High-Throughput Biological Data Analysis

A STEP TOWARD UNDERSTANDING CELLULAR REGULATION

PHOTO BY JASON A.K. BERNSTEIN

THANURA R. ELVITIGALA, ASHOKA D. POLPITIYA, WENXUE WANG,
JANA STÖCKEL, ABHA KHANDELWAL, RALPH S. QUATRANO,
HIMADRI B. PAKRASI, and BIJOY K. GHOSH

Living cells are complex dynamic systems, showing a remarkable ability to adapt to different environmental conditions for their survival. Central to this characteristic trait is an underlying regulatory mechanism that controls various biological processes. Unraveling the basic principles of cellular regulation is a fundamental challenge and is the main subject of this article. In the medical sciences, this understanding can lead to new treatments for diseases, such as cancer and diabetes, while facilitating drug discovery. In agriculture, an understanding of basic cellular regulation promises the development of new varieties of crops with higher yields and higher levels of nutrients that tolerate adverse climactic conditions. Furthermore, controlling the dynamics of a living cell may help humanity to address problems of global warming and find alternatives to fossil fuels.

Digital Object Identifier 10.1109/MCS.2010.938100

1066-033X/10/$26.00©2010IEEE

Central Dogma of Molecular Biology

The central dogma of molecular biology defines the main regulatory mechanism involved in cellular regulation based on dynamic interactions among DNA, RNA, and proteins (see Figure S1) [S1]. According to the central dogma of molecular biology, the genetic information that controls a cell is stored in DNA strands. A DNA strand consists of long sequences of nucleotides, which are denoted by A, T, G, and C. A primary task in genome-sequencing is to obtain the sequence of a DNA strand. Organisms differ in the number of these DNA strands, which are also known as chromosomes. For many organisms including humans, mammals, bacteria, and eukaryotes, genomes are completely sequenced.

It turns out that different subsequences of the DNA strands have different functional roles. Subsequences that have the capability to generate specific RNA molecules, through a process known as transcription, are called genes. A typical cell consists of thousands of genes. Although the genes on a DNA strand are virtually stationary, the transcribed RNAs are not. One type of RNA, known as the messenger-RNA (mRNA), gives rise to corresponding proteins through translation [S2] (see "Glossary of Biological Terms"). Proteins also move inside the cell and participate in various biochemical reactions. These biochemical reactions are also influenced by the availability of nutrients, variations in temperature, and fluctuations in light. Cells are able to sense these changes in their surroundings. Some of the proteins can control transcription of other genes by either enhancing or inhibiting them.

Although the central dogma of molecular biology is considered to be the main mode of regulation, numerous findings suggest that the control mechanisms in a living cell are more complex than what is presented by the central dogma [S3], [S4]. Many post-transcriptional and post-translational modifications alter the operation of central dogma. Reverse transcription and microRNAs are modes of regulation that are not captured by the central dogma of molecular biology.

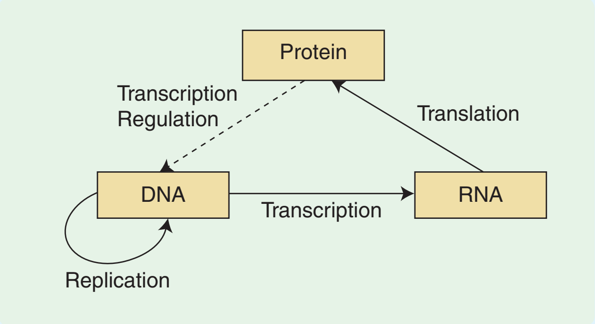

FIGURE S1 The interactions among DNA, RNA, and protein molecules in a living cell. Genetic information contained in DNA is transferred from one generation to the next through replication. Depending on the requirements of the cells, RNA molecules are produced in transcription and are subsequently translated into the corresponding proteins. Some proteins act as regulators to control the transcription process.

REFERENCES

[S1] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 8, pp. 561–563, 1970.

[S2] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, 4th ed. New York: Garland Science, 2002.

[S3] A. Shapiro, "Revisiting the central dogma in the 21st century," *Ann. NY Acad. Sci.*, vol. 1178, pp. 6–28, 2009.

[S4] M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder, "What is a gene, post-ENCODE? History and updated definition," *Genome Res.*, vol. 17, pp. 669–681, 2007.

CENTRAL DOGMA OF MOLECULAR BIOLOGY AND HIGH-THROUGHPUT EXPERIMENTS

Basic to the process of controlling the dynamics of living cells is the central dogma of molecular biology (see "Central Dogma of Molecular Biology"). The central dogma explains cellular regulation using the dynamic interactions among DNA, RNA, and proteins (see "Glossary of Biological Terms"). Cellular responses to various external environmental conditions, such as the availability of nutrients, and internal conditions, such as the presence or absence of essential proteins, are stored as genetic information in the form of DNA. Through transcription, RNA molecules are produced from the genetic information stored in the DNA. The DNA subsequences that have the ability to generate specific RNA molecules are called genes, thousands of which reside in a typical cell. One type of RNA molecule, known as messenger-RNA (mRNA), gives rise to corresponding proteins through translation (see "Glossary of Biological Terms"). Proteins perform crucial functions in a living cell. These functions include catalyzing chemical reactions, transporting compounds in and out of the cell, defining the structure of the cell, and cellular signaling. One of the functional roles of proteins is transcriptional control, a process where a protein synthesized by one gene binds to a specific DNA sequence and controls the transcription of one or more additional genes (see Figure 1).

The interactions among genes and proteins serve as the main regulatory mechanism in cells. In all organisms, the proteins that regulate the transcription of other genes are currently unknown for many genes. Uncovering the protein that promotes the activity of a specific gene requires simultaneous measurements of the abundance of proteins and cataloging the set of genes that are differentially expressed. Differential expression of a gene at each instant of time is typically assessed by measuring the quantity of

Unraveling the basic principles of cellular regulation is a fundamental challenge.

mRNA produced by a gene and comparing those mRNA abundances to a base level.

Proteins regulate the transcription level of other genes by physically anchoring onto regions of DNA called promoter regions (see Figure 1). Like a gene, a promoter region is also characterized by a DNA sequence. A gene and its corresponding promoter region are typically colocated on a DNA strand, where the promoter region is found upstream of the corresponding gene. However, the exact location of a promoter sequence with respect to a gene can vary significantly. By searching the entire upstream regions of related genes, it is possible to locate suitable candidates for a promoter region [1].

High-throughput techniques are used to expedite the process of identifying these regulator-target relationships among promoters and the genes they promote. Each high-throughput technique probes the central dogma at various levels. Genome-sequencing techniques [2] are used to discover genome sequences of organisms. Knowledge of the genome sequences is necessary for understanding the evolution of organisms [3] and for identifying genes responsible for various biological functions [4]. DNA microarrays [5] are utilized for the simultaneous measurement of transcription levels of genes, whereas high-throughput-proteomic techniques [6] are used to identify and quantify the proteins. For a description of the steps involved in performing DNA-microarray and proteomic experiments, see "Introduction to Microarray Technologies" and "Introduction to Proteomics Technologies," respectively. ChIP-Chip experiments [7], which combine chromatin immunoprecipitation (ChIP) with microarray

Glossary of Biological Terms

DNA: The hereditary material in a living cell. The DNA molecule is a blueprint in a cell, containing instructions needed to construct other components of a cell, such as RNA and protein molecules, as well as instructions on how a cell responds to various environmental conditions. DNA is structured as a long double helix polymer made from millions or billions of repeating units called nucleotides, namely, adenine (A), guanine (G), cytosine (C), and thymine (T). In the double helix, A is base-paired with T while C is base-paired with G.

RNA: A molecule consisting of four nucleotides, adenine (A), guanine (G), cytosine (C), and uracil (U). Uracil occurs in RNA, corresponds to the places of thymine in the DNA. Types of RNA molecules include messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA).

Complementary sequence: A nucleic acid sequence that is complementary to a given DNA or RNA sequence. Nucleic acid A is complementary to T (or U in the case of RNA), while C is complementary to G. Complementary sequences join each other to create a DNA-DNA or DNA-RNA double strand.

Codon: A sequence of three nucleotides. Four different nucleotides results in 64 different codons. Different amino acids are associated with one or more distinct codons.

Amino acids: Building blocks of various proteins. Cells typically contain 20 different amino acids. All amino acids consist of an amine group, a carboxylic acid group, and a side chain that varies from one amino acid to the other.

Peptides: Short sequences of amino acids. Using enzymes, such as trypsin, proteins can be fragmented into several constitute peptides.

Proteins: Main functional molecules in a cell. Proteins consists of long chains of amino acids.

Enzymes: Mostly proteins that catalyze chemical reactions in a cell. For example, restriction enzymes split DNA strands at positions containing specific nucleotide sequences.

Genes: DNA subsequences that have the capability to generate specific RNA molecules. Typically, DNA strands in a cell contain thousands of genes.

Transcription: A process of producing RNA molecules using DNA as a blueprint. Transcription is performed by a protein complex known as RNA polymerase.

Transcription control: Regulation of the level of transcription of RNA molecules. Transcription control is mostly mediated by regulatory proteins. However, some types of RNA molecules are now known to play a significant role in transcription control.

Reverse transcription: A process of producing DNA molecules using RNA as the blueprint. Reverse transcription is performed by the reverse transcriptase enzyme.

Translation: The process where protein molecules are synthesized using corresponding mRNA molecules. Translation occurs in ribosomes. Translation is started from start codon AUG and preceded by adding corresponding amino acids to the protein according to the codons found in the RNA molecule. Translation stops when RNA polymerase encounters one of the stop codons, UAA, UGA, or UAG.

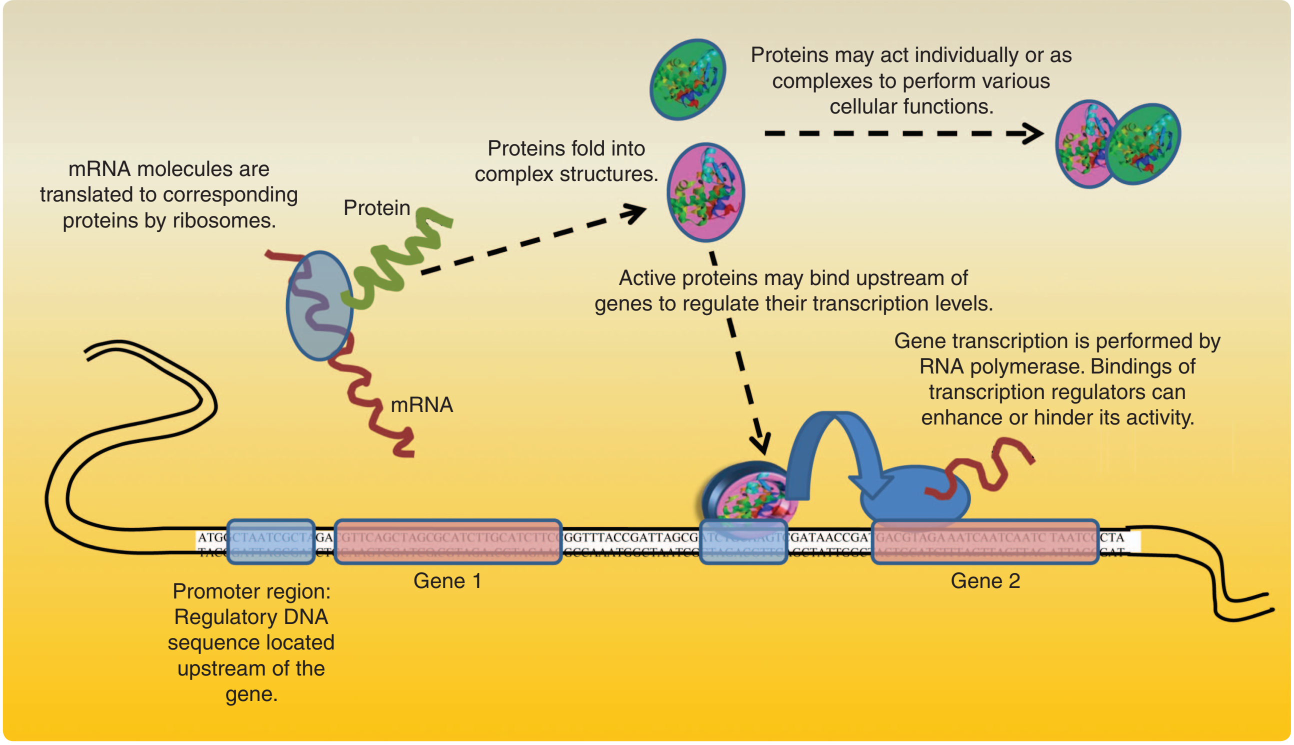
Proteins may act individually or as complexes to perform various cellular functions.
Proteins fold into complex structures.
mRNA molecules are translated to corresponding proteins by ribosomes.
Protein
Active proteins may bind upstream of genes to regulate their transcription levels.
Gene transcription is performed by RNA polymerase. Bindings of transcription regulators can enhance or hinder its activity.
mRNA
Gene 1
Gene 2
Promoter region: Regulatory DNA sequence located upstream of the gene.

