Annu. Rev. Genomics Hum. Genet. 2002. 3:341–69
doi: 10.1146/annurev.genom.3.030502.111004

First published online as a Review in Advance on June 6, 2002

GENETIC "CODE": Representations and Dynamical Models of Genetic Components and Networks

Alex Gilman[1] and Adam P. Arkin[1,2,3]
[1]Howard Hughes Medical Institute, [2]Departments of Bioengineering and Chemistry, University of California, Berkeley, [3]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720; email: agilman@lbl.gov, aparkin@lbl.gov

Key Words mathematical models, gene expression, regulation, transcription, translation

■ **Abstract** Dynamical modeling of biological systems is becoming increasingly widespread as people attempt to grasp biological phenomena in their full complexity and make sense of an accelerating stream of experimental data. We review a number of recent modeling studies that focus on systems specifically involving gene expression and regulation. These systems include bacterial metabolic operons and phase-variable piliation, bacteriophages T7 and λ, and interacting networks of eukaryotic developmental genes. A wide range of conceptual and mathematical representations of genetic components and phenomena appears in these works. We discuss these representations in depth and give an overview of the tools currently available for creating and exploring dynamical models. We argue that for modeling to realize its full potential as a mainstream biological research technique the tools must become more general and flexible, and formal, standardized representations of biological knowledge and data must be developed.

INTRODUCTION

The mathematical and computational modeling of biological systems is a subject of increasingly intense interest (see Appendix: A Brief Guide to Recent Reviews). The accelerating growth of biological knowledge, in concert with a growing appreciation of the spatial and temporal complexity of events within cells, tissues, organs, and populations, threatens to overwhelm people's capacity to integrate, understand, and reason about biology. The construction, analysis, and simulation of formal models is a useful way to manage such problems. Metabolism, signal transduction, genetic regulation, circadian rhythms, and various aspects of neurobiology are just a subset of phenomena that have been treated by modeling. In this paper we explore some recent modeling studies on systems that specifically include genetic components: genes and the concomitant phenomena involved in

their expression and regulation. Our focus is restricted to studies where the models are meant as a direct representation of the systems being investigated and, as much as possible, are compared to experimental data. Particular systems include the bacteriophage λ lysis/lysogeny decision, the bacteriophage T7 complete life cycle, the *lac* and *trp* operons and the response to phosphate starvation in *Escherichia coli*, and type-1 piliation in *E. coli*. There is a pronounced prokaryotic bias in this list, which is largely a reflection of the state of the affairs. The details of eukaryotic gene expression are still rather poorly understood, and so the few models that treat them are rather simple and general. We do not review these in depth but do briefly discuss the modeling approaches taken. We do examine models of networks of interacting genes involved in several eukaryotic developmental processes; these are interesting for treating complex phenomena without involving mechanistic detail. After reviewing the scope of the various prokaryotic and eukaryotic models and the biological questions they have been used to answer, we examine in detail how the genetic components and phenomena addressed by the models are represented conceptually and mathematically.

The detailed models we examine demonstrate the principal strength of modeling: It is a means to formulate all available knowledge about a system in as precise a manner as possible. In so doing, it allows a number of complex questions to be posed: (*a*) Is the available knowledge self-consistent? This question can often be answered during the formulation of a model. Contradictory information about the system needs to be resolved before a model can be completed. (*b*) Is the available knowledge of a system's components and their interactions sufficient to account for all the system's known behavior? If not, a model can often focus attention on areas that would most benefit from further investigation. (*c*) What are the consequences of various manipulations to the system (e.g., knocking out genes, modifying promoters, modulating various biochemical reactions with pharmaceuticals, etc.)? A model can provide predictions that are at the same level of detail as the model's formulation. A highly detailed mechanistic model can in principle predict a wide range of quantities, from the detailed timecourses of protein and nucleic acid levels, through the activation states of genes, all the way to a final phenotypic outcome. To date, questions like these have most often been treated informally, using drawn diagrams, conceptual thought, and logical argument. An excursion to the limits of such faculties, however, may be had with a glance at, say, a chart of metabolism or any of the maps at the Alliance for Cellular Signaling (http://www.afcs.org), or the contemplation of the mechanistic bases of quantitative trait loci, mutations with partial penetrance, "susceptibility" phenotypes, and mutants that are not gain/loss of function but disturbances of control of function. Proper intellectual treatment of such things will require increasingly complex hypotheses about how the systems we study work, and the more complex a hypothesis, the more difficult it is to check for internal consistency, to assess for explanatory power, and to reason from about consequences using unaided human thought.

It thus seems inevitable that computational models, together with formal representations of knowledge and data to support them, will play a greater and greater

role in mainstream biology. There are, however, formidable difficulties in creating and, for the nonspecialist, evaluating a model. We briefly discuss these difficulties in the final sections and suggest that the most useful thing to be done in overcoming them is to dramatically improve the tools available to the biological community for handling models and data.

THE MODELS

The first examples we review, the *trp* and *lac* operons and the *pho* regulon in *E. coli*, are important metabolic systems that have found extensive use as heterologous expression systems. Together they illustrate a number of important regulatory mechanisms, and questions about their control have commercial significance. We then review two bacteriophage systems, T7 and λ. The T7 model is the first "whole genome, whole life-cycle" model of an organism, and it addresses certain issues in genome organization and pharmaceutical strategies. λ, a temperate phage, has a choice of fates on infecting its host. The models explore how this fate is determined. Next are type-1 fimbriae, which play an important role in the pathogenicity of certain *E. coli* strains and, being phase variable, contribute to the heterogeneity of clonal *E. coli* cultures. Modeling is used to explore how the phase variation is regulated and the part it plays in infection. Finally we review some models of interacting gene networks involved in eukaryotic development.

The *trp* Operon

Santillán & Mackey study the *trp* operon of *E. coli* (67). This operon codes for five biosynthetic enzymes that convert chorismate to tryptophan. The biosynthesis of tryptophan in *E. coli* is subject to three modes of control: end-product inhibition by trp of the enzyme mediating the first step of the conversion, trp-dependent repression of the operon, and transcriptional attenuation. The model used by Santillán & Mackey is the first to treat all three of these modes. The authors carry out simulations and compare their results with previously reported experiments in which biosynthetic enzyme activity was followed after cultures of *E. coli* grown in rich medium to stationary phase were shifted to a minimal medium lacking trp. Two mutants, one with generally lowered transcriptional efficiency and one with enhanced transcriptional termination, are studied along with the wild type. The simulation results give only a fair match to experimental values. One qualitative feature, a transient overshoot of enzyme activity upon nutrient shift in the wild type, seems to be missed entirely. Two significant points must be noted here. First, the experimental results to be reproduced are widely spaced and appear to be quite variable, making meaningful detailed comparison difficult. Second, all of the parameters of the model, at least for describing the wild type, are estimated by the authors from the biochemical literature. No fitting to the target results is performed. This is not true for the mutants, however. Each mutant is modeled

by changing a single parameter assumed to be the key for the mutant phenotype. The new values are set by trial and error, presumably with reference to the mutant experimental results. Hence, an informal kind of fitting is performed, and the qualitative agreement of simulation with experiment does seem somewhat better for the mutant cases.

The *lac* Operon

Expression of the *lac* operon in *E. coli* is induced by intracellular lactose, which, after isomerization to allolactose, binds to the LacI repressor protein, disrupting its interaction with the operator site in the promoter. The operon is also involved in the global response to glucose. As glucose levels fall, cAMP builds up; the complex of cAMP and catabolite activating protein (CAP) stimulate transcription from *lac* and other metabolic operons. Glucose also counteracts the derepressive effects of lactose by interfering with its uptake into the cell.

Keasling and coworkers have produced two models of this operon (16, 84). The earlier one (84), which treats all three regulatory phenomena, successfully reproduced general experimental observations: In a medium containing both glucose and lactose, the glucose was consumed during an initial period of exponential growth. This was followed by a period of slow (diauxic) growth before exponential growth was renewed through lactose consumption. The model allowed the investigators to follow a large number of separate variables over time, including mRNA levels for the LacI repressor and lactose utilization enzymes. Repression and induction of the operon was clearly visible. However, no direct comparisons to experimentally observed kinetic profiles were done.

The other *lac* operon model from this group (16) does not include the effects of glucose or the metabolic consumption of the inducer. It thus more closely represents the conditions of a heterologous expression system driven by an artificial nonmetabolizable inducer [for instance, isopropyl β-D-1-thiogalactopyranoside (IPTG)]. The autocatalytic nature of the operon is retained: Induction of the operon leads to production of more of the inducer's transporter enzyme (LacY). The model also differs from the earlier one in explicitly treating the stochastic nature of gene expression at the level of single cells. Simulations showed that individual cells respond to inducer in an all-or-none manner, consistent with earlier experimental studies on single cells. The graded dose response of a bulk culture to inducer was reproduced when a large number of single-cell simulations were aggregated. Further, the model reproduced the phenomenon of "maintenance induction," that is, maintaining high expression of the operon after shifting an induced culture to a lower concentration of inducer.

