RULES FOR THE EVOLUTION OF GENE CIRCUITRY

M.A. SAVAGEAU
Department of Microbiology & Immunology, The University of Michigan Medical School, Ann Arbor, Michigan 48109-0629 USA

Cells possess the genes required for growth and function in a variety of contexts. In any given context there is a corresponding pattern of gene expression in which some genes are OFF and others ON. The ability of cells to switch genes ON and OFF in a coordinate fashion to produce the required patterns of expression is the fundamental basis for complex processes like normal development and pathogenesis. The molecular study of gene regulation has revealed a plethora of mechanisms and circuitry that have evolved to perform what appears to be the same switching function. To some this implies the absence of rules. However, simple rules capable of relating molecular design to the natural environment have begun to emerge through the analysis of elementary gene circuits. Two of these rules are reviewed in this paper. These simple rules have the ability to unify understanding across several different levels of biological organization -- molecular, physiological, developmental, ecological.

1. Introduction

Regulation of gene expression and its systemic manifestations are subjects of intense study. As a result of this effort we shall soon have identified all of the genes and proteins for a number of simpler organisms. Despite this enormous progress we are still at a loss to understand the integrated behavior of the organism. Our understanding is still fragmentary and descriptive. We are unable to predict changes in the organism's behavior when it is placed in a novel environment or when a change is made in one of its genes. Little is known about the forces that lead to the selection or maintenance of a specific mechanism for the regulation of a given set of genes in a particular organism. Is this process random, or is it governed by rules? The answer to this question is important. It will help us to understand the evolution of gene regulation; it also will help us to develop judicious methods of redirecting normal expression for biotechnological purposes or of correcting pathological expression for therapeutic purposes.

Our goal is to understand the integrated structure and function of organizationally complex systems in relation to their underlying molecular determinants. Moreover, we are particularly interested in identifying the rule-like properties of these systems that would allow for some algorithmic compression in their representation, and not simply a compilation of all the molecular details.

In pursuit of this goal we have developed a canonical nonlinear formalism that has desirable properties for the representation and analysis of organizationally complex systems (1). This formalism has been used to characterize alternative

modes of gene control and various forms of coupling among elementary gene circuits. The results allow us to identify a set of rules, or design principles, that govern the natural selection of gene circuits. Here we shall review the relevant biological background and then present results from our analysis of gene circuitry.

2. Biological Background

The common metaphor of the genome as a blueprint for construction of the organism masks the difficult task of relating structure and function of the intact organism to its underlying genetic determinants (2). The behavior of an intact biological system can seldom be related directly to its underlying molecular determinants. There are several different levels of hierarchical organization that are relevant. For our present purposes it will be sufficient to consider four different levels -- genome sequence, transcriptional unit, elementary gene circuit, environmental context.

2.1. The DNA sequence constitutes the genome

The recent sequencing of the complete genome for a number of simpler organisms, and the projected completion of the sequence for the human genome by the year 2005, illustrate the power of modern molecular biology to resolve complex systems into their simplest elements. The four bases -- A, T, G, and C -- are strung together in sequences that are mind-numbing in their simplicity; yet, these sequences provide the potential for incredible complexity. Whether it be the versatile metabolism of free-living microbes that can adapt to nearly any environment, or the sophisticated structures of multicellular organisms that can be seen in near endless variety, the physical basis for this complexity is the context-dependent expression of the organism's genome.

2.2. Information is encoded in transcriptional units

The mapping from DNA level to organismal level requires a deeper understanding of how information is encoded in the genome. DNA sequences are organized into functional units that consist of structural genes flanked by a start sequence at which transcription begins and a termination sequence at which it ends. In addition, there are a number of regulatory sites capable of binding specific transcription factors that interact with the transcription machinery to modulate the rate of transcription initiation or termination (Fig. 1).

