# 9

# YEAST AS A PROTOTYPE FOR SYSTEMS BIOLOGY

Goutham Vemuri[1] and Jens Nielsen[2]

[1] *Center for Microbial Biotechnology, Denmark Technical University,*
*Biocentrum-DTU, Lyngby DK2800, Denmark*

[2] *Department of Chemical and Biological Engineering, Chalmers University of Technology,*
*Kemivägen 10, SE-412 96, Gothenberg, Sweden*

## 9.1 INTRODUCTION

The bakers yeast, *Saccharomyces cerevisiae*, is arguably one of the earliest microorganisms to be domesticated for early biotechnological applications such as brewing and baking. Gradually, this yeast has established itself as the primary model eukaryote in the field of genetics and molecular biology. Currently, *S. cerevisiae* is also the most commonly used eukaryote for bioprocess applications, owing to its flexibility in aerobic and anaerobic modes of metabolism and its amenability to genetic manipulations. With the advent of the high-throughput omics technology, it is not surprising that *S. cerevisiae* served as the platform for deciphering the molecular details of chromosomal activity such as DNA replication and transcription as well as physiological activity such as translation and metabolism at a global level. The wealth of information on the cellular components and their structural components has greatly facilitated the progress of yeast biotechnology and metabolic engineering. Several aspects of *S. cerevisiae* fundamental metabolism have been modified and improved to meet the end bioprocess objectives, but more complex aspects such as expanding the range of consumable substrates and the mechanism of glucose repression still remain obscure. The primary reason impeding progress is the lack of knowledge on how the different

*Systems Biology and Synthetic Biology* Edited by Pengcheng Fu and Sven Panke

cellular components, such as genes, proteins, and metabolites interact with each other to impart the phenotype. The lack of detailed knowledge of regulation of cellular processes is impeding progress in metabolic engineering, despite rapid progress of technology in genomics, transcriptomics, proteomics, and metabolomics areas.

The cellular processes comprise of the transfer of mass in the metabolic pathways and the transfer of information in the regulatory and signal transduction pathways. The regulation of these processes as a result of environmental or genetic changes is the key step in imparting the phenotype. The information flow commences at the genome level with the DNA. For a given microbial strain, the sequence of the genome remains unvarying. The variation that begins with the primary step in the flow of information is transcription, the process of making mRNA. From this step forward, the abundance of the cellular components highly depends on the environment. For example, in the presence of high glucose concentrations, those genes whose products are required for rapid glucose consumption are transcribed to a greater extent. The next step in the information pipeline is the translation of mRNA into proteins. Proteins are the structural as well as functional entities, carrying out all the cellular functions such as adaptation, regulation, and even catalysis. The transfer of mass from the substrate to the product occurs depending on the protein availability and, therefore, proteins are the link between the information transfer and mass transfer. A simplified schematic of these two cellular pipelines is depicted in Figure 9-1.

Until the late twentieth century, the focus of most traditional enquiries was limited to in-depth analysis of only a small number of cellular components (usually genes, proteins, or signaling pathways) in relative isolation from the remaining system. Although this reductionist approach has been extremely useful in providing detailed description of the individual cellular components, it is to be noted that these components do not function in isolation in the system. Therefore, their biological role has to be elucidated in the context of the remaining components in the system. This line of thinking is the inspiration for modern systems biology, and will be the main focus of this chapter. This chapter begins with a historical perspective of the research that led to the current notion of systems biology and the experimental and computational tools available. The chapter focuses on the development and applications of systems biology in the context of *S. cerevisiae*.

## 9.2 INTEGRATIVE PHYSIOLOGY AS THE BIRTH OF SYSTEMS BIOLOGY

Although systems biology has entered the popular lexicon only after the millennium, the idea is not new. The natural confluence of systems science with biology and the representation of biological entities as systems were described as early as 1929 [18] in Walter Cannon's homeostasis theory that described the human body as dynamic control system. In 1963, Jacob and Monod followed this line by implementing the concept of control theory to the operation and regulation of the *lac* operon [109]. Subsequently, the concept of a holistic "systems approach," was developed, and it was in 1968 that the term "systems biology," was first used by Mesarovic [107] to indicate

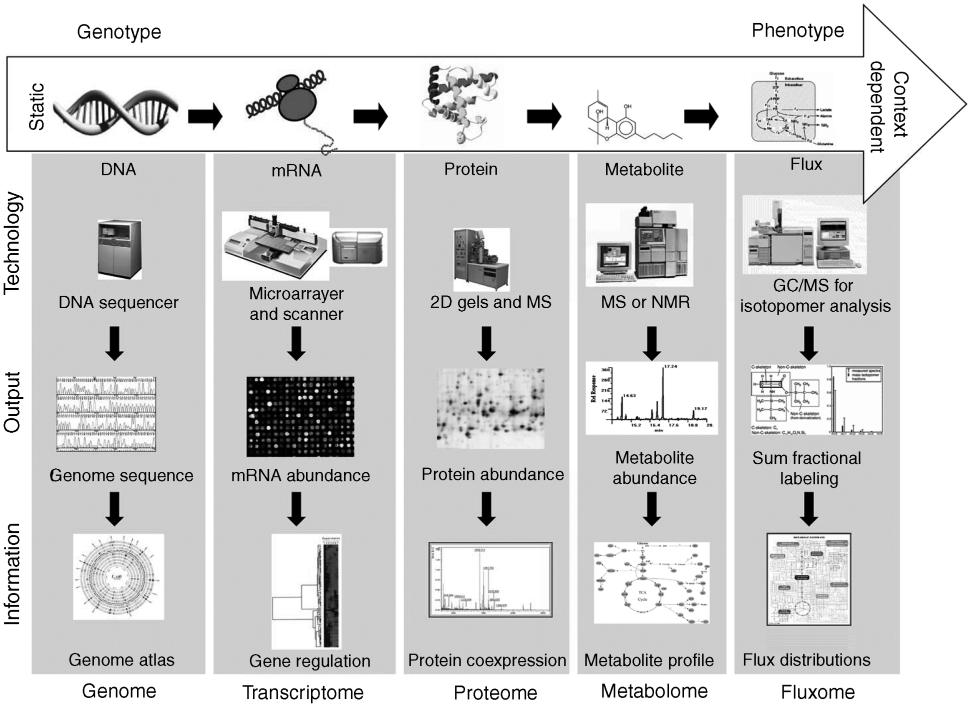


**Figure 9-1** The transfer of information from DNA that defined the genotype to metabolic fluxes, which quantify the phenotype. The genotype of a strain is a static entity, but the expression of subsequent components is context dependent. The technology used to quantify the various components in the information pipeline is depicted schematically along with the typical output. It requires careful analysis to extract useful biological information out of the data. Integration of the high-throughput data from these different stages of hierarchy in a context-dependent manner will reveal the interactions between the various components, leading to a holistic understanding of how the phenotype is linked to the genotype.

the application of the techniques of systems scientists (who were conventional control engineers, physicists, and mathematicians) to experimental biology. This was one of the early invitations for biologists to study vital biological phenomenon from a systems perspective. Development continued into the 1970s when researchers developed biochemical systems theory and metabolic control theory to create simplified mathematical models of biological systems, enticing nonbiologists to work with biological systems. This concept was immediately picked up by several researchers who reported many exemplary applications of the control systems theory to life sciences in the following decade. For example, in 1969, Yates pointed out the similarities in the conventional mechanical and electrical control systems and adrenal glucocorticoid control system in humans [178]. Goldbeter and Segel developed the kinetic theory of enzyme action in microorganisms in 1977 [52]. This concept was further developed by Iberall in 1977 using the laws of irreversible thermodynamics to describe the hierarchy in physiological systems. The results from this paper were subsequently used to describe several other regulatory phenomena in living systems [71]. Gradually, these concepts were developed to describe modeling of the structure, control, and optimality of metabolic networks [61].

Along with systems theory, cybernetics played a key role in drawing parallels between the information transfer in electronic systems and biological systems. Moreover, with the development of a formal framework for studying the design of biological networks in terms of error correction, feedback and feedforward control loops, and other circuit concepts, the confluence of the two fields became even more obvious. Together with systems theory and cybernetics, another field that contributed to early systems biology was the field of reaction engineering. The focus of reaction engineering is on the properties of complex reaction networks while monitoring the individual reactants. Although in classical reaction engineering it is possible to calculate important thermodynamic and kinetic parameters, the description of biological systems is far from such quantification. As the application of the concepts of control systems and reaction engineering in biology gained popularity, there was also a simultaneous progress in the development of experimental techniques and high-throughput methods, particularly the ability to sequence complete genomes. The availability of complete genome sequences provides abundance of information, but without any rules pertaining to how the cell processes the genomic information. In addition, the RNA microarray-based expression technology expedited the progress in understanding the transfer of information from the genome to proteins. The paradigm shift in the approach to study biological systems from a reductionist approach to an integrative approach provides a new meaning to the integrative approaches developed thus far and has given birth to the modern meaning of systems biology. Systems biology, in a very general way, can be defined as the integration of genomic, proteomic, transcriptomic, and metabolomic data using computational methods for a holistic understanding of systemic functions. In the context of metabolic engineering, this definition can be interpreted as the unification of information from the flow of mass and energy (in metabolic pathways) and the flow of information from the DNA (transcriptional regulatory pathways and signal transduction pathways). Understanding how the flow of information and the flow of mass occur in tandem is the fundamental tenet of systems biology.

## 9.3 SYSTEMS BIOLOGY AS AN ENGINEERING DISCIPLINE

We are currently at a crossroads in proceeding with the study of biology. The paradigm shift in the study of biology from a descriptive science to a well-defined quantitative discipline reflects the need to incorporate the theories and principles developed in other disciplines, particularly engineering sciences. The Human Genome Project validated the discovery-driven approach to systems biology for augmentation of the previous purely hypothesis-driven paradigm. With the completion of the human genome and the genomes of various other species, we are now introduced to a number of genes we have never even known existed before. At the same time we are also troubled with the disturbingly finite size of this gene list, and we quickly learned that the diversity of the genes could not approximate the diversity of functions within an organism. The key to this discrepancy is in the combinatorial use of the gene products to impart the diversity. This section will bring out the engineering concepts that are highly applicable in the progress of systems biology.

Systems are central to engineering. The traditional concepts of analysis, synthesis, and design that form the core of the engineering discipline are unified in the systems approach, as shown in Figure 9-2. The system is first decomposed into well-defined subsystems and each subsystem is analyzed for its components and functionality. This defines the analysis component, represented by the left arm in the figure. The knowledge gleaned from the components of the subsystems is assembled into larger and larger subsystems, until the complete system is synthesized. The methodology described here is also known as the bottom-up approach. As applied to biological systems, all the information of individual genes, proteins, metabolites, and so on is gathered, followed by assembling these components in the context of the observed phenotype. Therefore, each level of information processing shown in Figure 9-1 serves as one subsystem. Such a model designed by the bottom-up approach should be capable of describing exactly how the cell functions in response to a certain genetic or environmental alteration. Other less commonly used approach in systems analysis are

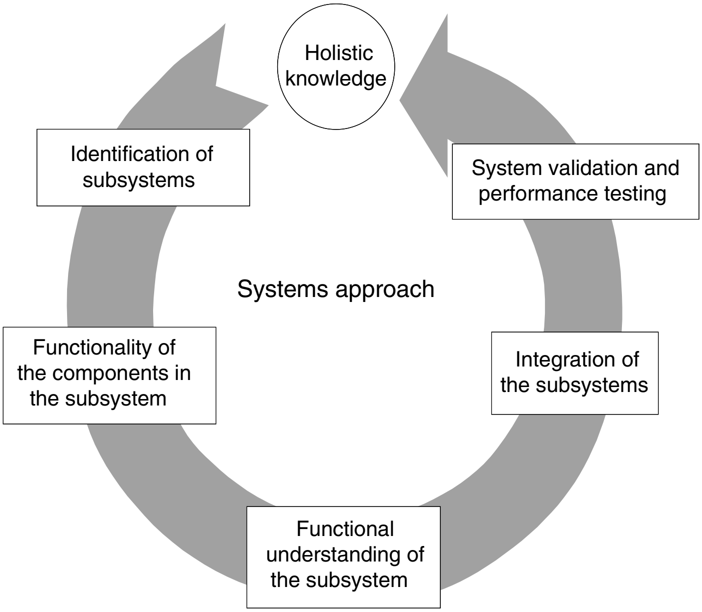

Holistic knowledge
Identification of subsystems
System validation and performance testing
Systems approach
Functionality of the components in the subsystem
Integration of the subsystems
Functional understanding of the subsystem

**Figure 9-2** One iteration in the cycle of analysis and synthesis using the bottom-up approach to systems biology. As in other disciplines, the system is first defined, followed by the identification of subsystems it comprises of. The components that make up the subsystem are studied in detail for their functionality to understand their role in the context of the whole system. The knowledge from these components is assembled to synthesize bigger subsystems until the complete system is put together and functionally evaluated. This approach reflects the implementation of the classic engineering principles of analysis and synthesis in the context of biological systems.

the top–down approach where the rules are defined for all the individual components identified by the analysis, allowing them to freely interact with each other.

There are three fundamental concepts that an engineer uses to understand a system: emergence, robustness, and modularity. An inherent property of complex systems is that they are larger than the sum of their individual parts, a property known as "emergent property.," The properties of a cell cannot be deduced based on the properties of DNA, RNA, or proteins. It takes a holistic understanding using systems level analyses for a comprehensive understanding of these emergent properties. The robustness of a mechanical or an electronic system is judged on its ability to maintain its functionality despite perturbations. Similarly, a biological system maintains its phenotypic robustness in the event of environmental and genetic perturbations and is a strong determinant in evolution. The feedback and feedforward control loops that comprise a biological or nonbiological system impart robustness to these systems. The third concept central to systems is their modularity. An engineer would define modularity as a subsystem, as shown in Figure 9-2. It is a collection of components that perform a distinct function through interactions and has clear inputs, control processes, and outputs. In biology, modularity refers to a set of components that have close interactions and share a common function. An example of a module of a subsystem in a biological system is the respiratory chain, which is composed of several genes, proteins, cofactors, and regulators that work together in the transport of electrons to oxygen with concomitant energy generation. From an evolutionary perspective, modularity contributed to robustness by restricting the change (malfunction) to the subsystem, thereby decreasing the severity of system failure.

## 9.4 HIGH-THROUGHPUT EXPERIMENTAL TECHNIQUES

Although the development of the systemic concepts and applying them to biology appealed to a large community of researchers, the lack of experimental techniques to verify the results of the analogy limited the progress of systems biology in 1980s and 1990s. The recent exponential increase in the availability of biological information in the form of genome sequences, RNA, and protein abundance and metabolic flux analysis to quantify physiology transformed systems biology into one of the most exciting scientific developments. Having provided an overview of the concept of systems biology and its evolution to the present-day notion, we devote this section to the experimental techniques that contributed to the advancement of the field from integrative physiology to systems biology.

### 9.4.1 Yeast Genome Sequencing

The main catalyst behind the rapid progress is the ability to sequence complete genomes. Since the completion of sequencing of the genome of the first independently living organism, *Haemophilus influenzae*, genome sequencing became a routine procedure with the genomes of several microorganisms, including *S. cerevisiae* becoming available. The landmark invention that triggered this explosion was the invention of nucleotide sequencing method by Frederick Sanger [140] and subsequent automation of the process. Sequencing of the yeast genome was the offshoot of a broad international consortium, acting upon a consensus reached in 1988 [1], that was committed to working on a 15-year massive effort to sequence the human genome, supported by a $3 billion funding. The recommendation of this consensus was that the genome sequences of some other eukaryotes should be determined alongside the human genome. The "model," eukaryotic genomes specifically chosen were those of yeast (*S. cerevisiae*), a nematode worm (*Caenorhabditis elegans*), and a fruitfly (*Drosophila melanogaster*). New and faster DNA sequencers were developed, following this initiative and sequencing individual genes became a routine process in yeast. Another landmark result in eukaryotic sequencing was the determination of the complete sequence of a whole chromosome (chromosome II in *S. cerevisiae*) in 1992 [117]. Subsequently, there were several reports of sequencing large fragments of the individual chromosomes in *S. cerevisiae*, which provided the foundation for completion of the chromosome sequencing. Figure 9-3 shows the time line when

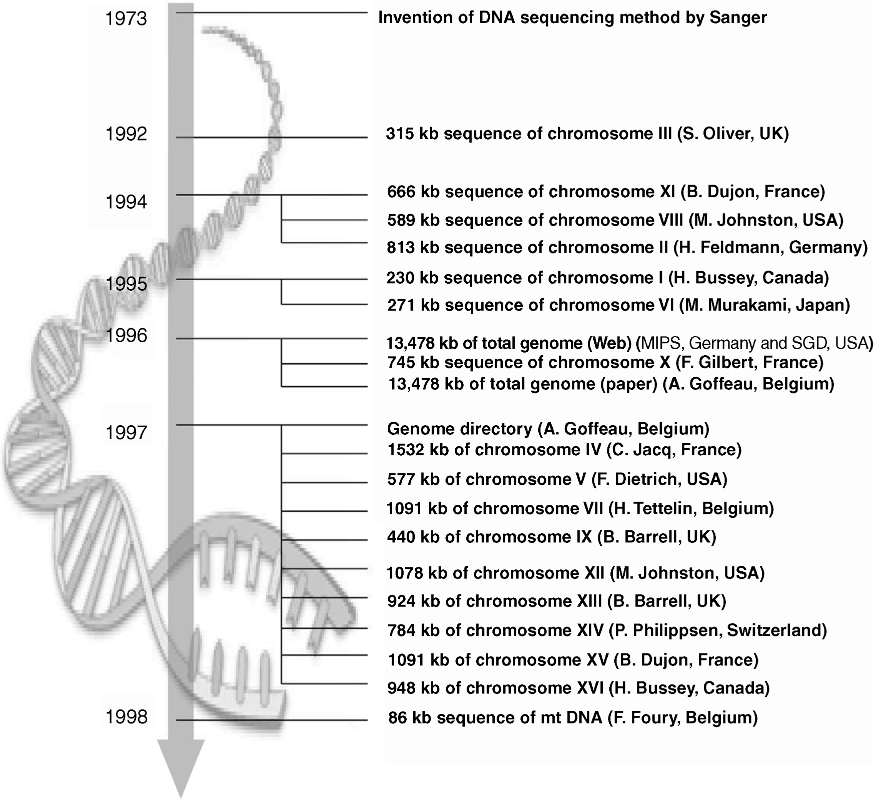

1973
Invention of DNA sequencing method by Sanger
1992
315 kb sequence of chromosome III (S. Oliver, UK)
1994
666 kb sequence of chromosome XI (B. Dujon, France)
589 kb sequence of chromosome VIII (M. Johnston, USA)
813 kb sequence of chromosome II (H. Feldmann, Germany)
1995
230 kb sequence of chromosome I (H. Bussey, Canada)
271 kb sequence of chromosome VI (M. Murakami, Japan)
1996
13,478 kb of total genome (Web) (MIPS, Germany and SGD, USA)
745 kb sequence of chromosome X (F. Gilbert, France)
13,478 kb of total genome (paper) (A. Goffeau, Belgium)
1997
Genome directory (A. Goffeau, Belgium)
1532 kb of chromosome IV (C. Jacq, France)
577 kb of chromosome V (F. Dietrich, USA)
1091 kb of chromosome VII (H. Tettelin, Belgium)
440 kb of chromosome IX (B. Barrell, UK)
1078 kb of chromosome XII (M. Johnston, USA)
924 kb of chromosome XIII (B. Barrell, UK)
784 kb of chromosome XIV (P. Philippsen, Switzerland)
1091 kb of chromosome XV (B. Dujon, France)
948 kb of chromosome XVI (H. Bussey, Canada)
1998
86 kb sequence of mt DNA (F. Foury, Belgium)

**Figure 9-3** The time line of significant events in the sequencing of the *S. cerevisiae* genome. The sequencing of the individual chromosomes paved the path for determining the complete genome sequence. The final draft of the genome sequence was completed in 1996. This is a collaborative effort that required several laboratories across the globe.

the sequences of the other chromosomes became available. The availability of the individual chromosome sequences finally led to determining the nucleotide sequence of the first eukaryotic genome, *S. cerevisiae*, in 1996 [50]. Thus, the yeast genome became the pioneer eukaryotic genome and yeast research community was the primary beneficiary of this knowledge of the complete sequence.

A dramatic transformation of yeast research ensued that presaged similar transformations in the approach to research on other model organisms, as their sequences, became available. This transformation began with technical improvements that accelerated research involving DNA cloning and recombinant techniques. The same consortium that played a key role in yeast genome sequencing undertook another major project of producing deletion mutants of every yeast open reading frame (ORF) [48,173], which led to the development of a whole class of genome-scale genetic methods. The estimated number of ORFs in *S. cerevisiae* is 6034, spread over the 16 chromosomes. A comparative analysis of the complete genome of yeast with the genomes of other model organisms and humans validated the conservation of sequence and function in evolution, despite the difference in the size and number of genes. This observation of "grand unification," has particularly useful ramifications in functional genomics. It became clear that a similarity in sequence is an important factor in assuming functional similarity. Therefore, comparative genomics permits the elucidation of a gene or protein in one organism to be applied to the same in another "lesser," known organism. Since yeast is still the most tractable eukaryotic system, much of the annotation of basic cellular functions in other eukaryotes, including humans, have functional identity in yeast. The availability of the entire genome sequence permits the asking of new kinds of research questions that can be answered only when one has truly comprehensive information about an organism. As previously mentioned, the sequence is the static entity and stable part in the organism under all conditions. It is the interplay between flow of information from the DNA sequence and flow of material in the metabolic network that imparts the variation. The subsequent sections will describe the high-throughput methods that have made the quantification of this variation possible and contributed to systems biology.

