Annu. Rev. Genomics Hum. Genet. 2001. 2:343–72

A NEW APPROACH TO DECODING LIFE: Systems Biology

Trey Ideker[1,2], Timothy Galitski[1], and Leroy Hood[1,2,3,4,5]
Institute for Systems Biology[1], Seattle, Washington 98105; Departments of Molecular Biotechnology[2], Immunology[3], Bioengineering[4], and Computer Science and Engineering[5], University of Washington, Seattle, Washington 98195; e-mail: tideker@systemsbiology.org, tgalitski@systemsbiology.org, lhood@systemsbiology.org

Key Words biological information, discovery sciences, genome, proteome

■ **Abstract** Systems biology studies biological systems by systematically perturbing them (biologically, genetically, or chemically); monitoring the gene, protein, and informational pathway responses; integrating these data; and ultimately, formulating mathematical models that describe the structure of the system and its response to individual perturbations. The emergence of systems biology is described, as are several examples of specific systems approaches.

INTRODUCTION

Perhaps the most important consequence of the Human Genome Project is that it is pushing scientists toward a new view of biology—what we call the systems approach. Systems biology does not investigate individual genes or proteins one at a time, as has been the highly successful mode of biology for the past 30 years. Rather, it investigates the behavior and relationships of all of the elements in a particular biological system while it is functioning. These data can then be integrated, graphically displayed, and ultimately modeled computationally. How has the Human Genome Project moved us to this new view? It has done so by catalyzing a new scientific approach to biology, termed discovery science; by defining a genetic parts list of human and many model organisms; by strengthening the view that biology is an informational science; by providing us with powerful new high-throughput tools for systematically perturbing and monitoring biological systems; and by stimulating the creation of new computational methods.

Discovery Science

The Human Genome Project was one of the first modern biological endeavors to practice discovery science. The objective of discovery science is to define all of the elements in a system and to create a database containing that information. For

1527-8204/01/0728-0343$14.00

example, discovery approaches are providing the complete sequences of the 24 different human chromosomes and of the 20 distinct mouse chromosomes. The transcriptomes and proteomes of individual cell types (e.g., quantitative measurements of all of the mRNAs and protein species) also represent discovery projects. Discovery science lies in contrast to hypothesis-driven science, which creates hypotheses and attempts to distinguish among them experimentally. The integration of these two approaches, discovery and hypothesis-driven science, is one of the mandates of systems biology.

Genomic Sequences in Humans and Model Organisms

The complete genomic sequences of human (78, 125), nematode (121), fly (2), arabadopsis (81), yeast (54), *Escherichia coli* (19), and a host of microbes and parasites are now available; others, including mouse, are in the pipeline. These sequences offer a number of powerful opportunities.

GENETIC PARTS LIST Software and global experimental techniques are now becoming available to identify the gene locations, and even coding regions, embedded in a sequenced genome (110). Comparative analysis of these coding regions reveals a lexicon of motifs and functional domains (essential to solving the protein-folding and structure/function problems). Moreover, genomic sequence provides access to the adjacent regulatory sequences—a vital component to solving the regulatory code (34)—and opens access to polymorphisms, some of which are responsible for differences in physiology and disease predisposition. Combined, these components make up the elements in the periodic table of life. With these components now in hand, the immediate challenge is to place them in the context of their informational pathways and networks.

MODEL ORGANISMS ARE THE ROSETTA STONES FOR DECIPHERING BIOLOGICAL SYSTEMS The genomic sequences of humans and model organisms have elegantly confirmed a basic unity in the strategy of life. Informational pathways in yeast are remarkably similar to those in fly, worm, and humans, and many orthologous genes can be identified across these species. Thus, it is feasible to use genetically and biologically facile model organisms (yeast, fly, worm) to infer the function of human genes and to place these genes in the context of their informational pathways. Alternatively, comparison of different model genomes offers the possibility of comparing and contrasting the logic of life between organisms. Differences in logic provide fundamental insights into the mechanisms of evolution, development, and physiology.

Biology is an Informational Science

The Human Genome Project has propelled us toward the view that biological systems are fundamentally composed of two types of information: genes, encoding the molecular machines that execute the functions of life, and networks of

regulatory interactions, specifying how genes are expressed. All of this information is hierarchical in nature: DNA → mRNA → protein → protein interactions → informational pathways → informational networks → cells → tissues or networks of cells → an organism → populations → ecologies. Of course, other macromolecules and small molecules also participate in these information hierarchies, but the process is driven by genes and interactions between genes and their environments. The central task of systems biology is (*a*) to comprehensively gather information from each of these distinct levels for individual biological systems and (*b*) to integrate these data to generate predictive mathematical models of the system.

Biological information has several important features:

- It operates on multiple hierarchical levels of organization.
- It is processed in complex networks.
- These information networks are typically robust, such that many single perturbations will not greatly effect them.
- There are key nodes in the network where perturbations may have profound effects; these offer powerful targets for the understanding and manipulation of the system.

Perturbation of Biological Systems

The development of systems biology has been driven by a number of recent advances in our ability to perturb biological systems systematically. Three technological trends have emerged in this respect. First, techniques for genetic manipulation have become more high-throughput, automated, and standardized by several orders of magnitude. Second, the availability of complete genomic sequences has stimulated the development of several systematic mutagenesis projects to complement more traditional efforts involving random mutagenesis. Third, technologies for disrupting genes *in trans* allow the application of genetic perturbations to a wide range of eukaryotic organisms.

HIGH-THROUGHPUT GENETIC MANIPULATION A number of recent and ongoing technological developments are making it possible to rapidly and systematically manipulate genomic material. To illustrate these developments, consider some of the tools available for the budding yeast *Saccharomyces*. Expanding yeast's already-formidable genetic toolkit is a series of plasmids that has greatly facilitated PCR-based gene replacement (83, 127). These versatile vectors render the gene-insertion process simple, standardized, and applicable to any gene or genomic region. First, forward and reverse PCR primers are synthesized with ~40 bp of DNA homologous to a gene of interest and another ~20 bp designed to flank a plasmid-encoded module. These primers are then used to PCR-amplify the module from the plasmid template. PCR products are directly transformed into yeast cells, and homologous recombination occurs with the desired gene at high efficiency. A wide variety of readymade sequence modules are available to disrupt, replace, or

modify essentially any genomic sequence using this technique. These include modules for gene knockouts, promoter fusions, protein fusions, and epitope tags, and because the 20-bp sequences flanking the modules are standard, the same primers can be used for multiple constructions involving several of these different module types. In order to select for successful recombination events while minimizing the formation of undesired recombinant types, modules usually include a marker gene that is non-native to yeast.

The demands of manipulating many genes via many different constructs, then observing the results of these manipulations in many strains simultaneously, are driving the development of ever-more facile and standardized plasmid-construction systems. For example, the GATEWAY recombinational-cloning system [Life Technologies (128)] allows for automated, high-throughput generation of an unlimited array of constructs derived from genes of interest. This procedure exploits the advantages of the bacteriophage lambda integration/excision reaction to transfer genes or other sequences of interest to a virtually unlimited number of clones in separate but identical in vitro reactions. PCR-based cloning into a single entry vector allows transfer of the gene to a variety of destination vectors without the need for restriction endonucleases. This procedure is efficient, standardized, precise, and directional and involves minimal investment in clone isolation or confirmation.

SYSTEMATIC GENE MUTATIONS Presently, there is a transition in gene mutation methods from the random generation of mutant alleles to a systematic approach in which genes are specifically targeted for mutation. The systematic approach has several distinct advantages. With random mutageneses, one employs a selection process or a screen to visually identify mutants of interest. Typically, some genes are identified multiple times, whereas others are not found. In contrast, with a systematic approach, the response of all generated genotypes is documented and the coverage of the genome is unambiguous.

For instance, one recent development is the completion of a collection of deletions of essentially all yeast genes (109, 134). This collection allows the systematic characterization of all yeast gene knockouts for phenotypes that one can assay in a high-throughput screen. In addition, this mutant collection uniquely identifies each deletion mutant genotype with a 20 base-pair "barcode" that can be used to quantify the relative numbers of each genotype in a pooled population of mutant strains (using a microarray of probes against the barcodes). Thus, one can quantitatively assess the fitness of each deletion strain in a given condition in a single experiment. With all single-gene deletion strains now available, ongoing and future research is attempting to characterize phenotypes among genes in combination through the use of double-deletion strains.

GENE DISRUPTION *IN TRANS* Many of the powerful genetic approaches available to yeast and other model organisms are not practicable in higher eukaryotes. However, recent studies suggest that genes in these higher organisms may be perturbed *in trans* using antisense inhibitors of mRNA translation or technologies based on

RNA-mediated interference (RNAi). Gene disruption *in trans* is inducible, thus allowing transient interrogation of practically any gene.

For example, modified oligonucleotides have proven very effective as antisense inhibitors of mRNA translation (44, 91). The modifications (e.g., morpholino groups or phosphoramidate linkages) render the oligonucleotides resistant to nucleases, and the modified oligonucleotides can be delivered by transfection or microinjection.

RNA-mediated interference (RNAi) is another recently discovered mechanism to silence genes in organisms ranging from mice to trypanosomes [recently reviewed by Bass (14)]. The introduction of double-stranded RNA (dsRNA) corresponding to a particular mRNA results in the specific and rapid destruction of that mRNA in cells. The current model (14) proposes that dsRNA is cleaved into 21-25 bp dsRNA molecules that then serve as a template for an RNA helicase that exchanges the sense strand for the mRNA, which is followed by cleavage of the mRNA. The cleavage destroys the mRNA and regenerates the 21-25 bp dsRNA. RNAi was discovered in the nematode worm *C. elegans* (46); systematic investigations of gene function in worm development have used libraries of bacterial clones (fed to the worms) expressing dsRNA (48, 55). These RNAi screens have multiplied the numbers of genes with functional assignments.

Quantitative High-Throughput Biological Tools

Just as the Human Genome Project has led to improvements in our ability to systematically perturb cells, it has also provided us with new technologies for systematically characterizing their cellular response: DNA sequencers, microarrays, and high-throughput proteomics. Because these tools can carry out global (or nearly global) analyses, they become the methods of choice for rapid and comprehensive assessment of biological system properties and dynamics. Typically, the development of these tools goes through three distinct stages: (*a*) proof-of-principle, (*b*) development of a robust instrument, and (*c*) the creation of an automated production line. For example, the automated DNA sequencer was first demonstrated in 1986 (112), made feasible by about 1990 (67), and has enjoyed widespread use in genome sequencing centers since about 1999 (98). The current production-line instrument includes 96 capillary sequencers with a front-end interface of automated sample preparation and a back-end process for tracking the data. A single 96-capillary sequencer can produce ~500,000 base pairs of raw DNA sequence data per day. From 1985 to the present, there has been a 2000-fold increase in the throughput of sequencing, with corresponding increases in the quality of sequence information and simultaneous decreases in cost—all achieved by incremental improvements in chemistry, engineering, and software. High-throughput DNA sequencing production lines may analyze genomic DNA and cDNAs as well as identify and type polymorphisms [either simple sequence repeats or single nucleotide polymorphisms (SNPs)]. Powerful new applications of sequencing technology are also emerging. For example, the biotechnology company Lynx

Therapeutics, Inc. has developed a technique that allows up to 500,000 different sequences to be determined simultaneously for 16 to 20 residues (22). Moreover, this is a powerful discovery approach for determining complete transcriptomes in individual cell types from organisms whose genome has been sequenced.