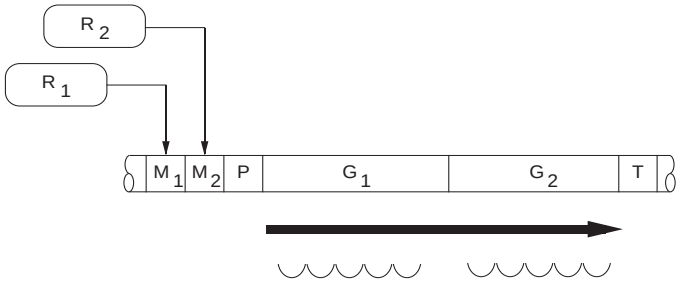

Figure 1. Unit of transcription. Structural genes (G_1 and G_2) are bounded by a promoter sequence (P) and a terminator sequence (T), and preceded by upstream modulator sites (M_1 and M_2) that bind regulators (R_1 and R_2) capable of altering transcription initiation. The solid arrow represents the mRNA transcript and the scalloped lines indicate the protein products encoded by genes G_1 and G_2.

2.3. Expression is organized into elementary gene circuits

Transcription of DNA is but one step in a cascade of information flow that constitutes the expression of a gene (Fig. 2). Each stage of such a cascade is a potential site at which expression can be regulated in a context-dependent fashion. The context is provided by the life cycle of the organism, and the interlocking mechanisms of gene regulation interpret that context.

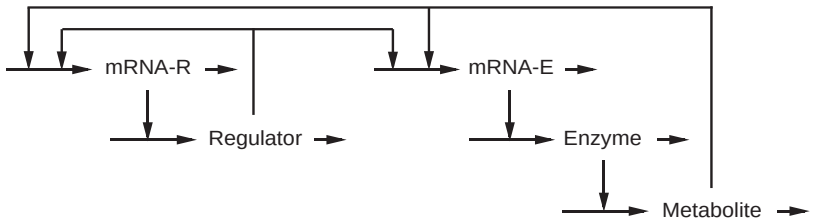

Figure 2. Cascade of information flow from DNA to RNA to protein to metabolite. The processes of synthesis and degradation are represented by horizontal arrows, whereas the catalytic and regulatory influences are represented by vertical arrows. An effector circuit is shown on right and a regulator circuit is shown on the left.

2.4. Physiology and ecology are reflected in the organism's life cycle

The life cycle of some organisms is largely programmed development from egg to embryo to mature adult and back to the egg (3). In other organisms it is dominated by random events involving a pathogen's ability to encounter one host, to exploit or

colonize that host for a period of time, to escape into a secondary environment, and to survive there until an encounter with a subsequent host (4). In each case, specific genes function in some phases of the organism's life cycle but not in others. Differential patterns of expression are exhibited as the context changes from one phase to the next and one set of genes is switched OFF while another set is switched ON in a combinatorial fashion.

Gene regulation -- the ability to switch gene expression ON and OFF in appropriate temporal and spatial patterns -- is central to modern biology. The inability to express a gene when it should be ON, or the inappropriate expression of a gene when it should be OFF, is usually dysfunctional and often lethal. The determination of what constitutes appropriate expression requires knowledge of the molecular mechanism, the physiological function it realizes, and the environmental demand for that function.

Organisms regulate expression of their genome by means of a diverse repertoire of molecular mechanisms. Most of the well-characterized examples have come from the study of prokaryotes. Although the situation is typically more complex in eukaryotes and there are undoubtedly some aspects of regulation unique to higher organisms, the general themes are much the same in both and most mechanisms that were originally thought to be unique to eukaryotes have subsequently been observed within the prokaryotic realm. For our analysis, we have abstracted the generic features of gene regulation that are thought to be common to both, but for testing our predictions we have turned to the more numerous and well-characterized prokaryotes systems. The extent to which the results might differ for eukaryotes remains to be determined.