The *Pho* Regulon

Another paper from the Keasling group (80) gives a model of part of the phosphate starvation response of *E. coli*. In this response, low extracellular phosphate concentration is signaled to a transcription factor (PhoB) that activates transcription

of a number of other genes, including the genes for alkaline phosphatase and components of the Pst phosphate transporter. The model is similar in character to the earlier of the lac operon models. Simulations reproduced the sharp induction of phosphate-responsive genes at a precisely defined phosphate concentration. The model also allowed the authors to examine the role of phosphate transport on the expression of the regulon under phosphate starvation. They found that changing the parameters of the transport portion of the model did not significantly affect expression. The model also predicts that expression of a heterologous protein from a PhoB-dependent promoter will become less efficient as copy number increases, being limited by the availability of PhoB.

Life Cycle of Bacteriophage T7

T7 is a lytic phage of *E. coli* [see (58) for review]. Its 56 genes are divided into three temporal classes, based on their ordered entry into the host cell. The genes in Class I, the first to enter, are transcribed by the host RNA polymerase. Among these genes is the viral RNA polymerase, gp1, which transcribes the viral genes of Classes II and III. Transcription is regulated by differential promoter strengths, by inhibition of host RNA polymerase by two viral enzymes, and by inhibition of gp1 by viral lysozyme. Endy et al. (22) have constructed an elaborate model spanning nearly the entire T7 genome and life cycle. The model explicitly treats 52 gene products. Of these, 15 play a direct role in viral processes treated by the model. Simulations give the time evolution of all 52 gene products, illustrating the shift from host to viral polymerase activity and allowing one-step growth curves (the build-up of progeny phage in an infected cell over time) to be derived for comparison with experiment. In a subsequent paper (24), the model is used to explore the ramifications of the organization of the T7 genome. The authors predict the effect on viral growth of relocating certain genes and construct the rearranged viral strains in the laboratory. Agreement with experiment is only fair. Although the problem may lie with the model, the experiments disrupt certain genomic sequences thought to be inessential but whose role has not been clearly established. The discrepancy between model and experiment may thus indicate aspects of T7 biology in need of further investigation.

The model's explicit representation of viral mRNA allows the use of antisense RNA as an antiviral agent to be explored. In these simulations, targeting viral structural genes slows viral growth but targeting certain others is predicted to accelerate it. These are the genes whose products exert an inhibitory influence on the host and viral polymerases. Targeting gp1 in fractional amounts is also predicted to enhance viral growth, likely by slowing the expression of the inhibitory gene products. This aspect of the study is continued in another paper (23), where the sense-antisense-mRNA binding constants are allowed to vary in order to simulate the effect of the phage mutating under pharmaceutical pressure. In concordance with the earlier result, simulations show that, for stoichiometric targeting of gp1, small decreases in RNA binding affinity would lead to slower-growing virus. Thus, small

mutations would be selected against, in principle slowing the emergence of resistance. Studies like this, practical only through modeling and simulation, hold great potential for drug-development applications.

Bacteriophage λ

λ is a temperate bacteriophage of *E. coli* [see (62) for review]. A crucial aspect of its physiology is the selection between the lytic and the lysogenic pathway after infection. Once established, lysogeny is maintained by the phage repressor protein CI, which represses all the other viral genes while stably maintaining its own expression. CI expression can be repressed, however, by the protein Cro. The arena for the competition between these regulators is the O_R operator, which lies between two divergent promoters, P_{RM}, which drives CI expression, and P_R, from which Cro is transcribed, and contains three binding sites. Both CI and Cro can bind the three sites in sequence, but in opposite order. Downstream and in the opposite direction of P_R is the promoter P_{RE}. Transcripts from this promoter include an antisense Cro message along with a normal CI message. P_{RE} transcription is stimulated by the viral proteins CII and CIII. The *cII* gene lies in the reading frame of P_R downstream of *cro*, but cotranscription is hampered by a leaky terminator. The terminator's effect is counteracted in the presence of the viral antiterminator protein N. The *cIII* gene is similarly downstream of a leaky but more efficient N-dependent terminator. CIII confers stability on CII, which is degraded five times faster if CIII is absent. The establishment of lysogeny upon infection requires a transient initial burst of CII expression. With sufficient CII early on, enough CI can be expressed from P_{RE} to activate its own autocatalytic expression from P_{RM} and thus secure its ascendancy. Otherwise, Cro will repress CI expression, and lysis will ensue. Lysis can also be induced in a lysogenized phage by mechanisms that activate the degradation of CI by the host protein RecA.

Early modeling of this system was carried out by Thomas (76) in an abstract framework in which the presence and activity of a gene were treated as Boolean variables (i.e., restricted to values of 0 or 1), and the regulation of the system was encoded by logical functions (OR, AND, NOT, and their combinations). Although such an approach can yield some insights in the absence of detailed biochemical information [see also Huang & Ingber (36) and the discussion of the *endo16* models below], the modifications required to accommodate increasing knowledge (75) dilute the simplicity of the original approach. The introduction of multiple-state logical values, semi-quantitative thresholds, and delay times complicates analysis but falls short of the detailed predictive capability of a more physically based approach.

Modeling efforts incorporating quantitative biochemical information began with Ackers and coworkers (5). In this paper the authors presented a quantitative equilibrium model for the cooperative action of the CI repressor on the activity of the P_{RM} and P_R promoters. The model of promoter activity was enlarged by Shea & Ackers (70) to include the effects of Cro. The authors also included the

synthesis of the CI and Cro proteins, making the model fully dynamical. Although experimental protein time courses were not available for comparison, the model gave the expected behavior for stable maintenance of lysogeny as well as for the induction of lysis when degradation of CI by the host protein RecA was included. Reinitz & Vaisnys (66) extended the dynamical model further by adding degradation of Cro. These authors found an inconsistency between the concentration of CI they measured and the level that would lead their model to exhibit two stable steady states corresponding to lysis and lysogeny, respectively. This inconsistency is likely a by-product of omitting the other proteins important for the decision.

The most extensive model in terms of explicitly represented viral genes is given by McAdams & Shapiro (53). Along with CI and Cro, they include the other proteins of the lysis/lysogeny switch, CII, CIII, and N. They also include genes that become active after the fate of the phage is determined. The model is interesting in that, where appropriate as shown by experimental knowledge, the detailed biochemistry is replaced by logical calculations. Simulations track biomolecule concentrations and cell fate over time and correctly predict the increased tendency toward lysogeny under conditions of increasing multiplicity of infection (MOI). [A model of essentially the same scope but formulated in the specialized mathematical framework of hybrid Petri nets has subsequently appeared (48).]

Arkin et al. (7), in a demonstration of the principles explored by McAdams & Arkin (49), give a model of the lysis/lysogeny switch based on a stochastic representation of transcription, translation, and reactions between proteins (Figure 1). The fraction of lysogens as a function of MOI predicted by their simulations closely agrees with experimental measurements. The authors also make quantitative predictions of this function for mutants that have not yet been studied experimentally. Further, the detailed time courses given by the simulations clearly illustrate how identical cells in identical conditions infected with the same number of phage can still meet different fates owing simply to chance.

Type 1 Fimbriae in *E. coli*

Fimbriae, or pili, are hair-like appendages extending from the bacterial cell. In *E. coli*, type 1 fimbriae [reviewed in (14)] confer the ability to adhere to mannose-containing surfaces. Pathogenic *E. coli* strains exploit this property to adhere to cell-surface receptors of epithelial tissues and ultimately invade the epithelial cells. The fimbriae provoke an immune response, however, so an invading population must strike a balance between piliated and unpiliated members and must control the degree of piliation in response to changing conditions. Piliation is thus phase variable. The members of a bacterial colony may change their piliation state randomly. The structural genes required for type 1 fimbriae, *fimACDFGH*, lie in an operon whose promoter is contained in an invertible region of DNA known as the fim switch. When the invertible region is in the proper orientation (the switch is on), the structural *fim* genes are expressed and the cell becomes piliated. In the inverted orientation (the switch is off), no promoter is available to drive Fim

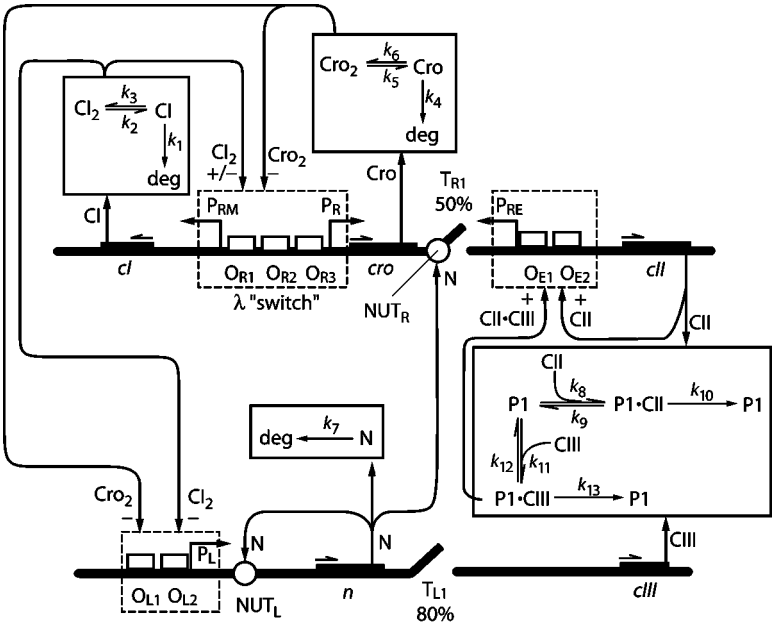


Figure 1 The bacteriophage λ lysis/lysogeny decision circuit, as modeled by Arkin et al. (7). Shown are five genes (*cI*, *cro*, *cII*, *cIII*, and *n*) and their products, four promoters (P_{RM}, P_R, P_{RE}, and P_L) with operator sites ($O_{R1\text{-}3}$, $O_{E1\text{-}2}$, and $O_{L1\text{-}2}$), two terminators (T_{R1} and T_{L1}) and their efficiencies, and two N-utilization sites (NUT_R and NUT_L). Arrows terminating on the operator regions set off by dashed boxes are labeled with the binding species and its effect on transcription. Arrows lying over the genes show the direction of transcription. The solid boxes contain nongenetic reactions between the protein components of the system. Deg indicates a degradation reaction. P1 is the host protease HflB. An additional protease for CII and CIII appears in the model but is not shown here. See (7) for details.