DNA arrays represent a second kind of powerful discovery tool. Two types of arrays are in common use: cDNA microarrays (102) and oligonucleotide arrays (47, 66). cDNA microarrays consist of double-stranded cDNA or PCR products spotted on a glass slide. If, indeed, the human genome only contains 30,000 to 40,000 genes (78, 125), this approach will easily allow interrogation of complete human transcriptomes. Oligonucleotide arrays are synthesized (66) or spotted (58) on glass slides at densities that can exceed 50,000 spots/slide. In principle, they are more specific than the cDNA microarray and make it possible to distinguish single-nucleotide differences. Using the oligonucleotide array, the mRNAs from individual members of multigene families can be distinguished, alternatively spliced genes can be characterized, alternative forms of SNPs can be identified and typed, and whole stretches of DNA can be resequenced. Clearly, DNA array technology is less mature than sequencing; although the technology is now robust, it is just entering the production-line stage at some companies.

Proteomics, the characterization of the many proteins within a cell type, involves analysis of different types of information corresponding to each protein species: protein identity, abundance, processing, chemical modifications, interactions, compartmentalization, turnover time, etc. Perhaps the major challenge of proteomics is to deal with the enormous dynamic range of protein abundances found in a single cell type—from 1 to 10^6 copies or greater.

For organisms whose genome has been sequenced, mass spectrometry is an especially powerful tool for identifying and quantifying large numbers of proteins (42), identifying and typing SNPs, and analyzing protein modifications. For example, Dr. Ruedi Aebersold and colleagues have recently developed a technique, termed isotope coded affinity tags (ICAT), for measuring the relative expression levels of proteins between two different cell populations (59). In brief, the ICAT reagent is a molecule with three functions: a biotin tag, a linker sequence containing either eight deuterium atoms (heavy reagent) or eight hydrogen atoms (light reagent), and a group reactive to cysteine residues. Proteins from the first cell population are labeled with the heavy reagent, whereas those from the second cell population are labeled with the light reagent. Equal quantities of each protein sample are combined and digested with trypsin, and cysteine-labeled peptides are isolated with an avidin column. Mass spectrometry is used to analyze the paired atomic masses for each peptide (light vs. heavy peptides differ by eight mass units) and, after further fragmentation, to determine their amino-acid sequences. Thus, peptides can be quantitated (to an accuracy of $\sim$20%) and the corresponding genes identified. Aebersold and colleagues have created an automated high-throughput production line for this procedure, capable of analyzing 1000 proteins per day; they are currently developing a next-generation facility to analyze up to 1 million proteins per day.

Because of their ability to separate many different cells and cell types at high speed, multiparameter cell sorters are another technology critical to systems biology. Although microarray and proteomics experiments typically measure average levels of mRNA or protein within a cell population, in reality these levels can vary from cell to cell; this distribution of expression levels contains important information about the underlying control mechanisms and regulatory network structure. Dr. Ger van den Engh and colleagues have developed a new cell sorter capable of separating 30,000 elements per second against 32 different parameters.

Computation for Systems Biology: Databases and Models

Biology is unique among the natural sciences in that it has a digital code at its core. Together with colleagues in computer science, mathematics, and statistics, biologists are developing the necessary tools to acquire, store, analyze, graphically display, model, and distribute this information. An enormous challenge for the future is how to integrate the different levels of information pertaining to genes, mRNAs, proteins, and pathways.

THE INCREASING IMPORTANCE OF COMPUTER DATABASES Computer databases first rose to prominence in molecular biology as central repositories for the plethora of data generated by large-scale sequencing projects. Although databases of nucleic-acid and amino-acid sequences are still the largest, most utilized, and best maintained, there has been a sudden explosion of interest in databases to store other types of molecular data. Such interest is primarily in response to demands placed by functional genomics and other emerging systems approaches. For instance, the Database of Interacting Proteins (137), BIND (10), and MIPS (90a) contain searchable indices of known protein-protein interactions; TRANSFAC (133) and SCPD (142) catalog interactions between proteins and DNA (i.e., transcription-factor interactions), and databases of metabolic pathways have also recently been established [e.g., EcoCyc (73), KEGG (92), and WIT (106)]. A growing number of databases are also under development for storing the now sizeable number of mRNA-expression data sets (1, 43, 63, 96, 116); companies, such as Affymetrix, Rosetta, Spotfire, Informax, Incyte, Gene Logic, and Silicon Genetics, market gene-expression databases commercially. A comprehensive review of recent developments in the molecular biology databases is available elsewhere (15).

This recent explosion, in both the variety and amount of information of interest, poses two challenges to database users and developers alike. First, the information must be maintained systematically in a format that is compatible with both single queries and global searches. Often, the desired information is present in the database but is not annotated consistently for all entries: For instance, an EST sequence or expression profile may have been derived from cancer cells, but in the absence of an enforced annotation style, this information may be recorded using many different keywords (e.g., cancer, tumor, metastatic, carcinoma, etc.). Alternately, the information recorded in the databases may be incomplete: For

instance, a protein-interaction database may correctly document an interaction between two proteins but may fail to include related, highly informative data, such as the strength of binding or the result of interaction on the functional activity of each protein.

The second challenge, and perhaps the more difficult one, is keeping the databases updated against the ever-increasing body of biological knowledge. In this regard, computer scientists working in the field of natural-language processing have made promising advances in computer programs that can parse textual passages, extract the key concepts, and catalog these concepts systematically (6, 18, 33, 38, 95, 117, 122). Thus in the near future, there is hope of updating biological databases automatically with the thousands of relevant results published each month in the primary biological literature, reducing the dependence on humans to perform this tedious, error-prone, and time-consuming task.

Why are databases so important to the future of systems biology? Although individual researchers may amass a great deal of knowledge about the genes, molecular interactions, and other biological information underlying one particular pathway, no single biologist can be familiar with the extremely large and complex number of interactions in an entire cell. The databases track all of these, provided that the analytical approaches are available to help the biologist to access, display, and interpret the information. In the end, however, biology cannot be done solely in silico. Biologists must employ their insights to bring coherence to the massive data sets.

THE INCREASING IMPORTANCE OF GLOBAL ANALYSES Given the recent accumulation of expression profiles, molecular interactions, and a variety of other global data in the biological databases, the immediate task is to develop powerful analyses and experimental strategies to integrate and analyze these data to make biological discoveries. To date, methods to analyze patterns of gene expression have received the most attention. In the most straightforward approach, gene-expression data are used to identify genes involved in a particular biological process, by selecting genes with clear changes in expression over different biological conditions or over time. Depending on the experiment, genes of interest have been implicated in cancer (129), development (131), aging (79, 107), or a specific cellular response (28, 29, 35, 50, 71, 114). In most of these cases, expression levels of tens, hundreds, or even thousands of genes changed over the conditions examined, often expanding the known number of changes by an order of magnitude.

Genes with similar responses over multiple conditions are often clustered together to form functional groups or to reveal coordinated patterns of expression. Several clustering methods have been proposed: Most are excellent and have associated software packages that are publicly available (4, 23, 25, 36, 39, 57, 62, 118). Some analyses achieve more specific and/or accurate functional predictions by integrating gene-expression clusters with complementary types of global data: for example, searching for shared regulatory sequences in the promoters of coexpressed genes (29, 89, 97, 111, 114, 119). Identification of these regulatory

sequences provides evidence that the associated genes are under the control of common transcription factor(s). Alternatively, gene-expression data may be combined with information on protein-protein interactions and protein phylogenetic profiles (85) or augmented with the genomic location of each gene to find shared regulatory elements (32). Finally, deciding whether an expression level changes significantly can be a difficult problem and is also an active area of research (27, 57, 65, 69, 84). Methods for analyzing expression data, including gene-expression clustering and its extensions, have been extensively reviewed elsewhere (21, 30, 64, 108, 141).

Where do we go from here? Although these analyses have certainly been informative, global data sets undoubtedly provide additional information that remains untapped. Ultimately, it would be highly desirable to analyze expression levels and other global measurements in a way that validates our current knowledge of a cellular process and isolates discrepancies between expected and observed levels. To achieve this level of analysis, we believe that it will be necessary to compare and incorporate global data with a well-defined model of the biological process of interest.

THE INCREASING IMPORTANCE OF COMPUTER MODELS Conventionally, a biological model begins in the mind of the individual researcher, as a proposed mechanism to account for some experimental observations. Often, the researcher represents their ideas by sketching a diagram using pen and paper. This diagram is a tremendous aid in thinking clearly about the model, in predicting possible experimental outcomes, and in conveying the model to others. Not surprisingly, diagrammatic representations form the basis of the vast majority of models discussed in journal articles, textbooks, and lectures.

Despite the useful simplicity of these conventional models, advances in systems biology are prompting some biologists to forego "mental" models, or pen-and-paper diagrams, for more sophisticated computer representations. Although the notion of modeling a biological process computationally is almost as old as the electronic computer itself [e.g., see biological models proposed by Turing (124)]; such models are gaining in importance for several reasons. First, it is now apparent that the magnitude and complexity of interactions in a cell are simply too vast for an unaided human mind to process and organize (132). Second, as DNA microarrays, sequencers, and other large-scale technologies begin to generate vast amounts of quantitative biological data, a paradigm shift is occurring in biology away from a descriptive science and toward a predictive one (52, 77). Computer systems are required to store, catalogue, and condense the rapidly accumulating mass of data, and automated tools are needed that, by assimilating these data into a network model, can predict network behaviors and outcomes that may be tested experimentally. It is encouraging that recent computer simulations of partial or whole genetic networks have demonstrated network behaviors, commonly called systems properties or emergent properties, that were not apparent from examination of a few isolated interactions alone (5, 16, 74).

Computer modeling tools have already achieved widespread acceptance within the engineering and physical sciences. For example, computer-aided design packages, such as SPICE, VHDL, or Prolog, are heavily used to simulate and test electronic circuitry (93). In contrast, a relative paucity of software and methods exists for analyzing biological circuits. Several useful tools are available for simulating small networks of chemical reactions [e.g., Gepasi (90) or the Chemical Reaction Network Toolbox (45)], but larger-scale simulations are still emerging. For example, ongoing projects, such as E-CELL (123) and the Virtual Cell (100, 101), attempt to model all molecular interactions in the cell as an integrated, computational process. Other efforts, such as BioJake (99) and a collection of guidelines set forth by Kohn (75), are working to define a standard graphical environment in which biologists may interactively define and simulate genetic circuit models. A widespread, standard notation (and/or software environment) is attractive because systems biologists working on diverse systems and at different institutions would be able to directly exchange their fully detailed models.