FIGURE 1 Steps in gene regulation. Gene transcription produces mRNAs that are translated into corresponding proteins by a protein complex known as ribosomes. Proteins undergo various structural changes before becoming functional. Functional proteins act individually or as complexes in performing various cellular activities. Regulatory-factor proteins bind to the promoter regions of a gene to regulate the gene's transcription activities. The promoter regions are DNA sequences located upstream of the genes.

technology (chip), are used to identify bindings between proteins and DNA.

In this article, salient steps involved in processing data from microarray and proteomics experiments are discussed. Various levels of the required data processing are highlighted in Figure 2. Central to microarray and proteomic experiments are i) *experimental design* for minimizing technical variations, ii) *quality assessment* for consistency during multiple runs and removing outlier runs, iii) *normalization techniques* for removing systematic variations, and iv) *identification and categorization* of genes that are hypothesized to be relevant to the experiment. Systems-level information about an organism that includes interactions among genes and proteins is identified utilizing techniques such as correlation measurements, probabilistic and Bayesian networks, as well as linear and nonlinear dynamic system models. To validate and explore the biological relevance of the preliminary findings regarding interactions among genes and proteins, the obtained results are further analyzed using promoter-sequence analyses.

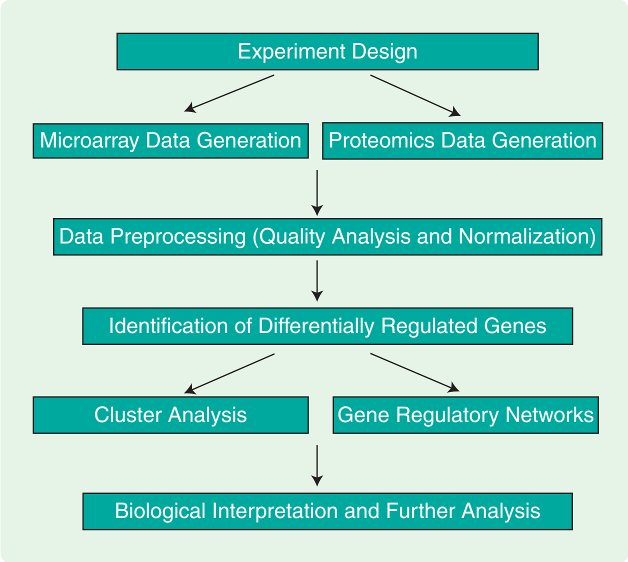
Experiment Design
Microarray Data Generation
Proteomics Data Generation
Data Preprocessing (Quality Analysis and Normalization)
Identification of Differentially Regulated Genes
Cluster Analysis
Gene Regulatory Networks
Biological Interpretation and Further Analysis

FIGURE 2 Typical steps involved in generating and analyzing microarray and proteomics data. Irrespective of the experimental objective, microarray and proteomics experiments involve the steps outlined.

EXPERIMENTAL DESIGN, QUALITY ASSESSMENT, AND DATA NORMALIZATION

The goal of experimental design is to reduce the undesirable effects caused by variations that are not the principal focus of the experiment. This variability arises from both biological and technical factors. Differences in the biological materials are caused mainly by differences in growth conditions, nutrients in the growth media, and cell densities of the cultures. Therefore, two to three biological replicates

Introduction to Microarray Technologies

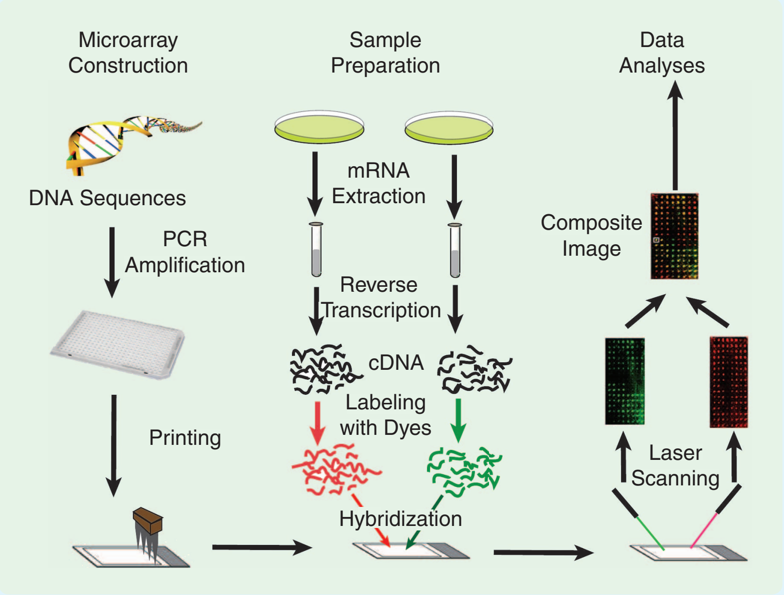

FIGURE S2 Steps involved in a two-color microarray experiment. DNA sequences corresponding to individual genes are amplified using polymerase chain reactions and printed on glass slides. During the experiment, mRNA extracts from two experimental conditions, identified as the target and the control, are converted into the corresponding complementary DNA (cDNA) sequences through reverse transcription. These cDNA from the two samples are labeled separately using two dyes and hybridized onto microarray chips. After washing away nonspecific bindings, the chips are scanned with lasers of two colors, and the scanned images are combined to obtain a composite image. Individual gene expressions are extracted from these images.

Microarray technology [5] adds a new dimension to the way biological experiments are conducted. Instead of monitoring the activities of a few selected genes, microarrays facilitate the simultaneous measurement of the activities of thousands, and often tens of thousands, of genes representing a significant part of the genome of an organism.

Various types of microarray technologies are available, although the underlying science is similar. Usually microarray chips are made on glass plates. DNA sequences corresponding to different genes are printed onto separate locations of the chip, which are referred to as probes. During an experiment, mRNA is extracted from a biological sample, and complementary DNA (cDNA) sequences are obtained using reverse transcription. Then the complementary sequences are labeled with dyes and applied on the microarray chip. On the chip, DNA sequences tightly bind (also known as hybridize) to corresponding complementary DNA sequences, so that it is possible to remove the looser nonspecific bindings. Nonspecific bindings occur when a particular cDNA molecule is bound to a probe that is not exactly complementary to its sequence. These nonspecific bindings are removed by washing. The resulting microarrays are scanned using lasers, and the levels of each dye attached to different spots are quantified by measuring reflected beam intensities. The amount of dye at each spot is proportional to the relative abundance of that particular gene product in the total mRNA extraction.

Many of the differences in microarray technologies are related to the length of the DNA sequences printed on the chip, the number of various sequences embedded on a single chip, the number of replicates for a given sequence, and the type of dye used to label the mRNA. The cDNA microarrays use longer DNA sequences, usually in the range of 300–400 nucleotides, while the oligonucleotide microarrays use shorter sequences, usually in the range of 15–75 bases. For example, Affymetrix chips are an oligonucleotide-type microarray that use sequences that are 15–18 bases long. However, to achieve gene specificity, several sequences from a given gene are included. The samples are labeled using a single dye, thus a global level of data scaling is required for comparison among microarrays.

On the other hand, Agilent microarrays, another oligonucleotide-type chip, uses sequences of about 60 bases but contains just one or two sequences for a given gene. The mRNA from the control and the target experiments are labeled separately with red and green dyes and hybridized onto the same chip so that differences in gene activities under two experimental conditions can be directly computed. In the scanned images of the microarray chips, red and green spots correspond to genes with different mRNA concentrations between two conditions, while yellow spots correspond to genes with similar concentrations of mRNA. This two-channel microarray technique introduces variability to the data due to differences in the dyes, and this variability needs to be taken into account during the experimental design and data processing. Figure S2 illustrates the steps involved in conducting a two-color microarray experiment.

are typically included in a single microarray experiment to take these differences into account. Technical variability can occur at all steps, from the extraction of mRNA to the scanning of microarrays, and is caused as a result of inconsistencies in sample preparation or the instruments used. Technical variabilities can also be introduced by microarray printing techniques. A principal source of variability, unique to the two-color microarrays, results from nonuniform behaviors of the dyes used. This variability is referred to as dye bias. To address the problem of dye bias,

Introduction to Proteomics Technologies

Proteomics is the logical continuation of transcriptional measurements by microarrays [S5]. While DNA microarrays quantify the abundance level of various mRNAs (transcriptome), proteomics measures the global protein (proteome) abundances. Combining global proteome with transcriptional measurements is a vital component of modern systems biology approaches, where the goal is to characterize the systems-level behavior rather than the behavior of single components. Measuring mRNA levels alone, as in DNA microarrays, does not always reveal much about the levels of the corresponding proteins in a cell and their regulatory behavior since many proteins are subjected to post-translational modifications.

While genome-wide microarrays are ubiquitous, proteome microarrays do not exist due to the fact that proteins do not share the same hybridization properties of nucleic acids. In other words, the detection of mRNA molecules is a straightforward process since each mRNA (and corresponding DNA) molecule is bound to its complementary DNA sequence, while amino acid sequences do not possess these binding properties. Mass-spectrometry (MS)-based methods are effectively used to characterize proteins and now are the platform of choice for analyzing complex protein samples. In this article, the bottom-up proteomics, an approach based on MS for identifying proteins on the basis of their digested peptides, is discussed. Proteins are typically digested into constituent peptides using special enzymes. For example, trypsin splits proteins at the places where amino acids arginine and lysine are present. The underlying assumption in bottom-up proteomics is that a peptide sequence, whose length is approximately six or more amino acids, usually maps uniquely to a protein, thus enabling the protein's identification by searching for the peptide sequence in a database of protein sequences. However, in practical applications, some peptides are shared by more than one protein. For example, a proteomics data set for cyanobacterium *Cyanothece* sp. ATCC 51142 contains about 100 peptides shared by more than one protein [55]. Compared to more than 6000 total peptides detected in the data set, this number represents 1.5% peptides with no unique mapping to single proteins.

The work flow of a standard bottom-up experiment has the following steps: a) extraction of proteins from a sample, b) fractionation to remove contaminants and proteins that are not of interest, c) digestion of proteins into peptides using an enzyme, such as trypsin, d) post-digestion separations to obtain a more homogeneous mixture of peptides, and e) analysis by MS (see Figure S3). Although informatics tools can process the resulting data from the MS, identification and quantization of proteins in a sample are fundamental challenges.

High-performance liquid chromatography (HPLC) methods are popular in separating proteins and peptides. The basic principle is based on a soluble sample that is separated in the liquid phase through a thin tube packed with particles of specific surface chemistry, also known as a column [S6]. Based on the chemical and physical interactions of proteins and peptides with this solid phase, these molecules take different times to traverse through the column. This time, called the retention time, relates to the quantity of a particular peptide or protein present in a sample, where the peak volume of the retention profile of a peptide provides a measure of its abundance.

The identification of peptides is done using MS. The peptides that are eluting from the liquid chromatography column are analyzed by MS to determine the mass-to-charge ratio values for peptides and their fragments. In some applications, two MS runs (MS/MS) are used in sequence to obtain a higher level of separation compared to that of a single MS run. When peptide ions collide with natural gas atoms in the collision cell in MS, the kinetic energy they absorb induces fragmentation and produces b-ions and y-ions. The most significant feature of the b-ions and y-ions observed in an MS/MS spectrum is that they are unique to the peptide sequence. The identification of peptides from the peptide MS/MS spectra can be done in two ways. The first method is de novo interpretation of the spectrum. This approach is usually hard because it is difficult to unambiguously interpret the data from the observed b-ion and y-ion series. The second method is to match the observed MS/MS spectra with peptide sequences in a database. The matching criteria can either be a cross-correlation value [S7] or a probability-based method [S8].

microarray experiments usually include a dye-swap, where two dyes for labeling the samples are switched on replicate arrays. As a result, each experiment typically includes four to six microarrays. The exact combination of samples used for each microarray varies depending on the design. In experimental designs, a distinction is made between *repetition*, the use of the same samples in multiple microarrays to reduce technical variability, and *replication*, the use of the multiple biological samples to reduce biological variability [8]. Sampling can be performed so that every sample is compared to one specific sample or samples are compared as a ring, where each sample is compared to an adjacent sample. Experimental design can consider various factors, such as variability in the measurements, the number of spots in a single microarray, and the cost of producing a single microarray chip [8].