expression. Inversion of the fim switch is accomplished by two recombinases, FimE and FimB. FimE exhibits an orientational bias in its activity; it is much more active in flipping the switch from on to off. FimB's inversion activity is the same regardless of switch orientation. In addition, expression of FimE depends on the switch orientation. With the switch in the off position, FimE is not expressed. There are also contributions to the behavior of the switch from global regulatory factors. Integration host factor (IHF) is necessary for the switch to function at all; it is thought to participate in bending the DNA into the proper shape for inversion to occur. Leucine responsive protein (Lrp) enhances the switching rate at moderate concentrations but tends to reduce it at higher concentrations unless leucine is present. H-NS represses the expression of the two Fim recombinases and Lrp in a temperature- and nutrient-dependent way.

Wolf & Arkin (83) give a stochastic model of the fim switch, the only model to our knowledge that treats DNA inversion. The model directly incorporates the effects of IHF and Lrp on the behavior of the switch. The effects of H-NS are handled indirectly through changes in concentrations of other proteins. The model allows the probability of piliation to be calculated from concentrations of the recombinases and global regulators, leading directly to the piliation statistics of a population. Switching behavior can be predicted under a wide range of temperatures and nutritional states, as reflected in the activities of the global regulators. The authors use the model to demonstrate the enhanced control properties of the switch provided by the action of two recombinases where it might be thought that one is sufficient. The model also shows how temperature tuning is accomplished by Lrp, upregulating the degree of piliation at host body temperature and downregulating it at temperatures indicative of an inflammatory response. Finally, the putative course of infection is followed, and the model predicts piliation behavior well in line with expectations for adaptive response.

Eukaryotic Development

All the models we discuss in this section have a distinctively different character from the above models. Developmental processes tend to be extremely complicated, all the more so because many cells are involved and their relative locations in space are important. Certain regulatory proteins are secreted and may diffuse through developing tissues; others are confined to the interiors of individual cells or nuclei or to the interfaces between cells. Moreover, the regulatory roles of many proteins important in development have not been clearly elucidated, let alone the mechanisms by which such functions are carried out. Thus, the principal focus of modeling in this area has tended to be for explanatory purposes, to try, by fitting a model to experimental data, to discover the rules governing regulation from the fitted model parameters.

Reinitz and coworkers have used time-dependent gene-expression levels inferred from micrographs of developing *Drosophila* embryos to fit dynamic gene network models. In one paper (65), the authors attempt to clarify the roles of four gap genes in establishing the striped pattern of *eve* gene expression. In a subsequent paper (64), the authors examine the roles of *bicoid* (*bcd*) and *hunchback* (*hb*), maternally expressed genes, in forming the domain pattern of some gap genes. After fitting to wild-type data, the model is used to predict the effect of changing the copy number of *bcd*. Predictions of how the location of a positional marker changes as the *bcd* copy number is increased are in good qualitative agreement with experiment. The model also allows investigation of the role of maternal *hb* expression by simulating the effect of "turning off" this expression. The result is a significant shift of the *bcd* dose-response curve.

Marnellos et al. (46) use a simple model to study lateral inhibition by intercellular Delta-Notch signaling [see (9) for a review] in development of *Xenopus* embryonic epidermis. Fitting is done not to a specific dataset but to the more general

experimental finding that, after differentiation, two thirds of the cells are epidermal (high levels of Delta expression and low levels of Notch) and one-third are ciliated (high Notch, low Delta), more-or-less homogeneously dispersed among the epidermal cells. The fitted model is then used to simulate two gene-injection experiments, one of which produced a highly variable response in different embryos. The model has a small number of parameters, and a large number of adequate fits emerge in the study. The authors identify this variation with naturally occurring genetic variability between different embryos. Simulations of the experiment using the various parameter sets give variable results, which the authors claimed as good qualitative agreement. Significantly, the various fitted parameter sets all gave a similar result in simulations of the other experiment. In an earlier paper (47), a more abstract version of this approach, not using individual genes but aggregates called proneural and epithelial, was applied to *Drosophila* neurogenesis. More recently, a model for Delta-Notch signaling using a hybrid automaton approach has been given by Ghosh & Tomlin (28).

Kyoda & Kitano (42) study the interaction of eight developmental genes in a model of leg formation in *Drosophila*. The authors include known patterns of repression and activation in a threshold-based treatment of regulation, along with one hypothesized interaction, the repression of *dpp* by CI. The parameters of the model are hand tuned to give good qualitative agreement with observed gene-expression patterns and protein-localization patterns in the *Drosophila* leg disc. Hence, the introduced interaction is a prediction of the model. Results of simulations are further used to support a particular view of proximal-distal axis formation over competing views.

Yuh et al. have studied the *endo16* gene in sea urchin (85, 86). This gene is differentially expressed in different embryonic tissues. Expression is regulated by means of a long and complicated cis-regulatory region that has been conceptually divided into modules. A series of reporter constructs were made combining various modules and sites within modules, both mutated and unmutated, and expression was followed over time. A procedural model summarizing a large number of experimental findings was constructed (85). The model is essentially phenomenological. It is devoid of biochemical detail, much of which is completely unknown. Instead, it takes as input the presence or absence (deletion or mutational inactivation) of various modules and sites, along with the time-dependent gene expression measured for other sites, and applies logical and arithmetic computations to reproduce experimentally observed behavior. The model is extended in a subsequent paper (86) to include more detailed experiments on module B and its interaction with module A (Figure 2). It is notable that the model allows the authors to make quantitative predictions about the expected levels of *endo16* expression in a number of mutants not previously examined. This is a strong example of the integrative function of modeling leading to predictive power. It will be interesting to see how the computation implied by the model is actually implemented by biochemical mechanisms. [Of note, a model along similar lines but covering much more of sea urchin development (19) appeared as this paper was going to press.]

von Dassow et al. (81) model the *Drosophila* segment polarity network. Known gene and intercellular interactions are collected into a model that attempts to reproduce the spatial expression patterns of segment polarity genes in rows of embryonic cells. Because specific data on the various interactions are sparse, the authors carry out a search for suitable parameters. For the initially formulated model, parameter sets that led to the correct expression patterns were very rarely produced in a random search. When additional interactions, not yet conclusively demonstrated experimentally but plausible given current knowledge, were added, parameter sets producing correct behavior became much more common. Further, varying individual parameter values within some range tended not to disrupt the model's behavior significantly. Thus, the model exhibited a degree of robustness to variation. On the basis that such robustness is a principal characteristic of certain biological networks, the hypothesized interactions that led to robustness are predictions of the model.

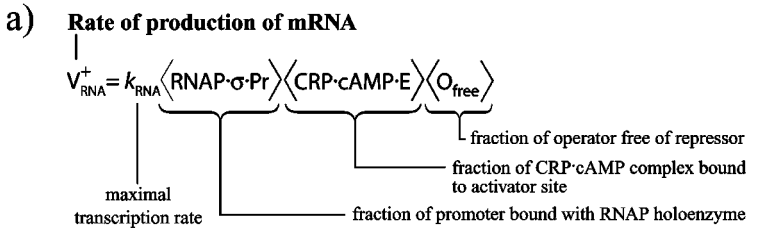
REPRESENTATIONS OF GENETIC PHENOMENA

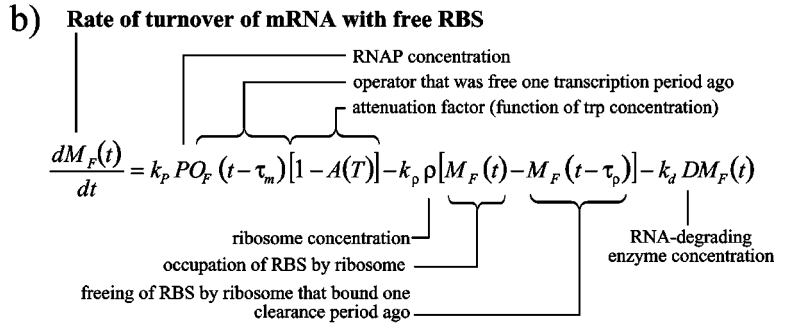
The construction of a quantitative model involves two principal considerations. The first is the model's conceptual structure. This lays out which molecular entities appear as explicit participants, which aspects of their behavior are treated by the model and which are neglected, and which entities or effects enter implicitly. The second consideration is how the conceptual structure is to be represented mathematically: which types of mathematical objects stand for the molecular players and which equations or algorithms are used to implement their behavior (Figure 3). There is a fair degree of latitude between these aspects of a model but also quite a lot of interplay. In this section we illustrate in detail how these aspects of model building are dealt with in the examples described above, as well as in others.