## 9.4.2 Transcriptomics

***9.4.2.1 Microarrays*** The genome basically defines the phenotypic space an organism can operate within and all phenotypic changes ultimately originate at the transcription level. The availability of genome sequences induced the development of technologies to quantify transcriptional activity on a genome scale and to identify the nature of information flow at the transcription level. Once the entire genome sequence became known, it became possible, for the first time, to study expression of all the genes at once; earlier one could study genes only a few at a time. The very idea of what constitutes "specificity," has been changed by the ability to study expression of all the genes without exception. It is now a routine procedure to simultaneously measure the abundance of mRNA species of every ORF in the genome using two-dye spotted arrays [141] or GeneChips [102] in an organism that respond to a specific stimulus or stress. One of the earliest genome-wide transcription characterizations

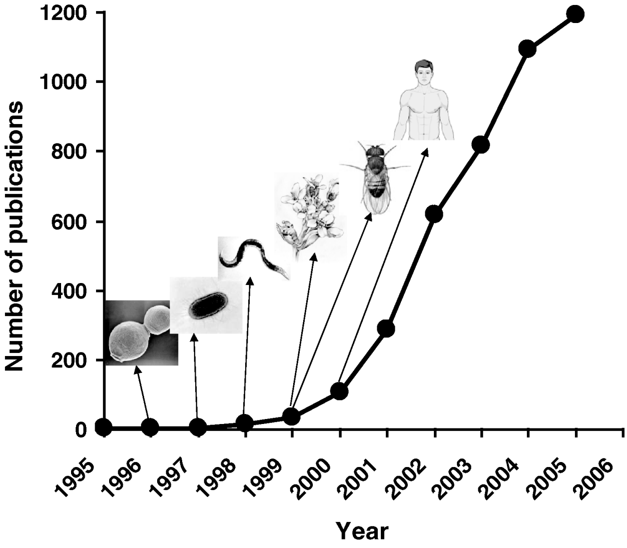


**Figure 9-4** The number of publications in peer-reviewed journals that contained the words/phrases "microarray," "global gene expression," or "oligonucleotide array" in the title or in the abstract. The exponential increase in the interest in quantifying global gene expression stems from the availability of the genomes of the various model organisms: *S. cerevisiae* (1996), *E. coli* (1997), *C. elegans* (1998), *Arabidopsis thaliana* and *Drosophila melanogaster* (1999), and, finally, human (2000). This technology has arguably revolutionized the concept of systems biology.

was to study the expression changes of all the genes in *S. cerevisiae* in response to metabolic shift from growth on glucose to diauxic shift to growth on ethanol [28]. Identifying similarities in the transcriptional profile, the role of many previously uncharacterized genes was predicted, based on the assumption that coexpressed genes are coregulated. Since then, there has been tremendous interest in quantifying gene expression in response to various conditions and, consequently, the number of publications using gene expression microarrays has exponentially increased over the past 10 years (Fig. 9-4).

Prior to the availability of complete sequences, cDNA clones from cDNA banks were PCR amplified and robotically printed onto glass slides, which were used to study gene expression [103,141]. On the other hand, in the photolithography technique, which is popularized by Affymetrix, synthetic linkers are adhered to a glass surface using photosensitive groups, and a light mask is used to direct light to specific areas on the glass to remove the exposed groups. A new mask is used to direct coupling at other sites, and the process is repeated until the desired sequence and length of the oligonucleotide is synthesized. This method, which is very similar to the production of computer chips, is very efficient in high-throughput generation of identical arrays. However, the method is quite expensive in the design phase and the DNA array generated using this method is not flexible when the new genes need to be added. A completely new array needs to be redesigned if new features are to be added. A slightly modified version of this method that resolves this issue is the ink-jet printing of

60mer oligonucleotides [10,68]. This method can generate new arrays or modify the gene content by reprogramming the synthesis of the new set of oligonucleotide sequences. The availability of the genome sequences for several model organisms has facilitated several researches to PCR-amplified genes (either a selected few or the entire list of open reading frames) from chromosomal DNA or design oligonucleotides to develop arrays that are very specifically suited for their purpose. Transcriptional profiling by global gene expression technology is a paradigm of the convergence of several technologies, such as DNA sequencing and amplification, synthesis of oligonucleotides, and fluorescence biochemistry. Transcriptional profiling is based on the fundamental base pairing ability of the nucleotides. The conventional terminology is to refer to robotically printed sets of PCR products or conventionally synthesized oligonucleotides on glass slides as microarrays [30,141], whereas high-density arrays of oligonucleotides that are synthesized *in situ* using photolithography are referred to as GeneChips [101,102], although here we refer to both as microarrays. Numerous reviews have been published that describe the methodologies and analytics behind these methods.

For *S. cerevisiae*, extensive applications of microarrays have been reported, and there are many examples of analysis of genome-wide responses to several environmental and genetic perturbations. These initial transcription applications relied on existing knowledge to confirm some of the results as a means of validating new discoveries. For example, the application of microarrays to the classical study of aging and cell cycle identified several previously known genes in addition to discovering several new ones. Although the cell division cycle in yeast is known to regulate the expression of several histone genes [63], the transcriptional changes in the genome were followed in synchronized yeast cells during various stages of the cell cycle [21,147]. About 7 percent of the genome oscillated with the cell cycle, and every chromosome contained at least one cell cycle-dependent gene. By correlating the expression of the oscillating gene with the stage of the cell cycle, hundreds of transcripts were discovered that exhibited rhythmic expression trends exhibiting close periodicity to the cell cycle. Based on the cell cycle stage, these genes were grouped into different clusters, and analyzing the upstream sequences of genes from the same cluster revealed binding sites for several known as well as unknown transcription factors, indicating the involvement of additional transcription factors in regulating gene expression during the cell cycle. Considering that a large number of human proteins have high homology to yeast proteins, this research could have important applications in understanding human aging.

Microarrays have been extremely useful in understanding regulation and metabolism in yeast. For example, studying the transcriptional responses of *S. cerevisiae* to growth limitation by carbon, nitrogen, phosphorus, sulfur, or oxygen enabled the identification of gene clusters that are involved in sensing nutrient limitation and trigger alternate pathways to minimize stress [11,150]. The stress response induced by nutrient limitation during steady-state growth is apparently different from that observed in normal stationary phase cultures, and that some aspect of starvation, possibly a component of stress response, may therefore be required for triggering metabolic reprogramming associated with diauxic shift as demonstrated by transcription profiling [14,138]. Another important feature of yeast metabolism that was characterized by microarrays is that of glucose repression. Yeast preferentially metabolizes glucose while repressing the genes required for the uptake of several other carbon sources. It also induces the expression of genes required for glucose metabolism such as the glucose transporters and those in glycolysis. Although some of the key players in this complex signal transduction cascade were known for a long time, the true complexity involved in this process is just beginning to be gauged. The central components in the glucose repression pathway are *MIG1*, a DNA-binding transcriptional repressor, and its homologue, *MIG2*, a protein kinase, *SNF1* and its associated regulators such as *SNF4*, and a protein phosphatase, *GLC7* and its regulatory subunit *REG1*. The binding of *MIG1* upstream of many of the genes seems to be the most prevalent mechanism by which the repression pathway acts. In the presence of glucose, Mig1 is localized in the nucleus where it is dephosphorylated and represses gene expression. Upon removing glucose, Snf1 phosphorylates Mig1 and transports it out of the nucleus, resulting in derepression of the genes sensitive to glucose. It is believed that AMP (or even more likely the AMP:ATP or ADP:ATP ratio) signals the phosphorylation/dephosphorylation of Mig1. Several excellent reviews have been written detailing the state-of-the-art knowledge about this mechanism [87,134,155]. The glucose induction pathway is triggered by a completely different mechanism, which induces the *HXT* and *HXK* genes for glucose uptake and phosphorylation, respectively. In this pathway, the key components are a transcriptional repressor, *RGT1*, a protein complex, SCF (SCF complexes are named for their constituent proteins: Skp1, Cdc53 and Cdc34, and an F-box-containing protein), and membrane-bound glucose sensors, *SNF3* and *RGT2*. Upon sensing glucose, Rgt1 binds to the glucose sensors and generates a signal that causes the SCF complex to inactivate the Rgt1 repressor, thereby enabling glucose uptake and metabolism. In the absence of glucose, Rgt1 binds to the HXT promoters and represses their transcription. In this process, Grr1 (glucose repression resistant) plays a key role through ubiquitinating the proteins involved in the signal transduction pathway. The expression of *GRR1* is independent of the carbon source and both the mRNA and protein are constitutively expressed in *S. cerevisiae* in low amounts, thus supporting the role of Grr1p being a regulatory protein [121,122].

Microarrays played a key role in elucidating the regulatory role of *GRR1* in glucose induction. Upon comparing the transcription profiles in Δ*grr1* with its isogenic control strain, we observed large transcriptional changes spread out over different parts of the metabolism [172]. Several genes of the TCA cycle, respiration, and oxidative phosphorylation were induced while many transporters and amino acid biosynthetic genes were repressed in Δ*grr1*. Since Grr1 has also been implicated to play a key role in regulating glucose transport, profiling the transcriptional response of the hexose transporters in *S. cerevisiae* indicated strong repression of the low-affinity transporters (*HXT1*, *HXT3*) and one high-affinity transporter (*HXT4*) while inducing another high-affinity transporter (*HXT8*) and *HXT16*, a hexose permease. These results indicate differential regulation of even the different high-affinity transporters. Analysis of the sequence upstream of these genes revealed the binding sites for the transcription regulators, suggesting a key role for Rgt1 in the repression mechanism. Similarly, upon the identification of a second homologue for *MIG1*, YER028, microarray experiments revealed its glucose-dependent transcription repressing nature [105]. Subsequent DNA binding assays revealed that the binding affinities of *MIG1*, *MIG2*, and YER028 are different, although they recognize the same binding sequence. Transcription profiling also revealed that about 50 percent of the genes that responded to *MIG1* or *MIG2* were of unknown function. High-throughput experiments such as these could help identify genes that could serve as indicators to sense nutrient limitation and so on for inverse metabolic engineering applications.

One such example of transcriptome-guided inverse metabolic engineering was that of designing a strain with enhanced galactose uptake capability [119]. Overcoming glucose control over galactose metabolism has industrial interest in prompt utilization of galactose that is present in lignocelluloses and beet molasses along with glucose. The *GAL* system that contains genes responsible for the uptake of galactose is subjected to dual regulation of glucose repression and galactose induction. Galactose induces the *GAL* system by an ATP-dependent mechanism where the transducer protein Gal3 interacts with Gal80 [149,177]. In forming a complex with Gal80, Gal4 binds to the activator sequences in the *GAL* system, expressing the structural genes, *GAL2*, *GAL1*, *GAL7*, and *GAL10* [174]. These genes code for galactose permease, galactokinase, galactose-1-phosphate uridylyltransferase, and UDP-glucose 4-epimerase, respectively (Fig. 9-5) and are responsible for galactose uptake and its

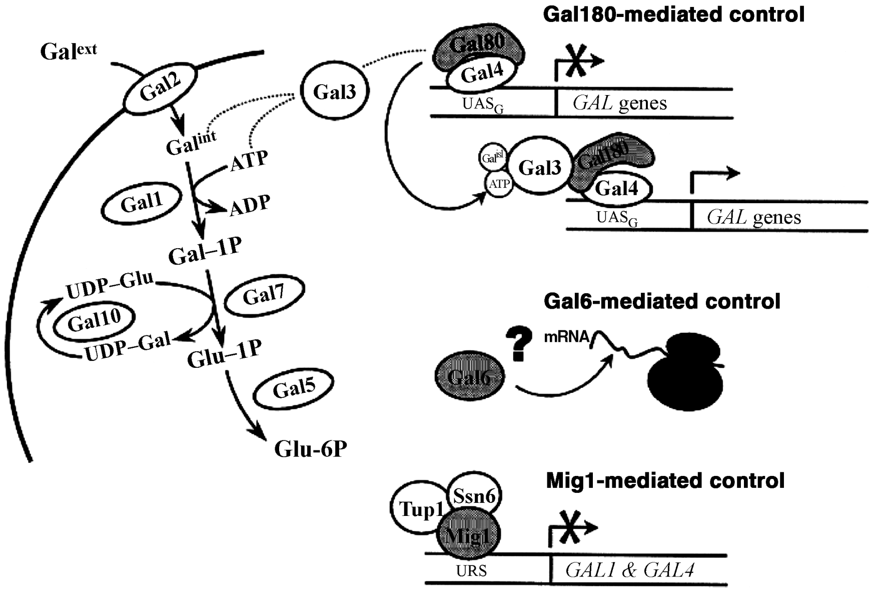


**Figure 9-5** Despite the similarity between galactose and glucose, yeast consumes galactose almost three times slower than it can consume glucose. The uptake and metabolism of galactose in the pathway is shown up to its entry in the glycolysis, where its subsequent metabolism is identical to that of glucose. Using inverse metabolic engineering strategies and global transcription analysis, we increased the rate of galactose consumption.

subsequent conversion to glucose-1-phosphate in the Leloir pathway. Therefore, deleting Gal6, Gal80, and Mig1 increased galactose uptake rate by about 40 percent and upon comparing the transcription profile in this strain and in a Gal4 overexpressed strain with a reference strain, there was no clear reason for the enhanced galactose uptake rates. Besides *GAL4*, *GAL6*, and *GAL80*, only the *PGM2* transcript, encoding the major isoform of phosphoglucomutase, exhibited a statistically significant change (of about 1.5 fold) and overexpressing *PGM2* resulted in a 70 percent increase in galactose uptake rate [15]. This case study presents a microarray-guided approach for inverse metabolic engineering, where the targets for metabolic engineering are identified by screening various strains.

Engineering the redox metabolism is an attractive target for metabolic engineering applications since it plays a crucial role in determining growth efficiency and product formation. Despite the importance of redox in metabolism, little knowledge exists about the transcriptional changes that emulate following a redox perturbation. As a fundamental study to identify redox-sensitive genes, a *S. cerevisiae* strain with cofactor modifications in the glutamate generation pathway was compared with the reference strain [16]. Gdh1 (encoding glutamate dehydrogenase) is one of the principal NADPH-consuming pathways in biomass synthesis, consuming more than half of NADPH generated. The sustenance of the Δ*gdh1* strain is ensured by substituting this pathway with the glutamate synthase reaction (GS-GOGAT), encoded by *GDH2*, which uses NADH as the cofactor (Fig. 9-6). Therefore, the switching of Gdh1 with Gdh2 perturbs the redox balance without disturbing biomass generation. Not surprisingly, comparing the transcript levels between the mutated strain with those in the

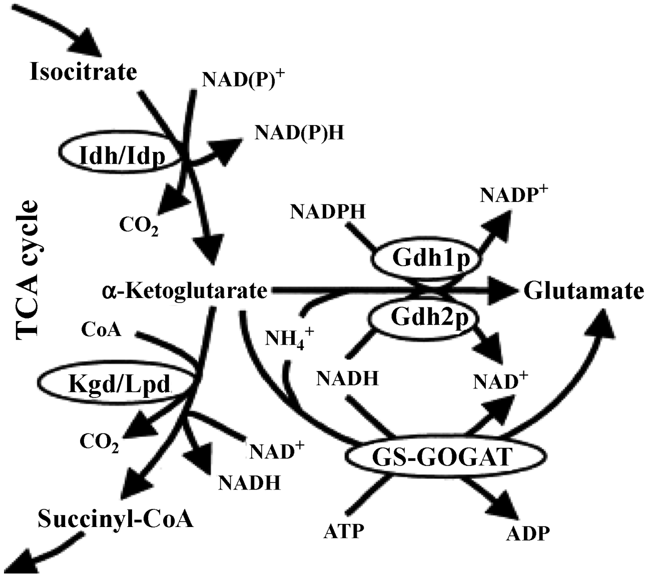


**Figure 9-6** Engineering cofactor utilization by replacing the *GDH1* gene with *GDH2* in *S. cerevisiae* [16].

reference revealed that several genes responsible for the regeneration of NADPH have altered expression. This study indicates a possible redox-dependent regulation among these genes, as revealed by the gene expression analysis.

The current use and application of microarray technology is tremendously valuable to systems biology. After more than 10 years of its conception, the rate-limiting step in microarray technology is not in the technical aspects, but rather in the data handling. Currently, only a small fraction of the data generated in a microarray experiment is being used to make inferences for functional testing. This significantly undermines the potential of microarrays in the ability to investigate the genomic response when only a handful of genes undergo further study. The next technological leap for microarray technology lies not in the study of model organisms, but in the interrogation and analyses of uncharacterized or even mixed cell samples whose complete genome may not be available.

***9.4.2.2 Serial Analysis of Gene Expression*** The genome sequences of all eukaryotes are large and contain enormous number of genes, the functions of most of which are yet to be elucidated. In the transcription stage of information transfer, each ORF synthesizes widely varying number of mRNA species. Techniques based on subtractive hybridization and differential display have been useful in identifying the differences among transcripts. However, these methods provide only a partial picture and may miss transcripts expressed at low levels. Oligonucleotide arrays were useful in comparing the expression of thousands of genes in a variety of tissues, including small cell populations, but they are limited to analyzing only previously identified transcripts. In contrast serial analysis of gene expression (SAGE) allows quantification and simultaneous analysis of a large number of transcripts without the prior knowledge of the genes [160]. The SAGE technique can be used in a variety of applications, including analysis of the effect of drugs on tissues, identification of disease-related genes, and elucidation of disease pathways. This method produces 9–10 base sequences or “tags,” that uniquely identify one mRNA species. These unique tags are concatenated serially into long DNA sequences for high-throughput sequencing. The frequency of each tag in the sequence quantifies the abundance of the corresponding transcripts. The resulting sequence data are analyzed to identify each gene expressed in the cell and the levels at which each gene is expressed. This information forms a library that can be used to analyze the differences in gene expression between cells. The frequency of each SAGE tag in the cloned multimers directly reflects the transcript abundance. Therefore, SAGE results in an accurate picture of gene expression at both the qualitative and the quantitative levels.

The transcriptome, as defined above, was described for the first time in *S. cerevisiae* [161]. To maximize the representation of the genes involved in normal growth and cell cycle, SAGE libraries were generated from three stages of cell cycle: exponential phase, S-phase arrested, and G2/M phase arrested. The number of SAGE tags required to define the yeast transcriptome depends on the desired confidence to detect low-abundance mRNA species. Employing this method to determine the complete set of yeast genes expressed under a given set of conditions (the transcriptome), 4665 genes (approximately 75 percent of the predicted protein-coding ORFs) were detected, with most genes being expressed at low level. Among the highly expressed genes were those corresponding to well-defined metabolic functions and energy generation under the three stages [161]. Using the SAGE method, metabolic modules that are subject to suppressed translation under normal conditions (translation on demand) were identified in *S. cerevisiae* [8]. This study, which investigated the relation of transcription, translation, and protein turnover on a genome scale, demonstrated significant posttranscriptional control of protein levels for a number of different compartments and functional modules in eukaryotes using *S. cerevisiae* as a model organism, a concept that is missed when exclusively focusing on transcript levels.

Like other genome-wide analyses, SAGE analysis of the yeast transcriptome also has several limitations. For example, a small number of transcripts that lack the appropriate site for tagging could not be detected by this method. Second, there is a basal level of frequency only above which the transcripts could be detected. Despite these limitations, the SAGE method has established itself as a more accurate method to quantify global and local snapshots of gene expression.

**9.4.2.3 Chromatin Immunoprecipitation** Genome sequencing and microarrays have provided the ability to simultaneously quantify the expression of the entire genome to study transcription. The transcription of genes highly depends on the environment, and the level of expression of a particular gene is controlled by transcription factors (TFs), which bind to the specific DNA sequences upstream of the gene either inducing it or repressing it. Almost every gene in eukaryotic cell is regulated by several positive and negative TFs that recognize the specific binding sites. TF–DNA interaction in a living cell is a complex process with most TFs interacting with other sequence-specific binding proteins and general transcription machinery. These protein–protein interactions (PPI) may affect DNA binding characteristics of the TF of interest. A comprehensive understanding of where enzymes and their regulatory proteins interact with the genome *in vivo* would greatly increase our understanding of the mechanism and logic of critical cellular events. Detailed *in vitro* studies of DNA–protein interactions have provided, and will continue to provide, useful information; it is clear that studies of TF–DNA interactions are critical to understanding the cause and effect relationship between transcription and environment. However, these traditional methods of investigation have failed to create high-resolution, genome-wide maps of the interaction between a DNA binding protein and DNA. For example, the DNA binding properties of a protein determined by *in vitro* oligo selection or gel–shift assays are often poor predictors of a factor's actual binding targets *in vivo*. The study of TF–DNA interactions has undergone a major revolution by overcoming this limitation, owing to the development of combining chromatin immunoprecipitation (ChIP) with DNA microarray analysis (ChIP–chip analysis).

The first ChIP-to-chip experiments were reported at more or less the same time by the Young and Brown groups [77,131]. Both studied TF–chromatin binding in *S. cerevisiae*. Yeast is a good model system for TF–DNA interaction studies for many reasons, one of which is that its genome is much smaller than that of mammals, allowing genome-wide microarrays. DNA fragments from cells grown under controlled experimental conditions that are bound to the transcriptional regulators are recovered by a ChIP assay using an antibody specific to the protein of interest and are hybridized to DNA microarrays that contain the complete set of intergeneic regions. The strength of hybridization intensity signal of a particular gene reflects binding of the transcriptional regulator to the promoter site of that gene. Ranging from yeast to cultured mammalian cells, there is surprisingly little variation in published ChIP–chip protocols. This second generation application of microarrays reveals the network of genes that are bound by one or more transcriptional regulators and presents a very powerful experimental methodology into revealing the first step in transcriptional regulation by identifying gene sets that are bound by the same transcription regulators.

The ChIP–chip technique was first applied successfully to identify binding sites for individual transcription factors in *S. cerevisiae* [77,96,131]. Later, also in yeast, a c-Myc epitope protein tagging system was used to map the genome-wide positions of 106 transcription factors [93]. Other applications including the study of DNA replication, recombination, and chromatin structure have also been reported in *S. cerevisiae* providing a wealth of information on the transcriptional regulation governing these mechanisms. In these experiments, microarrays containing ~1 kb PCR products representing ORFs, intergeneic regions, or both were used in conjunction with a two-color experimental scheme. The PCR products in these arrays were "tiled," across the genome, meaning the PCR products were directly adjacent to one another along the genome, with little or no DNA sequence between arrayed elements. The relatively compact and nonrepetitive nature of the simple genome harbored by yeast made such an approach feasible.

Based on known regulatory information gleaned from biochemistry, gene expression, and ChIP results, it was demonstrated that the strength of interactions between transcription factors and genes is context dependent in *S. cerevisiae* [104]. Studying the changes in gene expression patterns in response to changes in cell cycle, sporulation, diauxic shift, DNA damage, and stress, it was concluded that a few transcription factors are always involved in regulation whereas others depend on the stimulus, thus constantly reprogramming the regulatory network. Only a few target genes are expressed under a specific condition. One of the ramifications of this conclusion, based on over 7000 interactions between genes and transcription factors in *S. cerevisiae*, is that one must use caution when extrapolating the interactions and regulatory mechanisms identified under condition to another.

Recently, an *in vitro* DNA microarray technology for genome-scale characterization of the sequence specificities of DNA–protein interactions was reported based on the ChIP–chip protocol [110]. This technology, known as the protein binding microarray (PBM), allows rapid determination of *in vitro* binding specificities of individual transcription factors by assaying the sequence-specific binding of those individual transcription factors directly to double-stranded DNA microarrays spotted with a large number of potential DNA binding sites. A DNA binding protein of interest is expressed with an epitope tag, purified, and then bound directly to a double-stranded DNA microarray. The PBM is then washed to remove any nonspecifically bound protein and labeled with a fluorophore-conjugated antibody specific for the epitope tag. The PBM technology was used to compare the binding site specificities of the three yeast TFs, Abf1, Rap1, and Mig1 *in vitro* and *in vivo*. The PBM-derived binding site sequences are reportedly more accurate in identifying *in vivo* binding sites. In addition to previously identified targets, Abf1, Rap1, and Mig1 have been reported to bind to several new target intergeneic regions, many of which were upstream of previously uncharacterized open reading frames. Comparative sequence analysis indicated that many of these newly identified sites are highly conserved across five sequenced *sensu stricto* yeast species and, therefore, are probably functional *in vivo* binding sites that may be used in a condition-specific manner [110].