Quality Assessment

Microarray data analysis starts with an assessment to ensure that the collected data are of sufficiently good quality for

Proteins in a sample are inferred from the observed peptides by using Bayesian methods [S9].

Quantitative proteomics techniques primarily evolve under two categories, stable-isotope-labeling methods and label-free methods [6]. The stable-isotope-labeling techniques are analogous to the two-channel microarrays in transcriptome analysis. Samples from each experiment are analyzed using isotopes of N, O, or C. These isotopes are introduced metabolically, chemically, or enzymatically to the sample from one experimental condition. The two samples are then mixed and analyzed in a single cycle. Since the chemical properties of isotopes are the same, the isotope-labeled and native peptides differ by only their mass and are separately detected. The relative intensities of a given peptide under two conditions are determined by measuring the abundance of native and isotope-labeled forms.

Label-free quantification methods are analogous to single-channel microarrays. No labeling is involved, and the two samples are analyzed separately. While these techniques are free of the complexities related to labeling, the measurements are more prone to variations caused by the use of equipment in multiple runs. Peptide abundances are computed based on the peak volume of LC retention profiles as described above or as spectral counts. The spectral count method is the simplest approach, where the number of times a peptide is detected by MS is counted for each peptide, and those counts are accumulated over all the peptides for a given protein. This count gives a value proportional to the abundance of the protein [11].

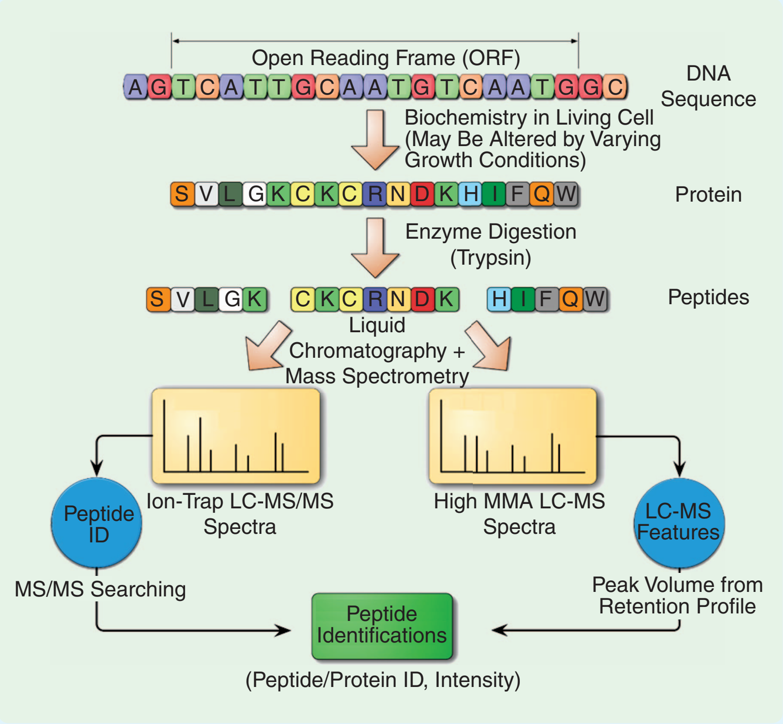
Open Reading Frame (ORF)
A G T C A T T G C A A T G T C A A T G G C
DNA Sequence
Biochemistry in Living Cell (May Be Altered by Varying Growth Conditions)
S V L G K C K C R N D K H I F Q W
Protein
Enzyme Digestion (Trypsin)
S V L G K
C K C R N D K
H I F Q W
Peptides
Liquid Chromatography + Mass Spectrometry
Ion-Trap LC-MS/MS Spectra
High MMA LC-MS Spectra
Peptide ID
LC-MS Features
MS/MS Searching
Peak Volume from Retention Profile
Peptide Identifications
(Peptide/Protein ID, Intensity)

FIGURE S3 Salient steps involved in bottom-up proteomics analysis. Protein extracts from biological samples are digested using enzymes, such as trypsin, to obtain corresponding peptides. Peptides are analyzed using liquid-chromatography-based tandem mass spectrometry (LC-MS/MS) to identify peptides, and the intensities are obtained from the area of the retention profile.

REFERENCES

[S5] M. R. Wilkins, J. C. Sanchez, A. A. Gooley, R. D. Appel, I. Humphery-Smith, D. F. Hochstrasser, and K. L. Williams, “Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it,” *Biotechnol. Genet.* Eng. Rev., vol. 13, pp. 19–50, 1996.

[S6] I. Neverova and J. E. Van-Eyk, “Role of chromatographic techniques in proteomic analysis,” *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, vol. 815, pp. 51–63, 2005.

[S7] A. J. Link, J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik, and J. R. Yates, III, “ Direct analysis of protein complexes using mass spectrometry,” *Nat. Biotechnol.*, vol. 17, pp. 676–682, 1999.

[S8] R. Craig and R. C. Beavis, “TANDEM: Matching proteins with tandem mass spectra,” *Bioinformatics*, vol. 12, pp. 1466–1467, 2004.

[S9] A. I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold, “A statistical model for identifying proteins by tandem mass spectrometry,” *Anal. Chem.*, vol. 75, pp. 4646–4658, 2003.

quantitative analyses. One statistic employed is the coefficient of variation, which is the ratio between the standard deviation and the mean of intensities of individual pixels in each spot on the array. When a microarray is scanned, feature-extraction software assigns each pixel either to the signal, which is the area where mRNA is bound, or to the background. The final intensity value given to each spot and used for further analyses is average intensity for all pixels determined to be from the signal area for a given spot. The coefficient of variation is used to quantify the intensity distribution of the individual pixels categorized as the signal. A lower coefficient of variation for the signal suggests a lower intensity variation among the pixels included as the signal. Another statistic that is taken into consideration is the overall signal-intensity distribution of the spots. Under a 16-bit-resolution scanner, the intensity of a pixel can vary between zero and 65,535. It is desirable that an array have a wide spread of intensities for spots within this allowable range. A dense distribution toward the lower range indicates an insufficient quantity of mRNA and thus is

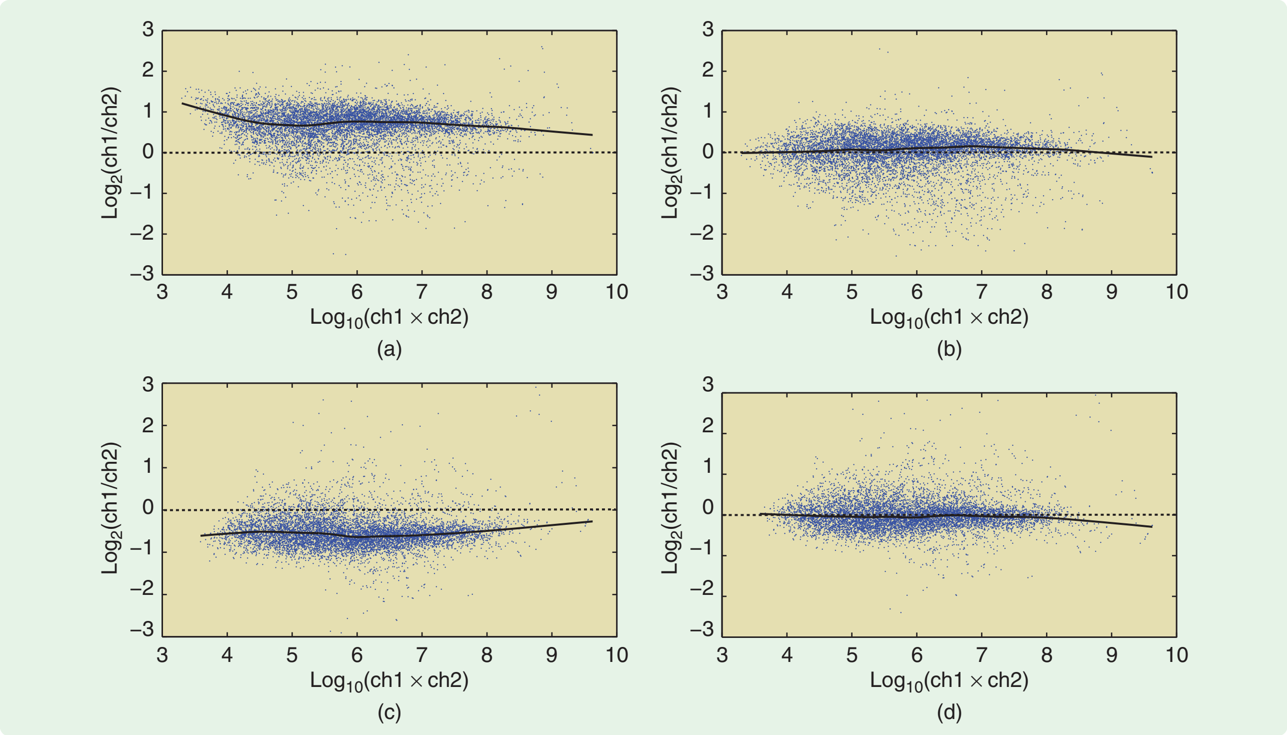

FIGURE 3 Distribution of intensities of spots in microarrays, from two channels, shown as product-ratio plots, where $\text{Log}_2[\text{ch1/ch2}]$ is plotted against $\text{Log}_{10}[\text{ch1}\times\text{ch2}]$. The intensity-based trend observed in two-channel microarray data shown in (a) and (c) is reduced through the local weighted linear regression normalization shown in (b) and (d).

likely to produce poor separation between the background and the signal. On the other hand, too many spots in the higher range are an indication of technical problems, such as insufficient sample dilution or inadequate washing of microarray chips after hybridization. These spots cause contamination of the neighboring spots, resulting in incorrect intensity readings. When a chip contains more than 25 spots with saturated pixels, a flaw in the experimental procedure is likely. Various matrices can be applied at either the array level or spot level to assess the quality of microarray data [9].

Data Normalization

Microarray experiments can be conducted using samples obtained at a single time point or samples obtained over several time points. The main objective of a non-time-series microarray experiment is to compare transcriptomic levels of an organism under two or more conditions. In a time-series microarray experiment, studying the behavior of genes over time is also of interest. To make these comparisons effective, the observed data need to be normalized so that technical variations present are either removed or minimized. A critical observation, typical in two-color microarray experiments, is the nonuniform behavior of dyes at different intensity levels. Since a majority of genes are not differentially regulated under a given condition, log-ratios are expected to be spread around zero. However, microarray data usually reveal a shift and an intensity-based trend due to differences in the dye behaviors. This behavior is observed by plotting the intensity-ratio graph for log values of the product and ratio of the intensities of the two colors for each spot. Figure 3(a) and (c) shows two intensity-ratio plots, where intensity-based shifts from zero are visible. Data normalization is used to remove these biases present in the data. In addition, normalization is used to correct for biases due to the printing technology, introduced during microarray slide preparation.

Analysis of variance (ANOVA) can be used to normalize microarray data generated by different experiments [10]. This ANOVA-based normalization does not treat the microarray data merely as proportions of differential expression in the form of log ratios but instead takes the view that every measured intensity represents an abundance of mRNA. The measured intensity is subject to experimental sources of variation that have to be modeled and accounted for. The ANOVA-based approach utilizes a pair of interconnected linear models to account for both across and within gene variability. However, one main drawback of this model is its inability to remove intensity-dependent trends observed in microarray data, as shown in Figure 3(a) and (c). The local weighted linear regression (LOWESS) is appropriate for normalization of microarray data, due to its ability to remove trends in the data. The intensity values for each gene in the microarray are plotted in a two-dimensional space, represented by the log values of the product and ratio of the intensities of the two colors. Linear regression in LOWESS is performed locally, considering only those points located inside a sliding window. The local weights w_{i1} corresponding to each point (x_i, y_i) within a selected window are calculated using the tricubic function

$$w_{i1} = \left(1 - \left|\frac{x_c - x_i}{d(x_c)}\right|^3\right)^3,$$

where x_c and $d(x_c)$ correspond to the value of the independent variable at the center of the selected window and the width of the selected window, respectively. Linear regression is then performed incorporating those weights, and the resulting trend line is used to remove the bias in the data. Local computations allow LOWESS to capture intensity-dependent trends present in microarrays.