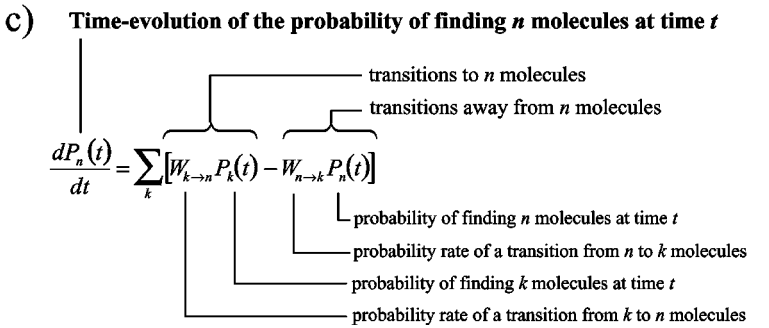
Of the above examples, four in particular take very detailed views of their subjects. These are the phage λ models by McAdams & Shapiro (M&S) (53) and by Arkin, Ross, and McAdams (ARM) (7); the induced *lac* operon model by Carrier & Keasling (C&K) (16); and the *trp* operon model by Santillán & Mackey (S&M) (67). Although these four works conceptualize gene expression in similar ways, each is rendered by a different mathematical formulation. ARM and C&K use stochastic algorithms, although the latter's is ad hoc, intended to approximately illustrate the effects of random fluctuations on gene expression in their system, whereas the former's is based on a rigorous derivation from microscopic physical principles (Figure 3*c*). M&S and S&M both use continuous, deterministic mathematics. But, as discussed below, S&M's conceptual scheme, while including the same level of process detail, does not include as many players as the M&S model, and this leads to their use of differential-delay equations (Figure 3*b*) as opposed to the straight differential equations employed by M&S. After following these four works through the stages of gene expression at very high degrees of detail, we use the remainder of this section to discuss some more abstract, simpler approaches.

Transcriptional Initiation

The most elaborate quantitative treatment of transcriptional initiation was introduced by Ackers and coworkers (5, 70) and is employed by ARM, along with others discussed later. The approach begins by enumerating the possible configurations of transcription factors and RNA polymerase (RNAP) bound to promoter and operator sites. Each distinct configuration is associated with an initiation rate, which is allowed to be zero if the configuration does not include a productively bound RNAP. The overall rate of transcript initiation is then calculated as a weighted average of







the "microscopic" initiation rates. The weight for each rate is simply the relative abundance of each configuration at equilibrium, a fraction determined by a thermodynamic calculation using the free energies of binding for each configuration. To have a compact way to refer to this approach, we propose the acronym BEWARE: binding equilibrium weighted average rate expression. BEWARE is a very versatile approach. It allows any number of transcription factors to be included and accommodates various effects they may have on initiation. For example, repressors that function by preventing RNAP binding give noninitiating configurations. Repressors that still allow RNAP binding but act to retard initiation in some way give configurations with low microscopic initiation rates. Activators that increase RNAP's affinity for the promoter give configurations with more favorable binding free energies than those with RNAP but without the activator and hence increase the overall abundance of initiating configurations. Cooperative transcription-factor binding, the effect this approach was originally used to investigate, is rendered with a favorable free-energy term added to the contributions of individual binding.

ARM incorporate BEWARE into their stochastic model of phage λ by treating the output as the instantaneous probability of an initiation event. However, the mathematical validity of this treatment has not been rigorously established. C&K use a similar but somewhat simpler approach. Initiation is equated with the binding of RNAP to the promoter, and this binding is a random event. Its probability increases with increasing inducer concentration. Overall, this approach is conceptually similar to BEWARE but cannot be rigorously derived from equilibrium considerations.

S&M also equate initiation with binding of RNAP to a free promoter site, but they let this binding be competitively inhibited by repressor. The term describing this inhibition is derived by using a stationary-state assumption that is justified by experimental measurements of repressor-binding kinetics. M&S also treat initiation events explicitly but take a hybrid approach. For some of the promoters in their

←

Figure 3 Mathematical representations of some genetic processes. (*a*) Equation giving the rate of production of *lacZYA* mRNA in the model by Wong et al. (84). The quantities set off by angle brackets are equilibrium binding fractions of the DNA sites that interact with RNA polymerase holoenzyme, catabolite regulatory protein-cAMP complex, and the LacI repressor, respectively. Note that the three interactions are implicitly independent under this representation. (*b*) Equation from the *trp* operon model by Santillán & Mackey (67). This equation treats the production, ribosome occlusion, clearance, and degradation of mRNA. Two delay times appear: τ_{m} is the time required to produce a full transcript once transcription has initiated. τ_{ρ} is the time required for a translating ribosome to clear the RBS (but not to complete translation). T is the (time-dependent) tryptophan concentration. Note that the rate constants are not explicitly labeled. (*c*) A general form of the chemical master equation, the basis of the stochastic simulations by Arkin, Ross, and McAdams (7). The W terms give the probability of transitions per unit time.

λ model initiation is determined in a way similar to BEWARE, but for others a simpler calculation approximating the function of a Boolean logic gate is performed. A positive output from this calculation means that initiation has occurred.

One of the few specifically eukaryotic models of transcription, a model by Wang et al. of the synergetic activation of Epstein-Barr virus genes by the viral protein ZEBRA (82), is a BEWARE-type model. Equilibrium is assumed between ZEBRA, the promoter, and the transcriptional machinery, with cooperative interactions between ZEBRA and the general transcription factors. Another eukaryotic example, an equilibrium model for the cooperative binding of regulatory proteins to nucleosomes to expose transcriptional units, has been given by Polach & Widom (61).

Promoter Clearance and Transcriptional Elongation

In the models by ARM, M&S, and C&K, elongation is completely explicit. These models track the position of the elongation complex within the coding region. The elongation complex moves at a constant rate in the M&S model. In the ARM model, the advance of the elongation complex over each consecutive nucleotide is a stochastic step with an exponential waiting time distribution, the stochastic equivalent of a constant-rate process. Elongation in the C&K model also has a stochastic character, but it appears through a random choice of elongation rates associated with different initiation events. This raises the possibility that a faster-moving elongation complex might overrun a slower one. The model explicitly prevents this from happening. The ARM model carries out a similar surveillance.

Because RNAP that has not moved clear of the promoter is thought to prevent any further transcriptional initiation, there is potentially a strong coupling between elongation and initiation. Both the C&K and the ARM models treat promoter clearance explicitly, requiring that an initiating RNAP move at least the length of its footprint before allowing any possibility of subsequent initiation. In S&M's model, there are no notions of an elongation complex or its position, but there is the notion of a free promoter. Time delays are used to represent the occlusion of the promoter after initiation and the time necessary to complete the transcript (Figure 3*b*).

The elongation phase of transcription is also where various control mechanisms such as termination, antitermination, transcriptional attenuation, and convergent transcription come into play. Termination and antitermination play crucial roles in phage λ physiology. These processes, at least as they occur in leaky terminators, are explicitly present in the models by M&S and ARM and are handled similarly, aside from the deterministic treatment of the former and the stochastic treatment of the latter. In the deterministic case, a leaky terminator prevents a fixed fraction of elongation complexes from passing; complexes that have previously been antiterminated are exempt. In the stochastic case, there is a fixed probability that the elongation complex will dissociate upon reaching the leaky terminator. This probability is zero for antiterminated complexes. Termination that occurs when the elongation complex reaches the end of a coding region is not given particular

treatment by any of the models we discuss. Typically, the completed transcript is simply released.

Transcriptional attenuation appears in the S&M model. The details of the process are entirely abstracted into a mathematical term that reduces by large amounts the fraction of initiating transcripts that are actually completed, unless the trp concentration is very low (Figure 3*b*).

The ARM model further makes allowances for convergent transcription from the facing P_R and P_{RE} promoters. When two elongation complexes originating at these promoters collide in the intervening sequence, one or both of the transcripts are terminated prematurely. The presence of convergent promoters also raises the possibility of antisense RNA interactions, but such effects are not treated by the ARM model. Handling of these interactions is further described below.

Translation

Translation occurs in stages much like those of transcription: initiation, elongation, and termination. However, only three of the four detailed models we discuss treat stages of translation explicitly. The model by M&S uses a lumped representation of translation and is discussed below. Initiation of translation is typically represented by the binding of a ribosome to a ribosome binding site (RBS) on the transcript. In all three models, this binding is allowed to occur before the transcript is complete. C&K and ARM, in tracking the growing transcript, also check for initial RBS availability and its clearance of transcribing ribosomes. S&M employ additional time delays to represent the time needed for an RBS to be transcribed initially and the time it takes an initiating ribosome to clear the RBS. Translational elongation is treated in a very similar way to transcriptional elongation—S&M utilize time delays, and ARM and C&K use explicit tracking of elongating ribosomes. Much like transcriptional termination, translational termination is implicit in all the models—the completed peptide is simply released. Other phenomena like translational stalling or premature termination have not been modeled. Even though translational stalling due to insufficient charged trp-tRNA is crucial to the mechanism of transcriptional attenuation of the *trp* operon, S&M chose not to treat it explicitly (see above). A highly detailed model of attenuation in the *trp* operon is given by Koh et al. (40), who use deterministic chemical kinetics for each step of translational elongation to derive the pause duration of a ribosome stalled at the *trp* codons in the attenuator. The degree of attenuation is then determined by how much transcription occurs during the stall and hence whether the terminator structure will form in the nascent transcript. Although the authors suggest that this model may be combined with an earlier model they give for repression of the *trp* operon (41), this has not been done.