Although the ChIP–chip method can only map the probable protein–DNA interaction loci within 1–2 kb resolution, it also fails to distinguish between positive and negative regulation. The development of the ChIP–chip assay has provided an extraordinarily powerful tool for the analysis of DNA–protein interactions in living cells or tissues on a global scale. In the near future, further advances in microarray construction and the increased availability of useful antibodies will increase the utility of this approach even more. Genomic profiling of transcription factor binding sites, histone modifications, and so on will almost certainly emerge as a central tool in understanding the systems biology of gene regulation in eukaryotic cells. In addition, studies of the genomic distribution of nuclear proteins that are not sequence-specific DNA binders, such as general transcription machinery, the proteasome and its component pieces, DNA replication and repair complexes, and so on will shed new light on fundamental aspects of basic genome function and maintenance. Already, the realization that the majority of transcription factors examined to date are localized outside the promoter sequences has contributed significantly to our growing realization of the importance of abundant noncoding small RNAs in the cell.

### 9.4.3 Proteomics

Even though global changes in gene expression provide deep insights into understanding transcriptional control, proteins have to be recruited to perform the process since they are the actual functional units. Therefore, knowledge of protein abundance reveals the extent to which regulatory proteins and transcription binding factors participate in the resulting change in gene expression profile. Since gene function is heavily associated with proteins, analysis of proteins will divulge more information on protein function and the pathways they act on. Moreover, although proteins are the end products of gene transcription, there is no one-to-one correspondence between the number of proteins and the number of genes. Therefore, mere transcriptome analysis does not reflect the functional profile at the protein level. This section will outline the emerging quantitative proteomic techniques that are often first developed and tested in *S. cerevisiae*. The focus will primarily be on the two major proteomic technologies that are commonly in use, 2D gel electrophoresis and liquid chromatography coupled to mass spectrometry (LC–MS). The applications of these technologies to investigate protein expression levels of yeast grown under different growth conditions and its implications on systems biology are also discussed.

***9.4.3.1 2D Gel Electrophoresis and LC–MS*** The most common trend in analyzing proteomes employs two-dimensional gel electrophoresis to separate proteins, followed by mass spectrometry to identify proteins. On a 2D gel, proteins are separated using isoelectric focusing (separation based on isoelectric point of proteins) in the first dimension and sodium dodecyl sulfate polyacrylamide gel electrophoresis (separation based on molecular mass of proteins) in the second dimension. The separated proteins can be visualized using a variety of staining methods such as Coomassie blue dye, silver staining, or fluorescent dyes. Generally, in the first dimension, the proteins are brought on a strip that contains an immobilized pH gradient. By applying an electric field over this strip, the proteins will migrate over the strip until they reach the pH area on the strip where they will be neutral. Each protein therefore will be separated and focused on the strip at the position of its isoelectric point. In the second dimension, proteins are separated on their size/mass. On the resulting two-dimensional gel each protein is present at a position that reveals its approximate pI and mass. Although the concept of 2D gel electrophoresis was introduced more than 30 years ago, its application to proteomics has really taken off since the development of MS-based techniques that enabled high-throughput protein identification.

As with other high-throughput technologies where *S. cerevisiae* was one of the first organisms in which these methods were tested and were subsequently used to conduct genome-level interrogations, some of the early proteomic studies in this context were performed in *S. cerevisiae*. These early large-scale separation and visualization of protein resulted in yeast reference maps, which can be used to locate and identify proteins. The digitalized image maps from these experiments were established by which annotated proteins can be localized and identified directly from the image. For example, the SWISS-2D PAGE yeast database at http://www.expasy.org/ch2d/2d-index.html presents the 2D protein pattern of yeast in the pH range 4–9 with 101 spots identified and localized in this area so far. The yeast protein map at http://www.ibgc.u-bordeaux2.fr/YPM/ contains a protein pattern of pH 4–7 with 410 proteins identified. Depending on the protein staining method, approximately 1000 proteins can be visualized on such gels. Also, subproteome reference maps of, for example, yeast mitochondria, have been generated [116,146]. Similar 2D reference maps have been constructed for important industrial yeast strains, such as an ale-fermenting strain, a wine strain, and a lager-brewing strain. These annotated reference maps are useful tools for yeast researchers because they can be used for 2D gel comparisons; however, because of poor gel-to-gel reproducibility and strain variation, protein spot identities should always be confirmed using MS.

Although the conventional trend in analyzing proteomes using two-dimensional gel electrophoresis has had a good turnover of information, the greatest drawback in this method is that it is heavily biased toward proteins expressed at high concentrations [55]. It is also extremely labor intensive and is often hampered by poor gel-to-gel reproducibility. Different staining methods have been developed to improve the accuracy and the sensitivity of protein detection and quantification [125,157], yet proteins expressed at low concentrations may not be detected accurately. Therefore, mass spectrometers are used to detect and identify proteins on a 2D gel. Nowadays, an ordinary mass spectrometer can precisely determine the masses even of large proteins (approximately 1 Da precision at 50 kDa). Since determining only the mass of a protein does not give a direct clue about its identity, two MS techniques are commonly used for protein identification. In the first method, a peptide fingerprint of a protein is recorded, usually by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry. In the second, slightly more complicated method, short amino acid sequences, the so-called sequence tags, are determined by tandem mass spectrometry.

In the first of the two approaches, the protein spot to be identified is cut out of the gel and digested (in-gel) with a protease, most often trypsin. The resulting peptide mixture is eluted from the gel and analyzed by MALDI–TOF mass spectrometry or alternatively by using electrospray ionization mass spectrometry. Collectively, these peptide masses form a fingerprint, which is indicative for the protein concerned. This fingerprint is then compared to theoretically expected tryptic peptide masses for each protein entry in the database. Generally, peptide finger printing is still the most rapid and efficient method for protein identification. As identification occurs via consultation of protein and genome databases, it may be apparent that their increasing comprehensiveness greatly aids in protein identification. In the second method, the peptide of interest is fragmented in the MS and mass analysis of the resulting fragments allows determining the amino acid sequence from the peptide. Although these fragmentation patterns maybe quite complicated, they generally allow the determination of partial sequences. With this partial sequence, possibly in combination with the peptide fingerprint already obtained, the chance of a unique hit in the database is considerably enhanced. With one or two of these short sequence tags (often no more than five amino acids), it is often possible to unambiguously identify a protein. These strategies, as with other technologies, come with inherent drawbacks. For example, only those proteins that can be visualized on a 2D gel can be analyzed. The 2D gels are incapable of handling large proteins and in general have bad reproducibility and are extremely cumbersome.

In recent years, proteomics methods that employ LC–MS have proven to provide strong alternatives. LC-based technologies have several advantages compared with 2D gel-based techniques. LC–MS, which can be automated, combines high-speed, high-resolution, and high-sensitivity separation of extremely complex peptide mixtures. Several 2D gel independent LC–MS–MS approaches have been introduced to overcome some of the inherent disadvantages of 2D gels. In one approach the proteins in the total proteome are only separated and resolved by molecular mass using 1D gels. Subsequently, this 1D gel is cut into pieces, all proteins in such a band are digested, and the mixture of peptides is analyzed by LC–MS and/or LC–MS–MS [129]. This approach provides an intermediate form between analyzing the very complex large peptide mixture obtained when digesting all proteins of a lysate and the single protein digested when using a 2D gel. As advantages over the 2D gels, a 1D gel-based approach is less elaborate. In addition, very large and basic proteins are easier to handle using just one-dimensional gels. In a third approach, the whole cell lysate is digested chemically or by a protease. This generates a very complex set of peptides, beyond the separation capacity of 1D separation techniques. For the analysis of such complex mixtures, several multidimensional separation techniques have been introduced. An example of an innovative online 2D chromatographic approach is the MudPIT, the multidimensional protein identification technology [169]. In this approach, the complete cellular protein mixture is protealyzed and the resulting peptide mixture is then separated and analyzed using online 2D chromatography directly coupled to tandem MS, which enables the identification of proteins by peptide sequencing. In one of the first applications, MudPIT was used to analyze the yeast proteome and a total of 1484 could be identified [169]. The resulting data set identified proteins from all subcellular compartments, with wide-ranging isoelectric points and molecular weights. Moreover, low abundance proteins, such as transcription factors and protein kinases, as well as hydrophobic membrane proteins, were detected. More recently, the MudPIT method was improved by adding an additional reversed phase column to the biphasic column, resulting in an online 3D LC method [171]. Using this method, it was possible to identify 3109 yeast proteins, which is the most comprehensive proteome coverage reported to date.

Alongside the continuing efforts to develop reliable methods to quantify the proteome, an important advancement in our understanding of the function is the global identification of protein localization in the cell [47,69]. Information about the localization of a protein reveals its function, activation state, and its potential interactions with other proteins particularly in eukaryotic cell, which is compartmentalized. For example, in *S. cerevisiae*, 82 new proteins were discovered in the nucleolus and were predicted to be involved in ribosomal function and, in general, the localization results had 80 percent agreement with the data in the Saccharomyces Genome Database [69]. This study confirmed previously known protein–protein interactions in addition to identifying new ones such as those between cell structure and morphology. Localization of proteins depends on the cell signaling events and their state of activation, which depends on the environmental conditions. Such intercompartmental translocation of proteins triggers new signals. Among the various methods used to study protein localization, variants of GFP are commonly used to tag the protein for visualization using a light microscope [23,69]. The dynamic nature of protein synthesis and consequent modifications, identification, and quantification of proteins alone may not be sufficient. It is also necessary to identify complex formation *in vivo* to obtain a systems view of cellular functioning.

***9.4.3.2 Two Hybrid System*** An important goal of systems biology is the identification of functional interactions between different cellular components. Since microarrays and 2D gels cannot contribute to the knowledge of protein interaction, protein–protein interactions play a crucial role in elucidating the nature of these mechanisms. Recently, innovative methods for a comprehensive analysis of protein interaction events and signaling pathways have been implemented to provide additional information such as the high-throughput yeast two-hybrid (Y2H) system. The yeast system provided the perfect platform for this assay since it had the advantages of speed, sensitivity, and simplicity in addressing an important biological question when the identification of an interacting protein following its purification was difficult. The Y2H system detects interaction of two proteins by their ability to reconstitute the activity of a split transcription factor, thus allowing the use of a simple growth selection in yeast to identify new interactions. Although the test case for this assay was only a single example of yeast proteins previously known to interact, the results led to the suggestion that the approach might be applicable to the identification of new interactions via a search of a library of activation domain-tagged proteins. Subsequently, the Y2H also proved to be very applicable to study protein interactions in any organisms, although certain types of proteins such as membrane-bound or extracellular proteins were less amenable to this method. The Y2H assay was subsequently adapted to detect protein–DNA, protein–RNA, or protein–small molecule interactions as well as protein–protein interactions that depend on posttranslational modifications, that occur in compartments of the cell other than the nucleus, or that yield signals other than transcription of a reporter gene.

Since its introduction about 15 years ago [35], the assay largely has been applied to single proteins, successfully uncovering thousands of novel protein partners. In the last few years, however, two-hybrid experiments have been scaled up to the proteome scale to identify the complement of all the proteins found in an organism. In the first array-based Y2H of the whole proteome, 192 “bait,” proteins were used to survey interactions with 6000 yeast “prey,” proteins, resulting in 281 distinct protein pairs [156]. Using a similar strategy with more “bait,” proteins to search the yeast genome for protein interactions, 4549 interactions were deduced, out of which a subset of 841 protein pairs were classified as “core,” interactions, that is, highly reliable [75,76].

Despite its routine use, the classical Y2H suffers from the appearance of a large number of false positives, even though arrays and other confirmation experiments help to identify them. Two hybrid systems in other organisms such as bacteria or mouse have not been used for large-scale screens, making it difficult to identify if the reproducibility issue is specific to Y2H or if it is a general trait in all such assays.

***9.4.3.3 Protein Arrays*** Considering the pivotal functional role proteins play in defining the phenotype, it is important to quantify protein abundance as well as activity. In the lines of DNA microarrays, protein arrays are rapidly becoming powerful high-throughput tools to identify proteins, monitor their expression, and elucidate their function and interactions within them and, more importantly, the posttranslational changes that they undergo. Several properties of proteins make building protein microarrays more challenging than building their DNA counterparts. First, unlike the simple hybridization chemistry of nucleic acids, proteins demonstrate a staggering variety of chemistries, affinities, and specificities. Moreover, proteins may require multimerization, partnership with other proteins, or posttranslational modification to demonstrate activity or binding. Second, there is no equivalent amplification process like PCR that can generate large quantities of protein. Third, expression and purification of proteins is a tedious task and does not guarantee the functional integrity of the protein. Finally, many proteins are notoriously unstable, which raises concerns about microarray shelf life. Despite these challenges, the development of protein microarrays has begun to achieve some recent success. Currently, protein arrays come in two main formats. The first, abundance-based microarrays, seeks to measure the abundance of specific biomolecules using analyte-specific reagents such as antibodies. The second, function-based microarrays, examines protein function in high-throughput by printing a collection of target proteins on the array surface and assessing their interactions and biochemical activities. Although the applications of these arrays widely differ, they all function on the underlying principle of detecting interaction partners. Abundance-based microarrays include antibody microarrays and reverse protein microarrays. In antibody microarray, antibodies are immobilized and purified proteins and complex mixtures are screened for antibody characterization as well as to quantify protein abundance. Fractionated proteins or protein mixtures are immobilized in reverse protein microarrays and single antibodies are the target screen partners. Function-based microarrays include the standard protein microarrays, where the immobilized component is the protein itself and proteins, antibodies, DNA, or other chemicals are used as the screening partners in functional characterization of the immobilized proteins and to identify their interaction partners. By far the greatest obstacle in developing function-based protein microarrays is the construction of a comprehensive expression clone library from which a large number of distinct protein samples can be produced. In building a clone library, it is desirable to construct recombinant genes where fusion proteins can be produced for the purpose of affinity purification and/or slide surface attachment. Cloning the genes of interest with an inducible promoter allows individual proteins to be expressed in high abundance. High-throughput purification can be accomplished with the addition of C- or N-terminal tags, such as glutathione-s-transferase or the IgG binding domain of protein A. The incorporation of fusion tags also facilitates the verification of clone inserts by sequencing across the vector–insert junction. It is highly desirable to transform the expression vector into a homologous or related cell type, ensuring the proper delivery of the protein product to the secretory pathway and hence correct folding and posttranslational modification of each recombinant protein.

Using these protein microarrays for the first time, the binding activities of three known pairs of interacting proteins was investigated in *S. cerevisiae* [106]. One protein of each pair was printed in quadruplicate onto aldehyde slides, and the arrays were probed with the labeled partners. The most important outcome of this research was that the researchers were able to quantify the concentrations of the bound and solution phase proteins necessary to carry out the experiments. Thus, these experiments demonstrated the feasibility of arraying proteins in a standard microarray format and at feature densities comparable with those of DNA arrays. In a subsequent study, a yeast high-density (13,000 samples per array) proteome microarray was developed that contained full-length, purified expression products of over 93 percent of the organism's complement of 6280 protein coding genes [183]. A total of 5800 ORFs were cloned as glutathione-s-transferase::His6 fusions, and expressed in their native cells under a Gal-inducible promoter. This work represented the first systematic cloning and purification of an entire eukaryotic proteome as well as the first large-scale functional protein array comprising discrete functional proteins. Several different experiments were performed with the arrays, including a calmodulin binding survey to assess protein–protein interactions and a large-scale screen for phospholipid binding specificity [182]. More recently, these proteome chips were used to study global protein phosphorylation in yeast [128], and this study identified over 4000 phosphorylation events involving 1325 proteins from a wide range of biochemical functions and cellular roles. It was also found that these interactions even occur across different compartments, and have helped construct the first draft of a phosphorylation map for *S. cerevisiae*. These results are expected to provide valuable insights into the mechanisms and role of protein phosphorylation in many eukaryotes since several of these proteins are highly conserved.

In spite of these advances, the fundamental aspect that currently limits the advancement of proteomics (in contrast to genomics) is the lack of protein amplification mechanisms analogous to PCR. Therefore, only those proteins that are produced naturally in large quantities or by recombinant techniques can be analyzed. Nevertheless, protein microarrays have shown considerable promise in determining protein–protein, protein–lipid, protein–ligand, and enzyme–substrate interactions. Protein microarrays also have great potential in drug development and clinical diagnostics. We can expect protein microarrays for other organisms as well as for membrane proteins in the near future. Although there is no established proteomics technology to detect all the desired aspects of proteins, aggressive research in the area of proteomics reflects the pivotal role that proteins play in executing metabolic control. It is expected that proteomics will continue to be in the forefront of systems biology research.

### 9.4.4 Metabolomics

The cells control the concentrations of their intracellular metabolites very rigidly. There is normally a very low tolerance on the allowable variation in the metabolite concentrations for a given physiological state. Conversely stated, a change in the concentration of a metabolite beyond the tolerance level induces a change in the cell physiology. Since they are the intermediates of biochemical reactions, metabolites play a pivotal role in maintaining the connectivity in the metabolic network. Certain metabolites such as ATP or NADH, which are involved in a large number of reactions in the metabolic network, are capable of bringing about significant changes in large parts of the metabolism [115]. The level of the metabolites is a complex function of enzymatic properties and regulatory processes at different levels of information hierarchy. Therefore, similar to the transcriptome and proteome, the metabolome (global set of all the intracellular metabolites) also presents a snapshot of the physiological state of the cell and measuring the changes in the concentrations of intracellular metabolites would reveal an aspect of regulation (such as allosteric inhibition/activation, metabolite–DNA binding, and so on), which cannot be studied by any other omic approaches described. Indeed, metabolome profile presents a closer snapshot of metabolism than the transcriptome or the proteome, because the information flow at this level is the closest to the phenotype (Fig. 9-1). Metabolome profiling also presents a more complete representation of metabolism by defining the thermodynamic equilibrium of a reaction. Therefore, metabolite profiling is now considered an important part of systems biology, playing a complementary role to genomics and proteomics [153,154,170]. However, this field is still in its infancy, mostly due to the lack of analytical techniques. In comparison to more than 6000 protein-coding genes in *S. cerevisiae* [50], there are only about 600 metabolites in *S. cerevisiae* [118]. Thus, even though the goal of any metabolome experiment is to quantify the level of all intracellular metabolites in a cell, tissue, or an organism, there is no single analytical method that can measure all metabolites.

Although the technology to quantify and study the genome (consisting of 4 nucleotides as building blocks) and the proteome (consisting of 22 amino acids as building blocks) is developed based on the similarity in their structure, the metabolomics technology is vastly more complex owing to the highly diverse building blocks, ranging from carbohydrates and organic acids to volatile alcohols and ketones. Consequently, it is virtually impossible to simultaneously determine the complete metabolome with current technologies. Nevertheless, the phrase "metabolome analysis," is used to describe the experimental approaches employed to quantify or detect metabolites. Currently, it is possible to quantify about 50 metabolites (Fig. 9-7). Although metabolite profiling has long been applied for medical and diagnostic purposes as well as for phenotypic characterization, particularly in plants, it is only recently that efforts toward the development of high-throughput analyses are being undertaken [33,34,153]. Mass spectrometry and nuclear magnetic resonance (NMR) are the most frequently used methods of detection in the analysis of the metabolome. The NMR is very useful in determining the structure of unknown compounds, but comes with the drawback of expensive instrumentation. In addition, NMR has the advantages that it is nondestructive to samples and provides rich information on the structures of molecules in complex mixtures. On the other hand, MS is considerably more sensitive and comes with the identification of unknown and unexpected compounds. The combination of separating the metabolites using a gas chromatogram or liquid chromatogram coupled with the MS is transpiring to be the most promising technique for metabolite profiling, thus far. The reader is directed to a very comprehensive review for detailed description and analysis on the different analytical methods employed to identify and quantify metabolites [164]. The issues related with different sampling methods and subsequent processing of the samples, particularly from yeast cultures, are described in another paper [162].

We reported a novel derivatization method for metabolome analysis of yeast that enabled us to measure several metabolites in the central carbon metabolites as well as in the amino acid biosynthesis pathways. Using this methodology, we compared responses of the metabolite profile in a Δ*gdh1* (NADPH-dependent glutamate dehydrogenase) and *GDH2* (NADH-dependent glutamate dehydrogenase) overexpressed mutant and its isogenic reference yeast strains under aerobic and anaerobic conditions [165]. During aerobic growth, the level of all the TCA cycle intermediates increased in the mutant compared with the wild type, indicating a higher TCA cycle flux in this mutant (Fig. 9-8). An increased level of 2-oxoglutarate reflects an alteration in ammonium metabolism due to the thermodynamically less favorable glutamate synthesis using NADH as the cofactor. Moreover, an elevated level of all amino acids was observed, indicating a wide change in amino acid metabolism. More recently, we reported the identification of a pathway for glycine catabolism and glyoxylate biosynthesis in *S. cerevisiae* using metabolite profiling and combining it with pathway

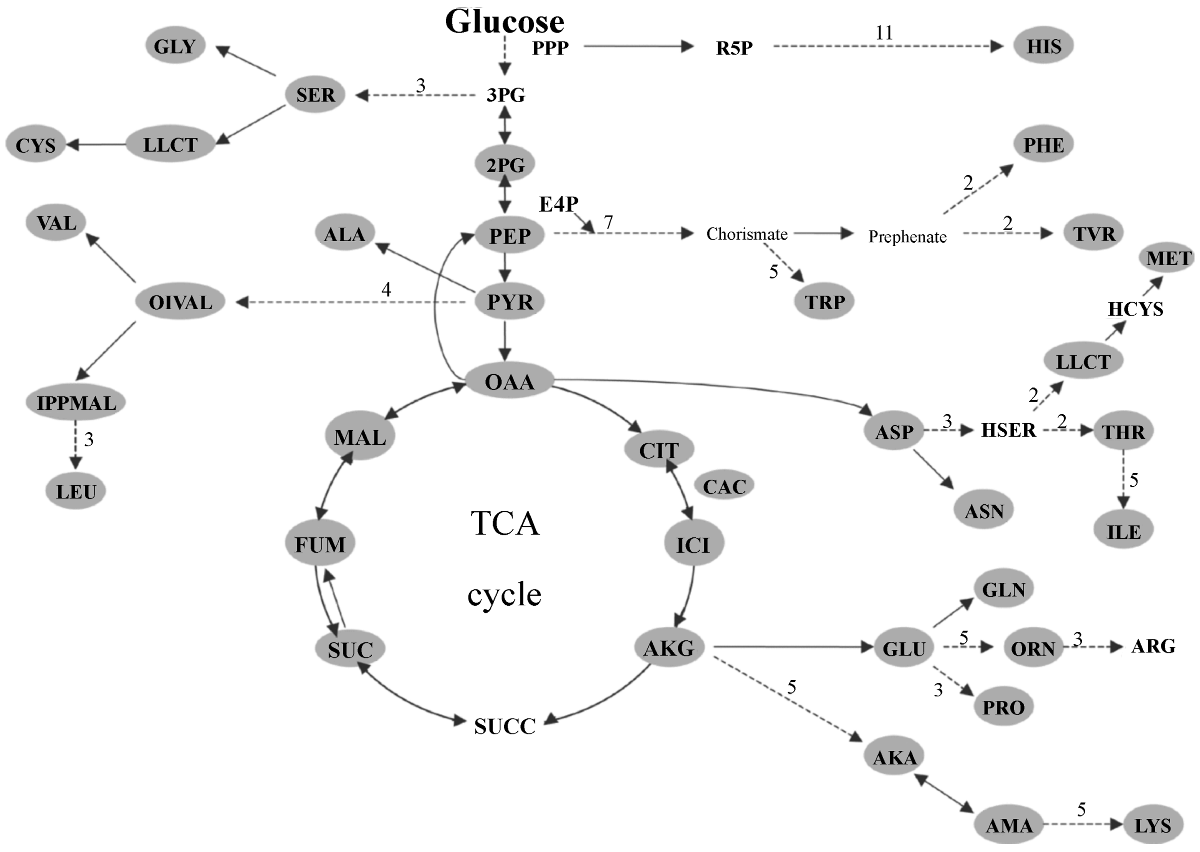


**Figure 9-7** Metabolites that can be quantified by the current state-of-the-art methods in metabolomics [162]. In addition to most of the amino acids, several sugar phosphates could also be quantified. However, there are several other metabolites that could be identified (but not quantified) on the chromatograms.