A robust version of LOWESS normalization that is more resistant to outliers than the standard LOWESS algorithm performs the smoothing through a two-step procedure. The smoothed curve obtained in the first step is used to find the residuals r_i for each data point, while a second set of weights w_{i2} are computed using

$$w_{i2} = \begin{cases} \left(1 - \left(\dfrac{r_i}{6\,\text{MAD}}\right)^2\right)^2, & \text{if } |r_i| < 6\,\text{MAD}, \\ 0, & \text{otherwise,} \end{cases}$$

to reduce the effects of outliers, where MAD denotes median absolute deviation. The final weights used to perform smoothing are the product of the two previously computed sets of weights. Usually a window size of 25–40% total points is selected.

As shown in Figure 3(b) and (d), the intensity-dependent trends in the original data shown in (a) and (c) are reduced by applying the robust LOWESS normalization.

Proteomics Data Processing

Mass-spectrometry (MS) coupled with liquid chromatography separations is the de facto platform for identifying and quantifying the protein content of an organism, known as the proteome. In bottom-up proteomics (see "Introduction to Proteomics Technologies"), the preprocessing steps for row data are significantly different from the corresponding microarray techniques. In a quantitative-proteomic experiment, protein abundances are inferred from the observed peptides (see "Glossary of Biological Terms"). One of the simplest approaches is to count the number of times a peptide is detected and accumulate those counts over all the peptides for a given protein. This count gives a value proportional to the abundance of the protein, where a higher abundant protein may have peptides that are observed more often [11]. A more accurate method for quantifying the abundance of a peptide is to calculate the peak volume or area across its retention profile, which is the retention-time versus peptide-intensity plot, obtained from the MS. The protein abundance is then inferred from the corresponding peptide abundances.

The quality assessments and normalization steps can be applied to proteomics data as well [12], [13]. One specific issue in MS-based proteomics is the extent of missing data points, which are largely due to the presence of species near the threshold for detection, leading to unbalanced data sets. A closer observation reveals that missing values result from two components. The first component relates to the intensity since lower intensity features have a higher rate of missing values. The second is a completely random component. These components can be statistically modeled, while testing for features that are significantly different among experimental conditions [14]. The resultant model can be used to infer missing data points. In contrast, microarray data contains only a few missing points.

Alternative attempts to impute missing values include techniques that are suitable for time-series data [15]. Depending on whether the missing data points are located at the end or in the middle of a time series, missing values are replaced by the closest observed data points or by interpolated values. Missing value imputation methods for microarray data can also be used for high-throughput-proteomic experiments [16]. These methods include substitutions with mean or median values, K-nearest-neighbor-based approaches where a missing point is computed as a weighted average of the observed values of the K-nearest neighbors, and singular-value-decomposition-based approaches, where missing points are determined using a linear combination of eigenvectors of the gene-expression matrix.

Inferring protein abundance from the observed peptide abundances in bottom-up experiments is another challenge unique to proteomics data. Even though peptides corresponding to a single protein are expected to show similar intensities, factors such as digestion efficiency of enzymes and electro-spray ionization properties of peptides can affect the identifications and signal intensities of peptides. Various algorithms are employed to augment the peptide intensities to corresponding protein-intensity values [12]. Outlier removal and adjustments for detection efficiencies are common to most of these techniques. For example, in the R-rollup method, peptides are first scaled based on the intensities of a reference peptide. Usually the peptide with the least amount of missing values and higher in abundance is selected as a reference. The overall abundance of the corresponding protein is then computed as the mean or median of the scaled peptide intensities. A tool that implements these methods and other data normalization steps relevant to proteomics data sets is discussed in [12].

Identification of Differentially Expressed Genes

Identification of differentially expressed genes is a classification problem, where the null hypothesis that a given gene does not show a change in its expression levels under two or more conditions is tested against the alternative hypothesis that the gene has differences in its expression levels. Depending on the design and objective of the experiment, various models of gene expressions are used for hypothesis testing. In this article, some approaches that are used to identify differentially expressed genes are presented.

STUDENT'S t-TEST

The student's t-test is a standard statistical test, which is used to identify differentially expressed genes between two distinct experimental conditions. The t-test is conducted either as a one-sample test, using log-ratios of expression values, or as a two-sample test, using absolute expression values under the two conditions.

One of the drawbacks of the t-test for identifying differentially expressed genes is that the t-test does not take into account the overall variability in measurements. Since the t-test is applied to individual genes separately and since only a few observations are available for an individual gene, the variance calculations are not stable [21]. Also the t-test does not control the false-positive error rate. To reduce the level of false positives, a threshold cutoff for the log-ratio values is typically applied together with the t-test.

ANALYSIS OF VARIANCE

Analysis of variance (ANOVA) models can be used for microarray data normalization as well as identification of differentially expressed genes [10]. ANOVA is required if the microarray study involves several factors instead of a single variable. Since ANOVA is also performed at individual gene levels, it suffers from the same drawbacks as the t-test.

FALSE DISCOVERY RATE

False discovery rate (FDR) [20] is a multiple comparison algorithm suitable for detecting differentially expressed genes from microarray data. FDR is defined as

$$\text{FDR} = \frac{\text{number of mistaken } H_0 \text{ rejections}}{\text{total number of } H_0 \text{ rejections}}.$$

The main strength of FDR is that it provides a control on the overall type-I error rate. Furthermore, compared to alternative multiple comparison procedures, such as Newman-Keuls, Bonferroni, and least significant difference, computing FDR is simpler. Additional implementations of FDR include positive-FDR (pFDR) and a statistical measure, known as the q-value, which is analogous to the p-value in FDR [20], [S10].

EXTRACTION OF DIFFERENTIAL GENE EXPRESSION

Extraction of differential gene expression (EDGE) [22] is intended to be used mainly with time-series data, though it can also be applied to non-time-series data sets. EDGE approximates data using a set of basis functions and fits a model, using either the least-squares [S11] or expectation-maximization [S12] algorithms. The null distribution of test statistics is calculated through a bootstrap procedure [S13]. Each gene is assessed either as differentially expressed or not using false-positive probability or false discovery rate [S14].

Since EDGE is optimized for detecting genes with an altered behavior over a time course, it does not pick a gene

IDENTIFICATION OF DIFFERENTIALLY REGULATED GENES

Once transcriptomics or proteomics data from relevant samples are normalized, data sets can be compared over two or more experimental conditions to identify genes with significant differential behaviors. Expression patterns of differentially expressed genes can be used to understand cellular responses under corresponding conditions. These genes are identified either at the transcriptional level using microarray measurements or at the translational level using proteomics measurements. In "Identification of Differentially Expressed Genes," some of the techniques used to identify differentially expressed genes are discussed and compared.

Depending on the design and objective of the experiment, some methods are more appropriate for detecting differentially expressed genes than others. The student's t-test (see "Identification of Differentially Expressed Genes") is used to detect differentially expressed genes between two experimental conditions. Since microarray experiments are often designed to study the effects of a single variable, such as a gene mutation, availability of a nutrient, or an environmental stress, the student's t-test is applicable to these data sets. A Bayesian probabilistic framework can be used to model gene expressions and combined with t-test to detect differentially expressed genes [17]. Where only a few samples are available, this regularized t-test identifies more genes as differentially expressed at a given false-positive error rate, compared to the standard t-test without the Bayesian framework [17]. In alternative approaches, statistical modeling of gene-expression levels uses mixture models, ANOVA [10], and linear models [18] to identify differentially expressed genes. These models quantify variances resulting from multiple factors, such as technical replicates, biological replicates, dye, and treatment. Genes with a significant variance due to treatment only are identified as differentially expressed, while the effects of all other factors are normalized.

that is up-regulated or down-regulated throughout the time course. On the other hand, since data are combined and processed as a series, EDGE can be applied to data sets with few replicates per time point.

FOURIER SCORE AND FDR

The combined Fourier-score-and-FDR approach is an example of an application-specific approach for detecting differentially expressed genes and can be used to identify genes with oscillatory behaviors including cell-cycle-regulated genes [S15] and diurnally regulated genes [24].

The Fourier score F of a signal $x(t)$, where $x(t)$ is a finite-dimensional vector representing measured expression levels of a given gene in a time-series experiment, is defined by

$$F = \sqrt{\left(\sum_t x(t)\sin \omega t\right)^2 + \left(\sum_t x(t)\cos \omega t\right)^2},$$

where ω is the angular velocity of the expected oscillations and t is the time associated with each measurement. To identify the main frequency components of a gene expression, fast Fourier transform [S16] can be performed on the mean deducted data. An oscillatory signal with the same frequency as the reference sinusoidal signal produces a larger Fourier score than a nonoscillatory signal or an oscillatory signal with a different frequency.

The significance of the Fourier scores is quantified using FDR. In FDR computations, the Fourier score for the original signal is compared with the Fourier scores of a large collection of random signals. These random signals are obtained using permutations of the original signal. An empirical FDR for a chosen threshold σ of the Fourier score can be defined as

$$\mathrm{FDR}(\sigma) = \frac{\sum_{j=1}^{M} \sum_{k=1}^{N} I(F_{j,\,k} \geq \sigma)/M}{\sum_{k=1}^{N} I(F_k^o \geq \sigma)},$$

where M, N, $F_{j,\,k}$, and F_k^o are, respectively, the number of permutations used in the null hypothesis, the total number of genes, the Fourier score for the jth random signal obtained using the expression of the kth gene, and the Fourier score for the original expression of the kth gene. The indicator function $I(x)$ takes the values

$$I(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The original expressions are scaled to have a unit standard deviation so that the Fourier scores of various genes are comparable.

REFERENCES

[S10] J. D. Storey, "A direct approach to false discovery rates," *J. R. Stat. Soc. Series B, R. Statist. Soc.*, vol. 64, pp. 479–498, 2002.

[S11] J. Wolberg, *Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments.* New York: Springer-Verlag, 2005.

[S12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. Ser. B (Methodological),* vol. 39, pp. 1–3, 1977.

[S13] A. C. Davison and D. Hinkley, *Bootstrap Methods and Their Applications.* Cambridge, U.K.: Cambridge Univ. Press, 1997.

[S14] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis, "Significance analysis of time course microarray experiments," *Proc. Nat. Acad. Sci.*, vol. 102, pp. 12,837–12,842, 2005.

[S15] M. E. Futschik and H. Herzel, "Are we overestimating the number of cell-cycling genes? The impact of background models on time-series analysis," *Bioinformatics,* vol. 24, pp. 1063–1069, 2008.

[S16] G. D. Bergland, "A guided tour of the Fast Fourier transform," *IEEE Spectr.*, vol. 6, pp. 41–52, 1969.

Since gene-by-gene hypothesis testing does not control false-positive error rate, multiple testing and p-values calculated based on permutations can be used to identify differentially expressed genes [19]. Similarly, multiple hypothesis testing algorithms are required when the objective of an experiment is to study the effects of multiple factors as opposed to a single factor. The Newman-Keuls, Bonferroni, and least-significant-difference procedures can be used for multiple comparisons [20]. Furthermore, false discovery rate (FDR) provides a practical approach for multiple comparisons [20]. Several multiple comparison procedures, including nominal p-values, family-wise error-rate control, FDR control, F-test, fixed-effects ANOVA, and mixed-model ANOVA, are reviewed in [21].

Specialized methods are available for analyzing time-series microarray data. The extraction of differential gene expression (EDGE) algorithm can be used for detecting differential behaviors from time-series microarray data [22]. Several approaches are available for identifying cell-cycle-regulated genes [23]. The Fourier-score-based criterion is combined with FDR in [24] to identify diurnally regulated genes. Diurnally regulated genes show an oscillatory behavior in their expression levels when light changes cyclically between day and night. Compared to 1445 genes selected in the original analysis [25], the Fourier-score-based criterion classifies 2138 genes representing 43% of the genome of an unicellular cyanobacterium *Cyanothece* sp. ATCC 51142, as diurnally regulated at an FDR of 2% [24].

Typically, differential behaviors of a small number of selected genes are verified by conducting experiments, such as northern blots [26], reverse transcription polymerase chain reaction (RT-PCR), and real-time polymerase chain reaction (qPCR) [27]. Depending on the results obtained from these experiments, thresholds used to identify differentially expressed genes are sometimes modified.

GENE CLUSTERING FROM HIGH-THROUGHPUT DATA

Gene Clustering

Clustering is a process of grouping data objects into disjoint classes called clusters, so that objects within a class are similar to each other, while objects in separate classes are dissimilar. Relevant to microarray and proteomics data, clustering problems can be divided into three broad groups, namely, gene-based clustering, sample-based clustering, and subspace clustering [28]. For gene-based clustering, which is the most common of the three approaches, genes are treated as data objects. Experimental conditions are considered as features used to classify the genes. For sample-based clustering, experimental conditions serve as data objects to be clustered, and genes play the role of features. Finally, subspace clustering treats genes and experimental conditions symmetrically such that either genes or samples can be regarded as objects or features.