TYPES OF COMPUTER MODELS A wide variety of cellular models have been proposed, each of differing complexity and abstraction. For example, chemical kinetic models attempt to represent a cellular process as a system of distinct chemical reactions. In this case, the network state is defined by the instantaneous quantity (or concentration) of each molecular species of interest in the cell, and molecular species may interact via one or more reactions. Often, each reaction is represented by a differential equation relating the quantity of reactants to the quantity of postreaction products, according to a reaction rate and other parameters. This system of differential equations is usually too complex to be solved explicitly, but given an initial network state, the quantity of each gene product or other molecular species can be simulated to produce a state transition path or trajectory, i.e., the succession of states adopted by the network over time. A variety of biological systems have been modeled in this way, including the networks controlling bacterial chemotaxis (5, 20), developmental patterning in *Drosophila* (24, 86), and infection of *E. coli* by lambda phage (88). Recently, it has been pointed out that transcription, translation, and other cellular processes may not behave deterministically but instead are better modeled as random events (87). Models have been investigated that address this concern by abandoning differential equations in favor of stochastic relations to describe each chemical reaction (8, 53).

In contrast to models involving systems of chemical reactions, another popular approach has been to model a genetic network as a simplified discrete circuit. Much like a neural network, this approach represents the network as a graph with nodes and arrows (i.e., directed edges), where a node represents the quantity or level of a distinct molecular species and an edge directed from one node to another represents the effect of the first node's level on that of the second. Also required is a function for each node, describing how all of the incoming effects should be combined to determine its level. Typically, nodes may assume one of two discrete levels, signifying whether the molecule is present or absent or whether a gene is

turned on or off. Given a starting state of levels for all nodes, the next level of each node may be determined directly from its function. In this way, the network state over all nodes evolves over a series of discrete time steps, where the state of the next step is computed from the current state.

Discrete circuit models have been investigated extensively (74, 113, 136), and simulation software is available (135). Clearly, such models are greatly simplified compared to a kinetic model. Proponents of discrete circuit models argue that they preserve the essential features of the underlying biology while greatly reducing network complexity and simulation time. A major criticism has been that they require the model to update simultaneously for all nodes, whereas molecular interactions within the cell are not synchronous. Also, a two-level representation of molecular species may not always be sufficient to capture the underlying biological behavior of the network.

CHOICE OF MODEL DETAIL Formulation of a model involves important choices about which genes, gene products, and other molecular species should be included in the network state. Genes may be regulated at the level of transcription or translation, and once translated, a protein may exist in one of several modified forms. Through alternative splicing, a single gene may encode several distinct mRNAs, and nongenetic molecules, such as metabolites, may also affect the network. Furthermore, some interactions are restricted to the nucleus, cell membrane, golgi apparatus, or other organelles. A complete model would therefore have to include molecular species, such as alternatively-spliced mRNA and modified protein products, and would have to restrict interactions between species located in different cellular compartments.

Increasing levels of detail are not always desirable, however, and deciding which information to include can be a difficult task. In general, one identifies the types of properties or behaviors that the model should be able to predict (e.g., mRNA levels, protein activation states, or growth rates) and includes only the components that impact these properties. The number of model parameters must be compatible with the amount and type of available data: If only mRNA-expression levels are measured, for instance, then detailed information about protein structures or compartmentalization may overload the model with too many hidden variables.

INFERENCE OF MODELS FROM GLOBAL PATTERNS OF GENE EXPRESSION Several methods have been proposed for inference of a genetic network from a measured time series of mRNA-expression profiles. These methods try to infer a discrete circuit model by looking for statistical correlations between expression levels (7), by training a neural network (130), or by information theoretic methods (80). Also under development are methods for inferring models from steady-state expression profiles, e.g., recorded over a battery of biological conditions or gene deletions (3, 70). In the future, several or all of these methods will almost certainly be expanded to take advantage of other types of global data, such as protein-expression levels, protein-modification states, or metabolite concentrations.

A FRAMEWORK FOR SYSTEMS BIOLOGY

Ultimately, one wishes to understand the underlying interactions, molecular or otherwise, that are responsible for the global changes observed in a system. In order to most directly address this goal, we argue that it will be necessary to integrate the various levels of global measurements together and with a mathematical model of the biological system of interest. Although these model-driven approaches may differ in the particulars of implementation, all follow a fundamental framework involving several distinct steps (as shown in Figure 1):

1. **Define all of the components of the system.** Use these components, along with prior biochemical and genetic knowledge, to formulate an initial model. Ideally, a global approach is the most powerful (i.e., defining all genes in the genome, all mRNAs and proteins expressed in a particular condition, or all protein-protein interactions occurring in the cell) because it does not require any prior assumptions about system components. Constructing a model by interrogating these components will ultimately accomplish two objectives: (*a*) to describe the structure of the interactions that govern the systems behavior and (*b*) to predict accurately relevant properties of the system given specified perturbations. If prior knowledge about the system is limited, the initial model may be rough and may involve purely hypothetical interactions.
2. **Systematically perturb and monitor components of the system.** Specific perturbations may be genetic (e.g., gene deletions, gene overexpressions, or undirected mutations) or environmental (e.g., changes in growth conditions, temperature, or stimulation by hormones or drugs). The corresponding response to each perturbation is measured using large-scale discovery tools to capture changes at relevant levels of biological information (e.g., mRNA expression, protein expression, protein activation state, overall pathway function). Once observed, data from all levels are integrated with each other and with the current model of the system. As in step 1, an approach in which all components are systematically perturbed and globally monitored is the most desirable.
3. **Reconcile the experimentally observed responses with those predicted by the model.** Refine the model such that its predictions most closely agree with experimental observations. Agreement between the observed and predicted responses is evaluated qualitatively and/or quantitatively using a goodness-of-fit measure. When predictions and observations disagree, alternative hypotheses are proposed to alleviate the discrepancies (maximize the good-ness-of-fit), resulting in a refined model for each competing hypothesis. If the initial model is largely incomplete or is altogether unavailable, the observed responses may be used to directly infer the particular components required for system function and, among these, the components most likely to interact. If the model is relatively well defined, its predictions may already

be in good qualitative agreement with the observations, differing only in the extent of their predicted changes.

4. **Design and perform new perturbation experiments to distinguish between multiple or competing model hypotheses.** Even for a moderate number of observations, the proposed refinements may result in several distinct models whose predictions fit equally well with the observations. These models are indistinguishable by the current data set, requiring new perturbations and measurements to discriminate among them. New perturbations are informative only if they elicit different systems responses between models, with the most desirable perturbations resulting in model predictions that are most dissimilar from one another. After choosing the set of new perturbations, repeat steps 2 through 4, thereby expanding and refining the model continually, over successive iterations. The idea is to bring the theoretical predictions and experimental data into close apposition by repeated iterations of this process so that the model predictions reflect biological reality.

Thus, systems biology requires that all of the elements of a system be studied (at multiple levels of the information hierarchy and in the context of their responses to perturbations), that these data be integrated and graphically displayed, and finally, that these responses be modeled mathematically to predict the structure and behavior of the informational pathway. Moreover, systems biology involves an iterative, strategic interplay between discovery- and hypothesis-driven science. Global observations (discoveries) are matched against model predictions (hypotheses) in an iterative manner, leading to the formation of new models, new predictions, and new experiments to test them.

EXAMPLES OF SYSTEMS BIOLOGY

A large number of recent and ongoing efforts are putting this systems biology framework into practice (as illustrated in Table 1). We now examine in detail four such studies, representing four distinct types of biological networks: (*a*) a *cis*-regulatory network, (*b*) a *trans*-regulatory network, (*c*) a signal-transduction network, and (*d*) a synthetic regulatory system engineered according to a predetermined network model.

Cis Gene Regulation in the Sea Urchin

Gene-regulatory networks are defined by *trans* and *cis* logic (34). *Trans* logic defines the interactions between protein transcription factors and the batteries of genes they control (e.g., other transcription factors as well as genes in the network periphery). Conversely, *cis* logic defines the precise relationships among promoter elements (DNA sequences) whose states (i.e., bound vs. unbound by transcription factors) are combined to produce the temporal and spatial patterns of expression for a particular gene. Both of these types of regulatory networks

TABLE 1 A sampling of systems-biology approaches

Model systems	Organisms	Approaches	References
Viral infection of *E. coli*	phage λ	Computer simulations via a mixed model (discrete and continuous); stochastic simulations	(8, 87, 88)
	phage T7	Time-differential equations	(41)
Bacterial chemotaxis	*E. coli*	Time-differential equations; stochastic simulations	(5, 20, 31, 138)
Embryo patterning or development	Sea urchin	Identification of *cis*-regulatory elements and interactions; computational modeling of the proposed *cis* network	(9, 139)
	Drosophila	Simulation via time-differential equations	(24, 86, 126)
Cross-talk between signaling pathways	Mammals	Simulation via time-differential equations	(16, 132)
Sugar metabolism	*S. cerevisiae*	Integrated physical-interaction network, Bayesian networks	(61, 68, 120)
Protein networks	*S. cerevisiae*	Directed-graph models	(68, 94, 105)
General metabolism	*E. coli*	Flux-balance analyses	(37, 103, 104)
Whole-cell model	*E. coli*, *neuron*	Simulation via time-differential equations	(100, 123)
Synthetic and circadian oscillators	*E. coli*, *Drosophila*, *Neurospora*, mice	Synthesis of a multigene network in vivo using a computer model as blueprint	(12, 40)
Synthetic flip-flop	*E. coli*	” ”	(51)
Cell cycle	Mammals, yeast	Molecular-interaction maps; Bayesian networks	(49, 75)

have input and output. For instance, network inputs may arise from exogenous signals (e.g., sperm penetrating the egg, steroid hormones, etc.) or from signal-transduction pathways. The output of the network, concentration of nuclear RNA, exhibits many possible levels of posttranscriptional control (e.g., RNA processing, alternative RNA splicing, protein processing, protein chemical modification, etc.).

The sea urchin is a powerful model for studying *cis* and *trans* regulation because its development is relatively simple (34) (the embryo has only 12 different cell types); enormous numbers of eggs can be obtained in a single summer (30 billion); the eggs can be fertilized synchronously and development stopped at any stage; and many transcription factors can be readily isolated and characterized, and their genes may be cloned using affinity chromatography, conventional protein chemistry, protein microsequencing, DNA probe synthesis, or library screening (60). A great

deal of developmental biology has been carried out on the sea urchin, and a modest genome effort has defined the general features of its genome (26).

The sea urchin *endo16* gene has the most completely defined *cis*-regulatory system to date (34, 139, 140). Strikingly, this system is highly analogous to a computer or other electronic circuit; multiple inputs from a wide range of transcription factors are integrated to send a signal to the RNA-synthesizing machine (the basal transcription apparatus) as to whether and how much transcript to synthesize. Endo16 is expressed in the endoderm of the embryo: It appears first in the vegetal plate (which gives rise to endodermal and mesodermal cell types), emerges later in the archenteron, and finally, intensifies in the midgut while diminishing in the fore- and hindgut (Figure 2*A*). Thus, the *cis*-regulatory apparatus must turn on gene expression in the appropriate cells, establish sharp boundaries of expression (expressed in endodermal but not mesodermal cells), and specify the cells of terminal expression.

