New technologies that analyze the behavior of thousands of genes in parallel are creating, for the first time, a foundation of data for building integrated models of cellular processes. This review discusses the general issues of utilizing genomic methods in fundamental and applied research settings, using the study of stress responses and improvement of secondary metabolite production as examples. A fusion of concepts from biological and nonbiological disciplines, including mathematics, computer science, physics, chemistry, and engineering, is required to address the theoretical and experimental challenges facing the field of genomics, and together promise great breakthroughs in our understanding and engineering of cellular systems.

Foreword
The availability of complete sequence information for many different organisms is driving a revolution in the biological sciences. For the first time, technologies that analyze thousands of genes in parallel are generating comprehensive, high-resolution, and quantitative information on the cellular states of living organisms. The widespread and routine use of functional genomic tools promises to shed light on important questions in virtually all arenas, from the fundamental issue of how cells grow, to the medical challenge of understanding cancer, and to the industrial goal of developing fermentation processes.

Numerous excellent reviews have been published on genome-wide technologies, including refs 1–5. This review is targeted toward the uninitiated and will refer to other reviews throughout the text. The primary technologies and their capabilities will first be presented, followed by a discussion of their application to (1) secondary metabolite fermentation development and (2) understanding cellular responses to stress.

Introduction
The multitude of genome projects in the past decade has generated vast amounts of DNA sequence information and, at the same time, highlighted our lack of understanding of how cells function. The genomes of even the simplest organisms are complex—more than half of the 6200 Saccharomyces cerevisiae genes had completely unknown functions upon completion of sequencing (6)—and the genomes of higher organisms are orders of magnitude more complex: >80% of sequenced human genes have no known function, and humans have an estimated 100,000 genes currently being sequenced. To address this knowledge deficit, genomic technologies have been developed that monitor thousands of genes in a single experiment. Their specific utilities are diverse, ranging from the rapid characterization of genes with poorly understood functions, to the identification of cellular responses under different conditions, to diagnostic and quality-control capabilities based on patterns of genomic data. Indeed, the progenitors of these technologies, such as the construction and analysis of mutant strains and northern blots to measure gene expression, have already proven their value to biology and biotechnology. By conducting thousands of such analyses in parallel, genome-wide experiments are generating large and complete datasets that are providing a foundation for integrated models of cellular behavior.

Genomic technologies can be broadly divided into two groups—those that alter gene structure and deduce information from altered gene function and those that observe the behavior of intact genes. These two groups will be discussed in sequence below.

Methods That Alter Gene Structure: Essential Genes and More. These methods involve the random or systematic alteration of genes across an organism’s genome to obtain functional information. As an illustration, a scheme to identify genes essential for a cellular process, e.g., growth under optimal conditions, is shown in Figure 1. A haploid cell population first undergoes a mutagenesis procedure so that each cell contains a different disrupted gene. The cell population must be large enough so that every gene disruption is represented at least once, i.e., no. of cells \( \geq \) no. of genes in the genome. The population is then subjected to a selection, such as growth in rich media, that distinguishes cells competent and dysfunctional in the cellular process of interest. Here, cells that contain disruptions in genes important for growth are depleted from the population due to, and in proportion to, their decrease in growth rate. Cells with nonessential gene disruptions have normal growth rates and remain in the population. Cells with beneficial gene disruptions have enhanced growth rates and are enriched in the population (not shown). After the selection, the genomes of the initial and final cell populations are assessed by a genomic method (see below) to quantitatively measure the relative enrichment or depletion of every disrupted gene.

Current genome-altering technologies are distinguished by their strategies for mutagenizing and analyzing cell populations in a genome-wide manner. For S. cerevisiae, two major strategies, which are being applied to other sequenced microorganisms such as Escherichia coli (7), have been adopted to identify essential genes. One strategy, called genetic footprinting, generates gene disruptions in a cell population using transposon mutagenesis, which inserts a 6 kb DNA fragment stochasti-
allow comparative fitness measurements between individuals of the population, e.g. of small differences in growth rates and nonlinear differences in survival. An example of a selection for secondary metabolite production is presented in the next section.