The main goal in gene-based clustering is to identify the principal behavioral patterns in the gene-expression data and to group genes into disjoint classes based on these patterns. Gene clusters make data handling easier and usually contain information of biological significance. For example, coregulated genes, whose activities are controlled by a common promoter, tend to show similar gene-expression patterns. Therefore, coregulated genes occur in a single cluster. In addition, some of these clusters are rich in genes of specific biological functions.

When a given biological pathway responds to an external or internal cue by changing the expression of its constituent genes, the expression profile of genes from that pathway tend to behave similarly and thus are clustered together. In a scenario where the information about all of the constituents of a cluster is not known, those unknown genes can be predicted to be from the same biological pathway as most of the genes in that cluster. This approach provides a useful means for assigning functions for genes [29], [30]. However, further experiments need to be performed to demonstrate the involvement of a given gene in a particular process. Gene-based clustering is also useful for gene-regulatory network modeling and for determining appropriate model classes for the networks. Additionally, when numerical algorithms are used to identify the possible interactions among genes, clustering results are imperative for reducing the search space. A few examples are discussed in the section "Identification of Gene Interactions Using Gene-Regulatory Networks."

Clustering methods focus on two questions, namely, how to measure the similarity between expressions of two genes and how to group similar genes together while separating dissimilar genes. Although it is possible to make observations about various clustering techniques, no one method is the best since no single criterion is suitable for measuring the goodness of the resulting clusters [31].

The similarity between two gene expressions is computed using various distance measures, such as the Euclidean distance, Pearson correlation, uncentered correlation, and Hamming distance. Usually when the data are in log-ratio values, as is common in two-channel microarrays, Euclidean distance is used. When the data are expressed in absolute values, as in the case of Affymetrix microarrays (see "Introduction to Microarray Technologies"), correlation or cosine distances are preferred. Finally, Hamming distance is limited to discretized data sets. These distance measures are compared in [31].

The method of measuring the intercluster distances and intracluster distances, also known as the *linkage function*, is selected next. Linkage functions include *single linkage*, which is the smallest distance between every pair of members of two clusters, *total linkage*, which is the largest distance between every pair of members of two clusters, *average linkage*, which is the distance between centroids of two clusters, and the average distance between each pair of members of two clusters.

Clustering techniques generate clusters using various approaches. One clustering method, known as k-means [32], requires users to define the number of clusters to be generated. First the initial cluster centroids are selected randomly, uniformly, or from a subset of genes. Then the remaining genes are distributed among the clusters based on the chosen linkage function. The k-means algorithm gives rise to different clusters each time it is used, since the choices of centroids vary among runs. As a result, the algorithm is run several times, and the clusters that give the minimum average distance are chosen. The k-means clustering algorithm has several limitations. The algorithm tries to distribute all the genes among the selected number of clusters, and thus genes with distinct expression patterns frequently end up being grouped together. Also, when a dominant expression pattern is present, which often occurs in gene-expression data, multiple seeds might be obtained from those genes, and the resulting clusters from those seeds show a similar pattern.

Another widely used clustering technique is the hierarchical clustering algorithm [33]. Here, the algorithm starts by considering all of the genes as separate clusters. Then, based on the distance, the closest two genes are joined to build a single cluster. Next, this step is repeated, but now the two genes clustered together are considered as a single node. This procedure is followed until all the genes are put into one cluster. The results are usually viewed as a tree diagram. By cutting the tree at different levels, different numbers of clusters are obtained.

The self-organizing map is another clustering technique that utilizes learning algorithms used in neural networks [34]. Based on a user-defined number, nodes are initialized randomly. An iteration proceeds by picking a gene randomly and moving the nodes toward the selected gene by amounts that depend on the distances among expressions of the selected gene and the nodes. The closest node is moved the most, while the furthest node is moved the least. This iteration is repeated multiple times (20,000–50,000), at the end of which the genes are organized as clusters.

Various graph-theoretical approaches, such as cluster-identification-using-connectivity-kernels (CLICK) and cluster-affinity-search-technique (CAST), along with model-based clustering using the expectation-maximization (EM) algorithm can also be used for gene-based clustering [28]. Additional algorithms for sample-based clustering include clustering-using-iterative-feature-filtering (CLIFF), while algorithms for subspace clustering include coupled-two-way-clustering (CTWC) and biclustering [28].

Determining the Number of Clusters

Deciding the number of clusters in microarray data is often a difficult task. However, the number of clusters is required as an input to every clustering algorithm. Cross-validation techniques, such as holdout cross-validation, k-fold cross-validation, or leave-one-out cross-validation, are often used to determine the number of clusters [35]. All validation techniques use a subset of the data for cluster identification and use the remainder to evaluate the performance. In the case of gene clustering, the average distance of the remaining genes to the closest cluster is used as a performance measure. The number of clusters, for which no significant improvement in average distance is noticed by increasing the cluster number by one, is used as a criterion for selecting the optimal number of clusters. Figure 4 shows results from the analysis of a microarray data set using k-fold cross-validation and self-organizing-maps. In Figure 4(a) of the number-of-clusters versus average-distance reveals that the improvement in the average distance is gradual and there is no point at which the gradient changes sharply. For this reason, the use of cross-validation to identify the optimal number of clusters could be challenging. Therefore, estimating the number of clusters in high-throughput biological data sets remains a challenging problem.

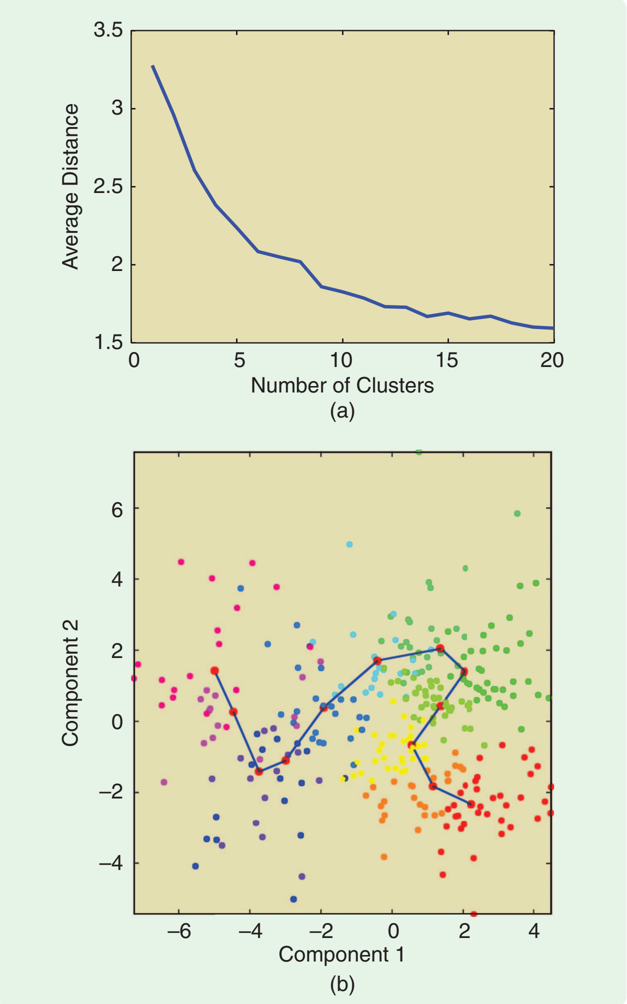

FIGURE 4 Determining the number of clusters present in a microarray data set. k-fold cross validation provides a guideline for determining the optimal number of clusters for classifying the data. The minimum number of clusters, where no significant reduction in average distance is achieved by increasing the cluster count, is selected as the optimal cluster number. For example, based on k-fold validation results shown in (a), corresponding microarray data are clustered into 12 groups using self-organizing maps. In (b), the resulting clusters are shown for a principal component space. Red points, connected by a line, identify the centroids of the resulting clusters.

Application-Based Clustering

Based on the experimental design and objective, users can define their own clustering algorithms that best serve their needs. For example, studies on diurnally regulated genes focus on genes with oscillatory expression patterns [36]. Since the oscillatory expressions in diurnal genes arise due to two main causes, an internal circadian clock and an external light, it is useful to classify diurnal genes to identify the biological processes that can be controlled by changing light-input patterns. Genes controlled by the circadian clock are not influenced by changes in light conditions, whereas genes controlled by light modify their expression patterns due to changes in frequency patterns of oscillations of the impinging light. Differences in the expression patterns under various light-input patterns can be utilized to classify the diurnal genes into two main categories, namely, circadian controlled and light controlled [24]. Using Fourier approximations, these two groups are then further clustered into subgroups based on their oscillatory periods and phases of the oscillations. In this way, genes that peak at the same time of day are grouped together. Analysis of the gene clusters reveals that genes belonging to similar gene functions peak at same time of the day. In Figure 5, various gene categories and their relative abundance levels are shown for one time point of the experiment. Figure 5

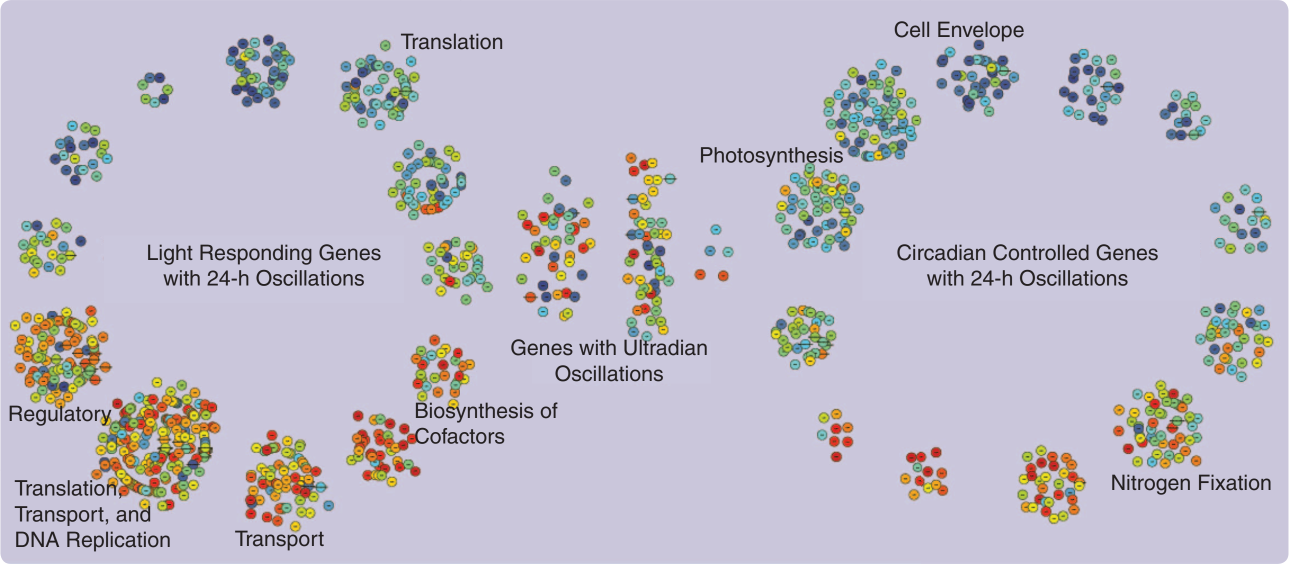
Translation
Cell Envelope
Photosynthesis
Light Responding Genes with 24-h Oscillations
Circadian Controlled Genes with 24-h Oscillations
Genes with Ultradian Oscillations
Biosynthesis of Cofactors
Regulatory
Translation, Transport, and DNA Replication
Transport
Nitrogen Fixation

FIGURE 5 Gene categories identified using the phase and frequency of the gene expressions. The circadian-controlled genes, on the right, show oscillations under both alternating and constant light, while the light-responding genes, on the left, oscillate under alternating light. These two groups, which are further clustered based on their phases of oscillations, are colored based on their activity levels, with red representing high activity and blue representing low, at a given point of time. Several genes with ultradian oscillations, that is, oscillations whose period is less than 24-h period, are also observed. This classification provides insight into how various biological processes are regulated under regular and altered light-input conditions.

shows that some of the gene clusters consist of genes from specific biological functions. This analysis also provides insight into the behavior of biological processes that are active at different times of the day.

IDENTIFICATION OF GENE INTERACTIONS USING GENE-REGULATORY NETWORKS

Although gene clustering provides insights into behavioral patterns in the gene-expression data, it does not directly identify the interactions among genes. To identify how genes affect the expression patterns of each other, the analysis must be extended to gene-regulatory networks.