A 2.3 kb sequence of genomic DNA contains all of the control elements necessary for normal endo16 expression patterns. Figure 2*B* depicts the 34 binding sites spread across this region, together with the 13 different transcription factors that bind them. The binding sites fall into seven regions of DNA sequence: six functional regions (modules A–G) and the basal promoter region where the transcriptional apparatus assembles. Each of these was defined by mutating one or more binding sites, attaching the resulting sequence to a reporter construct, placing this construct into transgenic sea urchins, then measuring the spatial and temporal gene-expression output. For example, this approach revealed that the module G is a positive booster, whereas the F and D modules are responsible for repressing gene expression in the adjacent ectoderm. Module A is the sole means of communication between the six functional regions and the basal transcription apparatus, integrating the positive or negative inputs from the G, F, E, DC, and B modules. In addition, module A mediates expression in the vegetal plate of the early embryo.

From these studies, a logic model was constructed delineating all the operations executed by these modules and their interactions (Figure 2*C*). This model clearly indicates that the output of module B runs through module A (which amplifies it) on the way to instructing the basal transcription apparatus. These interactions can be boolean (*dotted lines*), scalar (*thin solid lines*), or time-varying quantitative (*heavy solid lines*) inputs (Figure 2*C*). The important point is that the *cis*-control region behaves as a series of integrated electronic circuits (modules), each combining their environmental inputs (quantitatively changing levels of transcription factors) to send signals through module A to set the overall circuit output. Moreover, the model was derived according to the systems biology framework described earlier, having been developed after many iterative cycles of perturbation and gene-expression measurements.

The *cis*-regulatory regions of other genes and other organisms employ a very similar logic, although they certainly differ in detail (34). Thus, the responses of an organism, both developmental and physiological, are hardwired into its *cis*-regulatory circuitry. Almost certainly, the central driver of evolution is not changes

in individual genes, but rather changes in this circuitry. Clearly one of the major challenges for biology in the twenty-first century will be coming to understand the nature of the *cis*- and *trans*-regulatory networks that control an organism's development, physiological responses, and even its trajectory of evolution.

A Network Controlling Galactose Utilization in Yeast

As another example, we (T. Ideker & L. Hood) have recently used a systems approach to explore, expand, and refine the understanding of galactose utilization (GAL) in yeast (68). Like the previous example of sea urchin development, a regulatory network model is used to predict changes in gene expression resulting from a battery of directed perturbation experiments. Unlike the sea urchin example in which the components of the network are the *cis*-regulatory sequences controlling a particular gene, the network controlling galactose utilization consists of a large number of genes and gene products interacting in *trans*. Although the *trans* model includes less detail about the regulation of individual genes, it provides new information on how groups of genes interact to control a cellular process.

As shown in Figure 3, the yeast galactose-utilization system employs at least nine genes. Four encode the enzymes that catalyze the conversion of galactose to glucose-6-phosphate (*GAL1*, *5*, *7*, and *10*), whereas a fifth (*GAL2*) encodes a transporter molecule that sets the state of the system. If galactose is present in the yeast cell, the system is turned on; if galactose is absent, the system is turned off. A number of transcription factors regulate this on/off switch, including *GAL3*, *4*, *80*, and possibly *GAL6*.

GENETIC AND ENVIRONMENTAL PERTURBATION OF THE NETWORK We wished to determine whether the molecular interactions in the galactose network were sufficient to account for changes in gene expression resulting from extensive perturbations to the GAL pathway. Toward this goal, we constructed nine genetically perturbed yeast strains, each with a deletion of a different GAL gene (see above). These strains, along with wild-type yeast (no genes deleted), were grown to steady state in the presence (+gal) or absence (−gal) of 2% galactose. For each of these 20 perturbation conditions (10 strains $\times$ 2 media types), we used a whole-yeast-genome microarray to monitor changes in mRNA expression relative to wild type (+gal).

Nine hundred ninety-seven mRNAs (out of $\sim$6200) showed statistically significant concentration changes during one or more of these perturbations. The corresponding perturbed genes could be divided into 16 different clusters, where the genes within a cluster behaved in a similar manner through all perturbations. The striking observation was that the genes encoding various metabolic, cellular, and synthetic pathways tended to fall in individual clusters—thus beginning to reveal the network of interconnected informational pathways within the yeast cell.

CONSTRUCTION AND VISUAL DISPLAY OF THE NETWORK MODEL To reveal the nature of these informational pathways, we constructed a model of the known

molecular interactions connecting galactose utilization with other metabolic processes in yeast. For this purpose, we compiled a list of 3026 previously observed physical interactions, using all available entries from publicly available databases of protein-protein (105) and protein-DNA interactions (133, 142). The interactions in these databases came from several different sources relying on a variety of experimental approaches: Most were derived from biochemical association studies reported in the literature or through large-scale experiments such as the two-hybrid screen.

The interactions in these databases define a model of the molecular-interaction network, shown in Figure 4, for two small regions. Each node in the network represents a gene and is labeled with its corresponding gene name. An arrow directed from one node to another signifies that the protein encoded by the first gene can influence the transcription of the second by DNA binding (a protein $\rightarrow$ DNA interaction), whereas an undirected line between two nodes signifies that the proteins encoded by each gene can physically interact (a protein-protein interaction).

Expression data from each perturbation can be visually superimposed on the network. For example, Figure 4*a* shows the result of the *gal4*Δ deletion in the presence of galactose. In the figure, the grayscale intensity of each node represents the change in mRNA expression of its corresponding gene. When other types of information are available, they too can be superimposed on the network display. For example, we also measured changes in protein abundance for wild-type cells grown in the presence vs. absence of galactose (68). Using a procedure based on isotope coded affinity tags (ICAT) and tandem mass spectrometry (59), we detected a total of 289 proteins and quantified their expression-level changes between these two conditions. Strikingly, 30 proteins showed significant concentration changes, $\sim$15 of which showed no changes at the mRNA level. The implication is clear—these 15 proteins are regulated by posttranscriptional mechanisms—a compelling argument for the need to integrate both mRNA- and protein-expression changes to understand eukaryotic gene regulation. Figure 4*b* illustrates the addition of protein-abundance information to the visual display, focusing on the region of the network corresponding to amino-acid biosynthesis. By comparing the mRNA- and protein-expression responses displayed on each node, one can visually assess whether the mRNA and protein data are correlated and quickly spot genes for which they are remarkably discordant.

COMPARISON OF OBSERVED AND PREDICTED RESPONSES In keeping with the systems-biology framework outlined previously, we wished to determine if the expression changes observed across the 20 perturbations were consistent with changes as predicted by the molecular-interaction network. By itself, the network model makes only very coarse predictions. For example, a protein-DNA interaction involving protein A and gene B predicts that a change in expression of A could result in a change in expression of B. If A is additionally involved in a protein-protein interaction with C (C—A$\rightarrow$B), then a change in expression of C could also elicit a change at B, by first altering the activity of A. However, the network does not

dictate whether these interactions activate or repress transcription or, in the case that multiple interactions affect a gene, how these interactions should be combined to produce an overall change in expression. Similarly, the network does not specify whether a protein-protein interaction results in the formation of a functional protein complex or if, instead, one protein transiently modifies the other. Because none of these levels of information are encoded in the protein-DNA or protein-protein databases, they are also absent from the network model and the graphical display.

Interestingly, much of this information is known outside of the databases: In the case of the GAL genes, classic genetic and biochemical experiments have determined that Gal4p is a strong transcriptional activator and that Gal80p can bind to Gal4p to repress this function [see reviews by Johnston et al. (72) and Lohr et al. (82)]. Thus, we have supplemented the model with these and other results from the literature that address the effect of perturbing GAL genes on the expression of other genes in the cell. For example, when such information is known, we can indicate whether each relevant protein-DNA interaction serves to activate or repress gene expression. Likewise, we can indicate whether each protein-protein interaction can alter the activity of either protein and whether this change is positive or negative. Once incorporated into the model, these added levels of detail greatly increase its predictive power. Note that it is not necessary to integrate all previous evidence into the model, just evidence that bears on gene expression.

Figure 5 compares the observed to the predicted expression responses of the GAL genes, for each of the 20 perturbations. Although the observed response is obviously more complex than the predicted one, the two responses agree in many of their salient features. Not shown in the figure are the approximately 990 additional genes, outside of the core GAL pathway, whose mRNA expression levels were affected in at least one perturbation. Interestingly, very little is known about the molecular interactions that determine how the galactose-utilization pathway may influence these other genes. The majority of these expression changes, therefore, are not yet addressed by the model and will call for the addition of new interactions.

REFINING THE MODEL THROUGH ADDITIONAL PERTURBATIONS We used discrepancies between the predicted and observed expression responses to suggest possible refinements to the model. For example, the current model predicts that in galactose, perturbations to GAL enzymes (i.e., *gal1*Δ, *7*Δ, or *10*Δ) should not affect expression levels of other GAL genes. Although this is largely true for *gal1*Δ, the *gal7*Δ and *gal10*Δ deletions clearly affect expression levels of *GAL1*, *2*, *3*, *7*, *10*, and *80* (see Figure 5). Because both *gal7*Δ or *gal10*Δ deletions block the conversion of galactose-1-phosphate (Gal-1-P), leading to increased levels of this and other metabolites (76), one hypothesis is that one of these metabolites exerts control over GAL-gene expression. To address this hypothesis, we examined expression changes in a *gal1gal10*Δ double deletion strain grown in galactose. Although deletion of *GAL10* blocks the conversion of Gal-1-P, deletion of *GAL1* blocks a preceding step in the galactose-utilization pathway such that Gal-1-P levels are

greatly reduced. Thus, if metabolites are the cause of the change in GAL-gene expression observed in a *gal10*Δ mutant, these changes are predicted to disappear in the *gal1*Δ*gal10*Δ strain. In fact, when we measured the gene-expression profile of this double deletion using a microarray, this is exactly what happened, lending support for this revised model.

Thus, the systems approach to galactose utilization has given us new insights into how the pathway is regulated (e.g., it has generated many new, testable hypotheses) and how it is interconnected with other informational pathways in the yeast cell. Accordingly, this approach may be very powerful in elucidating the network of informational pathways in other biological systems and, ultimately, the interconnected networks of cells in metazoan organisms.

Bacterial Chemotaxis: A Robust Signal-Transduction Network

Bacterial chemotaxis, the process by which bacteria move toward or away from a chemical source, has intrigued researchers for 120 years. The last few decades of research (largely in *E. coli*) have focused on elucidating the molecular interactions responsible for the chemotaxic response, resulting in a large body of genetic, structural, physiological, and biochemical data [see recent reviews (17, 56)]. These large data sets have led to a number of recent attempts to model the chemotaxis network using systems approaches.

BEHAVIOR AND MOLECULAR BIOLOGY OF CHEMOTAXIS The physiology of the chemotactic response has been relatively well characterized (17). The bacterium moves in a chemical gradient by means of a biased random walk, alternating episodes of swimming straight (running) and random reorientation (tumbling). To run, the bacterium turns its flagellar motors counter-clockwise; a tumble ensues when the motors are reversed. As it moves, the bacterium senses gradients of chemical attractant (e.g., aspartate) or repellent (e.g., hydrogen peroxide) as changes in concentration over time. Increasing attractant or decreasing repellent results in an increase in the duration of runs in the desired direction.