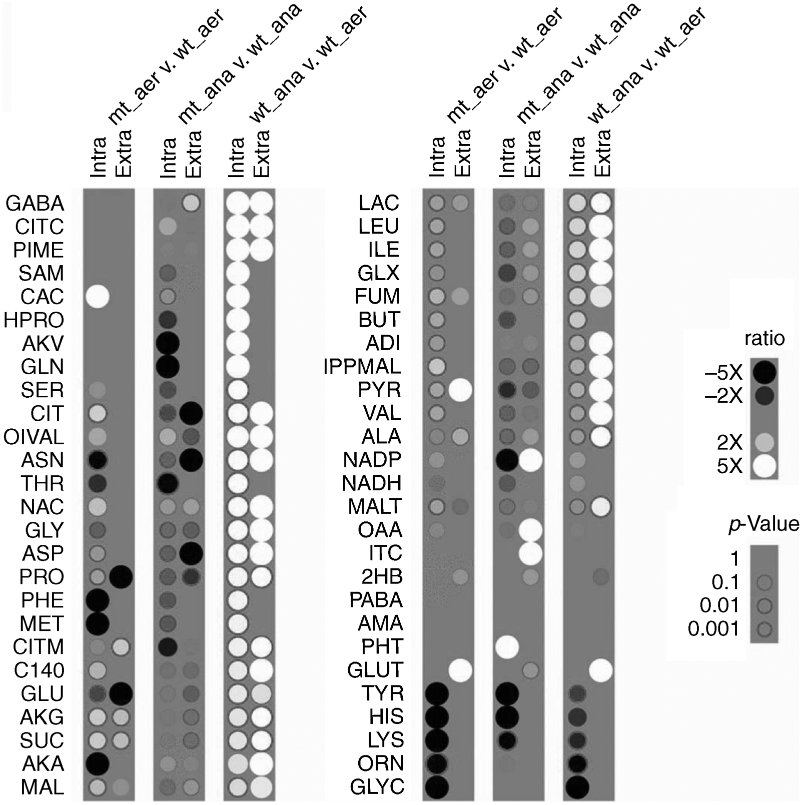


**Figure 9-8** Using the current metabolite profiling technology, we determined the differential levels of various central carbon metabolites [166] under various conditions, validating the method as well as laying the groundwork for an integrated transcription–metabolome studies in *S. cerevisiae.*

analysis [163]. Metabolic footprinting ability opens a new avenue in yeast systems biology research by providing results that neither gene expression nor proteome analysis could. Second, we demonstrated that the levels of specific metabolites could be quantified using this method, enabling the targeted and quantitative microbial metabolome analysis.

These examples demonstrate the immense potential of metabolite profiling in providing supplemental information to transcriptome and proteome analysis. However, there are a number of challenges for this nascent field. The fundamental problem arises due to the rapid turnover time of metabolite (in the order of 2–3 s), which makes it extremely difficult to capture a reliable snapshot of a metabolite profile. Second, the analytical methods for identifying and quantifying these metabolites are still in its infancy. Third, there is no robust data analysis methodology to integrate metabolite profile in the context of genome and proteome and interpret the physiological significance of an observed change in the metabolite level. Finally, the lack of standards in this field results in poor reproducibility. Recently, metabolomics ontology and experimental reporting standards have been proposed by the Metabolomics Society [100] to facilitate the establishment of credibility to the large amount of data that is being generated. Despite these challenges, there is growing belief in the scientific community that metabolomics holds the promise to expedite the progress of systems biology.

## 9.5 COMPUTATIONAL APPROACHES IN SYSTEMS BIOLOGY

The rapid advancement in the experimental approaches in measuring genome, transcriptome, proteome, and metabolome, as described in previous sections, gives rise to enormous data. It has now become clear that further discovery and progress in biological research will be limited not by the availability of data but by the lack of the right tools to analyze and interpret these data. Systems biology calls for the development of mathematical principles to integrate these high-throughput data. From its humble origins in control engineering and general systems theory, major challenges in the dynamic pathway modeling have been addressed and goals realized by (1) characterizing model structures that could realize the given stimulus–response relationship, (2) determining values for model parameters from experimental data and simulations, and (3) predicting the consequences of perturbations by introduction/removal of feedback or feedforward control loops. The new fields of genomics, proteomics, transcriptomics, and metabolomics are extremely essential not only to divulge information in the different levels of material processing in the cell but also to serve as precursors for realizing the larger objective of phenotypic characterization. Computational biology now plays a predominant role in the discovery process through automated genome reconstruction, flux balance analysis and metabolic networks, protein structure determination, and elucidation of regulatory networks. The synthesis arm of the systems biology cycle as depicted in Figure 9-1 heavily relies on the robust, reliable computational biology aspects. Since the ultimate cellular phenotype is the result of coordinated activity of multiple gene products and environmental factors, understanding the connectivity and interaction among these elements is pivotal.

### 9.5.1 Constraint-Based Genome-Scale Models

The key role of computational approaches in systems biology has been acknowledged and accepted. The development of mathematical models that can simulate the cellular phenotype by integrating high-throughput data forms the foundation of systems biology approaches. We view systems biology as an iterative process where mathematical models are built and developed based on the experimental data available. The goal of these models depends on the questions one is trying to address. For example, to describe and understand the metabolic dynamics, one uses a detailed kinetic model, or to study the mechanism of signaling cascades in a regulatory network, one formulates differential equations to describe the inputs to each module in the cascade and its response. The predictions and simulations from these models are then validated by experimental methods to complete one iteration in the process of understanding cellular properties. Any discrepancy between the model predictions and experimental observations will be addressed by incorporating the new experimental results in the model to start a new iteration. Since it is virtually impossible to determine the kinetics of all the steps in the system, these models have an obvious limitation. Among the several classes of mathematical models available to analyze cellular behavior, the constraint-based linear models are the only kind that can incorporate extensive biochemical data, genomic sequence data, and information from metabolic pathway databases into the context of simulating and predicting cellular metabolism and phenotype.

The development of linear metabolic models begins at the identification and functional annotation of the ORFs in the genome sequence. The Saccharomyces Genome Database (http://www.yeastgenome.org), Munich Information for Protein Sequences (http://mips.gsf.de/), and Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg) are the most commonly used databases to search for genes, their products, and their functional role in metabolism for *S. cerevisiae*. A comprehensive list of all the metabolic genes constitutes the genotype and an *in silico* representation of this subset of genes is the basis for creating of *in silico* strains. The gene products derived from the genes in the metabolic genotype carry out all the enzymatic reactions and transport processes that occur within the cell (Fig. 9-9). For example, the *in silico* representation of *S. cerevisiae* includes the genes involved in central metabolism, amino acid metabolism, nucleotide metabolism, fatty acid and lipid metabolism, carbohydrate assimilation, vitamin and cofactor biosynthesis, energy and redox generation, and macromolecule production (i.e., peptidoglycan, glycogen, and nucleotides). The reactions that are mediated by each of the gene(s) are represented as a linear equation, and all the stoichiometric coefficients from all the reactions are collected in the stoichiometric matrix, $S$, and the velocity (flux) of each reaction is collected in the velocity matrix $v$, which has the same number of rows as the number of reactions. Any exchange fluxes that are involved in material transfer across the systemic boundary are represented by the matrix, $b$. Under any given condition, the *in silico* metabolic network can then be represented as $S{\cdot}v = b$ [158]. The construction of the metabolic network is covered in greater detail in other chapters of this book. The fundamental drawback of these models is that they operate strictly on the stoichiometry and do not consider thermodynamic constraints and kinetics and, therefore, cannot resolve the directionality of the reaction. This inherent drawback is partly addressed by imposing constraints on the fluxes and defining their directionality and degree of utilization as $\alpha_i \leq v_i \leq \beta_i$, where $\alpha_i$ and $\beta_i$ are the lower and upper bounds, respectively, of the $i$th reaction. The values of the fluxes are estimated by imposing an objective function, often maximizing for biomass production rate, which is also expressed as an equation that is conceptualized by including the individual biomass precursors that contribute to the synthesis of biomass. Since typically the $S$ matrix is underdetermined (the number of unknown fluxes is far greater than the number of measured parameters), linear programming is the most commonly used procedure to estimate the unknown fluxes. Due to the existence of an infinite number of solutions in the feasible space and even the presence of several solutions of the flux vector that fulfils the objective

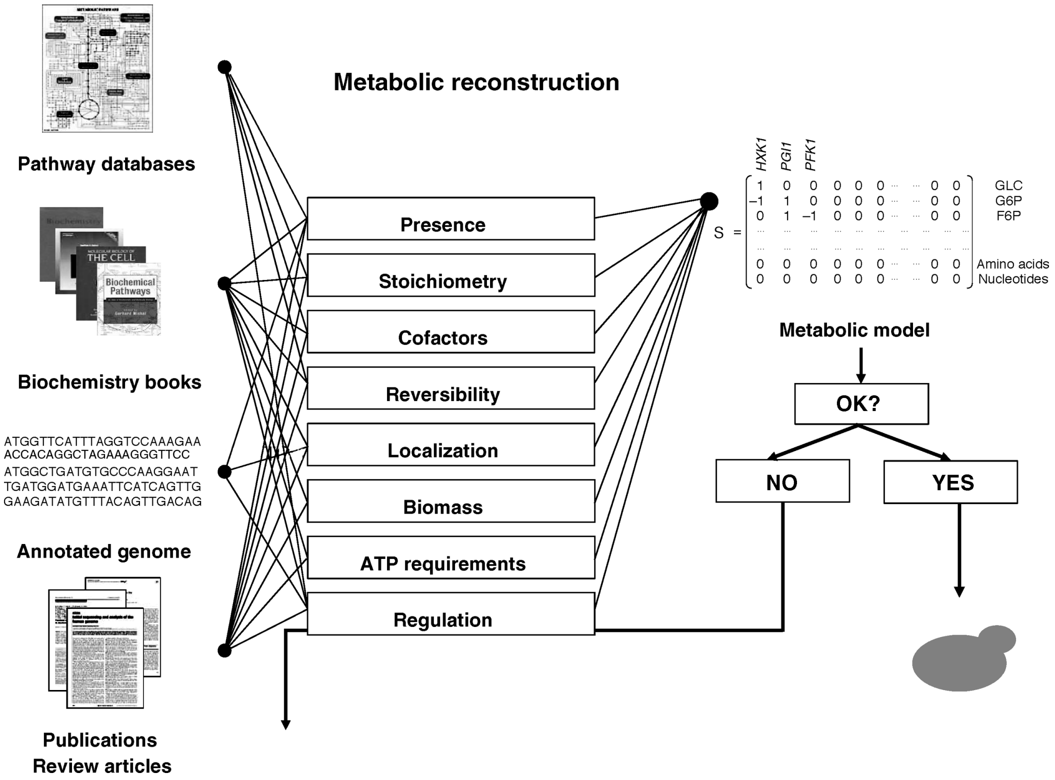


**Figure 9-9** A schematic methodology involved in constructing the genome-scale stoichiometric matrices. Information from various databases, knowledge from fundamental biochemistry, functional annotation of the genome of the organism, and information from various review and research articles are gleaned to assess the presence of a certain pathway and its stoichiometry. Additional information such as localization and reversibility are also incorporated in the matrix. Simulating the metabolism by flux balance analysis using this matrix and validating with experimental data will decide further fine-tuning of the model or its implementation to determine metabolic capabilities.

function, it is very likely that the estimated fluxes may not accurately represent biological reality [29]. Nevertheless, these models are extremely useful in characterizing the metabolic capabilities of the cell. The first genome-scale metabolic model of a eukaryote ever constructed was that of *S. cerevisiae* [39], which includes over 800 reactions and 500 metabolites. The details of this model can be viewed at http://www.cpb.dtu.dk/models/yeastmodel.html. Using this model, several aspects of *S. cerevisiae* metabolism such as biomass yields under various carbon sources, gene lethality and synthetic lethality, pathway utilization, and general network properties such as connectivity were successfully studied [39].

It is commonly observed that the performance of the *in silico* cell, such as the rates and yields of biomass and product formation, is far below the predicted theoretical maxima, particularly for strains that have undergone the first iteration of metabolic engineering. This phenomenon has detrimental impact on the utility of *S. cerevisiae* as a cell factory for commercial applications. This is due to the extensive adaptive mechanism of the cell to counteract any mutation. The stoichiometric models cannot predict this sluggish performance of the cell but rather provide the maximum cellular capabilities under the conditions of mutation. The discrepancy arises due to the assumption that maximizing biomass formation drives flux distribution. Even in response to a mutation, this approach assumes that the metabolic network could readjust to maintain optimal flux toward biomass. Recent evidence suggests that this could be achieved by selecting for fast growth [70]. However, in most industrial strain improvement scenarios, cells are subject to natural selection and a modification of the flux balance models using constraint-based linear programming approach is recently described to predict the sluggish metabolic phenotype [144]. This approach, known as minimization of metabolic adjustment (MOMA), assumes minimal response of the metabolic network to gene perturbations and suggests that the metabolic network has an inherent inertia to change and prefers to remain as close as possible to the original steady state (of the wild-type genotype).

### 9.5.2 Metabolic Pathway Analysis

The cell has multiple pathways at its disposal to attain its natural objective of survival. In the process of engineering these metabolic pathways, we attempt to manipulate these pathways to eliminate the ineffective ones or enhance the performance of the rate-limiting ones. Whether the goal is to delete pathways or overexpress them, it is necessary to develop an understanding of how the cell meets its metabolic objectives. This is the goal of metabolic pathway analysis. It is an integral analytical part in the discovery of meaningful routes in the metabolic networks, constructed as described in the previous section. By virtue of the complexity in the wide array of feasible metabolic pathways, it is not always intuitive which set of pathways are employed in reaching the cellular objective. The most commonly used mathematical tool that is used to analyze the set of all feasible pathways for robustness and efficiency is by elementary flux modes (EFMs) [142]. A flux mode is a steady-state flux distribution in which the proportions of the fluxes are fixed. If this steady-state solution is non-decomposable, then it is classified as elementary. In other words, an elementary flux mode is the minimal set of enzymes that could operate at steady state with all the reversible reactions assumed to proceed in the appropriate direction. Therefore, this concept assumes three conditions in determining an EFM: a pseudo-steady-state condition, a nondecomposability condition, and a feasibility condition.

Elementary flux modes are idealized representations of metabolism and it is very likely that any one EFM cannot represent biological reality. Instead, the real flux distribution is a linear combination of several EFMs, each of which has a fraction of contribution to the final flux. When *S. cerevisiae* grows on glucose, all the pathways that use other substrates are downregulated, even in the presence of a mixture of substrates. This is clearly demonstrated using DNA microarrays during the diauxic shift where 183 genes are induced and 203 genes are repressed at least fourfold [28]. This reflects a marked shift in the utilization of different metabolic modes, which are the likely superpositions of other EFMs. In fact, a very strong correlation was observed between the EFMs, as determined for yeast grown on different carbon sources, and the transcript measurements from microarray experiments [17]. Above all, the method of determining the control-effective fluxes to calculate the theoretical transcript values and correlating them with the experimentally derived transcript ratios demonstrates the importance of flexibility in metabolic networks. In this regard, the EFMs have a greater applicability over flux balance analysis. Moreover, since there is no objective function in this kind of analysis, unlike the flux balance analysis where the objective function is usually maximization of biomass formation, the system is not forced to behave in a particular manner. Since the metabolic reaction system is allowed full flexibility, it is free to choose all the possible routes toward product formation. Metabolic pathways analysis has also been used to assign function to orphan genes in *S. cerevisiae* based on convex analysis of its simplified metabolic network by combining metabolome analysis with metabolic pathway analysis [40]. Based on this analysis, a change in the pathway structure of deletion mutants could be combined with the different metabolite profile for that mutant to disclose the functionality of an orphan gene.

In many situations, the biosynthesis of a product is feasible by multiple routes and it is interesting to identify the pathways that give maximal yield. The optimal flux distributions, as predicted by the flux balance analysis, may not always be obtainable, thereby making it necessary to determine the suboptimal solutions using EFMs. The concept of EFM can also be used to predict the effects of an insertion or a deletion of a pathway, resulting in a pathway with new functional capabilities. This method allows a comparison of sets of admissible routes for product formation in wild-type cell and its engineered mutant. By comparing the elementary modes in the complete system with those in a deficient system, it can be shown whether or not an essential biological substance can still potentially be synthesized, via a bypass in the network system. Elementary flux modes essentially capture all the possible flux distributions (optimal as well as suboptimal) in the metabolic network as defined by the stoichiometric matrix, unlike flux balance analysis, which returns only one optimal solution. An important aspect of EFM that cannot be determined using the flux balance analysis is that of futile cycles. Futile cycles play an important role in regulation in eukaryotes and it is extremely important to identify them to avoid wasteful expenditure of cellular energy. Although such cycles are difficult to identify in large networks, they can be detected by calculating elementary modes, which include both cyclic and noncyclic metabolic pathways. This method is valuable for comprehending the complex architecture of cell physiology and together with other theoretical tools such as metabolic control theory, it can help to engineer living cells in a directed and rational way.

### 9.5.3 Gene and Regulatory Networks

The information pipeline in cells is extremely efficient and can robustly respond to multiple environmental and genetic signals. The mechanisms by which cells are able to achieve this are still not clear due to complex regulatory circuitry in the cell. To uncover the mechanisms that dictate the information processing, a modular approach is the most common approach [60,92]. The high degree of complexity involved in cellular response can be simplified by considering the large-scale genetic networks as composed of modules of simpler components that are interconnected through input and output signals, analogous to electrical circuits [130]. The analogy between genetic circuits and electrical circuits extends beyond just the superficial level. Just as electrical engineers construct circuits, genetic network engineers make use of the biological equivalents of inverters and transistors to manipulate living organisms by connecting these modules into gene regulatory networks that can control cellular function. Two landmark studies published in 2000 [31,43] clearly illustrate this concept, in which one describes a genetic circuit engineered into *Escherichia coli* cells that oscillates asynchronously with regard to the cell division cycle [31] and the other describes a toggle-switch circuit that can be switched between two stable states by transient external signals [43]. In both studies, the circuits' qualitative performance is consistent with the predictions of relatively simple differential equation models that characterize the dynamics of production, degradation, and genetic regulation.

The interactions between the functional modules in the gene regulatory networks involve proteins, DNA, RNA, and small molecules. For example, a simple module consists of a promoter, the genes expressed from that promoter, and the regulatory proteins that affect the expression of the promoter. The idea behind formulating gene networks and subnetworks is essentially to identify those genes that are commonly bound by the same transcription factor. Since the output of a microarray experiment is the end result of the interplay between transcription factors and genes, this aspect has been the focus of recent data analysis methods. Since one gene is under the control of multiple transcription factors, the amount of control from each transcription factor is not easy to quantify. Associating transcription with binding information for 106 transcription factors, Bar-Joseph et al. clustered coexpressed genes to reconstruct regulatory networks in *S. cerevisiae* [6]. They identified established interactions as well as discovered new interactions that they used to construct regulatory models. Liao et al. developed a similar approach called network component analysis to quantify the strength of interactions between genes and transcription factors [95]. The interactions were modeled as a two-layered network with transcription factors consisting of the first layer and the genes in the next layer and the interactions between the two layers as edges. Implementing this technique for glucose to acetate diauxic shift in *E. coli*, 16 transcription factors were found to be significantly involved in the transition. The biggest advantage of this method is that it does not assume independence or orthogonality of genes, unlike independent component analysis or the principal component analysis, respectively. Although these reports demonstrated the use of gene expression microarrays to study the regulation of specific pathways at the transcriptional level, they still do not account for regulatory effects brought about by proteins and metabolites interacting with DNA, and therefore such an approach would not be feasible in higher organisms with a greater level of complexity. As pointed out by Nielsen, the percentage of genes that are encoded for nonmetabolic functions (particularly for regulatory functions) increases with increasing cellular complexity [115]. To reveal regulatory phenomena based only on the changes in gene expression, detailed information about interactions between genes and their transcription factor proteins must be elucidated. Recently, it was demonstrated that a stochastic simulation algorithm can be efficiently implemented by using field programmable gate array devices to build a microelectronic circuit that simulates the kinetics of biochemical networks [139]. Such devices, built as an array of simple configurable logic blocks embedded in a programmable interconnection matrix, are ideally suited to implement highly parallel architectures comparable in complexity to biochemical networks. The parallel architecture of this logic-based programming can simulate the basic reaction steps in biological networks and since they can be scaled up efficiently, simulations of realistic biological systems should be possible.

We are still far from completely understanding the wiring of the regulatory circuit in a system, and the challenge lies in designing selection schemes that can be used to drive cells containing artificially engineered gene circuits for a robust, reliable, and noise-resistant behavior. The current paradigm for engineering regulatory circuits is to use computational methods to incorporate the desired changes in the cell. The engineered cells usually exhibit weak compliance with the desired objective and by using a directed evolution selection screen, more compliant mutants could be produced. The engineering of regulatory networks has immense applications in the production of industrial or medically important chemicals such as proteins and antibiotics and in the design of cells to perform complex multistage tasks such as conversions in bioremediations or cell-specific activity for gene therapy. A variety of relatively simple but useful types of biological circuits similar to switches, transducers, signal processors, sensors, and actuators are already being developed from the existing knowledge of the cellular components.