As noted above, microarray experiments are used to identify interaction among genes. Approaches to modeling interactions include Boolean and Bayesian networks, generalized-logical networks, nonlinear-differential equations, piecewise-linear dynamic equations, as well as partial-differential equations [37]. However, these models are typically applied to study interactions among a few genes. Limited availability of high quality data makes modeling of large-scale gene-interaction networks difficult.

Coexpression Networks

One of the simplest approaches to modeling large-scale-gene-interaction networks is to generate coexpression networks. These networks require the lowest number of data sets compared to dynamic-systems-based interaction networks and Bayesian networks. The assumption underlying coexpression networks is that genes that are coexpressed are likely to be coregulated. In a coexpression network, genes are connected in a network by drawing links between pairs of genes that are close in terms of their expressions. Closeness is measured using one of the distance measures discussed in the section “Gene Clustering.” Whether or not to make a link between two genes depends on the threshold selected for distance. Visualization software is typically used to view the gene network. The software Cytoscape [38] can format the network in addition to displaying it. As a result, it is possible to identify groups of genes, sometimes referred to as hubs, that represent genes that are more tightly connected to each other within the group than to those outside the group. These hubs are analogous to the clusters obtained from the clustering algorithms. Coexpression networks provide biologically relevant insights. For example, the coexpression network, which is generated to identify groups of genes with similar expression patterns under high light, consists of several biologically related gene groups [39]. Further analysis of this network identifies a stress response gene present in the plant *Arabidopsis thaliana*.

Dynamic-Systems-Based Gene Interactions

Although identification of a systems-level gene-interaction network using dynamic-systems-based approaches is usually not feasible due to the limited availability of time-series data, these models can be derived in certain applications. A dynamic-system model for gene-interaction networks can be constructed using feedforward loops. The feedforward-loop-type interactions are often observed in biological systems [40]. The underlying assumption is that the rate of change of mRNA is determined by its degradation and synthesis rates. Therefore, the dynamics of different gene products can be given by

$$\dot{Y}(t) = -\alpha_y Y(t) + \beta_y f(X(t), K_{xy}), \qquad (1)$$

where $X(t)$ and $Y(t)$ represent the expression levels of the genes X and Y, respectively, at time t. The activation function $f(X(t), K_{xy}) = (X(t)/K_{xy})^H/(1 + (X(t)/K_{xy})^H)$ has two parameters, H and K_{xy}. The parameter H, which controls the steepness of f, is selected to be one or two, depending on the gene group being modeled. The parameter K_{xy} depends on both X and Y and defines the expression level of gene X required to activate the transcription of gene Y. The regulator model (1) can be extended to

$$\dot{Z}(t) = -\alpha_z Z(t) + \beta_z g(X(t), Y(t), K_{xz}, K_{yz}), \qquad (2)$$

where the two genes X and Y act together to regulate the expression level of gene Z. The genes X and Y of (2) are assumed to act independently or additively, so that $g(t)$ is selected to have the form

$$g(t) = f(X(t), K_{xz}) f(Y(t), K_{yz})$$

or

$$g(t) = f(X(t), K_{xz}) + f(Y(t), K_{yz}),$$

respectively. The first term in the right-hand side of (1) and (2) corresponds to the degradation of the gene product, while the second term corresponds to synthesis. When (1) and (2) are used to model diurnally regulated genes showing an oscillatory behavior, a gene expression $X(t)$ can be approximated using

$$X(t) = a + bt + \sum_{n=1}^{N} \alpha_n \sin(n\omega t + \phi_n),$$

whose derivative is

$$\dot{X}(t) = b + \sum_{n=1}^{N} n\alpha_n \omega \cos(n\omega t + \phi_n),$$

where ω and N represent the angular velocity of the gene expression $X(t)$ and a positive integer, respectively. The nonlinear-least-squares method can be used to find the optimal parameters in (1), minimizing the error given by

$$F(\alpha_y, \beta_y) = \| \dot{Y}(t) + \alpha_y Y(t) - \beta_y f(X(t)) \|,$$

for combinations of genes. The possibility of interactions among genes is determined by computing the normalized error, given by

$$\mathrm{E} = F(\alpha_y, \beta_y)^2_{\text{opt}} / \| \dot{Y}(t) \|^2.$$

Figure 6 shows the possible interactions among diurnally regulated genes in *Cyanothece* sp. ATCC 51142, based on the interaction model (1) and (2) [36]. Various regulatory interaction patterns including autoregulation, coherent, and incoherent feedforward loops, as well as single- and multi-input regulations are identified among the genes shown in Figure 6. One of the advantages of a feedforward-loop-based model is that it accommodates possible time delays between the activation of a regulator gene and its effect on the target gene. These types of delays occur as a result of time differences between transcription and translation rates, the transport of proteins in the cell, and the post-transcriptional/translational modifications [see Figure 7(b)]. Although not explicitly visible, Figure 6 shows that the model captures time delays among interacting genes. The model also provides directionality among interacting genes explicitly identifying the regulators and their targets.

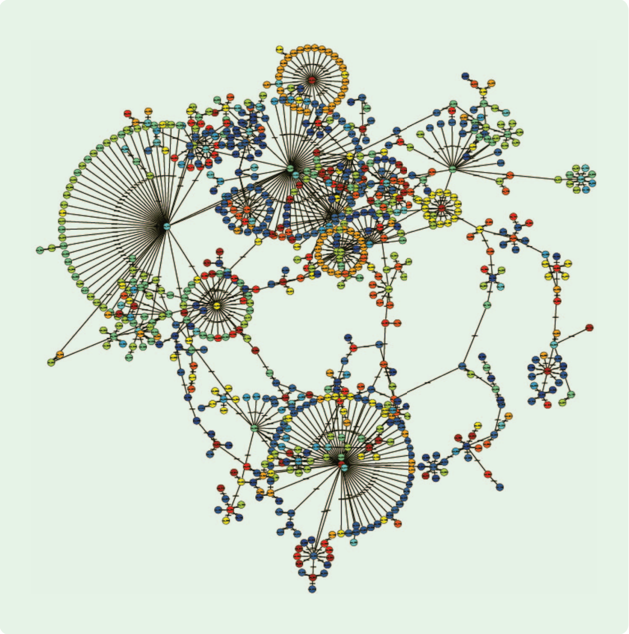

FIGURE 6 Gene-regulatory network showing possible regulatory links among diurnally expressed genes in *Cyanothece* sp. ATCC 51142. Genes are colored based on their relative expression levels at a given point in time. Regulatory links are identified using the dynamic-system model (1), (2). Various regulatory relationship structures, already characterized in other biological systems, are identified in the network. These regulatory relationships include autoregulation, coherent and incoherent feedforward loops, as well as single- and multi-input regulations [40]. The dynamic model can accommodate possible time delays among regulator and target genes. The network suggests a complex level of interaction among genes in a cell.

Stochastic Approaches: Bayesian Networks

Bayesian networks [41] can be used for identifying relationships among genes or gene functions from microarray data. One of the hurdles in applying Bayesian network modeling in biology is the small number of observations compared to the number of variables in the system. Also, learning the structure of a network with more than 20 variables is a computationally challenging problem.

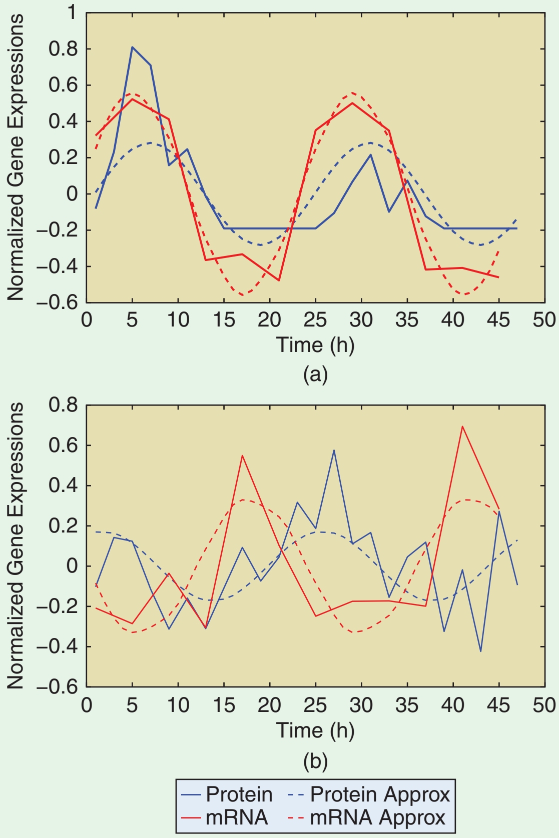

FIGURE 7 Two genes with oscillatory behavior in both mRNA and protein abundance levels. The peak times for mRNA and proteins can be shifted. The gene shown in (a) shows almost no time delay between the peak times of its mRNA and protein abundance expressions. In contrast, the gene shown in (b) has a significant delay in its protein expression profile, compared to the mRNA expression profile. In gene-interaction modeling, it is necessary to account for these behavioral differences. (Solid lines indicate actual observations, while dashed lines are approximate expressions based on sinusoidal functions.)

Bayesian networks are used to model interactions among genes in several organisms [37]. However these networks are limited to a few selected genes, since the search space for possible network structures grows exponentially with the number of nodes in the network [42]. A Bayesian network, containing 800 cell-cycle-regulated genes, is derived by limiting the search space using correlation measurements [43]. Instead of treating individual genes as network nodes, Bayesian networks can be derived at biological-process levels, which reduces the number of nodes in the network [44]. The strengths of connections among nodes in the network can be quantified using entropy measurements.

Due to its underlying probabilistic framework, Bayesian networks can be applied to data sets coming from multiple sources, experiments, and platforms. Further, non-time-series data can be combined with time-series data and used in the analysis. For these reasons, data representing multiple environmental conditions can be considered simultaneously in deriving relationships among genes, increasing the possibility of derived connections being biologically relevant. On the other hand, since a large amount of data is required to calculate the probability, use of Bayesian networks is limited to well-studied organisms.

From the point of view of systems biology, a model must mimic the molecular biological interactions in as much detail as possible. The chemical-master-equation models [45], [46] provide a stochastic description of dynamics of gene expressions. Estimating the parameters of these fine-scale models requires large time-series data sets made with cell-specific measurements as opposed to tissue-averaged measurements. A model of a tissue-averaged-measurement data can be provided by deterministic differential equations [37], [47]. If the network structure and the experimental data are limited, a Boolean network can provide an appropriate coarse model [48].

BIOLOGICAL RELEVANCE OF GENE CLUSTERS AND GENE-REGULATORY NETWORKS

The final objective of gene clustering and regulatory networks is identifying interaction patterns among genes using the transcription- and translation-level data. Genes that show coexpressions under few experimental conditions are not always coregulated. On the other hand, genes under the control of a single regulator are likely to show similar behaviors and thus are being coexpressed. As a result, coexpressed genes are a useful starting point for isolating true regulatory relationships.

One approach to reducing the number of false links from coexpression networks is to include data from as many experimental conditions as possible. When two genes are not interacting in a biological process, the probability that they are coexpressed is low. Gene-regulatory networks that have the ability to predict and simulate various cellular responses can be obtained from correlation-based approaches using large microarray data sets [49]. However, when data is obtained from multiple sources, the differences in the experimental conditions and microarray platforms can make the analysis difficult. Despite variations in the data sources, data from multiple sources can be combined using probabilistic algorithms to obtain biologically meaningful gene networks [44].

The use of existing biological knowledge of relationships among genes and metabolic pathways is another way to reduce the false positives in the coexpression networks and to identify the main players of gene regulation. Data mining techniques are increasingly being used to identify the relationships among genes reported in the literature [50]. For example, data mining is employed to identify a gene, RRTF1, with a crucial role in stress responses in the plant *Arabidopsis thaliana* [39].

Motif Identification

Transcription factors are proteins that regulate the transcription of genes by targeting specific regions of DNA also know as promoter regions (see Figure 1). These DNA sequences in the promoter region of a gene are referred to as binding-site motifs. Transcription factors induce or suppress gene expressions by binding to the promoter regions of genes they regulate. Although promoter regions are typically located upstream of a given gene, their exact locations on the DNA with respect to the corresponding genes vary. The presence of conserved sequences in the upstream region of a group of genes suggests that these genes might be regulated by the same transcription factor. In addition to using clustering or gene interaction modeling algorithms, the identification of conserved sequences is used as an independent criterion to suggest that a pair of genes is coregulated.

