Cellular and metabolic engineering: an overview. *Appl. Biochem. Biotechnol.* **1993**, *38*, 105–140.
- (21) Bailey, J. E. Mathematical modeling and analysis in biochemical engineering: past accomplishments and future opportunities. *Biotechnol. Prog.* **1998**, *14*, 8–20.
- (22) Stormo, G. D.; Field, D. S. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* **1998**, *23*, 109–113.
- (23) Birnbaum, S.; Bailey, J. E. Plasmid presence changes the relative levels of many host cell proteins and ribosome components in recombinant *Escherichia coli*. *Biotechnol. Bioeng.* **1991**, *37*, 736–745.
- (24) Bailey, J. E. Host-vector interactions in *Escherichia coli*. *Biotechnology* **1993**, *48*, 29–52.
- (25) Langen, H.; Fountoulakis, M.; Evers, S.; Wipf, B.; Berndt, P.  $^{15}\text{N}$  and  $^{13}\text{C}$ -labeling of cells for identification and quantification of proteins on 2D gels. Presented at From Genome to Proteome: 3rd Siena 2D Electrophoresis Meeting, Siena, Italy, Aug 31–Sept 3, 1998.
- (26) Figeys, D.; Aebersold, R. High sensitivity analysis of proteins and peptides by capillary electrophoresis-tandem mass spectrometry: recent developments in technology and applications. *Electrophoresis* **1998**, *19*, 885–892.
- (27) Courchesne, P. L.; Jones, M. D.; Robinson, J. H.; Spahr, C. S.; McCracken, S.; Bentley, D. L.; Luethy, R.; Patterson, S. D. Optimization of capillary chromatography ion trap-mass spectrometry for identification of gel-separated proteins. *Electrophoresis* **1998**, *19*, 956–967.
- (28) Binz, P. A.; Bienvenut, W.; Fabbretti, R.; Gasteiger, E.; Bairoch, A.; Appel, R.; Sanchez, J. C.; Hochstrasser, D. F. The “Molecular Scanner:” an highly automated method for protein identification and 2-D page annotation. Presented at From Genome to Proteome: 3rd Siena 2D Electrophoresis Meeting, Siena, Italy, Aug 31–Sept 3, 1998.
- (29) Sanchez, J. C.; Rouge, V.; Pisteur, M.; Ravier, F.; Tonella, L.; Moosmayer, M.; Wilkins, M. R.; Hochstrasser, D. F. Improved and simplified in-gel sample application using reswelling of dry immobilized pH gradients. *Electrophoresis* **1997**, *18*, 324–327.
- (30) Lee, K. H.; Harrington, M. G.; Bailey, J. E. Two-dimensional electrophoresis of proteins as a tool in the metabolic engineering of cell cycle regulation. *Biotechnol. Bioeng.* **1996**, *50*, 336–340.

Accepted January 25, 1999.

BP990004B