As shown in Figure 6, the molecular biology of the chemotaxis system is also known in considerable detail. Five transmembrane attractant receptors, known as methyl-accepting chemotaxis proteins (MCPs), have multiple methylation sites whose modification state governs signal transduction via a phospho-relay system. This signaling network modulates the output of flagellar motors. It also adapts the sensitivity of the network to changing concentrations of attractant or repellent.

MODELING "ROBUSTNESS" AS A SYSTEMS PROPERTY The molecular interactions responsible for chemotaxis have been studied quantitatively. For instance, Spiro et al. (115) modeled the transitions of the various chemotactic signaling molecules among different states of ligand occupancy, phosphorylation, and methylation. Using the equations of mass-action kinetics, they succeeded in recapitulating the

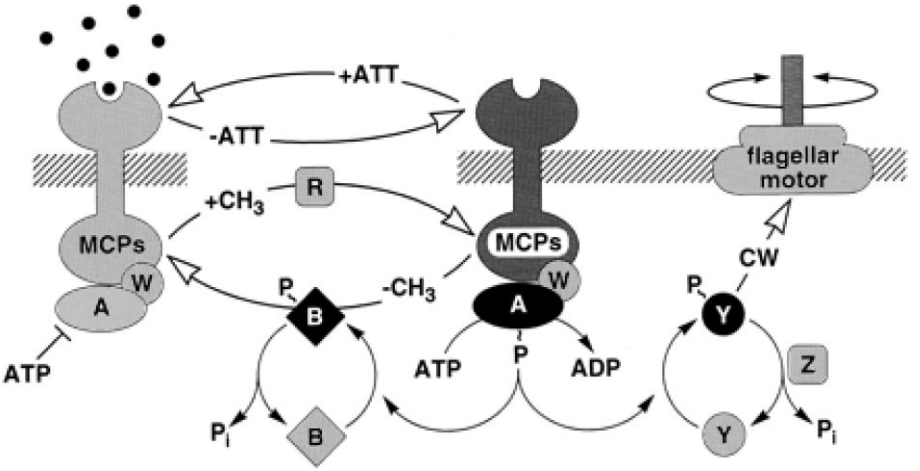


Figure 6 The bacterial chemotaxis system. Transmembrane attractant receptors, known as methyl-accepting chemotaxis proteins (MCPs), form a signaling complex with CheW and a kinase, CheA, that autophosphorylates and transfers phosphates to a response regulator, CheY. Phospho-CheY interacts with flagellar motor proteins to induce clockwise rotation (tumbling). The CheZ protein promotes CheY dephosphorylation. CheA transfers phosphates also to the CheB MCP methylesterase. The activated phospho-CheB demethylates MCPs; this demethylation diminishes the kinase activity of the MCP-CheW-CheA complex as part of the adaptation mechanism. CheR is a constitutive MCP methyltransferase. In the presence of increasing attractant, CheA autophosphorylation is inhibited, counterclockwise flagellar rotation (running) is extended, and subsequent MCP methylation allows adaptation to the higher attractant concentrations. When moving through decreasing attractant, CheA autophosphorylation is stimulated. This, in turn, promotes phospho-CheY induced tumbling and then adaptation through demethylation of MCPs. Figure from Spiro et al. (115).

response of these molecules to attractant gradients, step increases, and saturation. In doing so, Spiro et al. moved beyond the study of isolated components toward a quantitative and integrative reproduction of many interactions simultaneously. This complex web of interactions makes it possible to observe, model, and predict system properties (i.e., properties that are observed behaviorally but are not readily understood by studying any individual component of the system).

One interesting systems property of the chemotaxis network is its robustness of adaptation to different attractant concentrations (5, 11, 138). In this context, robustness means that the output of the system is insensitive to particular choices of its inputs or biochemical parameters (e.g., enzyme levels and rate constants). In particular, robust adaptation means that the system output responds to changes in the input, without depending on the overall input magnitude. Thus, *E. coli* cells respond to changes in attractant concentration (system input) by changing their tumbling frequency (output). However, owing to robust adaptation, a homogenous

solution of attractant will always result in the same tumbling frequency, regardless of the attractant concentration.

PRECISION OF ADAPTATION Modeling work of Barkai & Leibler (11) demonstrated that the robust adaptation of chemotaxis is a result of the structure of the underlying molecular signaling network. Assuming a simple two-state model in which the MCP-CheW-CheA protein complex is either active or inactive, they were able to capture a wide variety of behaviors that had been previously observed by experiment. In addition, they were able to make several striking predictions regarding the properties of adaptation. They defined the precision of adaptation of the chemotaxis network to mean its ratio of steady-state output before vs. after a change in input. The network model predicted that this ratio was always equal to one, thus indicating that precision of adaptation is a robust property; that is, the tumbling frequency is only transiently affected by an increase in the level of attractant and will eventually return to its initial steady-state value. In fact, one can change rate constants over several orders of magnitude and still maintain precise adaptation, regardless of whether Michaelis-Menten or cooperative kinetics are used for the model simulations. Interestingly, many other properties of chemotaxis were predicted to be nonrobust; for example, neither the adaptation time (the interval between the input change and the re-establishment of steady-state output) nor the tumbling frequency (which depends on enzyme levels) were robust properties under the model.

In a related study, Alon et al. tested these predictions directly, by varying enzyme concentrations over two orders of magnitude and observing the response of *E. coli* to the addition of saturating attractant (5). They observed steady-state tumbling frequency, adaptation time, and precision of adaptation. As predicted by the model, tumbling frequency and adaptation time were highly sensitive to changes in enzyme level, whereas precision of adaptation remained relatively constant (within the bounds of experimental error).

Returning to the model, the key structural feature responsible for this robust behavior appears to be a feedback-control loop involving modification of the MCP-CheW-CheA protein complex. Because the modification rate depends on the activity of the protein complex and not on the concentrations of its various modified forms, system activity tends to return to an initial steady state following a change in input.

Forward Engineering of Biological Networks

All of the examples discussed thus far have involved constructing a mechanistic model of a naturally occurring biological system. However, the systems-biology framework outlined above can equally be used to construct a synthetic system according to a predetermined model, with one key difference: In the former scenario, one alters the model to best fit the biological system (i.e., reverse engineering), whereas in the latter scenario, one alters the biological system to fit the model (i.e., forward engineering).

Although the idea of engineering biological systems to have desired properties or particular functions is not new, a series of ongoing research projects are putting these ideas into practice. Spearheading these efforts is work by Elowitz et al. (40), who constructed a gene-regulatory system in *E. coli* that functions as a synthetic oscillator, and work by Gardner et al. (51), who demonstrated a genetic toggle switch based on similar principles.

The basic network configuration of Elowitz et al. involves three transcriptional repressor proteins organized into a negative-feedback loop (Figure 7). To explore the potential oscillatory behavior of this configuration, the group constructed quantitative models (both kinetic and stochastic) describing the change in protein concentration over time for each of the three genes in the system. These models involved a number of biochemical parameters, including the overall rate of translation, the rates of mRNA and protein degradation, and the dependence of transcription rate on the concentration of the corresponding protein repressor. In simulations, a high protein-degradation rate (relative to mRNA degradation) tended to produce the desired oscillatory behavior. These simulations prompted the group to insert a carboxy-terminal tag at the 3′ end of each of the three repressor genes: These tags increased the degradation rate of each protein by targeting them for destruction by cellular proteases. In this way, parameters of the biological system were adjusted to match the desired parameters of the model.

To explore the behavior of the network in vivo, *E. coli* cells were transformed with two plasmids: one encoding the three repressor proteins and another containing green-fluorescent protein (GFP) under the transcriptional control of one of the repressors. By monitoring levels of fluorescence over time, the group showed that, as desired, individual cells exhibited oscillations with an average period of 150 minutes (three times that of the typical cell cycle, although there was some difficulty in synchronizing oscillations over an entire *E. coli* population).

Research efforts such as these in which novel biological networks are designed from a model will eventually converge and couple with efforts to study existing biological systems. In this scenario, one would not only possess predictive models but would also have the power to use these models to re-engineer cells. A range of potential modifications could be rigorously evaluated through model simulations then later verified directly in the biological system. This dualistic approach is one of the "holy grails" of biology and medicine in which a predictive model of a complex disease pathway is used to design and test cellular modifications that can, ultimately, ameliorate the disease response.

SUMMARY

In conclusion, what are the most striking challenges arising from systems biology?

- The inclusion of nongenetic molecules, small and large, into the systems picture. The cell contains thousands of distinct metabolic substrates and other

small and large molecules, a variety of which exert influence on gene expression (through direct interactions with proteins or DNA) and on allosteric enzymes. Methods to systematically measure levels of such molecules would be of enormous benefit.

- Further development of theoretical frameworks and tools for integrating the various levels of biological information, displaying them graphically, and, finally, mathematical modeling and simulation of biological systems.
- Systematic and detailed annotations of information in the public databases. As the databases become more advanced, so will our models of cellular processes. For instance, rather than simply provide a list of interactions, physical-interaction databases should specify if, and how, each interaction affects cell state.
- Education of cross-disciplinary scientists. Cross-disciplinary scientists should have a deep understanding of biology (their contributions will be proportional to their understanding). We believe that the solution to this problem is to teach biology as an informational science. This approach is conceptual, hierarchical, economical, and in the future mainstream of education in biology.
- The integration of technology, biology, and computation. Integration (also see points discussed below) presents one of the most striking challenges for systems biology, both for academia and industry.

In addition to these general challenges, the development and practice of systems biology involves a number of requirements that will pose particular difficulties for academic institutions. Among these requirements are:

- high-throughput facilities for global technologies, such as DNA sequencing, DNA arrays, genotyping, proteomics, and protein interactions;
- integration of different levels of biological information generated at each of these facilities;
- the integration of excellent biology with a strong computational infrastructure and analytic tools;
- the formation of teams of biologists, technologists, and computational scientists to attack the iterative challenges of systems biology;
- the integration of discovery- and hypothesis-driven science; and
- the development of diverse partnerships with academia and industry. Academia will provide new systems for exploration; industry will provide new technologies and resources to take on demanding problems.

These six challenges pose difficulties for most academic institutions—if they are to provide their biologists the opportunities to practice systems biology. Most academic institutions do not have the diverse scientific talent, funds, or space to initiate a self-sustaining systems biology effort. Although these resources may be available through industrial partnerships, it is difficult for academic laboratories

to form such partnerships, particularly when intellectual property is involved. Owing to severe salary constrains, recruiting scientists from high-demand fields, such as bioinformatics, proteomics, and engineering, can be equally problematic. Within an academic institution, individual departments often provide barriers for cross-disciplinary science—in geographical isolation, in the training of students, and in the constraints of what is expected from faculty (research projects, teaching, etc.). Finally, the demands of tenure force young faculty to carry out safe projects independently—at potentially the most creative phase of their careers—and penalize them for research performed as part of a team. Of course, some academic institutions may circumvent many of these limitations by the creation of special centers. In other cases, independent, nonprofit research institutes, such as our Institute for Systems Biology, can be fashioned to take advantage of these opportunities.