To subject an organism to these types of genome-altering analyses, they must satisfy several criteria. First, techniques must be available to generate gene alterations across the organism's genome, either by transposon-based or efficient recombinant DNA methods. Second, large cell populations of the organism must be easily manipulated, since population sizes that include alterations of every gene are required. Finally, these experiments are most conveniently conducted on haploid organisms, in which recessive mutations and essentiality are easily detected. Consequently, the current genome-altering strategies described above are primarily applicable to microorganisms and less so for plant and animal models. However, genes responsible for specific cellular phenomena in higher eucaryotes, such as genes responsible for human diseases, can be identified using genomic technologies that compare the genomes of two individuals (see next section) (15–18). While such approaches do not identify all essential genes for a defective cellular process, the identification of genetic defects in diseased individuals should accelerate progress toward understanding and treating human diseases.

**Methods That Observe Intact Genes: Gene Expression and More.** Several methods have been developed to efficiently monitor the behavior of thousands of intact genes. DNA chips are tools that fractionate a heterogeneous DNA mixture into known DNA components, while a complementary method, called SAGE (Serial Analysis of Gene Expression), has been important for identifying transcripts not predicted by sequence information alone (19).

The two commonly available DNA chips are oligonucleotide chips and DNA microarrays. Oligonucleotide chips are 1.3 cm x 1.3 cm arrays of 25-mer oligonucleotides, synthesized on a glass substrate using photolithographic techniques derived from the semiconductor industry (20–22). Each gene is analyzed by ~20 oligonucleotide pairs of perfect and single-mismatch sequences, which are used to distinguish specific vs non-specific DNA hybridization. Currently one chip contains oligonucleotides that represent up to 1700 genes (23). DNA microarrays are 2 cm x 2 cm arrays of DNA samples, typically PCR product, that are robotically printed onto a microscope slide (24–26). Each gene is analyzed by a full or partial-length target DNA sequence, and microarrays with up to 10,000 genes are now routinely printed commerically (27) and within academic labs (28). A primary technical difference between oligonucleotide chips and DNA microarrays is the size of their DNA targets: the 25-mer targets of oligonucleotides allow the fractionation of short DNA molecules, while the larger DNA targets of microarrays (~1000 bp) may provide more sensitivity in fractionating complex DNA mixtures. (While the larger microarray targets result in cross-hybridization between genes with greater than ~90% sequence identity, this problem is circumvented by PCR-amplifying and printing subregions of genes with lower homology.) The rest of this discussion will focus on DNA microarrays, but the principles outlined below are generally applicable to both systems. A series of reviews on various aspects of oligonucleotide chips and DNA microarrays is presented in ref 2.

The general features of a DNA microarray experiment are outlined in Figure 2. A microarray contains thousands of DNA spots that serve as hybridization targets...
for a heterogeneous DNA sample. Array experiments are comparative: Two DNA samples labeled with different fluorescent dyes (green Cy3 or red Cy5) are co-hybridized onto an array, after which the signal intensity of each fluor across the array is measured with a confocal laser scanning microscope. The quantitative ratio of red to green signal for each spot reflects the relative abundance of that particular DNA between the two experimental samples. The comparative nature of array hybridizations bypasses the need to measure absolute DNA concentrations in each sample—the exact concentrations of target DNAs on the array are unknown, partly due to the inherent heterogeneity of PCR reactions—and ensures internal consistency in experimental protocol, i.e. both samples are applied to the same array and experience precisely the same hybridization conditions, etc. This experimental design also means that only spot color—and not overall spot intensity—contains biological information, e.g. a bright yellow spot and a weaker yellow spot both represent a relative abundance ratio of 1. Finally, a series of samples can be compared with each other through separate co-hybridizations with a common reference sample. A more detailed discussion on the technical aspects of microarray experiments is presented in ref 3.