DNA Inversion

In the fim model by Wolf & Arkin (83), the flipping of an invertible DNA element is treated by a novel application of the BEWARE approach. Instead of factors bound

at a promoter, the authors enumerate the configurations of recombinases bound to the recombination sites of the fim switch and global regulatory factors bound to regulatory sites in the switch. Each configuration is associated not with a transcript initiation rate, but with the "flipping" rate for the switch, i.e., the rate of DNA inversion. The rate is a probabilistic one, the probability of switching per unit time. The effect of switch orientation on the availability of the FimE recombinase (orientational control) is treated by embedding the BEWARE model within a more complex stochastic model that explicitly constrains the FimE concentration based on the switch orientation. Although the precise mechanism of orientational control is not known, convergent transcription and masking by an antisense transcript have both been suggested. The abstract model can thus be used to represent the effects of either mechanism.

mRNA Interactions and Stability

Models that explicitly represent mRNA must include its degradation to be accurate. Moreover, the control of mRNA stability can be an important regulatory strategy in both prokaryotes and eukaryotes. Having an interest in the effects of mRNA stability on heterologous gene expression, C&K use the most detailed conceptual representation of the mRNA degradation process. In their model, an implementation of an idea proposed by Alifano et al. (5a), RNAse E binds to the 5′ end of an mRNA and attempts to cleave it at a randomly chosen internal site. The attempt may fail in two ways: if the chosen site is not a recognition site for the RNAse or if it is a recognition site but is protected by an elongating ribosome. In either case, the RNAse dissociates. A successful cleavage attempt prevents any further translation from initiating on the mRNA; any ribosomes downstream of the cleavage site are allowed to continue. The presence of two cistrons on the transcript in the *lac* operon model introduces additional complexity. RNAse cleavage within the upstream cistron still allows translation to initiate at the downstream cistron. The ARM model also includes protection of mRNA by ribosomes by prohibiting degradation when a ribosome is occupying the RBS. Most other models treat mRNA degradation more abstractly, as the action of a constant reservoir of RNA-destroying machinery, resulting in first-order decay processes.

The much more complicated issue of the fate of mRNA in eukaryotes (12, 56, 57) has recently been braved by detailed kinetic modeling by Cao & Parker (15), who include deadenylation, decapping, and exonuclease degradation processes of well-studied yeast mRNAs as series of first-order chemical reactions. Although tackling only a part of the thicket of eukaryotic gene expression, this model is a prime candidate for incorporation as a component into a larger gene-expression model. Components that could provide it with realistic input by simulating the production and transport of mRNA (treated as constant-rate processes by Cao & Parker) are, however, in short supply.

Before being degraded, some RNA molecules may participate in antisense interactions. Although the regulatory importance of such interactions in eukaryotes

is well established (13) and is increasingly being recognized in bacteria (6), only the M&S model contains one explicitly, where the production of the Q protein is inhibited in the presence of the antisense transcript. Targeted antisense interactions are explored in a more abstract way, proportionally removing the target transcript as it is formed, in the T7 growth model by Endy et al. (21, 22).

Lumped and Phenomenological Representations

Lumped (or lumped-parameter) representations are those in which details of various processes are aggregated into a single mathematical expression. The expression is formally derivable from a more detailed mathematical representation under given assumptions, and in the process, expressions and parameters of the more extensive representation are transformed and reduced in number. Lumped representations are to be distinguished from phenomenological representations, where the mathematical expression is chosen purely to reproduce observations, with no regard for underlying mechanism. Sometimes the distinction between these becomes blurred, especially in the absence of discussion about how various mathematical expressions were reached.

Lumped versions of transcription include only the end result, production of mRNA, in the conceptual scheme. The actions of various players in transcription do not appear in separate mechanistic steps of the process, but rather as immediate effects on the overall rate at which mRNA is produced. This is done in the deterministic *lac* operon model (Figure 3*a*) and the similarly constructed model of the phosphate starvation response by Keasling and coworkers (80, 84). Their approach starts with a basal rate of mRNA production that is then modified by "efficiency factors" that reflect binding equilibria between promoters and RNAP and between transcription factors and operator sites.

Endy et al. (22) use a similar approach, but the efficiency of T7 mRNA production in their model is affected by promoter strengths and the availability of polymerases. The latter depends both on the fraction of the viral genome that has entered the host cell (i.e., how many promoters are vying to bind polymerase) and the interaction of the viral and host polymerases with viral inhibitors, treated as equilibrium processes.

Lumped versions of translation tend to look similar to lumped versions of transcription, often even simpler. The deterministic models by Keasling and coworkers for the most part treat translation as protein appearing with a first-order dependence on its mRNA. Because the model by Endy and coworkers explicitly contains transcript lengths, translation rates are calculated from a constant elongation rate along a transcript, carried out by multiple ribosomes and scaled by the amount of transcript available.

When models do not explicitly include mRNA, the overall process of gene expression is given a lumped representation. A sophisticated lumped representation is used by Shea & Ackers (70) [and Reinitz & Vaisnys (66), whose model is a direct extension]. This model is conceptualized to include transcription and

translation but ends up having a lumped mathematical form because of certain assumptions: that transcription initiation is the rate-limiting step of expression and that each transcript produces a fixed (average) number of proteins. Any effects of mRNA degradation are assumed to be fixed and reflected in the average number of proteins per transcript. Thus a single production term lumps together contributions from transcription and translation. Less sophisticated lumped versions of gene expression have also been used and are, in fact, very common. The trivial example of a constant enzyme concentration can be interpreted as the simplest lumped representation of gene expression. Less trivial versions are the production of protein at a constant rate, a rate proportional to cell size, and a rate proportional to the concentration of some activator. Such versions of gene expression have appeared in the biochemically sophisticated models of the eukaryotic cell cycle given by Tyson and coworkers (17, 18), where the focus is on the interplay of enzymes, and gene expression is only of peripheral importance.

The majority of the models of eukaryotic systems we discuss employs phenomenological representations of gene action. The gene network models of eukaryotic developmental processes (46, 47, 64, 65, 69) all share the same basic mathematical form, which reflects nothing about the biochemistry behind genetic interaction. Rather, the inhibitory effect of one gene on another is represented as a negative entry in a matrix, whereas activation is represented by a positive one. This approach leads to a picture of the regulation within a set of genes that is clear but difficult to reason about and to extend with knowledge about the effects of experimental conditions or mutations. The *endo16* models by Yuh et al. (85, 86) are almost purely phenomenological also. The logical and algebraic formulae in these models describe informational rather than biochemical activities. They summarize extensive empirical knowledge, the mechanistic basis of which is still highly unclear. Even the segment polarity network model by von Dassow et al. (81), which is based on a fairly detailed conceptualization involving RNA synthesis and degradation and the interaction and transport of proteins, represents gene expression in an essentially phenomenological way. There, the inhibitory effects of proteins are represented by saturable, cooperative inhibition terms that modulate a maximal transcription rate. The actual inhibitory mechanisms are probably quite complex, but the phenomenological representation, although simple, gives dose-response curves that are reasonable given general biological considerations.

PROBLEMS, PLATFORMS, AND PERSPECTIVES

The discussion in the preceding sections should serve in bringing out the fundamental problem with modeling biological systems: It is very difficult to gauge what a model is worth. Trouble appears at both ends of a modeling study, the formulation of a model and its results. A model's formulation necessarily involves many simplifying assumptions, some of which may be deeply hidden within the model's mathematics. Every assumption raises the question of whether any important

effects are ignored, distorted, or introduced. Is the model pitched at the right level of conceptual detail, and does its mathematical formulation accurately reflect that level of detail? How accurate are the many parameter values that have not been measured directly and must therefore be estimated, sometimes from experiments done under different conditions, in a different strain, or even in a different but related organism than that under study?

On the other side are the results. In cases lacking quantitative experimental data, qualitative agreement with observations is taken as success. But evaluating qualitative agreement is stubbornly subjective. Even when quantitative data are available, there is no commonly accepted methodology for performing formal comparisons. It is certainly possible to do chi-squared tests to compare predicted dependencies to observed ones, for example, or to calculate least-squares distances between experimental and simulated data. Deeper techniques for using limited data to score and compare models are emerging from the study of Bayesian networks (25, 30, 33). It is thus possible to compare different models and assess which one gives closer agreement to experiment. But for individual models there is typically no context in which to place a quantitative comparison score, so subjectivity comes into play again. And if a model's output is deemed not to fit experimental data, do the faults lie somewhere in the typically massive edifice of assumption and estimation, are the experimental data used to parameterize the model erroneous, or are important pieces missing from the fundamental understanding of the system being modeled?

Creators of models go to great lengths to justify their formulations, appealing both to fundamental physical principles and to established biological knowledge, but questions of validity must ultimately be answered empirically. Direct experiments must be done to measure the numerical values of model parameters and to test the validity of various modeling assumptions. More of this work will get done as the gulf between modelers and experimentalists narrows. There is, however, an alternate empirical avenue, emphasizing correct prediction over correct formulation. The practical validation of modeling approaches according to their utility could be carried out efficiently if a large community of biologists could use models easily to make predictions and could explore various aspects and combinations of modeling components in attempts to best reproduce experimental results. Such a community effort would produce bilateral benefits. It would benefit mainstream biologists by illustrating the usefulness of modeling in a variety of contexts and more efficiently directing aspiring users of models to the most promising approaches. It would also benefit model builders, who would be prodded by a wider and more discerning audience to provide better, more useful models. An essential prerequisite for this situation, however, is the capability for nonspecialists to explore entire families of models for a given system, to vary not just parameter values but more substantial aspects of a model like the level of detail of the conceptualization and the mathematical treatment of its processes.