### 9.5.4 Protein–Protein Interactions

Proteins are the functional units in the cell and carry out most of the information processing such as intracellular communication, signal transduction, and even gene regulation via interaction with other proteins. Identification of protein–protein interactions on a proteome-wide scale is currently one of the main challenges of systems biology. Although the genome sequencing projects have identified the comprehensive set of genes and proteomic studies identified the protein abundance in several species, there are no thorough methods to identify interactions between proteins comprehensively. The current methods to identify PPIs include genetic and biochemical screens, which identify few interactions at a time but when applied in combination produce highly reliable results. Using a combination of expression profile reliability index to estimate biologically relevant fraction of protein interactions and paralogous verification method to score the interactions, over 8000 pairwise PPIs were detected in *S. cerevisiae* [25]. These interactions identified by such small-scale screens represent only a small fraction of the biologically significant interactions in yeast [53]. To identify the PPIs on a proteome scale, several high-throughput experimental methods have been developed such as the yeast two-hybrid assay described earlier [156], tandem affinity precipitation [44], and high-throughput mass spectrometry protein complex identification [64]. Although all these methods have the capability to detect thousands of interactions, their reliability is limited due to the high occurrence of false positives and false negatives. Besides these approaches, there also exist other approaches to infer PPIs based on indirect evidence, such as synthetic lethality [152], correlated expression of gene pairs [27], or identifying structural domains and subcellular localization [114]. Despite the relatively lower confidence of the interactions predicted by these methods, they are still very popular in elucidating PPIs. Recent reports indicated that integrating data from different levels of information hierarchy with these high-throughput methods significantly improve the reliability of the inferred interactions [51,80,180]. The computational aspect of predicting the interactions (e.g., between protein A and protein B) is usually based on the general criteria such as (1) they should have appropriate domains to facilitate interactions, (2) the expression levels of genes A and B should correlate, and (3) proteins A and B should be localized in the same compartment.

Using Y2H assays, proteins were assigned to functional classes on the basis of their network of physical interactions as determined by minimizing the number of protein interactions among different functional categories [159]. Such a functional assignment is proteome wide and is determined by the global connectivity pattern of the protein network. Using this approach, multiple functional assignments could be possible for a given protein, depending on its interaction with other proteins. This analysis is based on the concept that interacting proteins may belong to at least one common functional class, and thus knowledge of the functional classification of a subset of the proteins involved in the network may lead to an accurate prediction of the functional classification of the remaining subset of uncharacterized proteins. The idea behind this approach is to assign function to unclassified proteins based on its position in the interaction network, also known as the "majority rule," assignment [143]. The majority rule derives from the empirical observation that 70–80 percent of interacting protein pairs share at least one function. In most cases, only a few unclassified proteins interact with more than one protein of known function and often the interacting proteins with known functions do not generally share functionalities. In this respect, the majority rule assignment is inconclusive because the analysis does not include the links among proteins of unknown function. Therefore, much of the information contained in a reconstructed protein–protein interaction network is not used. A major concern in implementing network-based predictive methods is the topological accuracy of the protein interaction network. It is known that protein–protein interaction data obtained from two-hybrid experiments contain a certain number of false positive and negative results, as discussed earlier. These errors could compromise on the quality of the predictions by incorporating spurious connections into the network (false or missing edges).

## 9.6 INTEGRATING THE HIGH-THROUGHPUT DATA

The primary step in understanding any biological entity from a systems perspective is to identify its structural organization, such as the gene interactions and biochemical networks, followed by the dynamic interactions between them. Characterization of biological networks requires detailed maps elucidating proteins, RNAs, promoters, and other macromolecules. Toward this broad goal, metabolic networks [81], regulatory networks [93], and protein interaction networks [156] have already begun to be established. These maps are commonly represented as a static set of nodes to represent the components (RNA, proteins, macromolecules, transcription factors, etc.) of the network and edges to represent the interactions (activation/inhibition or induction/repression, etc.) between them. Human minds are incapable of inferring the emergent properties of a system from thousands of data points, but we have evolved to intelligently interpret an enormous amount of visual information. The data are therefore transferred to visualization programs. This is the initiation point for the formulation of detailed graphical or mathematical models, which are then refined by hypothesis-driven, iterative systems perturbations and data integration (Fig. 9-10). For example, using a bipartite graphical visualization, Patil and Nielsen showed similarities in metabolic network patterns and transcriptional responses that led to the identification of "reporter metabolites," in *S. cerevisiae*, which represent the hub of regulatory action [124]. Similarly, topological analysis of metabolism in 43 organisms revealed hierarchical modularity in the network organization [130]. Using the path of shortest length in graph-theory approach, Said et al. identified that the toxicity-modulating proteins in *S. cerevisiae* have more interactions with other proteins, leading to a greater degree of metabolic adaptation upon modulating the functioning of these proteins [137]. This result has direct implications on many human degenerative disorders such as cancer and even aging. The authors demonstrate that the protein interaction network is much more complex than the metabolic network, consistent with the knowledge that signaling pathways and regulatory networks have more complex organizational structure than the metabolic network. Although only protein interactions were studied, deeper regulatory aspects could have been revealed by also including protein interactions with DNA, particularly since the study focused on the recovery of *S. cerevisiae* from DNA-damaging agents. As opposed to the representation of biological networks as graphs that reflect only the static properties of system, de Lichtenberg et al. have recently reported the dynamics of protein interactions during the yeast cell cycle [24]. They used previously published gene expression data from different stages of the cell cycle [21,147] and integrating it with a network of

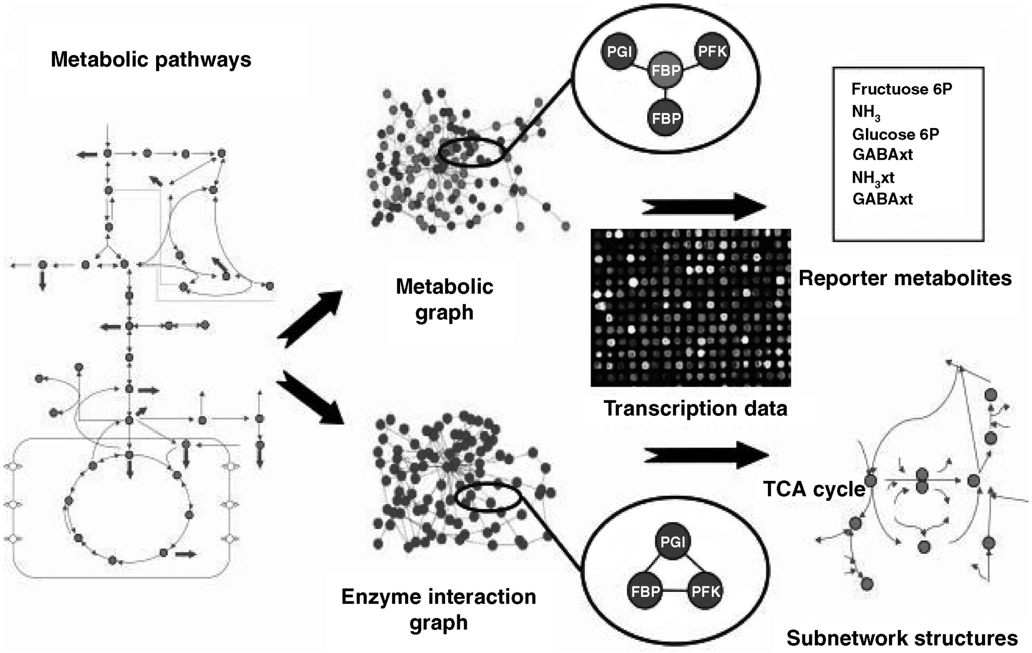


**Figure 9-10** The data integration methodology we have developed identifies reporter metabolites, which indicate the hubs of transcriptional regulation and the subnetwork structures [124]. Gene expression data from a particular experiment then is used to identify highly regulated metabolites (reporter metabolites) and significantly correlated subnetworks in the enzyme interaction graph.

physically interacting proteins from public databases such as MIPS discovered that most of the protein complexes are comprised of both constitutively and just-in-time expressed proteins. Currently, the mathematical models that represent cellular components and their interactions compromise either on the specificity or lack the sensitivity. This is due to several reasons, such as a limitation in biological information available and lack of mathematical rules to integrate the available information. Learning how the structure changes in response to various conditions and, more importantly, what makes the system respond in this fashion will enable identifying precise targets for metabolic engineering [86]. Established protocols are not immediately available to guide the merger of global information from various omes indicated in Figure 9-1. Ideker et al. [72] compared the global changes in the expression of mRNA and proteins in *S. cerevisiae* in response to a series of perturbations in the GAL regulatory system. They used the yeast galactose metabolic model as a prototype and studied the global responses to genetic and environmental perturbations. The key feature of this study that is missing from the previous comparisons was that the authors also considered protein interactions with other proteins and with DNA in their model. Not surprisingly, the expression of those genes that are linked by physical interactions exhibited a higher degree of correlation with corresponding protein levels. Information about protein–protein interactions in *S. cerevisiae* [143,156] facilitates the integration of the resulting mRNA and protein responses with known physical interactions to discover and/or refine gene functions. Since it is the proteins that actually execute the genetic program, mapping global interactions between proteins or "interactome," in single-celled [156] and multicellular [94] organisms is particularly valuable in revealing the signal transduction pathways, which play an integral part in overall regulation. These reports on transcriptome–proteome–interactome analysis communicate a unified theme, suggesting strong posttranscriptional as well as posttranslational control of metabolism.

Ihmels et al. [73] developed an integrated analysis methodology, called signature algorithm for *S. cerevisiae*, which analyzes patterns in gene expression changes over a large number of data sets with varying conditions to establish proximity between genes in terms of their expression under various conditions. Although this work did not incorporate changes in the metabolic profile as that of Ideker et al. [72] did, physiological changes were used to provide functionalities to genes, based on similarity profiles. The premise of organizing genes into transcription modules is that genes that are expressed similarly under a large variety of conditions are more likely to be coregulated than those clustered based on fewer conditions. This method was then used to study various cellular functions as well as the global transcription program. For example, applying this method to a *S. cerevisiae* data set, genes with previously unknown (or speculated) function such as YGR067C, YGL186C, and YJL1200C were identified with the regulation of the glyoxylate shunt, purine transport, and lysine biosynthesis, respectively [74]. An interesting discovery made by Ihmels et al. [74] was that only 63 percent of the isozyme pairs were not coregulated. An experimental validation of one such prediction of isozymes not being coregulated was that of the two glutamate dehydrogenases, encoded by GDH1 and GDH3. In a completely independent work, these isozymes were demonstrated to be nonredundant and their expression is carbon dependent [26]. This result agrees very nicely with the work of Kafri et al. [83] on identifying the nature of backup functions that genes perform. They argue that genes that are similarly expressed do not back up each other in the event of a mutation but rather through a transcriptional reprogramming mechanism that *S. cerevisiae* has evolved; paralogues for the mutated genes are activated only when the gene in question is inactivated. Although the authors did not discuss this aspect, this result might provide some clues to the nature of silent mutations. Hundreds of components in the cell are organized into modules and dynamically interact with one another. The consequent phenotype is a reflection of these dynamic interactions. Although there is no clear boundary between these modules, the probability of interaction of a component with $k$ other components, $p(k)$, has been shown to decrease according to the power law $k^{-2.2}$ [81]. However, few widely connected components such as ATP connect a large portion of the metabolism and result in an integrated module-free metabolic network. This dilemma has been resolved by demonstrating that metabolic networks are organized in highly connected modules that operate in conjunction with each other in a hierarchical manner [130]. Elucidating the principles that govern the nature and function of these individual modules may be possible with help from engineering, life sciences, and computer applications.

One of the several examples of such an integrative approach is that of identifying overlooked genes in *S. cerevisiae* [91]. Although the sequence information is extremely valuable, its ultimate utility lies in its accuracy and the completeness with which it is annotated. The yeast genome was sequenced and published to have 6274 genes, based on eukaryotic gene finding algorithms [108]. In the integrated approach, Kumar et al. [91] identified candidate genes by large-scale insertional mutagenesis using a modified transposon as a simple gene trap. The expression of each candidate gene is independently verified by microarray analysis. Only those gene sequences detected by both gene trapping and microarray analysis are classified as potential candidates. In this manner, they identified 137 previously overlooked genes in yeast, a majority of which are either short or overlap a previously annotated gene on the opposite strand. In yet another example of high-throughput data integration, the gene expression profiling and protein–protein interaction maps were integrated to compare the interactions between proteins encoded by genes that belong to common expression-profiling clusters with those between proteins encoded by genes that belong to different clusters [45]. The clusters derived from transcription profiling experiments were organized in a matrix, with each element of the matrix representing all pairwise combinations of genes either in a single cluster (diagonal or intracluster squares) or between two different clusters (nondiagonal or intercluster squares). This kind of a correlation approach suggested that the interactome data could help identify expression clusters with greater biological relevance. This study provides evidence that genes with similar expression profiles are more likely to encode interacting proteins and establishes a platform to integrate other functional genomic and proteomic data, both in yeast as well as in higher organisms.

The fundamental tenet of systems biology is capturing and integrating global data sets from biological systems from as many hierarchical levels as necessary. These include the static DNA sequences, context-dependent mass flow measurements in the form of RNA and protein quantifications, regulatory measurements such as protein–protein or protein–DNA interactions, and information flow measurements such as signaling pathways. The data collected from these measurements are transferred to a database where it is warehoused and analyzed for emergent properties systemic properties. The visualization methods described earlier permit a means to integrate the phenotypic features of the system directly to protein and gene regulatory networks. Cycles of iteration will result in a more accurate model to explain the subsystem or even the complete system (Fig. 9-2). Once the model has achieved sufficient level of accuracy and detail, it will allow biologists to accomplish tasks that remained elusive until now: predict the systemic response to a perturbation and redesign the regulatory networks to create new emergent systems. The second aspect of the systems biology will be addressed in greater detail in the next section. Therefore, fundamentally, systems biology is a hypothesis-driven, global, iterative, integrative, and dynamic branch in biological engineering.

## 9.7 SYNTHETIC BIOLOGY: STATE OF THE ART

Synthetic biology is a new and emerging direction that engineering of biological systems has taken. It is the synthesis of complex, biologically inspired systems that exhibit novel functionality, which do not exist naturally. This engineering perspective may be applied at all levels of hierarchy of biological structures. Therefore, synthetic biology is the design of biological systems in a rational and systematic way. The realization that the way to understand the cellular complexity requires a lot more than just compiling a ‘‘parts list,,’’ as provided by the genome sequencing, for example, has precipitated into the origins of synthetic biology. Elucidating the interaction between the parts is central to systems biology and is providing the necessary conceptual tools needed for synthetic biology. This nascent offspring of systems biology will share a symbiotic relationship with the fundamental sciences to expand on the biological control mechanisms using engineering approaches. These approaches include, but are not restricted to, the design and synthesis of novel genes and proteins, modifying the genetic code, altering regulatory mechanisms and signal sensing and enzymatic reactions, constructing multicomponent modules that impart complex phenotype, and even generating engineered cells.

The field of synthetic biology involves taking existing biological pieces, transforming them into micromachines, and creating artificial systems that mimic the properties of living systems. By creating systems that mimic what nature has created, scientists can discover the basic principles that rule living systems, manipulate these systems, and eventually find treatments for many diseases plaguing humanity. Today’s synthetic biologists are looking to channel genetic engineering from a hit-or-miss field of discovery to the type of discipline used by engineers to build bridges, computers, and buildings. This approach can translate into more specific anticancer therapies and antiviral drugs, as well as more efficient drug delivery systems that will have a significant impact on the health care industry.

### 9.7.1 Systems Biology and Synthetic Biology

Systems biology merely provides the analytical framework within which synthetic biology develops. The fundamental difference between systems biology and synthetic biology is that quite unlike systems biology, synthetic biology is not a discovery science (Fig. 9-11). It is a new way of constructing biology by adapting natural biological mechanisms to the requirements of an engineering approach. Similar to the mundane origins of systems biology described in Section 9.2, the first contribution of synthetic biology as defined above was made in 1964, when the first functional synthetic gene was made by a research team led by Khorana [84] as part of their work on elucidation of the genetic code. This gene, encoding tyrosine transfer RNA, was built from basic chemicals and was successfully tested in bacteria. Subsequently, this technology was automated and was used in making primers for polymerase chain reactions [111] and sequencing [140]. Since the simulation tools and models that are developed in systems biology could be used in synthetic biology, it is considered the design counterpart of systems biology. The design process demands sophisticated technology to target large number of components in addition to the high-throughput approaches. Therefore, synthetic biology will take some time before it matures to the status that systems biology is currently enjoying.

### 9.7.2 Synthetic Biological Circuits and Cascades

The discovery of signaling pathways controlling fundamental physiology [79] led to the application of nonlinear dynamics to understand gene regulation analogous to electric circuits and the development of the concept of a regulatory network. However, in the pregenomic era, the lack of sufficient experimental techniques precluded further expansion in this field. However, the recent explosion in the development of quantitative experimental methodology sparked interest in the elucidation of biological circuits and, more recently, introduction of synthetic circuits in biological systems. The simplest circuit is a transcriptional cascade, where genes are arranged in series and each gene product regulates the expression of one or more targets downstream in the series (Fig. 9-12). Although this concept has been optimized to perfection in natural systems using over evolution, in synthetic biology the networks are assembled from components that may not be related to each other. Therefore, the main obstacle in engineering synthetic circuits is to match the impedance of the individual elements such that they are kinetically functional in the context of the desired objective. There are two methods to optimize a synthetic circuit. The first one employs sensitivity analysis where randomly chosen kinetic rates are assigned to the functionality of the components and the contribution of each element's kinetics to the overall system behavior can be determined from analyzing the data from a large number of runs [32]. These data can be subsequently used to manipulate or fine-tune the system to achieve the end goal. The second approach is by directed evolution, which does not require detailed knowledge of the component kinetics. Directed evolution is most commonly achieved by subjecting a given component (usually a gene) in the circuit to a random mutation followed by a screening process to select the mutants that meet the

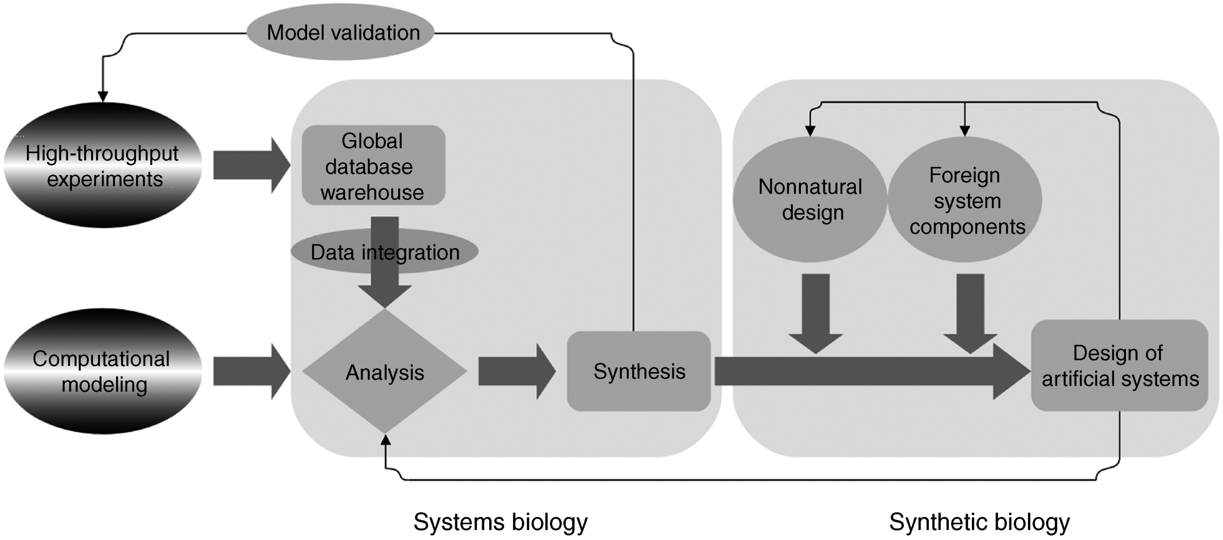
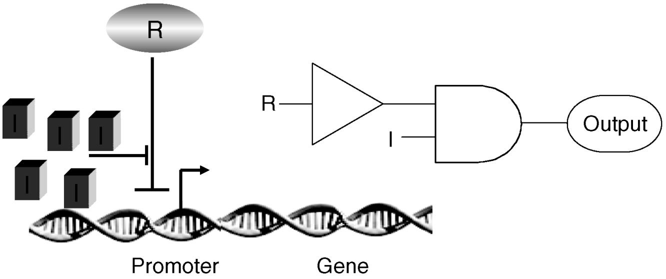


**Figure 9-11** We view synthetic biology as a specific case of the synthetic arm of systems biology. Information from the global X-omes is integrated with the aid of computational modeling to understand the system performance in systems biology. This will lead to new understanding of the system functioning, which can be modified by rewiring the system in a nonnatural way or replacing some of the systemic components with foreign components to impart new features that lead to novel functions. This approach has tremendous applications in industrial as well as medical biotechnology.



| R | I | Output |
|---|---|---|
| − | − | + |
| − | + | + |
| + | − | − |
| + | + | + |

**Figure 9-12** A simple depiction of how a gene regulatory network can be represented as an analogous electric circuit. The table on the right shows the conditions when the circuit will have an output. The circuit indicates how the output can be generated even in the presence of a repressor protein, making the gene circuit function like a switch.

desired criteria. This technique has been demonstrated in the optimization of a transcriptional cascade in *E. coli* [179] and cell-to-cell communication elements in *Vibrio fishcheri* [22]. The former strategy of rational design of synthetic circuits by computational approaches works well when properties dictating the component activity are well established, as is often the case with ribosome binding sites and operators. Directed evolution is more useful when the mechanism of the elements is not well known.

Signaling cascades are useful in elucidating the fundamental mechanisms of information flow in regulatory networks, and are usually characterized by having a steady-state output that is a monotonic function of the input. The steady-state behavior of most signaling cascades is similar to the digital logic with an ultrasensitive step-like dosage–response function, as illustrated in the case of mitogen-activated protein kinase in *S. cerevisiae* [67]. The response properties of transcription cascades also possess similar response characteristics. The dynamic and steady-state analysis of synthetic transcriptional cascades comprising one, two, and three repression stages has shown an ultrasensitive response to stimulus, and the sensitivity of the cascade increases as more elements are added to the cascade [66]. Synthetic transcriptional cascades have also been useful in studying noise propagation and in quantifying the contribution of intrinsic and extrinsic factors to phenotypic variations [126,135]. Most of the synthetic signaling cascades have been studied in prokaryotes, but long cascades are more common in eukaryotes, and therefore are more complicated. In contrary to the prokaryotic transcriptional cascades, eukaryotes exhibit a nonmonotonous response to stimulus, particularly in the presence of feedforward loops in the cascade.