A typical gene expression experiment, in which isolated mRNAs are reverse-transcribed into fluorescently labeled cDNAs, is outlined in Figure 3. While gene expression applications of DNA microarrays are well-known, this technology is applicable to any experiment that generates output information in the form of a heterogeneous DNA population (Figure 4A). For example, the number of ribosomes attached to a mRNA indicates how actively that gene is being translated into protein. Thus polysomal RNA can be isolated from cells, transcribed into fluorescently labeled cDNA, and hybridized onto an array to identify which genes are being actively translated. Similarly, membrane-bound polysomal RNA can identify genes that are being actively secreted. Furthermore, genomic DNA can be isolated and converted into fluorescent DNA by PCR and hybridized onto array to reveal deleted and duplicated genes between two genomes, for instance between related strains.
of an microorganism or malignant vs healthy mammalian cells (18). Microarrays can also be used to compare initial and selected cell populations in genome-altering experiments, given a method to amplify altered genes into fluorescently labeled DNA and thus measure their relative abundance between the two populations. The *Saccharomyces* Genome Deletion Project (see previous section) employs this strategy, utilizing universal primers to PCR-amplify the unique 20-bp markers of each gene deletion for cell population comparisons on oligonucleotide chips (12). Another example is a genome-altering experiment in which cell populations containing an overexpression plasmid gene library are subjected to a selection that differentiates competent and dysfunctional cells, e.g. inhibitory concentrations of a drug (Figure 4B). During such a selection, cells that contain beneficial or detrimental plasmid library genes are enriched or depleted, respectively, in the population. To identify these genes, plasmids from the initial and selected cell populations are isolated and their gene inserts PCR-amplified into fluorescently labeled DNA. After co-hybridization onto an array, red and green spots identify genes that confer selective advantage or disadvantage upon overexpression. This protocol is particularly useful for analyzing the biological effects of molecules, such as drugs, that inhibit cellular growth. Proteins naturally identified include direct drug targets and resistance proteins, which in conjunction with other genomic data can reveal drug mechanisms and predict susceptibilities to drug resistance mechanisms.

Two final features of DNA microarrays further enhance their versatility as a genomic tool. First, the thousands of target DNAs on an array need not only be genes but can be DNA elements from any part of a genome, such as intergenic regions and nonprotein coding genetic elements. Indeed, an organism’s genome also can be unsequenced before an array is constructed, e.g. bacterial clones of any library of DNA inserts, such as human ESTs, can serve as a source of target DNAs. Second, array experiments involve homogeneous cell populations, so that small numbers of cells provide sufficient quantities of RNA or DNA for array hybridizations (29, 30), and analyses can be conducted on diploid as well as haploid cells. As a result, DNA microarrays can be used to study virtually all organisms, from microorganisms to human, sequenced and unsequenced. These properties also make arrays a valuable diagnostic tool, since patterns of genomic data can provide comparative information without sequence verification of the DNA targets.

**Application of Genomic Technologies to Fundamental and Applied Biological Problems**

The following two discussions attempt to illustrate how genomic technologies can form new strategies for studying fundamental and applied biological problems.

**Fermentation Development: Improving Secondary Metabolite Production.** For fermentation processes, genome-wide technologies not only can serve diagnostic and quality-control functions but also can
influence the speed and efficiency of process development. Consider the large family of commercially important pharmaceuticals, the polyketide secondary metabolites, produced by the actinomycetes class of soil bacteria (31). Currently, the development of overproducing actinomycetes strains involves successive rounds of mutagenesis and screening, a procedure that can take several years. Furthermore, the genetic changes acquired by these strains are unknown and remain elusive due to lack of well-developed genetic tools for most actinomycetes. The application of genomic strategies toward actinomycetes with available genome sequences, e.g., Streptomyces coelicolor (32), will clearly facilitate the generation of overproducing descendant strains. Perhaps more importantly, they should reveal the genetic changes that lead to increased secondary metabolite production, allowing analogous mutations to be engineered in related producer strains that are not amenable to genome-wide analysis.

Several complementary strategies for improving secondary metabolite production are proposed in Figure 5. By diagnosing aberrant vs normal fermentation cultures through their data signatures, e.g., gene expression patterns, DNA chips can serve as high-resolution quality-control tools (not shown). Comparison of gene and protein expression patterns between high vs low producing states of a single strain, wild type vs mutant strains, and an existing overproducing strain vs wild type should also reveal cellular states that underlie the high-producing phenotype (Figure 5, top). Similarly, comparison of wild type and overproducing genomes should identify deleted or duplicated genes that may be responsible for metabolite overproduction. During the course of such studies, environmental factors and cellular states that reduce secondary metabolite production are likely to be discovered, possibly leading to engineered producer strains with higher product yields and more robust fermentation properties, for example, by engineering genetic profiles that mimic overproducing strains or desensitizing the organism to environmental cues that trigger its transition to a low-producing state.