Two factors currently make it difficult for even motivated nonspecialists to perform such explorations on existing models. First is the lack of integration with

data, an issue discussed further below. Second are the primitive ways in which models can currently be shared. In the examples we have discussed so far, the simulations have been hand crafted. Most of the required software was written by the various authors. Today there are three principal ways that such software is dispersed: (*a*) The authors encapsulate their model into a software tool specifically designed for use by others. (*b*) The authors make their in-house software available by request. (*c*) Interested researchers may implement their own version of the model based on published descriptions. Of these, only the first is of much interest to nonspecialists. Software tools are usually well documented and have a usable interface. However, because of their stand-alone nature, tools can be difficult if not impossible to modify, extend, or integrate with models in other forms. A potential solution is the use of shared simulation environments. A model implemented within an environment can readily be distributed by way of the files that configure the implementation. However, to make this effective, the environment, or at least the format for the configuration files, should be standard.

No such standard currently exists. Rather, a number of different simulation environments intended specifically for biological applications are currently available. These come in two basic types: environments that compute continuous deterministic equation systems and environments that do stochastic simulation. Packages in the former category include Gepasi (54), DBSolve (32), E-CELL (77), and Virtual Cell (45, 68). All these have a number of features in common. They all offer good facilities for chemical reactions, including predefined rate laws for various enzymatic mechanisms. They also allow user-defined rate laws, for which the necessary equations must be supplied. Facilities for various kinds of mathematical analysis, such as metabolic control analysis, linear stability analysis of steady states, parameter fitting, and bifurcation analysis, are also offered by Gepasi and DBSolve. Virtual Cell can do basic parameter sensitivity analysis. Although all the packages handle spatially homogeneous systems well, Virtual Cell is also particularly strong on spatially distributed systems with arbitrary geometries and subcellular compartments. It allows system geometries to be imported directly from microscopic images and provides image-like visualization of its simulation output. Gepasi handles multiple interacting compartments and has recently been expanded to handle a limited number of inhomogeneous spatial arrangements and random parameter distributions (55). The fully stochastic simulation environments include MCell (11) and StochSim (44). MCell is useful for low-level simulations and is particularly strong in its handling of spatial structure and diffusion. It allows elementary chemical reactions to be implemented as choices of outcomes for events where particles physically interact. Its visualization facilities are also striking. StochSim implements an ad hoc Monte-Carlo algorithm to simulate first- and second-order chemical reactions. It includes facilities for tracking the states of individual molecules, such as phosphorylation or methylation states of enzymes. Primarily intended for simulating spatially homogeneous systems, it has recently been extended to handle two-dimensional lattices with nearest-neighbor interactions (1).

Although these biological simulation environments are impressive efforts, they all suffer important limitations. First, although chemical reaction processes are well supported, more complex processes like those associated with genes are not [but see (39) for efforts to represent a simplified version of genetic regulatory interactions with DBSolve]. Those interested in modeling gene expression and regulation have two choices: decompose all of the processes of interest into chemical reactions or, in those environments that permit it, provide a purely mathematical representation in terms of equations.

Second, it is not possible to import many kinds of experimental data directly into these simulation environments. The exceptions are Virtual Cell for microscopic image data, as mentioned above, and DBSolve, which can import metabolic pathways directly into a model from the WIT database (59) and accepts tabular kinetic data entered by the user. The problem of data is a deeper one and cannot be solved by designers of simulation environments alone. This is discussed further below.

Finally, each environment uses a different format to store its models, parameters, and output, which limits their sharing. Efforts are in progress to improve this situation by creating a standard format for describing models of biological systems. The highest achievements of these efforts are currently the two formal description languages SBML (37) and CellML (2). Both languages accommodate the separation of models into conceptual structure and mathematical representation. However, the conceptual vocabulary is heavily skewed toward chemical reactions. SBML assigns chemical species to reactant or product roles in a reaction, whereas CellML expresses a slightly wider range of functional notions, including catalyst, activator, and inhibitor. Both languages contain the notion of a compartment, but CellML has a richer vocabulary for expressing topological relationships between compartments (e.g., containment or adjacency). In the mathematical realm, both languages allow the specification of arbitrary rate laws associated with reactions. SBML can also express global mathematical constraints on the chemical species in a model. CellML's handling of mathematics, in part relying on the mathematical dialect MathML (3), endows it with broad power to express abstract aspects of models, including the ability to specify purely abstract components of a model. That is, CellML can describe model components defined purely mathematically, with no reference to any biological concepts.

This dearth of conceptual expressiveness in the formal description languages and the simulation environments described above is one crucial factor impeding the integration of modeling into mainstream biological methodology. The primitive interactions between models and data is another. For modeling and simulation to become truly accessible to as wide an audience as possible, a software environment that provides flexible interaction between concepts, mathematics, and data must be developed. Such an environment must handle simulation components that implement models at a variety of levels of detail and with a variety of mathematical representations. At the same time, the environment must support the distinction between a model's conceptual structure and its mathematical representation. It must allow users to rapidly and intuitively construct a conceptual picture of their system

of interest, much like sketching a diagram, at whatever level of detail is deemed appropriate. Users must then be allowed to render this conceptual model into a simulable mathematical form by selecting among simulation components that provide various mathematical representations for the specified concepts. Throughout the process, transparent access to both qualitative and quantitative empirical knowledge must be maintained. Users should not only be able to locate experimentally determined quantities that can be used immediately to parameterize the simulation components they have chosen or data against which to check the output of their simulations. More generally, exploiting the integrative aspect of modeling, they should be able to use their models as information gateways, opening on things like evidence for and against the model's formulation, related models made by others, and alternate versions of the present model, together with reasons and results.

To make such an environment possible, developments must occur on several distinct fronts. First, to mitigate the shift of emphasis away from mathematics, modelers wishing to provide components for the environment will have to annotate their contributions exhaustively. This means a full description of the conceptual basis of the model, the physical and mathematical assumptions and when they would be valid, guidelines for appropriate use and permissible interaction with other components, and whenever possible, references to supporting literature and data.

Second, the simulation environment needs to provide a high degree of consistency checking on several levels at once. It should ensure conceptual consistency—an mRNA degradation process should not be fed from a lumped gene-expression process that does not treat RNA, for example. It should enforce mathematical compatibility—a module requiring continuous input should not be fed from one providing discrete output; stochastic and deterministic modules should be correctly combined, etc. And it should make sure that the assumptions behind different modules' formulations can all hold under the same conditions—an enzyme that is represented by a Michaelis-Menten V_{max} parameter in one component should not be allowed to be consumed by another component, for example.

Third, the capacity to link models to data must be greatly enlarged. The environment must allow transparent access to a range of information, from raw experimental data and simulation output to processed and edited statements of biological fact and hypothesis. Such ambitious capabilities can only arise through broader developments in how biological information is handled in general. A much wider variety of empirical information must be given standardized representations, in much the same way as sequence and structure data are today and microarray data will be soon (4, 26). Canonical, machine-readable representations are needed for data generated by methodology such as gene-expression assays, mass-spectroscopic studies, gel and blot assays, two-hybrid assays, genetic knockout experiments, and flow cytometry, just to name a few. These representations must include information describing the context in which the data were gathered, such as experimental conditions or detailed protocols, organisms, strains, genotypes, and even the hypotheses being investigated by the experiments. This contextual information must be standardized as well. An integral part of the standardization effort will be the requirement that data submitted for publication be provided in the standard formats.

Developments along these lines will ultimately result in very powerful tools useful to the entire biological community. In particular, they will enable the integration of modeling into everyday biological methodology, bringing to a much wider audience the advantages we discuss above. More generally, the standardization and formalization of biological knowledge resulting from these developments will further accelerate the pace of biological discovery and will inevitably make biology a more exact and powerful discipline.

CONCLUDING REMARKS

We take a narrow focus in this paper, discussing only those modeling studies that treat gene expression and regulation in real systems, referring whenever possible to actual experimental results. We steer clear of the more abstract models, as they often tend to put analysis before biology. Abstraction in itself, however, is not to be disparaged. It is only through abstraction that we can comprehend and formulate general principles of behavior of complicated systems. For example, virtually all the models we discuss rely on the abstraction of a regulated metabolism that supplies the gene-expression machinery with activated nucleotides and other biosynthetic intermediates and properly disposes of degradation products. Abstraction must be used with caution, however. By abstracting a part of a complex, interconnected system important connections may be severed. This is a particular danger for many biological systems, given our incomplete knowledge of how their various parts interact. Also, when a system behavior is given an abstract representation, access to the causal framework that gives rise to that behavior is foregone. Abstraction is best used when it can be grounded on a solid foundation of observations showing that important aspects have not been missed or on the careful consideration of the underlying details.