The identification of binding-site motifs is not a trivial task since the motifs are not always identical among genes. Several methods are available for identifying these motifs [51]–[54]. Many motif-search algorithms analyze upstream regions of a group of coexpressed genes, discovered using clustering or transcription networks, and search for conserved regions within them [1], [51]. Since an exhaustive search over all combinations is usually computationally intractable, these procedures employ heuristic search techniques, such as greedy algorithms [52]. In contrast, algorithms based on dictionary-building models can be used to search upstream regions of all genes and identify over-represented sequences and groups of genes that contain those sequences in their upstream regions [53].

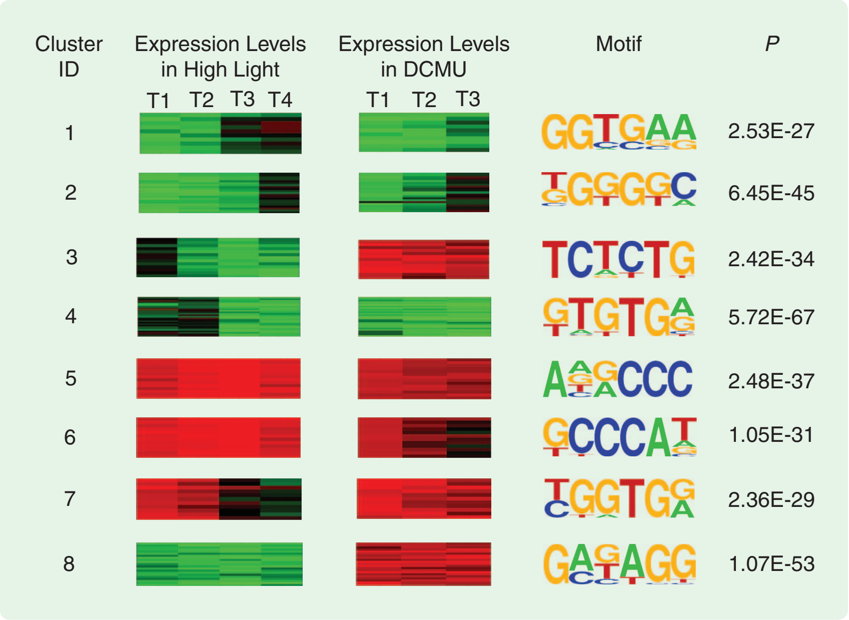

FIGURE 8 Promoter region motifs identified among coexpressed genes in *Arabidopsis thaliana*. Genes are first clustered based on their mRNA expressions under two experimental conditions, namely, high light and DCMU treatment. These two conditions affect the photosynthesis process in plants. Genes that behave similarly each other under multiple experimental conditions are more likely to be coregulated, but not just coexpressed, and serve as good candidates for motif search algorithms. The upstream regions of the genes in each cluster are analyzed using the consensus algorithm [1]. The third and fourth columns show the regulatory region motifs and corresponding significance values obtained for each cluster. Letter sizes in the regulatory region motifs represent the percentage of times each nucleotide is conserved at each location.

Consensus [1] uses a greedy algorithm to search and align conserved sequences in a set of upstream DNA sequences so that the final alignment matrix maximizes the information content. This method can be applied to find the conserved sequences in the upstream regions of the genes with similar expression patterns [39]. Figure 8 shows the expression levels of gene clusters under two experimental conditions and the conserved regulatory regions identified by the algorithm. These sequences are highly specific to corresponding genes, as observed by their low p-values. The discovered sequences can be verified experimentally by identifying binding sites of proteins of interest utilizing ChIP-Chip experiments [7].

Combining Proteomics Data with Microarray Data

The integration of data obtained from proteomics analysis further improves the understanding of cellular responses to various environmental conditions. Proteomics data sets are useful in improving the accuracy of the gene-regulatory networks derived using transcriptional data only. Proteomics data also provide insights into regulatory responses implemented by the cells. For example, only 5% of the genes in *Cyanothece* sp. ATCC 51142 are cyclic at the protein level [55], whereas more than 40% the genes show cyclic behavior at the transcriptional level [24].

Significant time delays between changes in mRNA and corresponding protein levels are also detected in *Cyanothece* sp. ATCC 51142 [55] [see Figure 7(b)]. In bacteria, transcription and translation occurs concomitantly, and time delays among mRNA and protein expressions are not anticipated. These time delays can arise possibly due to various post-transcriptional regulations, as well as changes in synthesis and degradation rates for corresponding mRNA and proteins. The dynamic-system model (1), (2) is able to capture some of these delays in the resulting network (see Figure 6). However, these interaction models, constructed from the microarray data, need to be improved to explicitly accommodate protein levels expressions.

CONCLUSIONS

This article presents a broad overview of techniques, both analytical and technological, that are used in systems biology to analyze large-scale data sets. While focusing on the computational aspects, the associated biology of cell regulation is also discussed. Discussions are centered around two widely used high-throughput technologies, namely, DNA microarrays and MS-based proteomics. The biological principles behind these techniques are presented. Typical steps involved in generation of high-throughput data to final biological interpretations of analyzed results are discussed. Various attempts to derive systems-level biological networks utilizing microarray and proteomics data are also presented.

There are several current developments in the fields of high-throughput microarray and proteomic technologies. Protein-binding microarrays are being developed to study DNA-binding properties of regulatory proteins [56]. Three-color microarrays, which can be used to study drug effects by probing healthy, sick, and drug-treated samples at the same time, are constructed mainly for medical applications [57]. Further, three-dimension microarrays are being developed to increase sensitivity and signal-to-noise ratio levels [58]. However, for various reasons, such as cost and complexity of the procedures required for manufacturing and data generation, three-dimensional microarrays are not as popular as their two-dimensional counterparts. In the field of proteomics, developments are focused on improving the sensitivity and the coverage of detected proteins. In addition, many data analysis tools are being developed to efficiently process proteomics data [59].

In addition to the two high-throughput technologies, several other technologies are used to produce data at different cellular levels. These technologies include genome-sequencing [2], metabolomics [60], and metabolic engineering [61]. Combining data generated by numerous technologies is significantly challenging, and control engineering plays a vital role in this challenge. Furthermore, various regulatory mechanisms beyond the central dogma of molecular biology, such as reverse transcriptase, RNA interference, and DNA methylation, are now identified [62], [63]. Gene-regulatory models need to be able to account for these regulatory mechanisms as well. Control engineering promises to play a crucial role in metabolic engineering. Currently, most large-scale metabolic models make steady-state assumptions and are solved using flux balance analysis [64]. Moving from steady-state to non-steady-state analysis requires theoretical and computational advances. Contributions of control engineering in the field of systems biology include development of artificial gene circuits [65], rapid modifications of genome sequences to study effects of mutations [66], and modeling circadian clocks [67]. However, these applications focus on a small set of genes, which is typically fewer than 20. Large-scale and genome-level computational models for cellular systems are still in the early stages. As a result, numerous modeling opportunities exist to contribute to the advancement of the field of systems biology.

ACKNOWLEDGMENTS

This work is part of a Membrane Biology EMSL Scientific Grand Challenge project at the W.R. Wiley Environmental Molecular Sciences Laboratory, sponsored by the U.S. Department of Energy's Office of Biological and Environmental Research (BER) program. This material is based on work supported in part by the National Science Foundation under grant 0425749. We would like to thank Jim Ballard, Gayle Geren, and Jeffrey Cameron of Washington University in St. Louis, for their help in proofreading this article.

AUTHOR INFORMATION

Thanura R. Elvitigala (telvitigala@wustl.edu) received the B.S. in electronic and telecommunication engineering from the University of Moratuwa, Sri Lanka, in 2006 and the M.S. in electrical engineering and the Ph.D. in systems science and mathematics from Washington University, St. Louis, Missouri, in 2006 and 2009, respectively. He is currently a postdoctoral research associate in the Department of Biology, Washington University, St. Louis. His research interests are in the areas of computational and systems biology, biological oscillations, and modeling of large biological systems. He can be contacted at Washington University in St. Louis, Campus Box 1137, One Brookings Drive, St. Louis, MO 63130 USA.

Ashoka D. Polpitiya received the B.S. in electrical and electronic engineering from the University of Peradeniya, Sri Lanka, in 1996 and the M.S. and D.Sc. in systems science and mathematics from Washington University, St. Louis, in 2000 and 2004, respectively. He was a postdoctoral research associate at the Department of Surgery, Washington University, School of Medicine, St. Louis, from 2004 to 2006. From 2006 to 2009, he was a senior research scientist with the biological separations and mass spectrometry group at Pacific Northwest National Laboratory in Richland, Washington. Currently, he leads the bioinformatics group for the Center for Proteomics at Translational Genomics Research Institute in Phoenix, Arizona. His research interests are in the areas of proteomics, systems biology, application of control theory to biological systems, and geometric control.

Wenxue Wang received the B.S. in automatic control from Beijing Institute of Technology in 1996, the M.S. in control theory from the Institute of Systems Science, Chinese Academy of Sciences, Beijing, in 1999, and the M.S. and the D.Sc. in systems science and mathematics from Washington University, St. Louis, in 2002 and 2006. He is currently a postdoctoral research fellow with the Institute for Collaborative Biotechnologies at the University of California, Santa Barbara. His research interests are in the areas of computational neuroscience, neural networks, systems biology, signal analysis and estimation, and application of systems science to biological systems.

Jana Stöckel received the M.S. in biology and Ph.D. in plant biology from the Friedrich-Schiller University in Jena, Germany. She is currently a research scientist in the Biology Department at Washington University, St. Louis. Her research interests are in the areas of molecular plant biology and systems biology.

Abha Khandelwal received the B.S. in chemistry from Benaras Hindu University, Varanasi, India, the M.S. in molecular biology and genetic engineering from G.B. Pant University of Agriculture and Technology, Pantnagar, India, and the Ph.D. from the Indian Institute of Science, Bangalore, India in 1995, 1997, and 2003, respectively. She was a research scientist at Washington University, St. Louis, from 2004 to 2009. She is currently a research scientist with Monsanto, St. Louis. Her research interests are in the field of plant biology, plant homologues of presenilin, cellular and molecular mechanisms of drought tolerance, and understanding redox homeostasis utilizing a systems approach.

Ralph S. Quatrano received the A.B. from Colgate University in 1962, the M.S. from Ohio University, and the Ph.D. in biology from Yale University in 1968. He was appointed the Spencer T. Olin Professor and chair of the Department of Biology at Washington University, St. Louis, in 1998. He is currently the dean of Engineering and Applied Sciences. He was the editor-in-chief of *The Plant Cell* (1998–2003) and was an editor for *Science* (1991–1998). He is a fellow of the American Association for the Advancement of Science, the St. Louis Academy of Science, and the American Society of Plant Biologists. His research interests include cellular and molecular mechanisms controlling the growth and development of seed plants. He is the author of over 165 research articles.

Himadri B. Pakrasi is the George William and Irene Koechig Freiberg Professor of Biology in Arts and Sciences and professor of energy in the School of Engineering and Applied Sciences, Washington University, St. Louis. He received undergraduate and graduate degrees in physics at the Presidency College and University of Calcutta. In 1984, he received the Ph.D. from the University of Missouri-Columbia, and he has been on the faculty of Washington University since 1987. His research interests include photosynthetic processes, in particular, membrane protein complexes in cyanobacteria and plant chloroplasts. He is currently the director of the International Center for Advanced Renewable Energy and Sustainability at Washington University.

Bijoy K. Ghosh received the B.Tech. from Birla Institute of Technology and Science, Pilani, India, in 1977 and the M.Tech. from the Indian Institute of Technology, Kanpur, India, in 1979, both in electrical and electronics engineering. He received the Ph.D. in engineering from Harvard University, Cambridge, Massachusetts, in 1983. From 1983 to 2006, he was a faculty member in the Department of Electrical and Systems Engineering, Washington University, St. Louis, and directed the Center for BioCybernetics and Intelligent Systems. Presently, he is a Dick and Martha Brooks Regent Professor in the Department of Mathematics and Statistics, Texas Tech University, Lubbock. His current research interests are in machine vision, computational neuroscience, and bioinformatics.

REFERENCES

[1] G. Z. Hertz and G. D. Stormo, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences," *Bioinformatics*, vol. 15, pp. 563–577, 1999.

[2] D. MacLean, J. D. G. Jones, and D. J. Studholme, "Application of 'next-generation' sequencing technologies to microbial genetics," *Nat. Rev. Microbiol.*, vol. 7, pp. 287–296, 2009.

[3] B. Snel, M. A. Huynen, and B. E. Dutilh, "Genome trees and the nature of genome evolution," *Annu. Rev. Microbiol.*, vol. 59, pp. 191–209, 2005.

[4] L. Stein, "Genome annotation: From sequence to biology," *Nat. Rev. Genet.*, vol. 2, pp. 493–503, 2001.