Genome-altering methodologies also have potential for advancing fermentation development (Figure 5, bottom). Given an effective selection, these approaches could unveil the genes essential for secondary metabolite production, thereby providing information on important cellular pathways. Moreover, these strategies should naturally identify genes that, when disrupted or overexpressed, result in increased secondary metabolite production. These automatic answers for increasing product yields can immediately be engineered into the sequenced organism or related producers strains.

While selections are not strictly required for the experiments just described, an effective selection for secondary metabolite production is required to fully capitalize on the power of genome-altering technologies. Traditional methods for screening a mutagenized cell population one-by-one generate a set of overproducing cells, but a selection that distinguishes competent and dysfunctional cells in secondary metabolite production provides a procedure that is internally controlled, quantitative, and easily repeated under many different conditions. To briefly discuss some possible selections for secondary metabolite production, consider the biosynthesis of the polyketide antibiotic erythromycin, by a natural or heterologous producer. Erythromycin-resistant bacteria possess a regulatory pathway that senses erythromycin concentration, which results in the activation of resistance genes in an erythromycin-dependent manner. If this erythromycin-sensing pathway could be functionally transferred into the actinomycetes producer along with the bacterial resistance gene promoter fused to a selectable marker, such as a kanamycin resistance gene or GFP, the extent of erythromycin production in a mutagenized cell could be reported—and selected—by the synthesis of the marker in an erythromycin-dependent manner. On the other hand, if the erythromycin-sensing pathway is not easily transferred to the producer strain, the erythromycin-resistant bacteria themselves could function as the erythromycin detector. In this case, each producer cell must be physically associated with several erythromycin-detecting bacteria, e.g., via agarose beads, and the erythromycin-sensing bacterial promoter fused to a reporter to allow sorting of bacteria—actinomycetes complexes. An important feature of both selections is that the secondary metabolite itself, erythromycin, triggers the synthesis of the selectable marker. This feature ensures that all cellular pathways important in erythromycin biosynthesis, i.e., not only enzyme expression but also enzyme folding and assembly, precursor biosynthesis, etc., are covered by the selection. Finally, the design of a successful selection system for one producer—metabolite pair would rapidly generate valuable information to be applied to other producer systems with no available selections.

**Fundamental Cell Biology: Understanding Cellular Responses to Stress.** The breadth of recently published microarray experiments (33-36) and work in progress (18, 37-40) underscores the impact genomic technologies will have on building integrated models of
cellular processes. However, progress toward this goal requires methods for extracting the biological information inherent in the wealth of genomic data. (Note that a single microarray can generate $\sim 10,000$ data points and a set of $n$ arrays, $n \times 10,000$ data points.) A recently developed method for DNA microarrays is cluster analysis (41), which (1) sorts genes into groups with similar behavior over a series of experiments and (2) maps numerical data onto a color scale for easier visualization of large bodies of data. A cluster diagram of several stress experiments is shown in Figure 6A (36, 39, 40). Array experiments are organized along the horizontal axis, with each column representing data from one array and each row data for a single gene. (Gene names are not shown.) All ratios are log transformed so that inductions and repressions of identical magnitude are numerically equal but opposite in sign. Log ratios of 0 (ratios of 1) are colored black, and increasingly positive or negative log ratios are colored reds or greens with increasing intensity, respectively. The clustering algorithm uses correlation coefficients to mathematically describe the similarity in data profiles between each pair of genes and subsequently clusters them in a hierarchical manner. The pairwise relationships of genes are represented by a dendrogram, in which branch lengths reflect the degree of similarity (Figure 6A). (Array experiments can also be clustered, as shown in the figure.) For gene expression experiments, such data manipulation results in blocks of color that represent genes with similar expression profiles and facilitates the visual detection of global as well as finer features of gene expression, i.e. the relative patterns of clusters and the identities of genes within a cluster. A detailed discussion on cluster analysis is presented in ref 41.