Detailed models constitute the strongest statements of the causal structure of biological phenomena and, hence, give the strongest predictions of the effects of perturbations (gene knockouts, enzyme activity mutants, antisense RNA, etc.). Many of the models we review reflect a rather high level of detail. This enables them to explore a wide variety of hypotheses, from the microscopic, like the precise concentrations of proteins and RNA at a given time, to the global, like whether an infecting λ phage will lysogenize. There are some serious disadvantages of modeling at high detail, however. First, the details may make important general principles difficult to discern. Second, simulating a detailed model requires significant resources of time and computational power. We expect the constraints on these resources to diminish progressively as simulation algorithms and computer hardware are further developed. Third, detailed models require large amounts of detailed data, such as individual binding constants, concentrations, reaction rates, half-lives, and probability distributions. Such data are difficult to obtain, particularly in relevant biological contexts (in vivo, in single cells rather than whole populations, etc.). Besides inevitable improvements in experimental methodology, we also expect that the integration of modeling into the biological mainstream will drive an increase in experiments specifically directed to gather such data.

Initial examples of this process will likely come from the study of recently constructed synthetic gene networks like the bistable genetic switches by Gardner et al. (27) and the "repressilator" by Elowitz & Liebler (20). Modeling played a key role in the design of both these systems, although the models are very abstract and provide little more than suggestions of feasibility. As these systems are composed of a small set of known parts and possess well-defined functionality with easily measurable outputs, they are good candidates to become "characterization platforms." That is, once all the parts and their interactions are characterized in sufficient biochemical detail, unknown parts can be substituted; the changes produced in the system output should then allow the relevant properties of the unknown parts to be readily calculated. It will be interesting to see how far such an approach can go.

Ultimately, the complexity of biological systems and the diversity of emphases of various research directions demand tools that offer the capability to make models at any level of abstraction, to simultaneously combine components at different levels of abstraction, and to move freely from one level to another. This capability is already available in mature engineering disciplines like electrical and mechanical engineering.

The heterogeneity in levels of abstraction, conceptual structure, and mathematical representation that we have tried to illustrate in recent hand-crafted models of genetic systems, together with the heterogeneity in biological knowledge, data sources, and experimental methodology, contributes to the difficulties inherent in evaluating the formulation and results of modeling studies. These difficulties and the lack of general tools that allow nonspecialists to implement and explore models on their own are slowing the widespread adoption of modeling by the biological community at large as an important complement to more traditional approaches to research. Modeling offers great advantages in integrating and evaluating information, providing strong predictions, and focusing experimental directions. Getting models and data together into the hands of mainstream biologists in forms that are standard and sophisticated but not mathematically overwhelming is therefore a critical necessity if these strengths are to be properly brought to bear on important biological problems.

Appendix: A Brief Guide to Recent Reviews

The review by Arkin (8) enumerates the various functions of models and points to examples of each. Smolen et al. provide a series of excellent reviews of abstract modeling of genetic regulatory systems (71–74). The authors focus on complex dynamical phenomena (multistability, oscillations, frequency selectivity, chaos) arising in various models. They also discuss the effects that changing the representation of biological transport or introducing stochastic fluctuations can have on system dynamics. Endy & Brent (21) discuss successes and practical problems of modeling and offer suggestions for overcoming the latter. Rao & Arkin (63) review regulatory motifs in biology from an engineering standpoint and provide

general discussion on modeling. Tyson et al. (78, 79) review modeling studies of the eukaryotic cell cycle. The latter paper also gives a brief tutorial on nonlinear dynamical systems.

McAdams & Arkin (50) give a strong statement of the circuit analogy to biological networks and discuss issues like promoter control models and stochasticity in gene expression. Stochastic effects are further reviewed by these authors in another paper (51), where ways for organisms to control fluctuations and guarantee nonrandom outcomes are also discussed. In a later paper (52), McAdams & Arkin give brief comments on genetic circuit engineering and discuss some recently constructed artificial genetic networks. Hasty et al. (35) review these systems in depth. Their paper also reviews abstract dynamical modeling studies of systems with transcription factors and gives a brief description of various modeling methodologies.

Recent reviews of modeling metabolic networks, a large and important field, are given by Gombert & Nielsen (31) and Giersch (29), the latter with an emphasis on plants. A more general perspective is presented by Palsson (60), who discusses the importance of constraints to modeling biological systems for which little detailed information is available.

The notion of modularity in biology is discussed by Hartwell et al. (34). Asthagiri & Lauffenburger (10) review complexity in cell signaling and discuss modules for modeling. Lauffenburger (43) also discusses modularity of function and biological control principles. Finally, Ideker et al. (38) review the emerging paradigm of "systems biology," in which modeling plays an important part. Recent advances in high-throughput genetic manipulation, large-scale data gathering, and biological databases are discussed.

ACKNOWLEDGMENTS

We would like to thank the Defense Advanced Research Project Agency (grant 6513841) and the National Institutes of Health (grant GMO-000919) for support during the writing of this manuscript. Thanks to Robin Osterhout for a critical reading of the manuscript.

The *Annual Review of Genomics and Human Genetics* is online at http://genom.annualreviews.org

LITERATURE CITED

1. 2001. StochSim. http://www.zoo.cam.ac.uk/comp-cell/StochSim.html
2. 2001. CellML. http://www.cellml.org
3. 2001. MathML. http://www.w3.org/Math/
4. 2002. The MGED Group. http://www.mged.org
5. Ackers GK, Johnson AD, Shea MA. 1982. Quantitative model for gene regulation by lambda phage repressor. *Proc. Natl. Acad. Sci. USA* 79:1129–33

5a. Alifano P, Bruni C, Carlomagno M. 1994. Control of mRNA processing and decay in prokaryotes. *Genetica* 94:157–72

6. Altuvia S, Wagner EG. 2000. Switching on

and off with RNA. *Proc. Natl. Acad. Sci. USA* 97:9824–26
7. 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
8. Arkin AP. 2001. Synthetic cell biology. *Curr. Opin. Biotechnol.* 12:638–44
9. Artavanis-Tsakonas S, Rand MD, Lake RJ. 1999. Notch signaling: cell fate control and signal integration in development. *Science* 284:770–76
10. Asthagiri AR, Lauffenburger DA. 2000. Bioengineering models of cell signaling. *Annu. Rev. Biomed. Eng.* 2:31–53
11. Bartol TM, Stiles JR. 2002. MCell. http://www.mcell.cnl.salk.edu
12. Beelman CA, Parker R. 1995. Degradation of mRNA in eukaryotes. *Cell* 81:179–83
13. Bernstein E, Denli AM, Hannon GJ. 2001. The rest is silence. *RNA* 7:1509–21
14. Blomfield IC. 2001. The regulation of pap and type 1 fimbriation in *Escherichia coli*. *Adv. Microb. Physiol.* 45:1–49
15. Cao D, Parker R. 2001. Computational modeling of eukaryotic mRNA turnover. *RNA* 7:1192–212
16. Carrier TA, Keasling JD. 1999. Investigating autocatalytic gene expression systems through mechanistic modeling. *J. Theor. Biol.* 201:25–36
17. Chen KC, Csikasz-Nagy A, Gyorffy B, Val J, Novak B, Tyson JJ. 2000. Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol. Biol. Cell* 11:369–91
18. Ciliberto A, Tyson JJ. 2000. Mathematical model for early development of the sea urchin embryo. *Bull. Math. Biol.* 62:37–59
19. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, et al. 2002. A genomic regulatory network for development. *Science* 295:1669–78
20. Elowitz MB, Leibler S. 2000. A synthetic oscillatory network of transcriptional regulators. *Nature* 403:335–38
21. Endy D, Brent R. 2001. Modelling cellular behaviour. *Nature* 409(Suppl.):391–95
22. Endy D, Kong D, Yin J. 1997. Intracellular kinetics of a growing virus: a genetically structured simulation for bacteriophage T7. *Biotechnol. Bioeng.* 55:375–89
23. Endy D, Yin J. 2000. Toward antiviral strategies that resist viral escape. *Antimicrob. Agents Chemother.* 44:1097–99
24. 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
25. Friedman N, Linial M, Nachman I, Pe'er D. 2000. Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7:601–20
26. Gardiner-Garden M, Littlejohn TG. 2001. A comparison of microarray databases. *Brief Bioinform.* 2:143–58
27. Gardner TS, Cantor CR, Collins JJ. 2000. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* 403:339–42
28. Ghosh R, Tomlin CJ. 2001. *Lateral inhibition through delta-notch signaling: a piecewise affine hybrid model*. Presented at Int. Workshop Hybrid Systems: Comput. Control, 4th, Rome, Italy
29. Giersch C. 2000. Mathematical modelling of metabolism. *Curr. Opin. Plant Biol.* 3: 249–53
30. Gifford DK. 2001. Blazing pathways through genetic mountains. *Science* 293: 2049–51
31. Gombert AK, Nielsen J. 2000. Mathematical modelling of metabolism. *Curr. Opin. Biotechnol.* 11:180–86
32. Goryanin I, Hodgman TC, Selkov E. 1999. Mathematical simulation and analysis of cellular metabolism and regulation. *Bioinformatics* 15:749–58
33. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. 2001. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Proc. Pac. Symp. Biocomput. 2001, Mauna Lani*, pp. 422–33
34. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. 1999. From molecular to modular cell biology. *Nature* 402:C47–52