### 9.7.3 Challenges for Synthetic Design

The construction of a functional synthetic network required assembling diverse genetic elements and getting them to work together. This process involves combining disparate components and tuning of biological parameters such as kinetic constants. Moreover, characterization of the circuit may not be valid under all conditions. To overcome some of these problems, several strategies have been suggested. First, the use of tunable elements such as transcription factors and promoters allow external control over some of these parameters. Second, the host cell in which the synthetic circuit has been integrated could be subjected to directed evolution in the laboratory and selected for optimized parameters. Another strategy is to implement a robust circuit design that is inherently insensitive to any kind of stimulus. These strategies have their basis in natural selection and are extremely useful in incorporating synthetic circuits in biological systems.

Another aspect of designing synthetic circuits is that of computational modeling. Simulation of synthetic models is essential for both the analysis of natural systems and also for engineering synthetic ones. Some of the problems that complicate the straightforward application of mathematical modeling to synthetic circuits include parameter sensitivity, lack of mathematical principles to model the complex biological circuits, and the inherent difficulty in distinguishing signal from noise in the circuits. On the positive side, synthetic circuits are simpler and, therefore, are better characterized than their natural counterparts; they will serve as ideal test systems to study their natural counterparts.

Currently, synthetic biology offers the ability to study cellular regulation and behavior using *de novo* networks. However, in the future, synthetic biology is expected to greatly contribute to the progress of medicine, biotechnology, and other areas of biology. The true potential of synthetic biology will be realized when the synthetic regulatory cascades mentioned above are interfaced with sensory inputs and biological response outputs. The inputs permit noninvasive monitoring of external environmental conditions and internal cell state and the outputs enable the engineered circuitry to control metabolism, cell cycle, and so on. Although the ability to program cell behaviors is still in its infancy, it is clear that the power to freely manipulate the set of instructions governing the behavior of organisms will have a tremendous impact on our quality of life and our ability to interact with and control the physical world surrounding us. One important difference between established quantitative engineering disciplines and synthetic biology is that state-of-the-art biological modeling tools still do not offer the same level of precision and predictive power. The future of synthetic biology looks very promising, with two goals clearly becoming obvious: understanding natural circuits by mimicking the natural systems and discovering what alternate nonnatural circuit designs are possible given the biological components. These hold the promise for immense potential in industrial as well as medical biotechnology.

## 9.8 COORDINATED RESEARCH IN YEAST SYSTEMS BIOLOGY

We are currently witnessing a transition in the approach to yeast physiology from traditional macroscopic procedures to a molecular approach and from a reductionist approach to an integrated approach (Fig. 9-2). Research in the field of systems biology and engineering is primarily driven by its end use and the quest for fundamental understanding. Truly comprehensive approaches to systems biology lie at the confluence of pure basic research and use-inspired basic research. Since such comprehensive approaches seem to be the future trend in studying physiology, it is necessary to establish a common platform to enable effective information exchange between different research groups. The generation of high-throughput global data that will be used in the integrated methodologies will prove to be an expensive venture and will undeniably require extensive knowledge about computer modeling, physiology, and metabolism, as well as excellent technical skills in measuring gene and protein expression and metabolic flux analysis. Although the current trend of generating high-throughput data is increasingly popular, we believe that there is extremely useful information that could still be extracted from the data that are already generated. Such a multidisciplinary approach paves the way to establishing strong symbiotic research collaborations. In this vein, there is also an increased government funding for systems biology. Notably, the U.S. National Institute of Health's roadmap for medical research provides $2.1 billion in funding over 5 years with heavy emphasis on systems biology, computational biology, and interdisciplinary programs. On a smaller scale, the U.S. National Science Foundation has launched a funding initiative entitled, "Quantitative Systems Biology FY 2004.," Also in the United Kingdom, BBSRC and EPSRC have launched a focused research program on systems biology resulting in the establishment of six national research centers tackling different aspects of systems biology. These and similar initiatives worldwide are catalyzing a renaissance in systems biology with special emphasis on producing a new generation of researchers trained in their core discipline and in complementary fields as well. We will focus on the European efforts in performing coordinated research in yeast systems biology.

### 9.8.1 European Functional Analysis

The European Functional Analysis (EUROFAN) network precipitated from the Yeast Genome Sequencing Network, which played a key role in yeast genome sequencing efforts. The goal of EUROFAN, which was established 2 years after the yeast genome sequence was published, is to provide a central repository of yeast mutants and characterize their transcriptome and proteome profiles to elucidate the biological function of novel genes revealed by the yeast genome sequence. The systemic functional analysis of the yeast genome is not intended to replace the regular biological enquiries that are conducted to answer specific questions. There are established approaches that permit the study of biological significance of every gene with increasing specificity. The approach implemented by EUROFAN is very efficient since it is not necessary to perform the global analyses on all the single gene disruption mutants. The ultimate idea of this project was to distribute the novel genes among the various laboratories in Europe, where additional information about its physiologic significance is evaluated. EUROFAN has now a well-curated database of gene function for most of the novel genes, an effort still in progress, and also serves as a genetic archive and stock center comprising yeast strains containing specific deletion mutants containing the individual genes as well as disruption cassettes allowing their manipulation in any laboratory or industrial yeast. This resource partly runs under the patronage of the Yeast Industrial Platform (YIP), which ensures rapid and efficient technology transfer to maintain European leadership in industrial yeast research.

The EUROFAN Project B0 has characterized more than 700 novel genes with respect to growth and morphology of deletion strains at several conditions of media and temperature. Project B1 has carried out quantitative phenotypic analysis of 564 deletion mutants with respect to 31 inhibitory chemicals and temperature shift [9]. Other EUROFAN projects have focused on other postgenomic technologies or have characterized the deletion mutants for phenotypes according to the specialties of the participating laboratories. The EUROFAN projects represent a major source of phenotypic data for the novel nonessential genes targeted by the European consortium. Details regarding the EUROFAN reports can be searched from the MIPS site (http://mips.gsf.de/proj/eurofan) and the deletion mutants, plasmids containing individual genes, and disruption cassettes are available at EUROSCARF (http://www.uni-frankfurt.de/FB/mikro/euroscarf/index-htlm) or Research Genetics (http://www.resgen.com). The MIPS primary gene query page has a link to the gene-specific EUROFAN data but the results of the EUROFAN functional analyses have not yet been linked to any yeast genome database and consequently, there is no single downloadable compendium of the EUROFAN data. However, the results from the B0 project have been curated into YPD. A resource such as EUROFAN laid the foundation for thorough high-throughput research using yeast to serve as a model as well as a tool.

### 9.8.2 Yeast Systems Biology Network

Systems biology is, by definition, multidisciplinary. It requires close collaboration of various laboratories specializing in experimental as well as theoretical disciplines to exploit the variety of methods to describe complex interactions in the yeast system. Thus far, it has not been possible for any single lab to possess the economic capability that is required to perform a variety of high-throughput experiments at the genome, transcriptome, proteome, metabolome, and the fluxome levels as well as the computational capability required to integrate data from these experiments to qualify for a true systems approach. Therefore, it is only through the coordination of activities in different labs such a systems approach can be realized. Despite the extensive government and private funding that the yeast systems biologists are enjoying worldwide, there is no concerted multilaboratory effort to coordinate and pool the individual competences toward studying yeast. The recognition for a unified effort for a symbiotic collaboration in yeast systems biology precipitated in launching the Yeast Systems Biology Network (YSBN) at the XXI International Conference on Yeast Genetics and Molecular Biology in Göthenburg, Sweden [65]. This alliance is expected to provide a platform for fostering collaboration between experimental yeast biologists and theoretical modelers in the "systems community.," The integrating platform for the alliance will be an internet-based functionality that generates a global virtual research community. The wider vision of the YSBN as part of both the yeast research and the emerging systems biology community is to work toward a comprehensive understanding of the function of the yeast cell, which will continue to serve as a paradigm for all eukaryotic cells. The other objectives of the YSBN are developing experimental and computational methods to iteratively improve the integration of experimental data in computational models and using the experimental approaches, in turn, to validate the predictions. This will facilitate easy data sharing for establishing standards for generating and documenting high-throughput data. Another important goal is to establish competence centers at regional as well as global levels for training of students and researchers. It is also necessary to spread the awareness of yeast systems biology among the general public and at the level of school education. These activities, in turn, are expected to increase the visibility of the YSBN to attract funding and financial support for yeast systems biology. An update on the progress of this international collaboration and its future activities and conferences are recently published [112].

## 9.9 SOCIETAL IMPACT OF YEAST SYSTEMS BIOLOGY

Systems biology, with its interdisciplinary approach to devising computational models of complex biological systems, may very well hold the key to unlocking the true value of the genome. There are vast commercial opportunities available for the pharmaceutical, human health, biotechnology, diagnostics, and agribusiness industries within systems biology. Market projections made by Research and Markets (http://www.researchandmarkets.com) for systems biology products and services are expected to grow at an annual compound rate of 66 percent to $785 million by 2008. *S. cerevisiae* has long served as a model eukaryote by virtue of the plethora of tools with which it can be manipulated genetically. In this section, we will illustrate some cases where similarities between yeast genes and human genes have been exploited to understand the mechanism of disease to improve human health and drug discovery in the pharmaceutical industry. We will also provide some case studies where novel metabolic engineering strategies in yeast have aided the bioprocess industry.

### 9.9.1 Human Health

The identification of several of the orthologues of human disease genes in this yeast has made it indispensable tool as a prototype system for medical research. Importantly, genetic dissections of yeast physiology serendipitously led to significant advances in our understanding of several human diseases, most notably cancer, through the exemplary studies on the regulation of the cell cycle performed by Hartwell et al. [59]. More recently, however, the genetic and biochemical tools available in yeast have been recruited for the purpose of directly examining the molecular basis and to aid in the treatment of several human diseases. High-throughput screening methods using the technology described earlier have been used to identify novel pharmacological targets produced in yeast or, through the two-hybrid screen, to obtain protein partners of medically relevant gene products. Moreover, the heterologous expression of proteins in yeast that lead to human disease has been used to uncover physiological responses to these proteins; yeast also encode homologues of several disease-causing proteins. In particular, the expression of specific proteins in yeast that fail to adopt their proper conformations or whose conformation lead to a pathological state in humans has helped us to understand how “conformational diseases,” arise and how eukaryotic cells respond to malconformed polypeptides [20,90].

***9.9.1.1 Mitochondrial Disorders*** Despite the extensive research conducted on the structure of mammalian respiratory complexes, our knowledge of mitochondrial biogenesis in humans relies on yeast genetics and biochemistry. Human cDNAs have been isolated based on their homology with newly discovered yeast genes and have been used to rescue yeast mutants deficient in the corresponding genes. This approach has led to isolation of human genes involved in mitochondrial protein import, expression, biogenesis, and assembly of the respiratory complexes. Although the complete sequence of the human 16 kb mitochondrial DNA circle was published in 1981 [3], the mitochondrial gene sequence in *S. cerevisiae* was achieved only in 1998 as a complement to the nuclear genome. In humans as well as in yeast, only a few polypeptides of the respiratory complexes and ATP synthases are mitochondrially encoded with the vast majority of the mitochondrial proteins encoded in the nucleus and imported into the mitochondria by sophisticated machinery. Among the diseases of mitochondrial origin, cystic fibrosis is the most common lethal, inherited disease in North America and Europe, the common problems being breathing disorders, pancreatic dysfunctioning, and male infertility. Although over 900 mutations have been identified in the gene encoding, the cystic fibrosis transmembrane conductance regulator (CFTR), a phenylalanine in the 508 position of the protein, accounts for more than 70 percent of all the disease-causing mutations as a result of poor folding (http://www.genet.sickkids.on.ca/cftr/). The mutant form of CFTR localizes in the endoplasmic reticulum instead of the plasma membrane. When expressed in yeast, CFTR expressed in the endoplasmic reticulum was degraded; however, the degradation was attenuated when expressed in yeast containing a rapid-acting thermosensitive allele of a cytosolic Hsp70 chaperone [181]. These results indicate that Hsp70 facilitates CFTR degradation. Moreover, based on the genome sequence, a new essential mitochondrial metabolic pathway was discovered in yeast that appears as a promising model to study human iron–sulfur clusters [85], since this pathway is conserved in the human mitochondria as well [97].

***9.9.1.2 Nutrient Sensing and Metabolic Response*** All organisms appear to have the nutrient sensing mechanism that can rapidly detect changes in the concentration of available nutrients, adjust flux through metabolic pathways, and networks accordingly. In single-celled organisms, certain nutrients can regulate their own uptake, synthesis, and utilization. By contrast, higher eukaryotes sense nutrient availability primarily through endocrine and neuronal signals (e.g., insulin, glucagon, epinephrine, and so on). However, research performed in the last decade has shown that many types of mammalian cells can directly sense changes in the levels of a variety of nutrients and transduce this sensory information into changes in flux through metabolic pathways. These signal transduction pathways appear to operate both independently from and coordinately with the hormonal pathways. Since several of these pathways are conserved from the unicellular yeasts to mammals, they must have originally evolved independent of hormonal control. This conservation has proven extremely useful in delineating these pathways. In this section, we will discuss the sensing and response to macronutrients, particularly glucose, with particular focus on the modulation of cellular energy and aging.

Yeast has also been the prototype in evaluating the onset and modulation of cellular energy metabolism with respect to glucose homeostasis and, therefore, plays an important role in elucidating the mechanisms of metabolic syndrome. The metabolic syndrome is characterized by insulin resistance, hyperinsulinemia, dyslipidaemia, and a predisposition to type-2 diabetes, hypertension, premature atherosclerosis, and other diseases such as nonalcoholic fatty liver. Patients with this syndrome are usually overtly obese or have more subtle manifestations of increased adiposity, such as an increase in visceral fat. This syndrome has reached an epidemic level in our modern society due to a number of environmental factors, in particular overnutrition and inactivity. A major collaborative effort between basic researchers, clinicians, dieticians, health care authorities, and the pharmaceutical industry is required to halt progression of this devastating clustering of diseases. The AMP-activated protein kinase (AMP kinase) plays a key role in the modulation of cellular energy metabolism by phosphorylating key metabolic enzymes in response to increased AMP levels (Fig. 9-13). AMP levels rise during states of low energy charge (i.e., reduced ATP/AMP ratios) that occur in a variety of normal processes such as exercise and possibly also in some pathological states such as diabetes. Activated AMP kinase phosphorylates key enzymes in both biosynthetic and oxidative pathways and differentially modulates their activities to promote a reestablishment of normal ATP/AMP ratios. Besides maintaining the energy balance within the cells, AMP kinase also plays a key role in sensing intracellular ATP levels. The discovery of naturally occurring mutations in AMP kinase that cause cardiac hypertrophy provides direct evidence that AMP kinase has a fundamental role in maintaining normal human

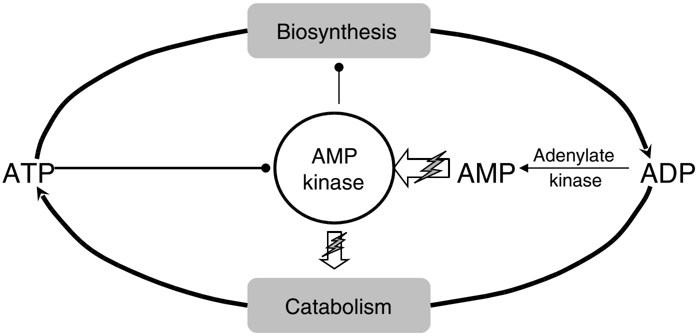


**Figure 9-13** The AMP-activated protein kinase in yeast serves as a sensor of cellular energetic state. If the rate of ATP consumption exceeds its production (rate of biosynthesis exceeds catabolism), the concentration of ADP will increase, stimulating adenylate kinase to convert ADP to AMP. The rise in the level of AMP along with the reduction in ATP levels activates AMP kinase, which then switches off ATP-consuming processes and stimulates catabolism. The exact mechanisms involved in the activation of AMP kinase and its subsequent action are not yet known.

physiology. Moreover, the recent discovery of an upstream kinase in the AMP kinase cascade could implicate the role of AMP kinase in cancer development [57].

AMP kinase is a heterotrimeric complex with a catalytic α-subunit and two regulatory β- and γ-subunits, and homologues of all these three subunits have been identified in all eukaryotes [58]. The identification of these subunits in yeast, catalytic α-subunit (*SNF1* in yeast), the regulatory γ-subunit (*SNF4*), and the scaffolding β-subunit (three partly redundant proteins in yeast: *GAL83*, *SIP1*, and *SIP2*), provided *S. cerevisiae* as an ideal platform to elucidate the regulation and control of AMP kinase in humans. This conservation suggests an essential role of this complex in the functioning of the kinase [58]. Detailed studies on *S. cerevisiae SNF1* complex revealed an intimate role of this complex in transcriptional activation of many genes that are sensitive to glucose repression [19]. Growth on sucrose requires the expression of invertase, whereas growth on nonfermentable carbon sources requires expression of mitochondrial genes needed for oxidative metabolism. The expression of all of these genes is repressed by glucose, and the *SNF1* and *SNF4* genes are required for their derepression. One mechanism by which this is mediated is the phosphorylation of the repressor protein Mig1 by the *SNF1* complex (Fig. 9-14). Phosphorylation causes Mig1 to bind to a nuclear export protein that promotes its removal from the nucleus [82]. Therefore, the primary role of AMP kinase in yeast appears to be in the regulation of energy metabolism by repressing ATP-consuming processes and stimulating ATP generation via control of glucose uptake and its catabolism. Upon activation, AMP kinase controls many metabolic processes, ranging from stimulating fatty acid oxidation and glucose uptake to inhibiting protein, fatty acid, glycogen, and cholesterol synthesis. Its central role as a metabolic glucose sensor is illustrated by

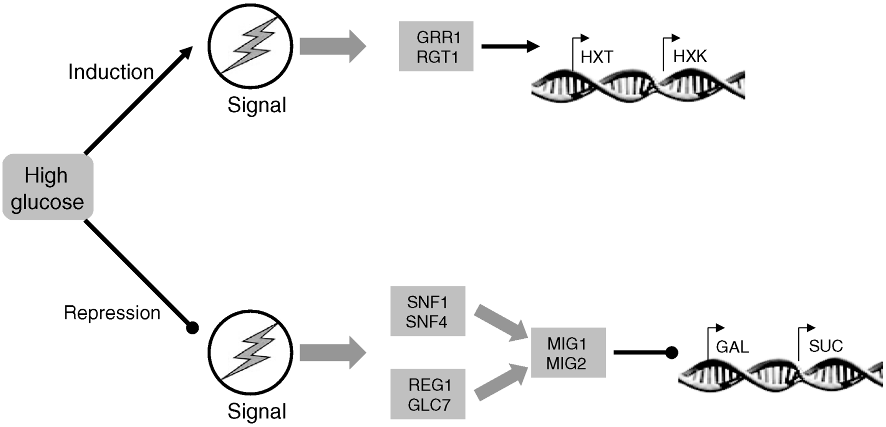


**Figure 9-14** An extremely simplified schematic depicted the induction and repression mechanisms that are triggered by the presence of high glucose concentrations. The genes responsible for glucose metabolism (e.g., hexose transporters and kinases) are induced, whereas those responsible for the metabolism of other sugars are repressed (e.g., galactose and sucrose). Currently, there are several unknown steps involved in both the pathways and we believe that at least some of these uncertainties could be solved by employing systems biology techniques.

recent studies showing that mice lacking one of the AMP kinase isoforms have abnormal glucose tolerance and are insulin resistant [167]. Upon the discovery that AMP kinase is the major target of the antidiabetes drugs metformin and rosiglitazone, there has been tremendous interest in understanding the kinetics and action of this enzyme [136].

The fact that several of the nutrient-sensing pathways are conserved between humans and yeasts has proven extremely useful in studying the control and utilization of these pathways. Another aspect of medical research where *S. cerevisiae* has been used as the model system is in the elucidation of the mechanism of aging. Traditionally, rodents have been used to study these phenomena, analogous to the human processes. Over the past 75 years, many studies have shown that caloric restriction extends life span in a wide variety of species, from invertebrates to rodents to mammals. So far, no long-term studies have been completed in primates or conducted in humans because of the sheer length of any proposed study (perhaps a century or more for human studies!). With the recent explosion in yeast biology, coupled with the identification of the cell cycle regulators that share high homology with the human genes, yeasts are taking over as the ideal system to study aging. Moreover, the short life span of yeast makes them the convenient and preferred hosts over rodents.

Aging in budding yeast is measured by the number of mother cell divisions before senescence. Genetic studies have linked aging in *S. cerevisiae* to the *Sir* (silent information regulator) genes, which mediate genomic silencing at telomeres, mating-type loci, and the repeated ribosomal DNA (rDNA) [54]. Sir2 determines life span in a dose-dependent manner by creating silenced rDNA chromatin, thereby repressing recombination and the generation of toxic rDNA circles. This protein also functions in a meiotic checkpoint that monitors the fidelity of chromosome segregation [98]. Glucose enters yeast cells via highly regulated glucose-sensing transporters (HXT) and is then phosphorylated by hexokinases (Hxk1, Hxk2, and Glk1) to generate glucose-6-phosphate. Limiting the glucose availability by mutating *HXK2* also significantly extended the life span [98]. This is brought about by the yeast $NAD^+$-dependent histone deacetylase Sir2 and it is shown to be required for life span extension by glucose restriction and low-intensity stress [2,99]. The function of Sir2 enzymes in longevity and cell survival appears to be conserved in higher organisms as well. Currently, it is not clear how calorie restriction stimulates Sir2 activity, whether by feedback regulation of nicotinamide, an inhibitory product of Sir2 itself, or by increasing either $NAD^+$ or the $NAD^+$:NADH ratio. Although it is possible to affect Sir2 activity by genetically manipulating $NAD^+$ metabolic pathways, it is not known whether $NAD^+$ is a bona fide regulator of Sir2 in normal cells. Sir2 represses transcription by removing acetyl groups from lysines of histone tails and certain transcription factors (e.g., FOXO and p53) [62]. These findings have led to the intriguing possibility that Sir2 acts as a metabolic sensor, via its $NAD^+$ dependence, that links caloric intake to a transcriptional program that modulates life span. The fact that Sir2 requires the central metabolic cofactor $NAD^+$ to catalyze protein deacetylation is surprising, since from the chemical perspective, deacetylation does not require the destruction of a high-energy cofactor. There is also no indication that the breakdown of $NAD^+$ during the deacetylation reaction is coupled to any form of protein

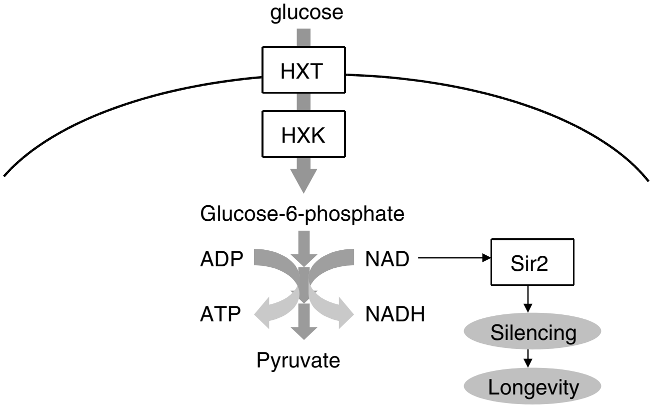


**Figure 9-15** In another example of nutrient sensing and metabolic response in yeast, the Sir2 protein plays a key role in the aging and longevity via calorie restriction. This protein requires NAD for its activity is believed to serve as a cellular redox sensor. When the NAD levels are high (low sugar uptake or calorie restricted conditions), Sir2 is activated that extends life span. This protein is also believed to have partial control on glucose uptake, depending on the intracellular NAD levels.

conformational change or other work. Instead, the $NAD^+$ requirement may serve to link the activity of Sir2 to the metabolic status of the cell. Mutation of the $NAD^+$-salvage pathway in yeast lowers the $NAD^+$ concentration and prevents the life span extension conferred by caloric restriction [99]. This is similar to what was seen for *Sir2* mutants and led to the suggestion that Sir2 activity might depend on the intracellular concentration of some component of the $NAD^+$ pathway (Fig. 9-15). Support for the idea that Sir2 acts as a sensor of the $NAD^+$/NADH ratio (or the concentration of some other component that would be influenced by this ratio) comes from a study on mammalian skeletal muscle cell differentiation [42]. These studies provide strong evidence that Sir2 might be functioning as a metabolic or a redox sensor. However, the difficulty in measuring the *in vivo* $NAD^+$/NADH ratio and the threshold value it triggers in the activation of Sir2 is inhibiting further insight into its sensory and regulatory role.