3. Rules for the Molecular Mode of Gene Control

One of the first variations in design to be well documented is that involving positive vs. negative modes of gene control (Fig. 3). For example, the lactose (*lac*) catabolic system in *Escherichia coli* is governed by a classical repressor protein (5), the negative mode of control. Induction of gene expression in this system is achieved by the addition of an inducer that removes the repressor protein to allow transcription. The maltose (mal) system in *E. coli*, by contrast, is governed by an activator protein (6), the positive mode of control. Induction in this case is achieved by the addition of an inducer that converts the activator protein into its functional form that facilitates transcription. What is the significance of this variation in design?

This difference in design was originally believed to have no functional significance. Subsequent analysis showed that mode of control is related to the of control showed that in most respects their behavior can be identical. However,

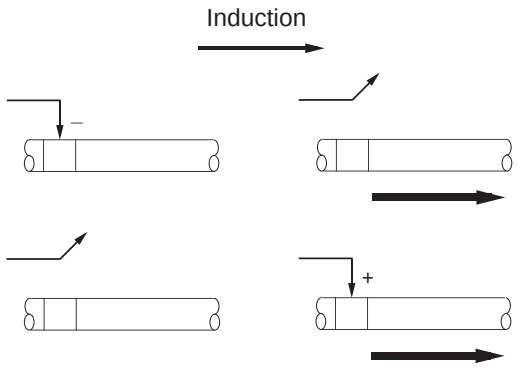

Figure 3. Alternative molecular modes of controlling gene expression.

demand for expression of the regulated gene in the organism's natural environment (7). The analysis of mathematical models with either the positive or negative mode they behave in diametrically opposed ways to mutations in the components of the regulatory mechanism itself. Mutants altered in the positive mechanism are unable to express the corresponding gene product despite the presence of inducer, whereas mutants altered in the negative mechanism express the corresponding gene product even in the absence of inducer. The relative growth of mutant and wild-type organisms was examined in high- and low-demand environments. The high-demand environment, in which high-level expression is frequently required for the organism's survival, leads to selection of the positive mode of gene control; the low-demand environment leads to selection of the negative mode. Thus, molecular mode of control is correlated with level of demand for expression of the regulated gene product in the organism's natural environment (Table 1). These qualitative predictions are well supported by experimental evidence (8).

Table 1. Predicted correlation between demand for expression and mode of control

Demand for expression	Mode of regulation	
	Positive	Negative
High	Regulation selected	Regulation lost
Low	Regulation lost	Regulation selected

In recent analysis we have examined the quantitative implications of this demand theory (in preparation). First, we define two key parameters: the cycle time C, which is the average time for a gene to cycle through the OFF state, the ON state, and back to the OFF state; and demand D, which is the fraction of the cycle time that the gene is ON. Second, a quantitative analysis involving mutation rates and growth rates reveals non-overlapping regions in the C vs. D space for which selection of wild-type regulatory mechanisms with the negative or the positive mode is realizable (Fig. 4).

The quantitative theory specifies more precisely what we mean by high and low demand. As can be seen in Figure 4, with the nominal values for the parameters of the lactose and maltose operons in *E. coli*, selection of the negative mode of control requires a demand less than 0.04, whereas selection of the positive mode requires a demand greater than 0.32.

Although these limits on demand are influenced by a number of parameters, by far the most influential parameter is the reduction in growth rate when there is excess expression of a gene whose function is not required. The nominal value for