35. Hasty J, McMillen D, Isaacs F, Collins JJ. 2001. Computational studies of gene regulatory networks: in numero molecular biology. *Nat. Rev. Genet.* 2:268–79
36. Huang S, Ingber DE. 2000. Shape-dependent control of cell growth, differentiation, and apoptosis: switching between attractors in cell regulatory networks. *Exp. Cell Res.* 261:91–103
37. Hucka M, Finney A, Sauro H, Bolouri H. 2001. SBML Level 1. http://www.cds.caltech.edu/erato/sbml/docs/papers/sbml-level-1/html/sbml-level-1.html
38. Ideker T, Galitski T, Hood L. 2001. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* 2:343–72
39. Juty NS, Spence HD, Hotz HR, Tang H, Goryanin I, Hodgman TC. 2001. Simultaneous modelling of metabolic, genetic and product-interaction networks. *Brief Bioinform.* 2:223–32
40. Koh BT, Tan RB, Yap MG. 1998. Genetically structured mathematical modeling of trp attenuator mechanism. *Biotechnol. Bioeng.* 58:502–9
41. Koh BT, Yap MG. 1993. A simple genetically structured model of trp repressor-operator interaction. *Biotechnol. Bioeng.* 41:1115–18
42. Kyoda K, Kitano H. 1999. Simulation of genetic interaction for drosophila leg formation. *Proc. Pac. Symp. Biocomput. 1999, Mauna Lani*, pp. 77–89
43. Lauffenburger DA. 2000. Cell signaling pathways as control modules: complexity for simplicity? *Proc. Natl. Acad. Sci. USA* 97:5031–33
44. Le Novere N, Shimizu TS. 2001. StochSim: modelling of stochastic biomolecular processes. *Bioinformatics* 17:575–76
45. Loew LM, Schaff JC. 2001. The Virtual cell: a software environment for computational cell biology. *Trends Biotechnol.* 19: 401–6
46. Marnellos G, Deblandre GA, Mjolsness E, Kintner C. 2000. Delta-notch lateral inhibitory patterning in the emergence of ciliated cells in Xenopus: experimental observations and a gene network model. *Proc. Pac. Symp. Biocomput. 2000, Honolulu*, pp. 329–40
47. Marnellos G, Mjolsness E. 1998. A gene network approach to modeling early neurogenesis in Drosophila. *Proc. Pac. Symp. Biocomput. 1998, Maui*, pp. 30–41
48. Matsuno H, Doi A, Nagasaki M, Miyano S. 2000. Hybrid petri net representation of gene regulatory network. *Proc. Pac. Symp. Biocomput. 2000, Honolulu*, pp. 341–52
49. McAdams HH, Arkin A. 1997. Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA* 94:814–19
50. McAdams HH, Arkin A. 1998. Simulation of prokaryotic genetic circuits. *Annu. Rev. Biophys. Biomol. Struct.* 27:199–224
51. McAdams HH, Arkin A. 1999. It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet.* 15:65–69
52. McAdams HH, Arkin A. 2000. Towards a circuit engineering discipline. *Curr. Biol.* 10:R318–20
53. McAdams HH, Shapiro L. 1995. Circuit simulation of genetic networks. *Science* 269:650–56
54. Mendes P. 1997. Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends Biochem. Sci.* 22:361–63
55. Mendes P, Kell DB. 2001. MEG (Model Extender for Gepasi): a program for the modelling of complex, heterogeneous, cellular systems. *Bioinformatics* 17:288–89
56. Mitchell P, Tollervey D. 2000. mRNA stability in eukaryotes. *Curr. Opin. Genet. Dev.* 10:193–98
57. Mitchell P, Tollervey D. 2001. mRNA turnover. *Curr. Opin. Cell Biol.* 13:320–25
58. Molineux I. 1999. T7 bacteriophages. In *Encyclopedia of Molecular Biology*, ed. TE Creighton, pp. 2495–507. New York: Wiley
59. Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E Jr, et al. 2000. WIT: integrated system for high-throughput genome

sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* 28:123–25
60. Palsson B. 2000. The challenges of in silico biology. *Nat. Biotechnol.* 18:1147–50
61. Polach KJ, Widom J. 1996. A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *J. Mol. Biol.* 258:800–12
62. Ptashne M. 1992. *A Genetic Switch*. Boston, MA: Blackwell
63. Rao CV, Arkin AP. 2001. Control motifs for intracellular regulatory networks. *Annu. Rev. Biomed. Eng.* 3:391–419
64. Reinitz J, Mjolsness E, Sharp DH. 1995. Model for cooperative control of positional information in drosophila by *bicoid* and maternal *hunchback*. *J. Exp. Zool.* 271:47–56
65. Reinitz J, Sharp DH. 1995. Mechanism of eve stripe formation. *Mech. Dev.* 49:133–58
66. Reinitz J, Vaisnys JR. 1990. Theoretical and experimental analysis of the phage lambda genetic switch implies missing levels of cooperativity. *J. Theor. Biol.* 145: 295–318
67. Santillán M, Mackey MC. 2001. Dynamic regulation of the tryptophan operon: a modeling study and comparison with experimental data. *Proc. Natl. Acad. Sci. USA* 98: 1364–69
68. Schaff J, Loew LM. 1999. The virtual cell. *Proc. Pac. Symp. Biocomput. 1999, Mauna Lani*, pp. 228–39
69. Sharp DH, Reinitz J. 1998. Prediction of mutant expression patterns using gene circuits. *Biosystems* 47:79–90
70. Shea MA, Ackers GK. 1985. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J. Mol. Biol.* 181:211–30
71. Smolen P, Baxter DA, Byrne JH. 1998. Frequency selectivity, multistability, and oscillations emerge from models of genetic regulatory systems. *Am. J. Physiol.* 274:C531–42
72. Smolen P, Baxter DA, Byrne JH. 1999. Effects of macromolecular transport and stochastic fluctuations on dynamics of genetic regulatory systems. *Am. J. Physiol.* 277:C777–90
73. Smolen P, Baxter DA, Byrne JH. 2000. Mathematical modeling of gene networks. *Neuron* 26:567–80
74. Smolen P, Baxter DA, Byrne JH. 2000. Modeling transcriptional control in gene networks—methods, recent results, and future directions. *Bull. Math. Biol.* 62:247–92
75. Thieffry D, Thomas R. 1995. Dynamical behaviour of biological regulatory networks. II. Immunity control in bacteriophage lambda. *Bull. Math. Biol.* 57:277–97
76. Thomas R. 1973. Boolean formalization of genetic control circuits. *J. Theor. Biol.* 42:563–85
77. 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
78. Tyson JJ. 1999. Models of cell cycle control in eukaryotes. *J. Biotechnol.* 71:239–44
79. Tyson JJ, Chen K, Novak B. 2001. Network dynamics and cell physiology. *Nat. Rev. Mol. Cell Biol.* 2:908–16
80. Van Dien SJ, Keasling JD. 1998. A dynamic model of the *Escherichia coli* phosphate-starvation response. *J. Theor. Biol.* 190:37–49
81. von Dassow G, Meir E, Munro EM, Odell GM. 2000. The segment polarity network is a robust developmental module. *Nature* 406:188–92
82. Wang J, Ellwood K, Lehman A, Carey MF, She ZS. 1999. A mathematical model for synergistic eukaryotic gene activation. *J. Mol. Biol.* 286:315–25
83. Wolf DM, Arkin A. 2002. Fifteen minutes of *fim*: control of type 1 pili expression in *E. coli*. *Omics* 6:91–114
84. Wong P, Gladney S, Keasling JD. 1997. Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose. *Biotechnol. Prog.* 13:132–43

85. Yuh CH, Bolouri H, Davidson EH. 1998. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 279:1896–902

86. 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

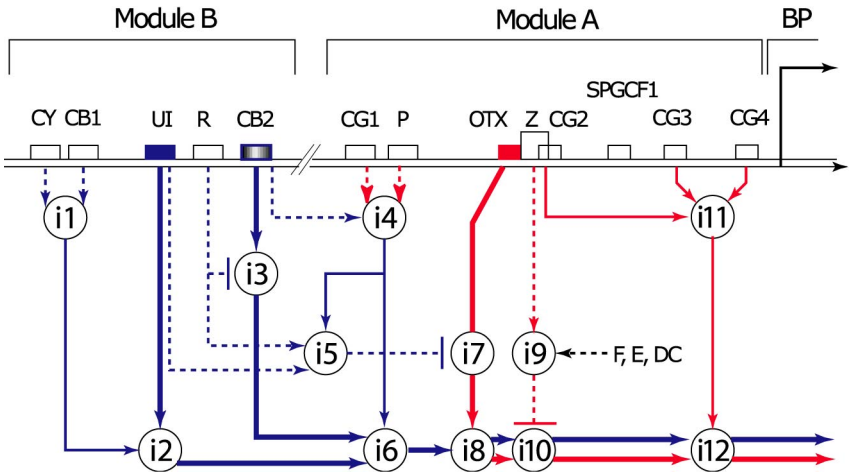


Figure 2 Computational model diagram of the BA region of the *endo16 cis*-regulatory system, redrawn from (86). The double line represents the DNA upstream of the *endo16* gene, and the labeled boxes are individual binding sites. The components of Module B and its effects are shown in *blue*. Those of Module A are in *red*. BP is the basal promoter. The labeled circles represent nodes where intermediate logical or algebraic computations are performed by the model. Heavy lines show the flow of time-dependent numerical information, dashed lines show Boolean information, and solid lines indicate time-invariant numerical information. See (86) for the specific functions calculated at each intermediate node and for further details.