[5] A. Schulze and J. Downward, "Navigating gene expression using microarrays—A technology review," *Nat. Cell. Biol.*, vol. 3, pp. E190–E195, 2001.

[6] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster, "Quantitative mass spectrometry in proteomics: A critical review," *Anal. Bioanal. Chem.*, vol. 389, pp. 1017–1031, 2007.

[7] M. J. Buck and J. D. Lieb, "ChIP-chip: Considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments," *Genomics*, vol. 83, pp. 349–360, 2004.

[8] G. A. Churchill, "Fundamentals of experimental design for cDNA microarrays," *Nat. Genet.*, vol. 32, pp. 490–495, 2002.

[9] T. Wilkes, H. Laux, and C. A. Foy, "Microarray data quality: Review of current developments," *OMICS*, vol. 11, pp. 1–13, 2007.

[10] R. D. Wolfinger, G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. S. Paules, "Assessing gene significance from cDNA microarray expression data via mixed models," *J. Comput. Biol.*, vol. 8, pp. 625–637, 2001.

[11] Y. Zhang, Z. Wen, M. P. Washburn, and L. Florens, "Effect of dynamic exclusion duration on spectral count based quantitative proteomics," *Anal. Chem.*, vol. 81, pp. 6317–6326, 2009.

[12] A. D. Polpitiya, W. J. Qian, N. Jaitly, V. A. Petyuk, J. N. Adkins, D. J. Camp, II, G. A. Anderson, and R. D. Smith, "DAnTE: A statistical tool for quantitative analysis of omics data," *Bioinformatics*, vol. 24, pp. 1556–1558, 2008.

[13] S. J. Callister, R. C. Barry, J. N. Adkins, E. T. Johnson, W. Qian, B. M. Webb-Robertson, R. D. Smith, and M. S. Lipton, "Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics," *J. Proteome Res.*, vol. 5, pp. 277–286, 2006.

[14] Y. Karpievitch, J. Stanley, T. Taverner, J. Huang, J. N. Adkins, C. Ansong, F. Heffron, T. O. Metz, W. J. Qian, H. Yoon, R. D. Smith, and A. R. Dabney, "A statistical framework for protein quantitation in bottom-up MS-based proteomics," *Bioinformatics*, vol. 15, pp. 2028–2034, 2009.

[15] X. Du, S. J. Callister, N. P. Manes, J. N. Adkins, R. A. Alexandridis, X. Zeng, J. H. Roh, W. E. Smith, T. J. Donohue, S. Kaplan, R. D. Smith, and M. S. Lipton, "A computational strategy to analyze label-free temporal bottom-up proteomics data," *J. Proteome Res.*, vol. 7, pp. 2595–2604, 2008.

[16] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, pp. 520–525, 2001.

[17] P. Baldi and A. D. Long, "A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, pp. 509–519, 2001.

[18] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Stat. Appl. Genet. Mol. Biol.*, vol. 3, no. 1, article 3, 2004.

[19] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Stat. Sin.*, vol. 12, pp. 111–139, 2002.

[20] D. Curran-Everett, "Multiple comparisons: Philosophies and illustrations," *Amer. J. Physiol. Regul. Integr. Comp. Physiol.*, vol. 279, pp. R1–R8, 2000.

[21] X. Cui and G. A. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biol.*, vol. 4, p. 210, 2003.

[22] J. T. Leek, E. Monsen, A. R. Dabney, and J. D. Storey, "EDGE: Extraction and analysis of differential gene expression," *Bioinformatics*, vol. 22, pp. 507–508, 2006.

[23] U. de-Lichtenberg, J. Jensen, A. Fausboll, T. S. Jensen, P. Bork, and S. Brunak, "Comparison of computational methods for the identification of cell cycle-regulated genes," *Bioinformatics*, vol. 21, pp. 1164–1171, 2005.

[24] T. R. Elvitigala, J. Stöckel, B. K. Ghosh, and H. B. Pakrasi, "Effect of continuous light on diurnal rhythms in Cyanothece sp. ATCC 51142," *BMC Genomics*, vol. 10, p. 226, 2009.

[25] J. Stockel, E. A. Welsh, M. Liberton, R. Kunnvakkam, R. Aurora, and H. B. Pakrasi, "Global transcriptomic analysis of Cyanothece 51142 reveals robust diurnal oscillation of central metabolic processes," *Proc. Nat. Acad. Sci.*, vol. 105, pp. 6156–6161, 2008.

[26] J. C. Alwine, D. J. Kemp, and G. R. Stark, "Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes," *Proc. Nat. Acad. Sci.*, vol. 74, pp. 5350–5354, 1977.

[27] S. A. Bustin, "Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays," *J. Mol. Endocrinol.*, vol. 25, pp. 169–193, 2000.

[28] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene-expression data: A survey," *IEEE Trans. Knowledge Data Eng.*, vol. 16, pp. 1370–1386, 2004.

[29] L. F. Wu, T. R. Hughes, A. P. Davierwala, M. D. Robinson, R. Stoughton, and S. J. Altschuler, "Large-scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters," *Nat. Genet.*, vol. 31, pp. 255–265, 2002.

[30] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, Jr., and D. Haussler, "Knowledge-based analysis of microarray gene-expression data by using support vector machines," *Proc. Nat. Acad. Sci.*, vol. 97, pp. 262–267, 2000.

[31] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Networks*, vol. 16, pp. 645–678, 2005.

[32] J. B. McQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, 1967, vol. 1, pp. 281–297.

[33] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley, 1990.

[34] S. Haykin, *Neural Networks—A Comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[35] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artificial Intelligence*, 1995, vol. 2, pp. 1137–1143.

[36] T. R. Elvitigala, H. B. Pakrasi, and B. K. Ghosh, "Dynamic network modeling of diurnal genes in Cyanobacteria," in *Emergent Problems in Nonlinear Systems and Control*, B. K. Ghosh, C. F. Martin, and Y. Zhou, E ds. vol. 393. New York: Springer-Verlag, 2009, pp. 21–41.

[37] H. D. Jong, "Modeling and simulation of genetic regulatory systems: A literature review," *J. Comput. Biol.*, vol. 9, pp. 67–103, 2002.

[38] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Res.*, vol. 13, pp. 2498–2504, 2003.

[39] A. Khandelwal, T. R. Elvitigala, B. K. Ghosh, and R. S. Quatrano, "Arabidopsis transcriptome reveals control circuits regulating redox homeostasis and the role of an AP2 transcription factor," *Plant Physiol.*, vol. 148, pp. 2050–2058, 2008.

[40] A. Uri, *An Introduction to Systems Biology: Design Principles of Biological Circuits*. London, U.K.: Chapman & Hall, 2006.

[41] D. Heckerman, "A tutorial on learning with Bayesian networks," in *Learning in Graphical Models*, M. I. Jorden, Ed. Dordrecht, The Netherlands: Kluwer, 1998.

[42] F. V. Jensen and T. D. Nielsen, *Bayesian Networks and Decision Graphs*. New York: Springer-Verlag, 2007.

[43] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *J. Comput. Biol.*, vol. 7, pp. 601–620, 2000.

[44] A. K. Singh, T. Elvitigala, J. C. Cameron , B. K. Ghosh, M. Bhattacharyya-Pakrasi, and H. B. Pakrasi, "Integrative analysis of large scale expression profiles reveals core transcriptional response and coordination between multiple cellular processes in a cyanobacterium," *BMC Syst. Biol.*, vol. 4, no. 105, 2010.

[45] D. T. Gillespie, "A rigorous derivation of the chemical master equation," *Physica A*, vol. 188, pp. 404–425, 1992.

[46] H. H. McAdams and A. Arkin, "Stochastic mechanisms in gene expression," *Proc. Nat. Acad. Sci.*, vol. 94, pp. 814–819, 1997.

[47] P. Smolen, D. A. Baxter, and J. H. Byrne, "Modeling transcriptional control in gene networks—Methods, recent results and future directions," *Bull. Math. Biol.*, vol. 62, pp. 247–292, 2000.

[48] I. Shmulevich, E. R. Dougherty, and W. Zhang, "From Boolean to probabilistic Boolean networks as models of genetic regulatory networks," *Proc. IEEE*, vol. 90, pp. 1778–1792, 2002.

[49] R. Bonneau, M. T. Facciotti, D. J. Reiss, A. K. Schmid, M. Pan, A. Kaur, V. Thorsson, P. Shannon, M. H. Johnson, J. C. Bare, W. Longabaugh, M. Vuthoori, K. Whitehead, A. Madar, L. Suzuki, T. Mori, D. Chang, J. DiRuggiero, C. H. Johnson, L. Hood, and N. S. Baliga, "A predictive model for transcriptional control of physiology in a free living cell," *Cell*, vol. 131, pp. 1354–1365, 2007.

[50] W. Dubitzky, M. Granzow, and D. P. Berrar, *"Fundamentals of data mining in genomics and proteomics."* New York: Springer-Verlag, 2007.

[51] A. F. Neuwald, J. S. Liu, and C. E. Lawrence, "Gibbs motif sampling: Detection of bacterial outer membrane protein repeats," *Protein Sci.*, vol. 4, pp. 1618–1632, 1995.

[52] G. D. Stormo and G. W. Hartzell, III, "Identifying protein-binding sites from unaligned DNA fragments," *Proc. Nat. Acad. Sci.*, vol. 86, pp. 1183–1187, 1989.

[53] G. Wang, T. Yu, and W. Zhang, "WordSpy: Identifying transcription factor binding motifs by building a dictionary and learning a grammar," *Nucl. Acids Res.*, vol. 33, pp. 412–416, 2005.

[54] C. Sabatti and K. Lange, "Genomewide motif identification using a dictionary model," *Proc. IEEE*, vol. 90, pp. 1803–1810, 2002.

[55] J. Stöckel, T. R. Elvitigala, M. Liberton, J. Jacobs, E. Welsh, B. K. Ghosh, and H. B. Pakrasi, "Diurnal rhythms result in significant changes in the cellular protein complement in the cyanobacterium Cyanothece sp. ATCC 51142," *Mol. Cell. Proteomics*, to be published.

[56] M. L. Bulyk, "Protein binding microarrays for the characterization of DNA-protein interactions," *Adv. Biochem. Eng. Biotechnol.*, vol. 104, pp. 65–85, 2007.

[57] M. J. Hessner, X. Wang, K. Hulse, L. Meyer, Y. Wu, S. Nye, S. Guo, and S. Ghosh, "Three color cDNA microarrays: Quantitative assessment through the use of fluorescein-labeled probes," *Nucl. Acids Res.*, vol. 31, p. e14, 2003.

[58] B. J. Cheek, A. B. Steel, M. P. Torres, Y. Yu, and H. Yang, "Chemiluminescence detection for hybridization assays on the flow-thru chip, a three-dimensional microchannel biochip," *Anal. Chem.*, vol. 73, pp. 5777–5783, 2001.

[59] J. C. Wright and S. J. Hubbard, "Recent developments in proteome informatics for mass spectrometry analysis," *Comb. Chem. High Throughput Screen.*, vol. 2, pp. 194–202, 2009.

[60] W. Weckwerth, "Metabolomics in systems biology," *Annu. Rev. Plant Biol.*, vol. 54, pp. 669–689, 2003.

[61] H. Kitano, "Systems biology: A brief overview," *Science*, vol. 295, pp. 1662–1664, 2002.

[62] S. Henikoff, "Beyond the central dogma," *Bioinformatics*, vol. 18, pp. 223–225, 2002.

[63] A. Razin and H. Cedar, "DNA methylation and gene expression," *Microbiol. Mol. Biol. Rev.*, vol. 55, pp. 451–458, 1991.

[64] J. S. Edwards, M. Covert, and B. Palsson, "Metabolic modelling of microbes: The flux-balance approach," *Environ. Microbiol.*, vol. 4, pp. 133–140, 2002.

[65] T. S. Gardner, C. R. Cantor, and J. J. Collins, "Construction of a genetic toggle switch in Escherichia coli," *Nature*, vol. 403, pp. 339–342, 2000.

[66] H. H. Wang, F. J. Isaacs, P. A. Carr, Z. Z. Sun, G. Xu, C. R. Forest, and G. M. Church, "Programming cells by multiplex genome engineering and accelerated evolution," *Nature*, vol. 460, pp. 894–898, 2009.

[67] H. P. Mirsky, A. C. Liu, D. K. Welsh, S. A. Kay, and F. J. Doyle, III, "A model of the cell-autonomous mammalian circadian clock," *Proc. Nat. Acad. Sci.*, vol. 106, pp. 11,107–11,112, 2009.