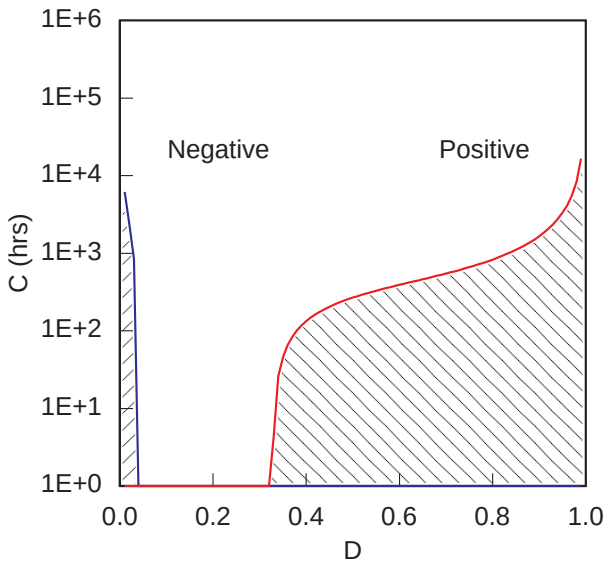

Figure 4. Thresholds for discriminate selection of wild-type regulatory mechanisms with negative or positive modes. There is maximum value of demand for selection of the negative mode and a minimum value of demand for selection of the positive mode. The values of cycle time C and demand D are based on parameter values for the *lac* and *mal* systems in *E. coli*.

this parameter was set at 5%, based on data for the lactose operon that suggest this value as a maximum for the reduction in growth rate of operator-constitutive mutants in a low-demand environment. In the case of the positive mode, the same value was used to characterize the reduction in growth rate of an up-promoter mutant in a low-demand environment. A 10% variation in this parameter yields a two-fold change in the limits of D for both the negative and positive mode. The remaining parameters have much less influence on the limits of D; approximately half exhibit a nearly linear influence, whereas the other half have a negligible influence.

4. Rules for the Coupling of Elementary Gene Circuits

A second variation in design is that involving the coupling of elementary gene circuits for regulator and effector genes. Early experimental studies (9) suggested that expression of regulator genes is invariant in some cases (classical regulation), such as in the *lac* system in *E. coli*, and coordinate with the regulated effector genes in other cases (autogenous regulation), such as in the histidine utilization (*hut*) system in *Salmonella typhimurium*. Our earlier work focused on the functional implications of these alternatives, which we now refer to as the completely uncoupled and perfectly coupled patterns of regulator and effector gene expression (10). However, inducible systems with other patterns of gene expression were subsequently reported, and these have become the stimulus to extend our earlier work.

Logically, there are three qualitatively distinct patterns of regulator and effector gene expression that can be exhibited by an inducible system (Fig. 5). These are the directly coupled, uncoupled, and inversely coupled patterns in which regulator gene expression increases, remains the same, and decreases with an increase in effector gene expression. Well-studied examples of direct coupling, uncoupling, and inverse coupling are provided by the D-serine deaminase (11), arabinose (6), and methionine (12) systems in *E. coli*.

The functional implications of direct coupling, uncoupling, and inverse coupling have been determined from an analysis of a generalized model capable of representing these different forms of coupling (Fig. 6). The fundamental equations that characterize this model are mass-balance equations that take the general form

$$dX_i/dt = V_{+i}(X_1, \ldots, X_8) - V_{-i}(X_1, \ldots, X_8) \qquad i = 1, \ldots, 5 \qquad (1)$$

The rate laws V_{+i} and V_{-i} describe mass fluxes due to synthetic and degradative processes. These rate laws can be represented as products of power-law functions according to the results of theoretical analyses (1) and empirical case studies (13). Thus, we can rewrite Eq. 1 to obtain the following system of equations:

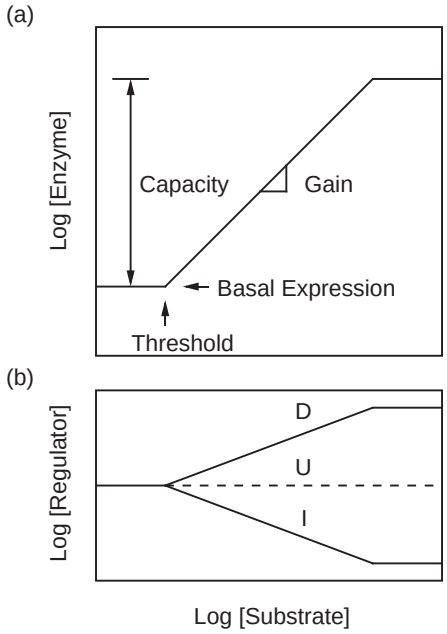

Figure 5. Expression characteristics for (a) effector and (b) regulator gene expression. Three distinct patterns of coupling are illustrated. Effector gene expression increases while regulator gene expression (D) increases (directly coupled), (U) remains unchanged (uncoupled), or (I) decreases (inversely coupled).

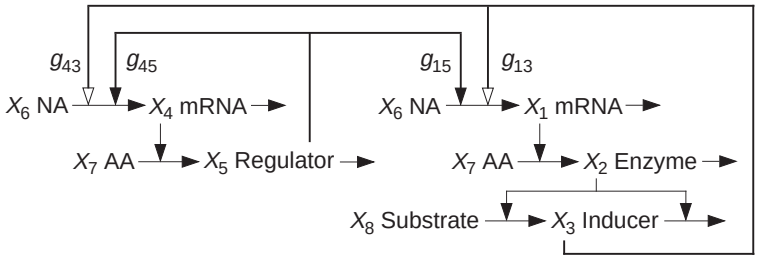

Figure 6. Coupled circuits for the expression of regulator and effector genes. Mass fluxes that characterize the state of the system are represented by horizontal arrows, whereas catalytic and regulatory influences are represented by vertical arrows. The influences of the regulator (closed arrowheads) are described by the kinetic orders g_{15} and g_{45}; the influences of the inducer (open arrowheads) are described by the kinetic orders g_{13} and g_{43} (see Eqs. 2-6).

$$dX_1/dt = V_{+1} - V_{-1} = \alpha_1 X_6^{g16} X_3^{g13} X_5^{g15} - \beta_1 X_1^{h11} \quad (2)$$

$$dX_2/dt = V_{+2} - V_{-2} = \alpha_2 X_7^{g27} X_1^{g21} - \beta_2 X_2^{h22} \quad (3)$$

$$dX_3/dt = V_{+3} - V_{-3} = \alpha_3 X_8^{g38} X_2^{g32} - \beta_3 X_2^{h32} X_3^{h33} \quad (4)$$

$$dX_4/dt = V_{+4} - V_{-4} = \alpha_4 X_6^{g46} X_3^{g43} X_5^{g45} - \beta_4 X_4^{h44} \quad (5)$$

$$dX_5/dt = V_{+5} - V_{-5} = \alpha_5 X_7^{g57} X_4^{g54} - \beta_5 X_5^{h55} \quad (6)$$

These equations are used to analyze systems with the positive or negative mode of control for each circuit. The effects of physicochemical limitations, which arise from the subunit structure of regulator proteins and place bounds on kinetic orders in this model (10), are also considered. The functional effectiveness of these various circuits has been compared on the basis of several properties (decisiveness, efficiency, selectivity, robustness, stability, and responsiveness) that represent possible criteria for natural selection. Of these, responsiveness has proved the most sensitive to variations in circuit design (14).

The results allow us to predict a correlation between the form of coupling and the capacity for induction (ratio of maximal to minimal level of effector gene expression). Negatively controlled systems with low, intermediate, and high capacities for gene expression are predicted to have direct coupling, uncoupling, and inverse coupling, respectively. Positively controlled systems, in contrast, are predicted to have inverse coupling, uncoupling, and direct coupling (Table 2).

These predictions are compared with data available in the literature for systems in which the pattern of regulator and effector gene expression is known (Fig. 7). They are found to be in reasonable agreement, given measurement error.