Regardless of these initial hurdles, it is clear that systems biology will necessarily be a leading academic and industrial thrust in the years to come. Its impact on medicine, agriculture, biological energy production, and many other areas will make biotechnology a powerful driving force as we move into the century of biology.

LITERATURE CITED

1. Aach J, Rindone W, Church GM. 2000. Systematic management and analysis of yeast gene expression data. *Genome Res.* 10:431–45
2. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–95
3. Akutsu T, Kuhara S, Maruyama O, Miyano S. 1998. Identification of gene regulatory networks by strategic gene disruptions and gene overexpressions. *Proc. ACM-SIAM Symp. Discrete Algorithms, 9th*. New York: ACM Press
4. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, et al. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* 96:6745–50
5. Alon U, Surette MG, Barkai N, Leibler S. 1999. Robustness in bacterial chemotaxis. *Nature* 397:168–71
6. Andrade MA, Valencia A. 1998. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* 14:600–7
7. Arkin A, Ross J. 1995. Statistical construction of chemical reaction mechanisms from measured time series. *J. Phys. Chem.* 99:970
8. Arkin A, Ross J, McAdams HH. 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics* 149:1633–48
9. Arnone MI, Davidson EH. 1997. The hardwiring of development: organization and function of genomic regulatory systems. *Development* 124:1851–64
10. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, et al. 2001. BIND—The biomolecular interaction network database. *Nucleic Acids Res.* 29: 242–45

11. Barkai N, Leibler S. 1997. Robustness in simple biochemical networks. *Nature* 387:913–17
12. Barkai N, Leibler S. 2000. Circadian clocks limited by noise. *Nature* 403:267–68
13. Deleted in proof
14. Bass BL. 2000. Double-stranded RNA as a template for gene silencing. *Cell* 101: 235–38
15. Baxevanis AD. 2001. The molecular biology database collection: an updated compilation of biological database resources. *Nucleic Acids Res.* 29:1–10
16. Bhalla US, Iyengar R. 1999. Emergent properties of networks of biological signaling pathways. *Science* 283:381–87
17. Blair DF. 1995. How bacteria sense and swim. *Annu. Rev. Microbiol.* 49:489–522
18. Blaschke C, Andrade MA, Ouzounis C, Valencia A. 1999. Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc. Int. Conf. Intell. Syst. Mol. Biol., 7th, Heidelberg, 1999*, pp. 60–67. Menlo Park, CA: AAAI Press
19. Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277:1453–74
20. Bray D, Levin MD, Morton-Firth CJ. 1998. Receptor clustering as a cellular mechanism to control sensitivity. *Nature* 393:85–88
21. Brazma A, Vilo J. 2000. Gene expression data analysis. *FEBS Lett.* 480:17–24
22. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* 18:630–34
23. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, et al. 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* 97:262–67
24. Burstein Z. 1995. A network model of developmental gene hierarchy. *J. Theor. Biol.* 174:1–11
25. Butte AJ, Kohane IS. 2000. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.* 5:418–29
26. Cameron RA, Mahairas G, Rast JP, Martinez P, Biondi TR, et al. 2000. A sea urchin genome project: sequence scan, virtual map, and additional resources. *Proc. Natl. Acad. Sci. USA* 97:9514–18
27. Chen Y, Dougherty E, Bittner M. 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* 2:364–74
28. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2:65–73
29. Chu S, Derisi J, Eisen M, Mulholland J, Botstein D, et al. 1998. The transcriptional program of sporulation in budding yeast. *Science* 282:699–705
30. Claverie JM. 1999. Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* 8:1821–32
31. Cluzel P, Surette M, Leibler S. 2000. An ultrasensitive bacterial motor revealed by monitoring signaling proteins in single cells. *Science* 287:1652–55
32. Cohen BA, Mitra RD, Hughes JD, Church GM. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* 26:183–86
33. Craven M, Kumlien J. 1999. Constructing biological knowledge bases by extracting information from text sources. *Proc. Int. Conf. Intell. Syst. Mol. Biol., 7th, Heidelberg, 1999*, pp. 77–86. Menlo Park, CA: AAAI Press
34. Davidson EH. 2001. *Genomic Regulatory Systems: Development and Evolution.* San Diego, CA: Academic
35. Derisi JL, Iyer VR, Brown PO. 1997. Exploring the metabolic and genetic control

of gene expression on a genomic scale. *Science* 278:680–86

36. Dysvik B, Jonassen I. 2001. Exploring gene expression data using Java. *Bioinformatics* 17:369–70
37. Edwards JS, Palsson BO. 2000. Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnol. Prog.* 16: 927–39
38. Eilbeck K, Brass A, Paton N, Hodgman C. 1999. INTERACT: an object oriented protein-protein interaction database. *Proc. Int. Conf. Intell. Syst. Mol. Biol., 7th, Heidelberg*, 1999, pp. 87–94. Menlo Park: AAAI Press
39. Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863–68
40. Elowitz MB, Leibler S. 2000. A synthetic oscillatory network of transcriptional regulators. *Nature* 403:335–38
41. Endy D, You L, Yin J, Molineux IJ. 2000. Computation, prediction, and experimental tests of fitness for bacteriophage T7 mutants with permuted genomes. *Proc. Natl. Acad. Sci. USA* 97:5375–80
42. Eng JK, McCormack AL, Yates JRI. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass. Spectrom.* 5:976–89
43. Ermolaeva O, Rastogi M, Pruitt KD, Schuler GD, Bittner ML, et al. 1998. Data management and analysis for gene expression arrays. *Nat. Genet.* 20:19–23
44. Faria M, Spiller DG, Dubertret C, Nelson JS, White MRH, et al. 2001. Phosphoramidate oligonucleotides as potent antisense molecules in cells and in vivo. *Nat. Biotechnol.* 19:40–44
45. Feinberg M. 1995. *The Chemical Reaction Network Toolbox*, Version 1.02, ftp.che.rochester.edu/pub/feinberg/. Rochester, NY
46. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, et al. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391:806–11
47. Fodor SPA, Dower WJ, Ekins RP, Flynn GC, Houghten RA, et al. 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* 251:767–73
48. Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, et al. 2000. Functional genomic analysis of cell division in *C. elegans* chromosome 1 by systematic RNA inteference. *Nature* 408:325–30
49. Friedman N, Linial M, Nachman I, Pe'er D. 2000. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7:601–20
50. Galitski T, Saldanha AJ, Styles CA, Lander ES, Fink GR. 1999. Ploidy regulation of gene expression. *Science* 285:251–54
51. Gardner TS, Cantor CR, Collins JJ. 2000. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* 403:339–42
52. Gilbert W. 1991. Towards a paradigm shift in biology. *Nature* 349:99
53. Gillespie DT. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Physics* 22:403–34
54. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, et al. 1996. Life with 6000 genes. *Science* 274:546, 563–67
55. Gonczy P, Echeverri G, Oegema K, Coulson A, Jones SJ, et al. 2000. Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature* 408:331–36
56. Grebe TW, Stock J. 1998. Bacterial chemotaxis: the five sensors of a bacterium. *Curr. Biol.* 8:R154–57
57. Greller LD, Tobin FL. 1999. Detecting selective expression of genes and proteins. *Genome Res.* 9:282–96
58. Guo Z, Guilfoyle RA, Thiel AJ, Wang R, Smith LM. 1994. Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports. *Nucleic Acids Res.* 22: 5456–65

59. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, et al. 1999. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17:994–99
60. Harrington MG, Coffman JA, Calzone FJ, Hood LE, Britten RJ, et al. 1992. Complexity of sea urchin embryo nuclear proteins that contain basic domains. *Proc. Natl. Acad. Sci. USA* 89:6252–56
61. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. 2001. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.* 6:422–33
62. Hartuv E, Schmitt A, Lange J, Meier-Ewert S, Lehrach H, et al. 1998. *An algorithm for clustering cDNAs for gene expression analysis*. Presented at Hum. Genome Meet. 1998. Torino, Italy
63. Hawkins V, Doll D, Bumgarner R, Smith T, Abajian C, et al. 1998. PEDB: the Prostate Expression Database. *Nucleic Acids Res.* 27:204–8
64. Heyer LJ, Kruglyak S, Yooseph S. 1999. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.* 9:1106–15
65. Hilsenbeck SG, Friedrichs WE, Schiff R, O'Connell P, Hansen RK, et al. 1999. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J. Natl. Cancer Inst.* 91:453–59
66. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, et al. 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* 19:342–47
67. Hunkapiller MW. 1991. Advances in DNA sequencing technology. *Curr. Opin. Genet. Dev.* 1:88–92
68. Ideker T, Thorsson V, Ranish J, Christmas R, Buhler J, et al. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292:929–934
69. Ideker T, Thorsson V, Siegel A, Hood L. 2000. Testing for differentially expressed genes by maximum likelihood analysis of microarray data. *J. Comput. Biol.* 7:805–17
70. Ideker TE, Thorsson V, Karp RM. 2000. Discovery of regulatory interactions through perturbation: inference and experimental design. *Pac. Symp. Biocomput.* 5:305–16
71. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, et al. 1999. The transcriptional program in the response of human fibroblasts to serum. *Science* 283:83–87
72. Johnston M, Carlson M. Regulation of carbon and phosphate utilization. 1992. *The Molecular and Cellular Biology of the Yeast Saccharomyces*, ed. E Jones, J Pringle, J Broach, Vol. 2. New York: Cold Spring Harbor Lab. Press
73. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, et al. 2000. The EcoCyc and MetaCyc databases. *Nucleic Acids Res.* 28:56–59
74. Kauffman SA. 1993. *The Origins of Order: Self Organization and Selection in Evolution*. New York: Oxford Univ. Press
75. Kohn KW. 1999. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell* 10:2703–34
76. Lai K, Elsas LJ. 2000. Overexpression of human UDP-glucose pyrophosphorylase rescues galactose-1-phosphate uridyltransferase-deficient yeast. *Biochem. Biophys. Res. Commun.* 271:392–400
77. Lander ES. 1996. The new genomics: global views of biology. *Science* 274:536–39
78. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
79. Lee CK, Klopp RG, Weindruch R, Prolla TA. 1999. Gene expression profile of aging and its retardation by caloric restriction. *Science* 285:1390–93
80. Liang S, Fuhrman S, Somogyi S. 1998. REVEAL, a general reverse engineering

algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.* 3:18–29

81. Lin X, Kaul S, Rounsley S, Shea TP, Benito MI, et al. 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* 402:761–68
82. Lohr D, Venkov P, Zlatanova J. 1995. Transcriptional regulation in the yeast GAL gene family: a complex genetic network. *FASEB J.* 9:777–87
83. Longtine MS, McKenzie A, Demarini DJ, Shah NG, Wach A, et al. 1998. Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. *Yeast* 14:953–61
84. Manduchi E, Grant GR, McKenzie SE, Overton GC, Surrey S, et al. 2000. Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics* 16:685–98
85. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* 402:83–86
86. Marnellos G, Mjolsness E. 1998. A gene network approach to modeling early neurogenesis in *Drosophila*. *Pac. Symp. Biocomput.* 3:30–41
87. McAdams HH, Arkin A. 1997. Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA* 94:814–19
88. McAdams HH, Shapiro L. 1995. Circuit simulation of genetic networks. *Science* 269:650–56
89. McGuire AM, Hughes JD, Church GM. 2000. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* 10:744–57
90. Mendes P. 1997. Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* 22:361–63