Cluster analysis has revealed patterns of data, in particular those involving more well-characterized genes, that report the status of cellular processes under particular experimental conditions, e.g. environmental stresses. For example, during the diauxic shift from fermentative to aerobic growth, yeast begins to exhaust its glucose supply and enter into stationary phase. Under these conditions, respiratory genes are induced, indicating that yeast cells are attempting to utilize its carbon sources more efficiently, and gluconeogenic and storage carbohydrate biosynthetic enzymes are up-regulated, suggesting that the cells are anticipating and preparing...
also provide comparative information, and array experi-

ences between experiments, genes with poorly defined functions provides a way to rationalize similarities and differences of--if not all--genes (Figure 6A). Such analyses are also useful for examining the cellular responses of mutants because aberrant behavior of global gene programs can be identified as well as specific cellular defects that underlie mutant phenotypes, e.g., a haploid yeast mutant with an overall expression profile similar to diploid sporulating cells may indicate that the haploid mutant is inappropriately trying to sporulate.

Finally, genomic technologies and the poor characterization of significant fractions of most genomes are creating opportunities for gene discovery and exploratory research in general (1). In Figure 6B, for example, the strong induction of a gene cluster across various stress conditions suggests its importance in helping yeast cells cope with stress. The lack of functional information for these genes, indicated by the abundance of “unknown” descriptors, makes these genes obvious candidates for follow-up mutational studies to decipher their roles in stress survival. While the presence of several genes with known functions may provide clues about the other genes, e.g. some of the “unknowns” may lie in oxidative stress pathways (Figure 6B), their strong coregulation also suggests the presence of common regulatory elements in their promoter regions, which can be searched for with sequence similarity algorithms (42, 43). Gene expression studies of the yeast sporulation response (35) and cell cycle (34) time course, which respectively identified various sets of sporulation induced and cell-cycle regulated genes, are two recently published examples of this kind of analysis.

This discussion has attempted to illustrate the general issues regarding the use of genomic technologies in fundamental and applied research. To fully capitalize on their potential, additional theoretical and computational approaches are needed to uncover the complex network of biological information residing in the large bodies of genomic data. For example, the ability to map data onto quantitative models of interconnected cellular pathways is a necessary first step toward building global and integrated pictures of the cell. However, attempts to model even the simplest pathways quickly highlight the need for tools that measure other cellular parameters on a genome-wide scale, such as protein concentrations, protein--protein interactions, subcellular localization of mRNAs and proteins, identities of small molecule substrates of proteins, small molecule fluxes, etc. Methods for observing post-translational states of proteins are especially important, since currently the effects of protein modifications as well as mutations that do not affect transcription levels of a gene (e.g., frameshifts) can be detected only indirectly through their influence on downstream cellular processes. These theoretical and experimental challenges will require the assimilation of concepts from biological and nonbiological disciplines, including mathematics, computer science, physics, chemistry, and engineering, which together promise to lead to great breakthroughs in our understanding of how cells and living organisms function.

Acknowledgment

The author thanks P. O. Brown, C. Khosla, and R. N. McDaniel for helpful comments on the manuscript. The author is a recipient of a postdoctoral fellowship from the Cancer Research Fund of the Damon Runyon-Walter Winchell Foundation.

Figure 7. Yeast gene expression patterns of some carbon and nitrogen metabolic enzymes in the diauxic shift response, sporulation, carbon starvation, and nitrogen starvation.
References and Notes


(14) http://yacmi.med.yale.edu/YGAC/home.html.


(23) http://www.affymetrix.com/.


(29) For a standard reverse transcription reaction, ~2 ng mRNA is required. However, an RNA amplification protocol has been developed to amplify samples as small as ~20 ng mRNA (30).


(32) http://www.sanger.ac.uk/Projects/S_coelicolor/.


(42) http://www.sdsu.edu/mem/website/

(43) http://alize.ulb.ac.be/~YRT.

Accepted February 24, 1999.

BP990027Q