Table 2. Predicted correlation between circuitry and capacity for regulation

Demand	Mode	Capacity	Circuit
High	Positive	Low	Inversely coupled
High	Positive	Intermediate	Uncoupled
High	Positive	High	Direct coupled
Low	Negative	Low	Directly coupled
Low	Negative	Intermediate	Uncoupled
Low	Negative	High	Inversely coupled

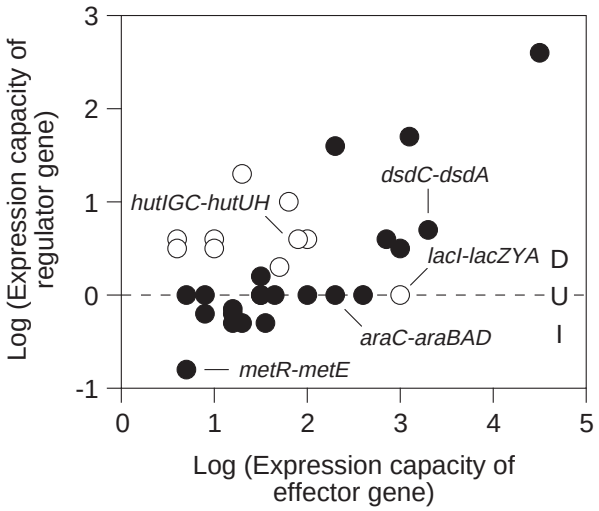

Figure 7. Patterns of regulator and effector gene expression in inducible systems of bacteria. Expression of each gene is measured in the presence of excess inducer and normalized with respect to its basal level in the absence of inducer. For the effector gene this is equivalent to its capacity; for the directly coupled regulator gene this also is equivalent to its capacity, but for the inversely coupled regulator gene this is equivalent to the inverse of its capacity. Estimates of capacity are based on published reports. Directly coupled (D), uncoupled (U), and inversely coupled (I) systems are represented above, on, and below the dashed line, respectively. Negatively regulated systems are shown as open circles; positively regulated systems are shown as closed circles.

5. Discussion

The genome of an organism evolves to realize a developmental program with specific gene circuitry that can be viewed as computing the solution to the environmental problem faced by the organism. This is a suggestive metaphor, but at present we have little understanding of the circuits and the computations they might perform. The large number of genes encoded in the DNA of even the simplest of organisms suggests that this circuitry might be very complex and exhibit a high degree of connectivity. If this were the case, then the task of elucidating the circuitry would be daunting.

However, a number of different lines of evidence suggest that although there may be a large number of gene circuits, they may have a minimal degree of connectivity. First, molecular analysis of gene regulation in bacteria has shown that most gene circuits are governed by a small number of regulators, usually one to

three. In eukaryotes the numbers are larger in some cases, but seldom more than a dozen regulators influence a given gene circuit. Second, the enumeration of regulators and their targets, based on sequence homologies, has shown the same results for bacteria; namely, one or two regulators affecting a given circuit (15,16). Third, computer simulations of large, randomly-connected circuits have been used to explore the question of connectivity. The most biologically-suggestive behaviors were found when each circuit was subject to two or three regulatory interactions, and less relevant behaviors were found with higher or lower degrees of connectivity (17).

Low degrees of connectivity suggest that a 'bottom-up' strategy of characterizing genome circuitry in terms of rules for elemental gene circuits is likely to prove fruitful. Indeed, this seems to be the case with our initial experience attempting to generalize on the basis of the few rules that we have uncovered to date. To give one example, consider the carbon regulation system in *E. coli*.

Carbon regulation in *E. coli* is manifested in large part through the action of the cyclic AMP receptor protein (CRP)-cyclic AMP (cAMP) system (18), which was among the first global regulators to be characterized. This system coordinates the utilization of diverse sources of carbon whose levels vary in both time and space. An application of demand theory indicates that all of the regulators in this system fit a self-consistent pattern. Because the CRP-cAMP regulator is an activator of transcription for the inducible catabolic systems, one can predict that at least some of these systems are in high demand in the organism's natural environment. Indeed, a number of the inducible systems for non-PTS substrates are controlled by specific activators (8). Conversely, one can predict that the PTS substrates, which repress the levels of CRP-cAMP, are seldom present in high concentrations in the natural environment. Indeed, all of the inducible systems for PTS substrates that have been examined involve control by a specific repressor (8), which again is what one would predict according to demand theory. Thus, at least the modality of all the regulators in this system seem to be self-consistent.