90a. Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, et al. 1999. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 27:44–48

91. Nasevicius A, Ekker SC. 2000. Effective targeted gene *knockdown* in zebrafish. *Nat. Genet.* 26:216–20
92. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 27:29–34
93. Pillage L. 1994. *Electronic Circuit and System Simulation Methods*. New York: McGraw Hill
94. Pirson I, Fortemaison N, Jacobs C, Dremier S, Dumont JE, et al. 2000. The visual display of regulatory information and networks. *Trends Cell Biol.* 10:404–8
95. Pulavarthi P, Chiang R, Altman RB. 2000. Generating interactive molecular documentaries using a library of graphical actions. *Pac. Symp. Biocomput.* 5:266–77
96. Ringwald M, Eppig JT, Kadin JA, Richardson JE. 2000. GXD: a Gene Expression Database for the laboratory mouse: current status and recent enhancements. The Gene Expression Database Group. *Nucleic Acids Res.* 28:115–19
97. Roth FP, Hughes JD, Estep PW, Church GM. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* 16:939–45
98. Rowen L, Lasky S, Hood L. 1999. Deciphering genomes through automated large-scale sequencing. In *Methods in Microbiology*, ed. AG Craig, JD Hoheisel, pp. 155–91. San Diego, CA: Academic
99. Salamonsen W, Mok KY, Kolatkar P, Subbiah S. 1999. BioJAKE: a tool for the creation, visualization and manipulation of metabolic pathways. *Pac. Symp. Biocomput.* 4:392–400
100. Schaff J, Fink CC, Slepchenko B, Carson JH, Loew LM. 1997. A general computational framework for modeling cellular structure and function. *Biophys. J.* 73:1135–46
101. Schaff JC, Slepchenko BM, Loew LM.

2000. Physiological modeling with virtual cell framework. *Methods Enzymol.* 321:1–23

102. Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–70
103. Schilling CH, Letscher D, Palsson BO. 2000. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* 203:229–48
104. Schilling CH, Palsson BO. 1998. The underlying pathway structure of biochemical reaction networks. *Proc. Natl. Acad. Sci. USA* 95:4193–98
105. Schwikowski B, Uetz P, Fields S. 2000. A network of protein-protein interactions in yeast. *Nat. Biotechnol.* 181257–61
106. Selkov E Jr, Grechkin Y, Mikhailova N, Selkov E. 1998. MPW: the Metabolic Pathways Database. *Nucleic Acids Res.* 26:43–45
107. Shelton DN, Chang E, Whittier PS, Choi D, Funk WD. 1999. Microarray analysis of replicative senescence. *Curr. Biol.* 9:939–45
108. Sherlock G. 2000. Analysis of large-scale gene expression data. *Curr. Opin. Immunol.* 12:201–5
109. Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW. 1996. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet.* 14:450–56
110. Shoemaker DD, Schadt EE, Armour CD, He YD, Garrett-Engele P, et al. 2001. Experimental annotation of the human genome using microarray technology. *Nature* 409:922–27
111. Sinha S, Tompa M. 2000. A statistical method for finding transcription factor binding sites. *Proc. Int. Conf. Intell. Syst. Mol. Biol., 8th, La Jolla, California, 2000,* 8:344–54. Menlo Park: AAAI Press
112. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, et al. 1986. Fluorescence detection in automated DNA sequence analysis. *Nature* 321:674–79
113. Somogyi R, Sniegoski C. 1996. Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. *Complexity* 1:45–63
114. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, et al. 1998. Comprehensive identification of cell cycle–regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9:3273–97
115. Spiro PA, Parkinson JS, Othmer HG. 1997. A model of excitation and adaptation in bacterial chemotaxis. *Proc. Natl. Acad. Sci. USA* 94:7263–68
116. Stoeckert CJ Jr, Salas F, Brunk B, Overton GC. 1999. EpoDB: a prototype database for the analysis of genes expressed during vertebrate erythropoiesis. *Nucleic Acids Res.* 27:200–3
117. Tamames J, Ouzounis C, Casari G, Sander C, Valencia A. 1998. EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics* 14:542–43
118. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, et al. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* 96:2907–12
119. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. 1999. Systematic determination of genetic network architecture. *Nat. Genet.* 22:281–85
120. Termonia Y, Ross J. 1998. Oscillations and control features in glycolysis: numerical analysis of a comprehensive model. *Proc. Natl. Acad. Sci. USA* 78:2952–56
121. The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282:2012–18
122. Thomas J, Milward D, Ouzounis C, Pulman S, Carroll M. 2000. Automatic

extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.* 5:541–52

123. Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y, et al. 1999. E-CELL: software environment for whole-cell simulation. *Bioinformatics* 15:72–84
124. Turing AM. 1952. The chemical basis of morphogenesis. *Proc. R. Philos. Soc. B* 237:37–72
125. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51
126. Von Dassow G, Meir E, Munro EM, Odell GM. 2000. The segment polarity network is a robust developmental module. *Nature* 406:188–92
127. Wach A, Brachat A, Alberti-Segui C, Rebischung C, Philippsen P. 1997. Heterologous HIS3 marker and GFP reporter modules for PCR-targeting in *Saccharomyces cerevisiae*. *Yeast* 13:1065–75
128. Walhout AJ, Temple GF, Brasch MA, Hartley JL, Lorson MA, et al. 2000. GATEWAY recombinatorial cloning: application to the cloning of large numbers of open reading frames or ORFs. *Methods Enzymol.* 328:575–92
129. Wang K, Gan L, Jeffrey E, Gayle M, Gown AM, et al. 1999. Monitoring gene expression profile changes in ovarian carcinomas using a cDNA microarray. *Gene* 229:101–8
130. Weaver DC, Workman CT, Stormo GD. 1999. Modeling regulatory networks with weight matrices. *Pac. Symp. Biocomput.* 4:112–23
131. Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, et al. 1998. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA* 95:334–39
132. Weng G, Bhalla US, Iyengar R. 1999. Complexity in biological signaling systems. *Science* 284:92–96
133. Wingender E, Chen X, Hehl R, Karas H, Liebich I, et al. 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* 28:316–19
134. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, et al. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285:901–6
135. Wuensche A. 1995. Discrete Dynamics Laboratory (DDLAB). http://www.ddlab.com/. Santa Fe, NM: Discrete Dynamics, Inc.
136. Wuensche A. 1998. Genomic regulation modeled as a network with basins of attraction. *Pac. Symp. Biocomput.* 3:89–102
137. Xenarios I, Fernandez E, Salwinski L, Duan XJ, Thompson MJ, et al. 2001. DIP: the database of interacting proteins. *Update. Nucleic Acids Res.* 29:239–41
138. Yi TM, Huang Y, Simon MI, Doyle J. 2000. Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc. Natl. Acad. Sci. USA* 97:4649–53
139. Yuh CH, Bolouri H, Davidson EH. 1998. Genomic *cis*-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 279:1896–902
140. Yuh CH, Bolouri H, Davidson EH. 2001. *Cis*-regulatory logic in the *endo16* gene: switching from a specification to a differentiation mode of control. *Development* 128:617–29
141. Zhang MQ. 1999. Large-scale gene expression data analysis: a new challenge to computational biologists. *Genome Res.* 9:681–88
142. Zhu J, Zhang MQ. 1999. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* 15:607–11

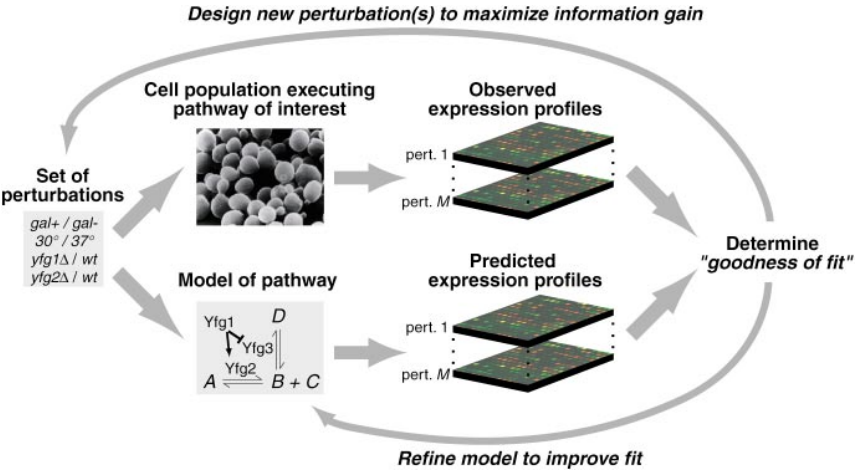
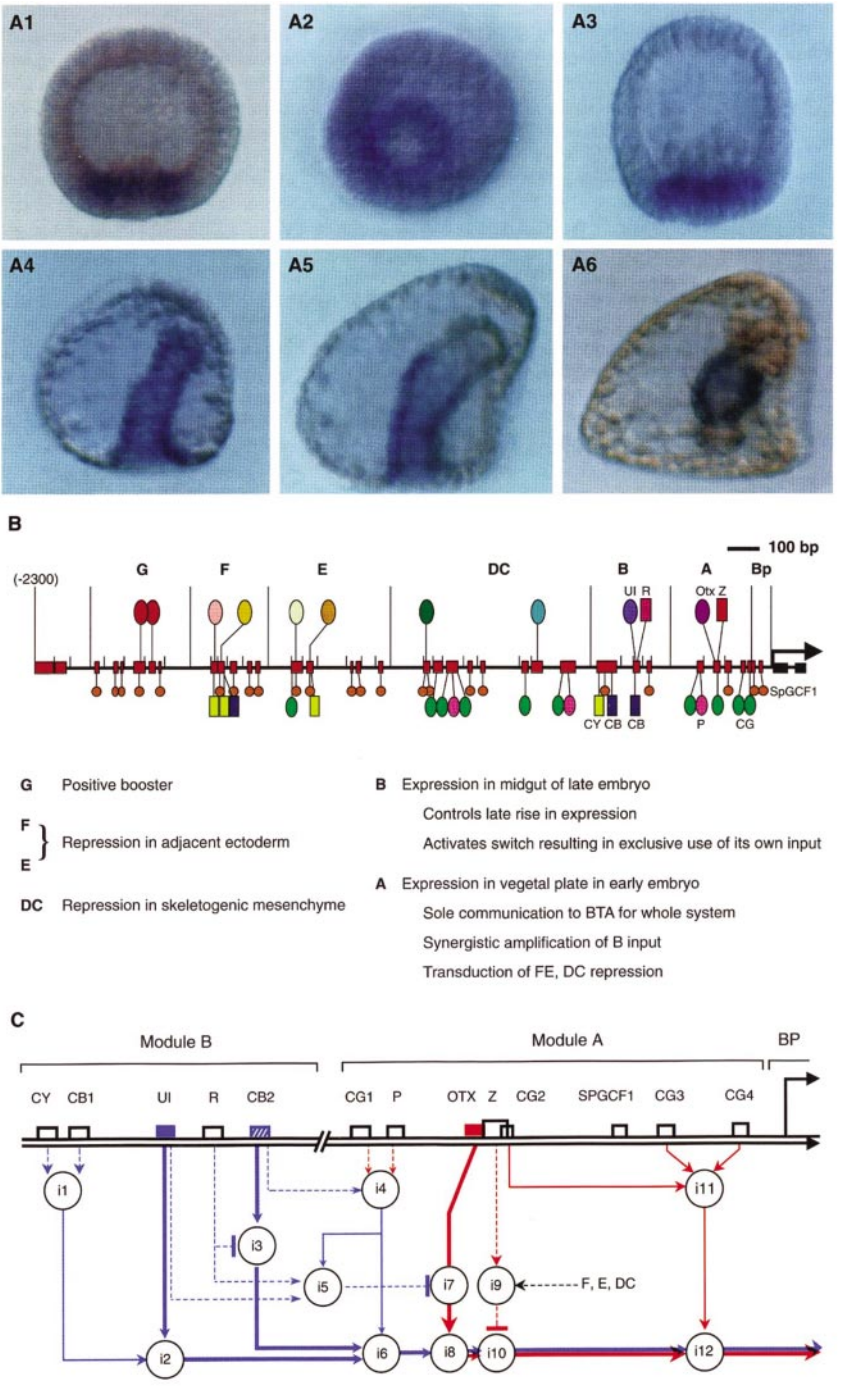