***9.9.1.3 Mechanism of Cancer*** Most human cancers are the consequence of some form of genome instability, and therefore maintaining the stability of the genome is critical to cell survival and normal cell growth. In general, these aberrations occur either due to increased rate of chromosome instability or due to increased rates of point mutations and frameshift mutations [89]. Mismatch repair is the process by which incorrectly paired nucleotides in DNA are recognized and repaired. Our understanding of mismatch repair in eukaryotes relevant to cancer research mostly comes from studies completed in *S. cerevisiae* and, to a lesser extent, in higher eukaryotes. This section will deal with some of the recent insights into these issues that have emerged from recent genetic studies in *S. cerevisiae*.

*S. cerevisiae* contains at least two genes (*MSH2* and *MSH1*), which function in mismatch repair in the nucleus and mitochondria, respectively. It was identified that mutations in *MSH2* caused high spontaneous mutation rate, a defect in the repair of base pair mismatches with 1–4 nucleotide insertion/deletions, along with a modulation of genetic recombination. This is consistent with the view that *MSH2* functions in the major mismatch repair pathway in *S. cerevisiae* [36]. The human MSH2 protein (hMSH2) was identified as a minor component of a protein fraction that was purified by virtue of its mismatch binding activity, providing evidence that hMSH2 protein also recognizes mispaired bases [123]. A considerable amount of evidence has accumulated indicating that mutations in this gene are the primary cause of hereditary nonpolyposis colon cancer (commonly known as colorectal cancer) in humans [37]. Colorectal cancer is the disorder where rapid cell proliferation occurs in the lining of the large intestine, and these aberrant cells invade other tissues. This disorder most often begins as a benign polyp, which subsequently develops into malignant cancer. Cancer in the colon is the second largest cause of cancer-related deaths in the United States, and if discovered in the early states, it is treatable. Even when the abnormal cell proliferation spreads into nearby lymph nodes, surgical treatment followed by chemotherapy has been demonstrated to be highly successful (information from Colorectal Cancer Alliance Website, http://www.ccalliance.org/). Mapping studies have shown that *hMSH2* mapped to the chromosome 2 colon cancer locus, and analysis of chromosome 2-linked colon cancer families revealed germline *msh2* mutations that cosegregate with colon cancer in these families [88]. By combining the approaches used to define the yeast and human *MSH* genes with methods for identifying the yeast *MLH1* gene in a database of cDNA sequences, the human *MLH1* (*hMLH1*) gene has isolated and demonstrated to map to the chromosome 3p colon cancer locus, and mutational analysis indicate cosegregation of *hMLH1* mutants with the colon cancer locus, providing evidence that inheriting *hMLH1* mutations also causes colon cancer [145].

In the case of cancers caused by mutations in mismatch repair genes, genome instability arises due to elevated mutation rate, although the cause behind this is not clearly understood. However, very little is known about the molecular mechanisms underlying the genome rearrangements, their suppression mechanisms, and the possible defects in the suppression mechanisms that could potentially lead to many cancers. The utility of *S. cerevisiae* to study genome rearrangements began more than 20 years ago, when an extra copy of a DNA sequence was inserted at a site on an unrelated chromosome, followed by selection for recombination [148]. This resulted in chromosomal translocations due to mitotic recombination, similar to those seen in leukemia. The checkpoints shown in the S-phase of the cell cycle were originally identified to promote cell cycle delay or arrest in response to DNA damage, providing the cell an opportunity to repair the damage (Fig. 9-16) [38]. The sensitivity of the checkpoint-defective mutants to killing by DNA-damaging agents suggested that these checkpoints might function in suppressing genome instability. A survey of the *S. cerevisiae* genome for these checkpoints revealed that mutations that disrupt the replication checkpoint (*RFC5-1*, *DPB11-1*, *MEC1*, *DDC2*, and *DUN1*) significantly increase the rate of genome rearrangements [113]. In contrast, mutations in the genes required for the classical G1 and G2 DNA damage checkpoints and the mitotic spindle checkpoints had little effect, suggesting that the DNA replication checkpoint in the S-phase plays a critical role in suppression of spontaneous genome instability.

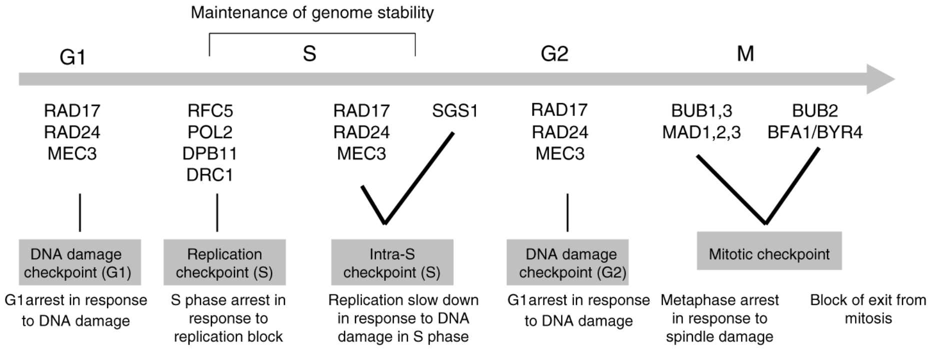

Maintenance of genome stability
G1
S
G2
M
RAD17
RAD24
MEC3
RFC5
POL2
DPB11
DRC1
RAD17
RAD24
MEC3
SGS1
RAD17
RAD24
MEC3
BUB1,3
MAD1,2,3
BUB2
BFA1/BYR4
DNA damage checkpoint (G1)
Replication checkpoint (S)
Intra-S checkpoint (S)
DNA damage checkpoint (G2)
Mitotic checkpoint
G1arrest in response to DNA damage
S phase arrest in response to replication block
Replication slow down in response to DNA damage in S phase
G1arrest in response to DNA damage
Metaphase arrest in response to spindle damage
Block of exit from mitosis

**Figure 9-16** Different stages of the cell cycle and the checkpoints for DNA damage, replication, and mitosis. The proteins that are believed to detect the faults at each checkpoint are indicated below the cell cycle stage. The effect of activating the checkpoint is shown below the proteins in a box. This figure is redrawn from Ref. [89].

Therefore, replication errors appear to be the cause for genome rearrangements. The function of the replication checkpoint in suppressing genome instability likely includes regulating cell cycle progression in response to replication errors, modulating DNA repair functions, ensuring the establishment of sister chromatid cohesion, and maintaining stalled replication forks in a state that allows them to restart DNA synthesis. All of the genome rearrangements seen when this checkpoint was inactivated involved deletion of a chromosome end coupled with *de novo* addition of a new telomere [89]. Although data-driven systems biology in its purest form has generated progress mainly in the area of basic research, the more general concept of combining global data of multiple types is already making significant contributions, especially in the areas of drug discovery and development.

### 9.9.2 Drug Discovery

The extraordinary advances in biological research over the last decade have failed to translate into successful applications in drug discovery. Indeed, a recent analysis reported a decline in the productivity of pharmaceutical R&D, despite a 13 percent annual growth in investment in biomedical research from industry and government [12]. Moreover, the pharmaceutical industry will lose nearly $80 billion in revenue by 2008 due to patent expiration, and the current drug pipeline will replace only a small fraction of this value [4]. The bottleneck in the drug development technology lies in our inability to visualize the complexity of biological systems. Three major issues are associated with identifying effective new drugs: first, discovery of a relevant drug target; second, identification of a drug that will appropriately perturb the target; and third, assessment of the possible side effects and pharmaceutical properties of the drug before its deployment in clinical trials. Systems biology offers powerful new approaches for dealing with these problems.

In the long run, systems biology approach to drug discovery holds the promise to have a profound impact on medical practice, allowing a detailed evaluation of underlying predisposition to disease, diagnosis of disease, and the progression of disease. However in the near future, as a consequence of vigorous biomedical research, systems biology will provide powerful means for validating new drug targets, improving the success with which pharmaceuticals are identified. Farther into the future, the same approaches will drive the development of early diagnostics, enabling disease stratification, individualized therapy, and ultimately preventive drugs, based on both genetic and environmental considerations. Although systems biology as currently envisioned does not have a direct impact on the chemistry of identifying drugs or pharmacological challenges of drug metabolism, it may provide rapid and useful assays for these in the future.

Yeast can contribute to the drug discovery pipeline at an early state in identifying potential drug targets and evaluating the physiological outcome of modulating the activities of these targets. Although there are obvious limitations to using a microorganism to identify potential human drug targets, several yeast proteins share a significant part of their primary amino acid sequence with at least one known or predicted human protein (around 2700 at BLAST with *e*-value less than $10^{-10}$ and around 1100 at BLAST *e*-value $<10^{-50}$). Among these are several hundred with sequence similarity to proteins implicated in human disease [7,13]. A large number of familiar drugs used against human targets specifically inhibit the orthologous proteins in yeast, providing a strong case for the use of yeast physiology to identify and study potential human drug targets. Among the conserved proteins that are uncharacterized, functional studies in yeast will shed light upon possible utility of the human counterparts as drug targets. Most of the proteins conserved between yeast and humans are involved in basic cellular processes such as small-molecule metabolism, protein synthesis, cell division, DNA synthesis and repair, secretion, and so on. Hence, target identification in yeast has proven especially relevant for cancer, which at the simplest level is a disorder of proliferation control caused by accumulated mutations. Many of the common mutations in human cancers including genetic and physical interactions between the mutated genes/proteins can be modeled in yeast, greatly simplifying and accelerating directed study. The concept of "synthetic lethality," a phenomenon where a combination of two innocuous genetic mutations renders the cell inviable, has shown great promise in identifying targets for anticancer therapy. Screening for mutation pairs that display synthetic lethality could lead to identifying drug targets that could selectively inhibit proliferation only in cells carrying a cancer-causing mutation. Such "gene-therapy," applications are presumably less detrimental than chemical or radiation therapy. Due to the obvious combinatorial problem associated with the experimental analysis of all the ordered pairwise mutations (even with 6000 genes), an automated system for creating and analyzing all pairwise combinations between a single mutant and all of the around 5000 viable single-gene deletion mutants has recently been described [151,152].

The increasing cases of fungal infections, particularly among immunity-compromised persons (those with AIDS and transplant patients), the need for safer and more effective antifungals is widely recognized. Although *Candida albicans* and *Aspergilli* have been used for the development of antifungals, *S. cerevisiae* presents a ready-made model system, particularly for azole-based antifungals. The discovery of pathogenic strains of *S. cerevisiae* that display invasive filamentous growth [49] or biofilm formation [132] provides excellent opportunity to examine the association between gene function and hyphal growth and infective capacity and biofilm formation, potentially leading to the identification of new antifungals. Screening for antifungals begins with a specific target with a known mechanism of action, since they could be used as templates for combinatorial modifications. An ideal antifungal should be required for the growth of yeast and should have minimal or no activity in humans (and therefore, not be conserved in humans). Among the 1100 essential genes in yeast, 350 do not have orthologues in humans and a subset of these genes would make an ideal target to screen for antifungals. However, some of the most successful antifungal compounds have properties far from the ideal criteria. For example, morpholines inhibit the Erg2 protein in the ergosterol pathway. Deletion of the *ERG2* gene is not lethal, and it shares sequence similarity with human sigma receptor protein [5]. The standard method used to screen for antifungals in particular and drug targets in general is the Y2H, which has been described earlier.

### 9.9.3 Food and Chemical Technology

White biotechnology (or industrial biotechnology) is an emerging field that specifically caters to the needs of the chemical and environmental industry [41]. It relies largely on using living cells like yeast as cell factories for sustainable production of biochemicals, biomaterials, and biofuels from renewable resources. A recent study conducted by McKinsey and Co predicts immense growth potential for white biotechnology in the future (http://www.mckinsey.com/clientservice/chemicals/pdf/BioVision_Booklet_final.pdf), with some of the large chemical companies such as BASF and DSM already replacing their chemical processes with cleaner, more efficient bioprocesses. An important component in developing yeast as a cell factory for an economically viable, efficient bioprocess is to optimize its metabolic network and systems biology has propelled the field of white biotechnology to new heights. In addition to the traditional use of yeast for baking purposes and ethanol production, it is also the system of choice for producing a variety of recombinant proteins such as insulin and various vaccines. There are a number of advantages of using yeast as a cell factory such as

- the availability of complete genome sequence
- its generally regarded as safe (GRAS) status
- well-defined cellular architecture
- established genetic manipulation techniques
- ease of scale-up of yeast bioprocesses
- availability of metabolic models

Besides its conventional applications in the brewing industry and as bakers yeast, *S. cerevisiae* is now used for a number of other industrial applications (Table 9.1).

**Table 9-1 Industrial applications of bakers yeast**

| | Nonproprietary Name | Trade Name | Company | Reference |
|---|---|---|---|---|
| Pharmaceuticals | Hepatitis surface antigen | Ambirix | GlaxoSmithKline | [46,168] |
| | | Comvax | Merck | [46,168] |
| | | HBVAXPRO | Aventis Pharma | [46,168] |
| | | Infanrix-Penta | GlaxoSmithKline | [46,168] |
| | | Pediarix | GlaxoSmithKline | [46,168] |
| | | Procomvax | Aventis-Pasteur | [46,168] |
| | | Twinrix | GlaxoSmithKline | [46,168] |
| | Insulin | Actrapid | NovoNordisk | [46,168] |
| | | Novolog | NovoNordisk | [46,168] |
| | | Levemir | NovoNordisk | [46,168] |
| | Hirudin/Desirudin | Refuldan | Aventis | [46,168] |
| | Urate oxidase | Elitex | Sanofi-Synthelabo | [46,168] |
| Fine chemicals | Epicedrol | — | — | [78] |
| | Lycopene | — | — | [175] |
| | β-carotene | — | BASF, Roche | [175] |
| | Artemesinin | — | Amyris Biotechnologies | [133] |
| | Flavanones | — | — | [176] |
| | Ascorbic acid | — | — | [56] |
| Bulk chemicals | Glycerol | — | — | [120] |
| | Lactic acid | — | — | [127] |

## 9.10 PERSPECTIVE

Systems biology offers an opportunity to study how the phenotype is generated from the genotype and with it a glimpse of how evolution has crafted the phenotype. One aspect of systems biology is the development of techniques to examine broadly the level of protein, RNA, and DNA on a gene-by-gene basis and even the posttranslational modification and localization of proteins. In a very short time we have witnessed the development of high-throughput biology, forcing us to consider cellular processes *in vivo*. Even though much of the data is noisy and today partially inconsistent and incomplete, this has been a radical shift in the way we address problems one interaction at a time. When coupled with gene deletions by RNAi and classical methods and with the use of chemical tools tailored to proteins and protein domains, these high-throughput techniques become still more powerful. It is evident that a wide range of experimental approaches are being developed for use in *S. cerevisiae* that will allow functional genomics to build up an integrative view of the workings of a simple eukaryotic cell. This should enable a deeper understanding of more complex eukaryotes, both by the identification of orthologous genes in the different species and also by the expression of foreign coding sequences in yeast for complementation or two-hybrid analyses. However, many of these techniques are sufficiently general that once they have been tried and tested in the experimentally tractable yeast system, they should be directly applicable to the study of the functional genomics of higher organisms.

## ACKNOWLEDGMENT

The authors wish to thank Michael L. Nielsen for his comments and suggestions on the contents of this chapter.

## REFERENCES

1. Albert B, Botstein D, Brenner S, Cantor CR, Doolittle RF, Hood L, McKusick VA, Nathans D, Olson MV, Orkin S. Mapping and Sequencing the Human Genome, 1988.
2. Anderson RM, Latorre-Esteves M, Neves AR, Lavu S, Medvedik O, Taylor C, Howitz KT, Santos H, Sinclair DA. Yeast life-span extension by calorie restriction is independent of NAD fluctuation. *Science* 2003;302:2124–2126.
3. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. Sequence and organization of the human mitochondrial genome. *Nature* 1981;290: 457–465.
4. Anonymous Pharmaceutical Development, http://www.newsrx.com 2004.
5. Bammert GF, Fostel JM. Genome-wide expression patterns in *Saccharomyces cerevisiae*: comparison of drug treatments and genetic alterations affecting biosynthesis of ergosterol. *Antimicrob Agents Chemother* 2000;44:1255–1265.
6. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* 2003;21:1337–1342.
7. Bassett DE Jr, Boguski MS, Spencer F, Reeves R, Kim S, Weaver T, Hieter P. Genome cross-referencing and XREFdb: implications for the identification and analysis of genes mutated in human disease. *Nat Genet* 1997;15:339–344.
8. Beyer A, Hollunder J, Nasheuer HP, Wilhelm T. Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. *Mol Cell Proteomics* 2004;3:1083–1092.
9. Bianchi MM, Ngo S, Vandenbol M, Sartori G, Morlupi A, Ricci C, Stefani S, Morlino GB, Hilger F, Carignani G, Slonimski PP, Frontali L. Large-scale phenotypic analysis reveals identical contributions to cell functions of known and unknown yeast genes. *Yeast* 2001;18:1397–1412.
10. Blanchard AP, Hood L. Sequence to array: probing the genome's secrets. *Nat Biotechnol* 1996;14:1649.
11. Boer VM, de Winde JH, Pronk JT, Piper MD. The genome-wide transcriptional responses of *Saccharomyces cerevisiae* grown on glucose in aerobic chemostat cultures limited for carbon, nitrogen, phosphorus, or sulfur. *J Biol Chem* 2003;278:3265–3274.

12. Booth B, Zemmel R. Prospects for productivity. *Nat Rev Drug Discov* 2004;3:451–456.
13. Botstein D, Chervitz SA, Cherry JM. Yeast as a model organism. *Science* 1997;277: 1259–1260.
14. Brauer MJ, Saldanha AJ, Dolinski K, Botstein D. Homeostatic adjustment and metabolic remodeling in glucose-limited yeast cultures. *Mol Biol Cell* 2005;16:2503–2517.
15. Bro C, Knudsen S, Regenberg B, Olsson L, Nielsen J. Improvement of galactose uptake in *Saccharomyces cerevisiae* through overexpression of phosphoglucomutase: example of transcript analysis as a tool in inverse metabolic engineering. *Appl Environ Microbiol* 2005;71:6465–6472.
16. Bro C, Regenberg B, Nielsen J. Genome-wide transcriptional response of a *Saccharomyces cerevisiae* strain with an altered redox metabolism. *Biotechnol Bioeng* 2004;85: 269–276.
17. Cakir T, Kirdar B, Ulgen KO. Metabolic pathway analysis of yeast strengthens the bridge between transcriptomics and metabolic networks. *Biotechnol Bioeng* 2004;86:251–260.
18. Cannon WB. *Bodily Changes to Pain, Hunger, Fear and Rage*, 2nd ed. New York and London: D. Appleton and Co, 1929.
19. Carlson M. Glucose repression in yeast. *Curr Opin Microbiol* 1999;2:202–207.
20. Carrell RW, Lomas DA. Conformational disease. *Lancet* 1997;350:134–138.
21. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 1998;2:65–73.
22. Collins CH, Arnold FH, Leadbetter JR. Directed evolution of Vibrio fischeri LuxR for increased sensitivity to a broad spectrum of acyl-homoserine lactones. *Mol Microbiol* 2005;55:712–723.
23. Davis TN. Protein localization in proteomics. *Curr Opin Chem Biol* 2004;8:49–53.
24. de Lichtenberg U, Jensen LJ, Brunak S, Bork P. Dynamic complex formation during the yeast cell cycle. *Science* 2005;307:724–727.
25. Deane CM, Salwinski L, Xenarios I, Eisenberg D. Protein interactions: two methods for assessment of the reliability of high-throughput observations. *Mol Cell Proteomics* 2002;1:349–356.
26. DeLuna A, Avendano A, Riego L, Gonzalez A. NADP-glutamate dehydrogenase isoenzymes of *Saccharomyces cerevisiae*. Purification, kinetic properties, and physiological roles. *J Biol Chem* 2001;276:43775–43783.
27. Deng M, Sun F, Chen T. Assessment of the reliability of protein–protein interactions and protein function prediction. *Pac Symp Biocomput* 2003;140–151.
28. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997;278:680–686.
29. Edwards JS, Ramakrishna R, Palsson BO. Characterizing the metabolic phenotype: a phenotype phase plane analysis. *Biotechnol Bioeng* 2002;77:27–36.
30. Eisen MB, Brown PO. DNA arrays for analysis of gene expression. *Methods Enzymol* 1999;303:179–205.
31. Elowitz MB, Leibler S. A synthetic oscillatory network of transcriptional regulators. *Nature* 2000;403:335–338.
32. Feng XJ, Hooshangi S, Chen D, Li G, Weiss R, Rabitz H. Optimizing genetic circuits by global sensitivity analysis. *Biophys J* 2004;87:2195–2202.

33. Fiehn O. Metabolomics: the link between genotypes and phenotypes. *Plant Mol Biol* 2002;48:155–171.

34. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L. Metabolite profiling for plant functional genomics. *Nat Biotechnol* 2000;18:1157–1161.

35. Fields S, Song O. A novel genetic system to detect protein–protein interactions. *Nature* 1989;340:245–246.

36. Fishel R, Kolodner RD. Identification of mismatch repair genes and their role in the development of cancer. *Curr Opin Genet Dev* 1995;5:382–395.