In conclusion, regulation of gene expression is clearly one of the most fundamental processes in the living world. Knowledge of gene regulation is a prerequisite for understanding function, adaptation and evolution, and such understanding will in turn be essential for the design and implementation of novel metabolic pathways by means of genetic engineering. The results of our studies suggest that although there is an enormous diversity of mechanisms, there also are well-established patterns that can be understood in terms of simple rules.

Acknowledgments

This work was supported in part by grants from the U.S. National Institutes of Health and Office of Naval Research.

References

1. M.A. Savageau, in *World Congress of Nonlinear Analysts 92*, Vol. 4, Ed. V. Lakshmikantham (Walter de Gruyter Publishers, Berlin, 1996), pp. 3323-3334
2. F.C. Neidhardt and M.A. Savageau, in *Escherichia coli and Salmonella; cellular and molecular biology*, Vol. 1, 2nd ed., Eds. F. C. Neidhardt, et al. (ASM Press, Washington, D.C., 1996), pp. 1310-1324
3. J.M.W. Slack, *From Egg to Embryo*, 2nd ed. (Cambridge University Press, Cambridge, 1992)
4. A.A. Salyers, *Bacterial Pathogenesis: A Molecular Approach* (ASM Press, Washington, D.C., 1994)
5. H. Choy and S. Adhya, in *Escherichia coli and Salmonella; cellular and molecular biology*, Vol. 1, 2nd ed., Eds. F. C. Neidhardt, et al. (ASM Press, Washington, D.C., 1996), pp. 1287-1299
6. R. Schleif, in *Escherichia coli and Salmonella; cellular and molecular biology*, Vol. 1, 2nd ed., Eds. F. C. Neidhardt, et al. (ASM Press, Washington, D.C., 1996), pp. 1300-1309
7. M.A. Savageau, *Proc. Nat. Acad. Sci. USA* **74**, 5647-5651 (1977)
8. M.A. Savageau, in *Theoretical Biology -- Epigenetic and Evolutionary Order*, Eds. B.C. Goodwin and P.T. Saunders (Edinburgh University Press, Edinburgh, 1989), pp. 42-66
9. R.F. Goldberger, *Science* **183**, 810-816 (1974)
10. W.S. Hlavacek and M.A. Savageau, *J. Mol. Biol.* **248**, 739-755 (1995)
11. E. McFall, in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, Eds. F.C. Neidhardt et al. (American Society for Microbiology, Washington, D.C., 1987), pp. 1520-1526
12. I.G. Old, S.E. Phillips, P.G. Stockley and I. Saint Girons, *Prog. Biophys. Mol. Biol.* **56**, 145-185 (1991)
13. E.O. Voit, *Canonical Nonlinear Modeling: S-System Approach to Understanding Complexity* (Van Nostrand Reinhold, New York, 1991)
14. W.S. Hlavacek and M.A. Savageau, *J. Mol. Biol.* **255**, 121-139 (1996)
15. J. Collado-Vides, B. Magasanik, and J.D. Gralla, *Microbiol. Rev.* **55**, 371-394 (1991)
16. J. Otsuka, H. Watanabe, and K.T. Mori, *J. Theoret. Biol.* **178**, 183-204 (1996)
17. S.A. Kauffman, *J. Theoret. Biol.* **22**, 437-467 (1969)
18. A. Kolb, S. Busby, H. Buc, S. Garges, and S. Adhya, *Annu. Rev. Biochem.* **62**, 749-795 (1993)