Figure 1 Overview of the systems biology approach, involving pathway verification and refinement through systematic, successive perturbations. The pathway of interest is perturbed genetically by gene deletion or overexpression and/or biologically by modulation of metabolite levels, temperature, or other pathway components. Gene expression profiles measured in response to each perturbation, obtained using microarrays or related technologies, are compared to those predicted by a model of the pathway mechanism. Perturbations are initially selected to target known pathway components and are thereafter chosen to distinguish between alternative models that are consistent with the present set of observations. All aspects of the process are amenable to automation (laboratory or computational), including model refinement and choice of perturbations.

Figure 2 (see figure on next page) A *cis*-regulatory network at the sea urchin *endo16* gene. (*A1–A6*) A developmental time course of *endo16* in situ expression patterns in sea urchin. The gene is expressed early in the vegetal plate (*A1*), although not in the early blastula (*A2*) nor at ingression of skeletogenic cells (*A3*). After gastrulation, *endo16* expression is observed throughout the archenteron (*A4*). Subsequently, expression is shut down in the foregut, the secondary mesenchyme (*A5*), and the hindgut (*A6*) but remains in the midgut. (*B*) A map of protein-DNA interactions in the 2300bp *endo16 cis*-regulatory sequences. Different colors represent different proteins. Repeated sites are marked with symbols below the line. Distinct modules with identifiable roles in *endo16* expression patterns are indicated and annotated. (*C*) Control logic model for modules A and B. Binding sites are indicated above the line. Below the line, circles indicate logical operations. Effects exerted by module A are indicated in red; those of module B are in blue. Interactions that can be modeled as boolean inputs are indicated as dashed lines, scalars as thin solid lines, and time-dependent quantitative inputs as heavy lines. Outputs indicated with an arrowhead exert positive effects; perpendicular bars represent negative effects. As an example, CY and CB1 interact synergistically to promote the output of the module B spatial-temporal control element U1. Originals reprinted with permission from Davidson (34). Copyright 2001 Academic Press.



See legend on previous page

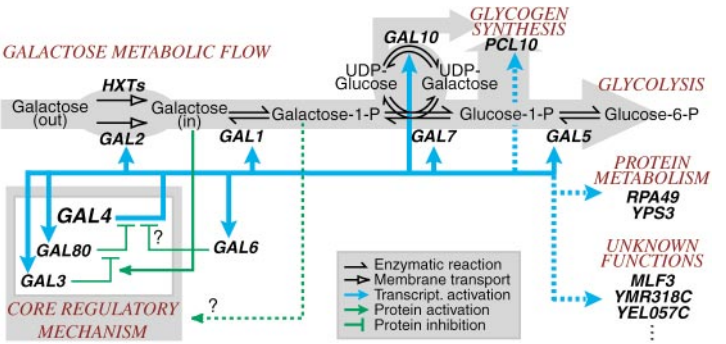


Figure 3 The galactose system. Yeast metabolize galactose through a series of steps involving the *GAL2* transporter and enzymes produced by *GAL1*, *7*, *10*, and *5*. These genes are transcriptionally regulated by a mechanism consisting primarily of *GAL4*, *80*, and *3*. *GAL6* produces another regulatory factor thought to repress the GAL enzymes in a manner similar to *GAL80*. Dotted interactions denote model refinements supported by our systems approach. Reprinted with permission from Ideker et al. (68). Copyright 2001 American Association for the Advancement of Science.

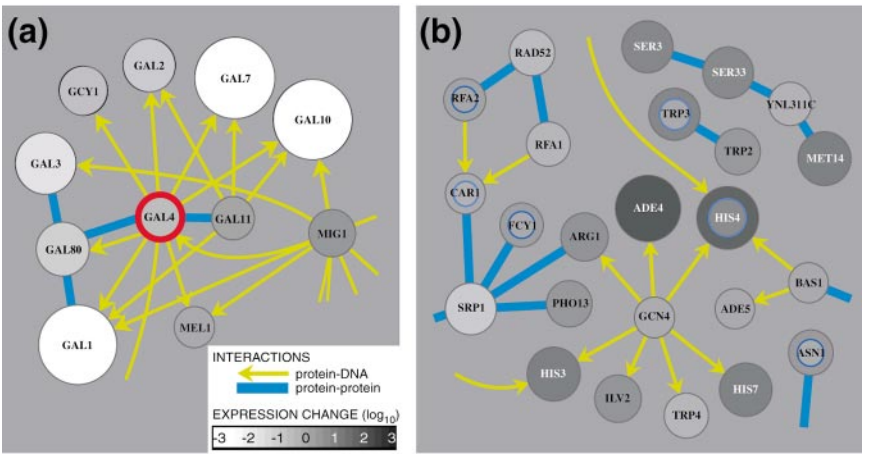


Figure 4 Sample regions of the integrated physical-interaction network, corresponding to (*a*) galactose utilization and (*b*) amino-acid biosynthesis. Each node represents a gene, a yellow arrow directed from one node to another represents a protein-DNA interaction, and a blue line between nodes represents a protein-protein interaction. The intensity of each node indicates the change in mRNA expression of the corresponding gene, with medium-gray representing no change and darker or lighter shades representing an increase or decrease in expression, respectively (node diameter also scales with the magnitude of change). Nodes for which protein data are also available (panel *b*) contain two distinct regions: an outer circle, or ring, representing the change in mRNA expression and an inner circle representing the change in protein expression. To signify that the expression level of *GAL4* has been perturbed by external means (panel *a*), it is highlighted with a red border. Reprinted with permission from Ideker et al. (68). Copyright 2001 American Association for the Advancement of Science.

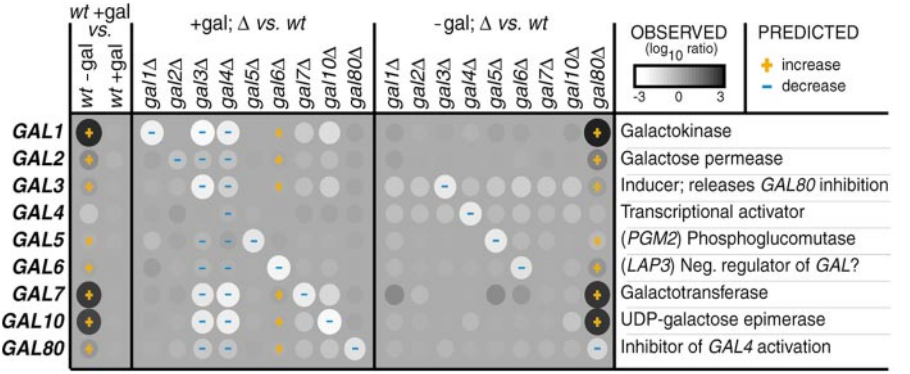


Figure 5 Matrix of observed vs. predicted gene-expression responses. DNA microarrays were used to measure the mRNA-expression responses of yeast cells undergoing steady-state growth in the presence of each of 20 perturbations to the galactose-utilization pathway. Each spot in the matrix represents the quantitative change in expression observed for a GAL gene (*rows*) in one of the perturbations (*columns*), according to the intensity scale shown at upper right. Superimposed on each spot are the corresponding (qualitative) predictions of the network model as shown in Figure 3, with the symbols + vs. − indicating a predicted increase vs. decrease in expression, respectively. Unannotated spots represent genes for which no expression change is predicted. Reprinted with permission from Ideker et al. (68). Copyright 2001 American Association for the Advancement of Science.

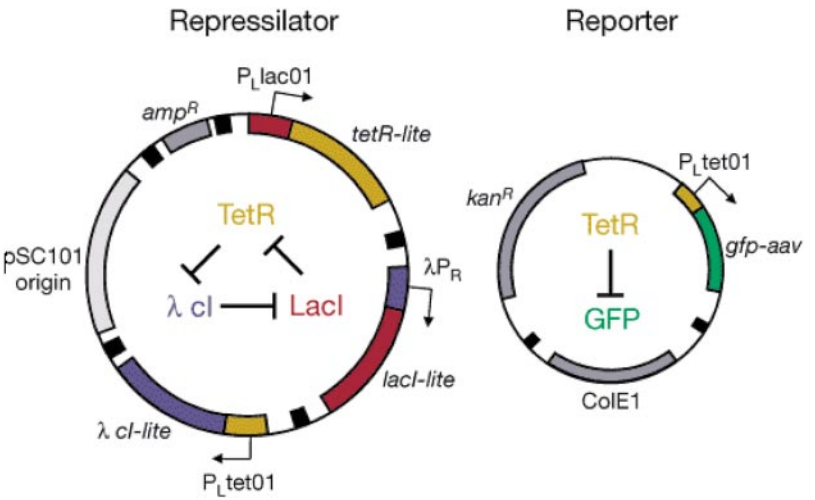


Figure 7 A synthetic, three-protein oscillatory network. The "repressilator" plasmid encodes three synthetic proteins, each fused to a heterologous promoter such that one protein represses the transcription of the next in a closed negative-feedback loop (TetR represses cI, cI represses LacI, and LacI represses TetR). A reporter plasmid is used to track oscillations in TetR concentration: it contains the TetR-binding sequence upstream of the gene encoding for green fluorescent protein. Reprinted with permission from Elowitz & Leibler (40). Copyright 2001 Nature.

Annual Review of Genomics and Human Genetics,
Volume 2

CONTENTS