37. Fishel R, Lescoe MK, Rao MR, Copeland NG, Jenkins NA, Garber J, Kane M, Kolodner R. The human mutator gene homolog MSH2 and its association with hereditary non-polyposis colon cancer. *Cell* 1993;75:1027–1038.

38. Foiani M, Pellicioli A, Lopes M, Lucca C, Ferrari M, Liberi G, Muzi Falconi M, Plevani1 P. DNA damage checkpoints and DNA replication controls in *Saccharomyces cerevisiae*. *Mutat Res* 2000;451:187–196.

39. Forster J, Famili I, Fu P, Palsson BO, Nielsen J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res* 2003;13:244–253.

40. Forster J, Gombert AK, Nielsen J. A functional genomics approach using metabolomics and *in silico* pathway analysis. *Biotechnol Bioeng* 2002;79:703–712.

41. Frazzetto G. White biotechnology. *EMBO Rep* 2003;4:835–837.

42. Fulco M, Schiltz RL, Iezzi S, King MT, Zhao P, Kashiwaya Y, Hoffman E, Veech RL, Sartorelli V. Sir2 regulates skeletal muscle differentiation as a potential sensor of the redox state. *Mol Cell* 2003;12:51–62.

43. Gardner TS, Cantor CR, Collins JJ. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* 2000;403:339–342.

44. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002;415:141–147.

45. Ge H, Liu Z, Church GM, Vidal M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat Genet* 2001;29:482–486.

46. Gerngross TU. Advances in the production of human therapeutic proteins in yeasts and filamentous fungi. *Nat Biotechnol* 2004;22:1409–1414.

47. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'shea EK, Weissman JS. Global analysis of protein expression in yeast. *Nature* 2003;425:737–741.

48. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke

JD, Snyder M, Philippsen P, Davis RW, Johnston M. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 2002;418:387–391.

49. Gimeno CJ, Ljungdahl PO, Styles CA, Fink GR. Unipolar cell divisions in the yeast *S. cerevisiae* lead to filamentous growth: regulation by starvation and RAS. *Cell* 1992;68:1077–1090.

50. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. Life with 6000 genes. *Science* 1996;274:546, 563–546, 567.

51. Goldberg DS, Roth FP. Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci USA* 2003;100:4372–4376.

52. Goldbeter A, Segel LA. Unified mechanism for relay and oscillation of cyclic AMP in *Dictyostelium discoideum. Proc Natl Acad Sci USA* 1977;74:1543–1547.

53. Grigoriev A. On the number of protein–protein interactions in the yeast proteome. *Nucleic Acids Res* 2003;31:4157–4161.

54. Guarente L. Diverse and dynamic functions of the Sir silencing complex. *Nat Genet* 1999;23:281–285.

55. Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc Natl Acad Sci USA* 2000;97:9390–9395.

56. Hancock RD, Galpin JR, Viola R. Biosynthesis of L-ascorbic acid (vitamin C) by *Saccharomyces cerevisiae*. *FEMS Microbiol Lett* 2000;186:245–250.

57. Hardie DG. The AMP-activated protein kinase cascade: the key sensor of cellular energy status. *Endocrinology* 2003;144:5179–5183.

58. Hardie DG, Carling D, Carlson M. The AMP-activated/SNF1 protein kinase subfamily: metabolic sensors of the eukaryotic cell? *Annu Rev Biochem* 1998;67:821–855.

59. Hartwell LH, Culotti J, Pringle JR, Reid BJ. Genetic control of the cell division cycle in yeast. *Science* 1974;183:46–51.

60. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature* 1999;402:C47–C52.

61. Heinrich R, Schuster S. The modelling of metabolic systems. Structure, control and optimality. *Biosystems* 1998;47:61–77.

62. Hekimi S, Guarente L. Genetics and the specificity of the aging process. *Science* 2003;299: 1351–1354.

63. Hereford LM, Osley MA, Ludwig TR, McLaughlin CS. Cell-cycle regulation of yeast histone mRNA. *Cell* 1981;24:367–375.

64. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 2002;415:180–183.

65. Hohmann S. The Yeast Systems Biology Network: mating communities. *Curr Opin Biotechnol* 2005;16:356–360.

66. Hooshangi S, Thiberge S, Weiss R. Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proc Natl Acad Sci USA* 2005;102:3581–3586.

67. Huang CY, Ferrell JE Jr. Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc Natl Acad Sci USA* 1996;93:10078–10083.

68. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley PS. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 2001;19:342–347.

69. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'shea EK. Global analysis of protein localization in budding yeast. *Nature* 2003;425:686–691.

70. Ibarra RU, Edwards JS, Palsson BO. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* 2002;420:186–189.

71. Iberall AS. A field and circuit thermodynamics for integrative physiology. I. Introduction to the general notions. *Am J Physiol* 1977;233:R171–R180.

72. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 2001;292:929–934.

73. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. Revealing modular organization in the yeast transcriptional network. *Nat Genet* 2002;31:370–377.

74. Ihmels J, Levy R, Barkai N. Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nat Biotechnol* 2004;22:86–92.

75. Ito T, Chiba T, Yoshida M. Exploring the protein interactome using comprehensive two-hybrid projects. *Trends Biotechnol* 2001;19:S23–S27.

76. Ito T, Ota K, Kubota H, Yamaguchi Y, Chiba T, Sakuraba K, Yoshida M. Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol Cell Proteomics* 2002;1:561–566.

77. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 2001;409:533–538.

78. Jackson BE, Hart-Wells EA, Matsuda SP. Metabolic engineering to produce sesquiterpenes in yeast. *Org Lett* 2003;5:1629–1632.

79. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 1961;3:318–356.

80. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M. A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* 2003;302:449–453.

81. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. The large-scale organization of metabolic networks. *Nature* 2000;407:651–654.

82. Johnston M. Feasting, fasting and fermenting. Glucose sensing in yeast and other cells. *Trends Genet* 1999;15:29–33.

83. Kafri R, Bar-Even A, Pilpel Y. Transcription control reprogramming in genetic backup circuits. *Nat Genet* 2005;37:295–299.

84. Khorana HG. Polynucleotide synthesis and the genetic code. *Fed Proc* 1965;24:1473–1487.

85. Kispal G, Csere P, Prohl C, Lill R. The mitochondrial proteins Atm1p and Nfs1p are essential for biogenesis of cytosolic Fe/S proteins. *EMBO J* 1999;18:3981–3989.

86. Kitano H. Computational systems biology. *Nature* 2002;420:206–210.

87. Klein CJ, Olsson L, Nielsen J. Glucose control in *Saccharomyces cerevisiae*: the role of Mig1 in metabolic functions. *Microbiology* 1998;144(Part 1):13–24.

88. Kolodner RD, Hall NR, Lipford J, Kane MF, Rao MR, Morrison P, Wirth L, Finan PJ, Burn J, Chapman P. Human mismatch repair genes and their association with hereditary non-polyposis colon cancer. *Cold Spring Harb Symp Quant Biol* 1994;59:331–338.

89. Kolodner RD, Putnam CD, Myung K. Maintenance of genome stability in *Saccharomyces cerevisiae*. *Science* 2002;297:552–557.

90. Kopito RR, Ron D. Conformational disease. *Nat Cell Biol* 2000;2:E207–E209.

91. Kumar A, Harrison PM, Cheung KH, Lan N, Echols N, Bertone P, Miller P, Gerstein MB, Snyder M. An integrated approach for finding overlooked genes in yeast. *Nat Biotechnol* 2002;20:58–63.

92. Lauffenburger DA. Cell signaling pathways as control modules: complexity for simplicity? *Proc Natl Acad Sci USA* 2000;97:5031–5033.

93. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 2002;298:799–804.

94. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, Goldberg DS, Li N, Martinez M, Rual JF, Lamesch P, Xu L, Tewari M, Wong SL, Zhang LV, Berriz GF, Jacotot L, Vaglio P, Reboul J, Hirozane-Kishikawa T, Li Q, Gabel HW, Elewa A, Baumgartner B, Rose DJ, Yu H, Bosak S, Sequerra R, Fraser A, Mango SE, Saxton WM, Strome S, Van Den Heuvel S, Piano F, Vandenhaute J, Sardet C, Gerstein M, Doucette-Stamm L, Gunsalus KC, Harper JW, Cusick ME, Roth FP, Hill DE, Vidal M. A map of the interactome network of the metazoan *C. elegans*. *Science* 2004;303:540–543.

95. Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci USA* 2003;100:15522–15527.

96. Lieb JD, Liu X, Botstein D, Brown PO. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein–DNA association. *Nat Genet* 2001;28:327–334.

97. Lill R, Muhlenhoff U. Iron–sulfur–protein biogenesis in eukaryotes. *Trends Biochem Sci* 2005;30:133–141.

98. Lin SJ, Defossez PA, Guarente L. Requirement of NAD and SIR2 for life-span extension by calorie restriction in *Saccharomyces cerevisiae*. *Science* 2000;289:2126–2128.

99. Lin SJ, Ford E, Haigis M, Liszt G, Guarente L. Calorie restriction extends yeast life span by lowering the level of NADH. *Genes Dev* 2004;18:12–16.

100. Lindon JC, Nicholson JK, Holmes E, Keun HC, Craig A, Pearce JT, Bruce SJ, Hardy N, Sansone SA, Antti H, Jonsson P, Daykin C, Navarange M, Beger RD, Verheij ER, Amberg A, Baunsgaard D, Cantor GH, Lehman-McKeeman L, Earll M, Wold S, Johansson E, Haselden JN, Kramer K, Thomas C, Lindberg J, Schuppe-Koistinen I, Wilson ID, Reily MD, Robertson DG, Senn H, Krotzky A, Kochhar S, Powell J, van der Ouderaa F, Plumb R, Schaefer H, Spraul M. Summary recommendations for standardization and reporting of metabolic analyses. *Nat Biotechnol* 2005;23:833–838.

101. Lipshutz RJ, Morris D, Chee M, Hubbell E, Kozal MJ, Shah N, Shen N, Yang R, Fodor SP. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* 1995; 19:442–447.

102. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* 1996;14:1675–1680.

103. Lou XJ, Schena M, Horrigan FT, Lawn RM, Davis RW. Expression monitoring using cDNA microarrays. A general protocol. *Methods Mol Biol* 2001;175:323–340.

104. Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 2004;431:308–312.

105. Lutfiyya LL, Iyer VR, DeRisi J, DeVit MJ, Brown PO, Johnston M. Characterization of three related glucose repressors and genes they regulate in *Saccharomyces cerevisiae*. *Genetics* 1998;150:1377–1391.

106. MacBeath G, Schreiber SL. Printing proteins as microarrays for high-throughput function determination. *Science* 2000;289:1760–1763.

107. Mesarovic MD, Systems theory and biology: view of a theoretician. *Systems Theory and Biology.* New York: Springer, 1968, pp. 59–87.

108. Mewes HW, Albermann K, Bahr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG, Pfeiffer F, Zollner A. Overview of the yeast genome. *Nature* 1997;387:7–65.

109. Monod J, Changeux JP, Jacob F. Allosteric proteins and cellular control systems. *J Mol Biol* 1963;6:306–329.

110. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet* 2004;36:1331–1339.

111. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* 1986; 51(Part 1):263–273.

112. Mustacchi R, Hohmann S, Nielsen J. Yeast systems biology to unravel the network of life. *Yeast* 2006;23:227–238.

113. Myung K, Datta A, Kolodner RD. Suppression of spontaneous chromosomal rearrangements by S phase checkpoint functions in *Saccharomyces cerevisiae*. *Cell* 2001;104:397–408.

114. Ng SK, Zhang Z, Tan SH, Lin K. InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res* 2003;31:251–254.

115. Nielsen J. It is all about metabolic fluxes. *J Bacteriol* 2003;185:7031–7035.

116. Ohlmeier S, Kastaniotis AJ, Hiltunen JK, Bergmann U. The yeast mitochondrial proteome, a study of fermentative and respiratory growth. *J Biol Chem* 2004;279:3956–3979.

117. Oliver SG, van der Aart QJ, Agostoni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, Antoine G, Anwar R, Ballesta JP, Benit P. The complete DNA sequence of yeast chromosome III. *Nature* 1992;357:38–46.

118. Oliver SG, Winson MK, Kell DB, Baganz F. Systematic functional analysis of the yeast genome. *Trends Biotechnol* 1998;16:373–378.

119. Ostergaard S, Olsson L, Johnston M, Nielsen J. Increasing galactose consumption by *Saccharomyces cerevisiae* through metabolic engineering of the GAL gene regulatory network. *Nat Biotechnol* 2000;18:1283–1286.

120. Overkamp KM, Bakker BM, Kotter P, Luttik MA, van Dijken JP, Pronk JT. Metabolic engineering of glycerol production in *Saccharomyces cerevisiae*. *Appl Environ Microbiol* 2002;68:2814–2821.

121. Ozcan S, Johnston M. Three different regulatory mechanisms enable yeast hexose transporter (HXT) genes to be induced by different levels of glucose. *Mol Cell Biol* 1995;15:1564–1572.

122. Ozcan S, Johnston M. Function and regulation of yeast hexose transporters. *Microbiol Mol Biol Rev* 1999;63:554–569.

123. Palombo F, Hughes M, Jiricny J, Truong O, Hsuan J. Mismatch repair and cancer. *Nature* 1994;367:417.

124. Patil KR, Nielsen J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc Natl Acad Sci USA* 2005;102(8): 2685–2689.

125. Patton WF. Detection technologies in proteome analysis. *J Chromatogr B Analyt Technol Biomed Life Sci* 2002;771:3–31.

126. Pedraza JM, van Oudenaarden A. Noise propagation in gene networks. *Science* 2005;307:1965–1969.

127. Porro D, Brambilla L, Ranzi BM, Martegani E, Alberghina L. Development of metabolically engineered *Saccharomyces cerevisiae* cells for the production of lactic acid. *Biotechnol Prog* 1995;11:294–298.

128. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, McCartney RR, Schmidt MC, Rachidi N, Lee SJ, Mah AS, Meng L, Stark MJ, Stern DF, De Virgilio C, Tyers M, Andrews B, Gerstein M, Schweitzer B, Predki PF, Snyder M. Global analysis of protein phosphorylation in yeast. *Nature* 2005;438:679–684.

129. Rappsilber J, Siniossoglou S, Hurt EC, Mann M. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal Chem* 2000;72:267–275.

130. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL. Hierarchical organization of modularity in metabolic networks. *Science* 2002;297:1551–1555.

131. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. Genome-wide location and function of DNA binding proteins. *Science* 2000;290:2306–2309.

132. Reynolds TB, Fink GR. Bakers' yeast, a model for fungal biofilm formation. *Science* 2001;291:878–881.

133. Ro DK, Paradise EM, Ouellet M, Fisher KJ, Newman KL, Ndungu JM, Ho KA, Eachus RA, Ham TS, Kirby J, Chang MC, Withers ST, Shiba Y, Sarpong R, Keasling JD. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 2006;440:940–943.

134. Ronne H. Glucose repression in fungi. *Trends Genet* 1995;11:12–17.

135. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB. Gene regulation at the single-cell level. *Science* 2005;307:1962–1965.

136. Rutter GA, Da, Silva X, Leclerc I. Roles of 5′-AMP-activated protein kinase (AMPK) in mammalian glucose homoeostasis. *Biochem J* 2003;375:1–16.

137. Said MR, Begley TJ, Oppenheim AV, Lauffenburger DA, Samson LD. Global network analysis of phenotypic effects: protein networks and toxicity modulation in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 2004;101:18006–18011.

138. Saldanha AJ, Brauer MJ, Botstein D. Nutritional homeostasis in batch and steady-state culture of yeast. *Mol Biol Cell* 2004;15:4089–4104.

139. Salwinski L, Eisenberg D. *In silico* simulation of biological network dynamics. *Nat Biotechnol* 2004;22:1017–1019.

140. Sanger F, Donelson JE, Coulson AR, Kossel H, Fischer D. Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. *Proc Natl Acad Sci USA* 1973;70:1209–1213.

141. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;270:467–470.

142. Schuster S, Dandekar T, Fell DA. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol* 1999;17:53–60.

143. Schwikowski B, Uetz P, Fields S. A network of protein–protein interactions in yeast. *Nat Biotechnol* 2000;18:1257–1261.

144. Segre D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci USA* 2002;99:15112–15117.

145. Shimodaira H, Filosi N, Shibata H, Suzuki T, Radice P, Kanamaru R, Friend SH, Kolodner RD, Ishioka C. Functional analysis of human MLH1 mutations in *Saccharomyces cerevisiae*. *Nat Genet* 1998;19:384–389.

146. Sickmann A, Reinders J, Wagner Y, Joppich C, Zahedi R, Meyer HE, Schonfisch B, Perschil I, Chacinska A, Guiard B, Rehling P, Pfanner N, Meisinger C. The proteome of *Saccharomyces cerevisiae* mitochondria. *Proc Natl Acad Sci USA* 2003;100:13207–13212.

147. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 1998;9: 3273–3297.

148. Sugawara N, Szostak JW. Construction of specific chromosomal rearrangements in yeast. *Methods Enzymol* 1983;101:269–278.

149. Suzuki-Fujimoto T, Fukuma M, Yano KI, Sakurai H, Vonika A, Johnston SA, Fukasawa T. Analysis of the galactose signal transduction pathway in *Saccharomyces cerevisiae*: interaction between Gal3p and Gal80p. *Mol Cell Biol* 1996;16:2504–2508.

150. Tai SL, Boer VM, ran-Lapujade P, Walsh MC, de Winde JH, Daran JM, Pronk JT. Two-dimensional transcriptome analysis in chemostat cultures: combinatorial effects of oxygen availability and macronutrient limitation in *Saccharomyces cerevisiae*. *J Biol Chem* 2004;280:437–447.

151. Tong AH, Boone C. Synthetic genetic array analysis in *Saccharomyces cerevisiae*. *Methods Mol Biol* 2006;313:171–192.

152. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 2001;294: 2364–2368.

153. Trethewey RN. Gene discovery via metabolic profiling. *Curr Opin Biotechnol* 2001; 12:135–138.

154. Trethewey RN, Krotzky AJ, Willmitzer L. Metabolic profiling: a Rosetta Stone for genomics? *Curr Opin Plant Biol* 1999;2:83–85.

155. Trumbly RJ. Glucose repression in the yeast *Saccharomyces cerevisiae*. *Mol Microbiol* 1992;6:15–21.

156. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000;403:623–627.

157. Unlu M. Difference gel electrophoresis. *Biochem Soc Trans* 1999;27:547–549.

158. Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl Environ Microbiol* 1994;60:3724–3731.

159. Vazquez A, Flammini A, Maritan A, Vespignani A. Global protein function prediction from protein–protein interaction networks. *Nat Biotechnol* 2003;21:697–700.

160. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995;270:484–487.

161. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW. Characterization of the yeast transcriptome. *Cell* 1997;88:243–251.

162. Villas-Boas SG, Hojer-Pedersen J, Akesson M, Smedsgaard J, Nielsen J. Global metabolite analysis of yeast: evaluation of sample preparation methods. *Yeast* 2005;22: 1155–1169.

163. Villas-Boas SG, Kesson M, Nielsen J. Biosynthesis of glyoxylate from glycine in *Saccharomyces cerevisiae*. *FEMS Yeast Res* 2005;5:703–709.

164. Villas-Boas SG, Mas S, Akesson M, Smedsgaard J, Nielsen J. Mass spectrometry in metabolome analysis. *Mass Spectrom Rev* 2004;24(5): 613–646.

165. Villas-Boas SG, Moxley JF, Akesson M, Stephanopoulos G, Nielsen J. High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. *Biochem J* 2005;388(Part 2):669–677.

166. Villas-Boas SG, Moxley JF, Akesson M, Stephanopoulos G, Nielsen J. High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. *Biochem J* 2005;388:669–677.

167. Viollet B, Andreelli F, Jorgensen SB, Perrin C, Flamez D, Mu J, Wojtaszewski JF, Schuit FC, Birnbaum M, Richter E, Burcelin R, Vaulont S. Physiological role of AMP-activated protein kinase (AMPK): insights from knockout mouse models. *Biochem Soc Trans* 2003;31:216–219.

168. Walsh G. Biopharmaceuticals: recent approvals and likely directions. *Trends Biotechnol* 2005;23:553–558.

169. Washburn MP, Wolters D, Yates JR III. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 2001;19: 242–247.

170. Weckwerth W. Metabolomics in systems biology. *Annu Rev Plant Biol* 2003;54: 669–689.

171. Wei J, Sun J, Yu W, Jones A, Oeller P, Keller M, Woodnutt G, Short JM. Global proteome discovery using an online three-dimensional LC–MS/MS. *J Proteome Res* 2005;4:801–808.

172. Westergaard SL, Bro C, Olsson L, Nielsen J. Elucidation of the role of Grr1p in glucose sensing by *Saccharomyces cerevisiae* through genome-wide transcription analysis. *FEMS Yeast Res* 2004;5:193–204.

173. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El-Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-Macdonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Veronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu K, Zimmermann K, Philippsen P, Johnston M, Davis RW. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 1999;285:901–906.

174. Wu Y, Reece RJ, Ptashne M. Quantitation of putative activator–target affinities predicts transcriptional activating potentials. *EMBO J* 1996;15:3951–3963.

175. Yamano S, Ishii T, Nakagawa M, Ikenaga H, Misawa N. Metabolic engineering for production of beta-carotene and lycopene in *Saccharomyces cerevisiae*. *Biosci Biotechnol Biochem* 1994;58:1112–1114.

176. Yan Y, Kohli A, Koffas MA. Biosynthesis of natural flavanones in *Saccharomyces cerevisiae*. *Appl Environ Microbiol* 2005;71:5610–5613.

177. Yano K, Fukasawa T. Galactose-dependent reversible interaction of Gal3p with Gal80p in the induction pathway of Gal4p-activated genes of *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 1997;94:1721–1726.

178. Yates FE, Brennan RD, Urquhart J. Application of control systems theory to physiology. Adrenal glucocorticoid control system. *Fed Proc* 1969;28:71–83.

179. Yokobayashi Y, Weiss R, Arnold FH. Directed evolution of a genetic circuit. *Proc Natl Acad Sci USA* 2002;99:16587–16591.

180. Zhang LV, Wong SL, King OD, Roth FP. Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics* 2004;5:38.

181. Zhang Y, Nijbroek G, Sullivan ML, McCracken AA, Watkins SC, Michaelis S, Brodsky JL. Hsp70 molecular chaperone facilitates endoplasmic reticulum-associated protein degradation of cystic fibrosis transmembrane conductance regulator in yeast. *Mol Biol Cell* 2001;12:1303–1314.

182. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, Mitchell T, Miller P, Dean RA, Gerstein M, Snyder M. Global analysis of protein activities using proteome chips. *Science* 2001;293:2101–2105.

183. Zhu H, Klemic JF, Chang S, Bertone P, Casamayor A, Klemic KG, Smith D, Gerstein M, Reed MA, Snyder M. Analysis of yeast protein kinases using protein chips. *Nat Genet* 2000;26:283–289.