On schemes of combinatorial transcription logic

Nicolas E. Buchler, Ulrich Gerland, and Terence Hwa*

Department of Physics and Center for Theoretical Biological Physics, University of California at San Diego, La Jolla, CA 92093-0319

Edited by Mark Ptashne, Memorial Sloan–Kettering Cancer Center, New York, NY, and approved March 4, 2003 (received for review January 17, 2003)

Cells receive a wide variety of cellular and environmental signals, which are often processed combinatorially to generate specific genetic responses. Here we explore theoretically the potentials and limitations of combinatorial signal integration at the level of cis-regulatory transcription control. Our analysis suggests that many complex transcription-control functions of the type encountered in higher eukaryotes are already implementable within the much simpler bacterial transcription system. Using a quantitative model of bacterial transcription and invoking only specific protein–DNA interaction and weak glue-like interaction between regulatory proteins, we show explicit schemes to implement regulatory logic functions of increasing complexity by appropriately selecting the strengths and arranging the relative positions of the relevant protein-binding DNA sequences in the cis-regulatory region. The architectures that emerge are naturally modular and evolvable. Our results suggest that the transcription regulatory apparatus is a ''programmable'' computing machine, belonging formally to the class of Boltzmann machines. Crucial to our results is the ability to regulate gene expression at a distance. In bacteria, this can be achieved for isolated genes via DNA looping controlled by the dimerization of DNA-bound proteins. However, if adopted extensively in the genome, long-distance interaction can cause unintentional intergenic cross talk, a detrimental side effect difficult to overcome by the known bacterial transcription-regulation systems. This may be a key factor limiting the genome-wide adoption of complex transcription control in bacteria. Implications of our findings for combinatorial transcription control in eukaryotes are discussed.

Biological organisms ranging from bacteria to humans possess an enormous repertoire of genetic responses to ever-changing combinations of cellular and environmental signals. To a large extent, this repertoire is encoded in complex networks of genes closely regulating the activities of each other. Characterizing and decoding the connectivity of gene regulatory networks has been an outstanding challenge of post-genome molecular biology (1–4). However, unlike integrated circuits, which process information through synchronized cascades of many simple and fast nodes and for which connectivity is the primary source of network complexity, a gene-regulatory network typically consists of only a few tens to hundreds of genes, the expression of which is slow and asynchronous (5). Yet these "nodes" are very sophisticated in their capacity to integrate signals: In eukaryotes, each node can be regulated combinatorially, often by four to five other nodes (1, 6), and the regulatory control function can be extremely complex (7). Here we focus primarily on one node of a gene-regulatory network and investigate quantitatively the power and limitations of combinatorial gene regulation in the context of bacterial transcription. We find that the bacterial transcription system is already capable of implementing many of the complex regulatory functions known for eukaryotes. At the end, we discuss factors limiting the genome-wide adoption of complex regulation for bacteria, and explore how they may be overcome by the eukaryotic transcription system.

Quantification of Combinatorial Transcription Control

The activity of a gene is regulated by other genes through the concentrations of their gene products, the transcription factors (TFs). This is accomplished mechanically by the interaction of the TFs with their respective DNA targets, with each other, and with the RNA polymerase (RNAP) complex in the regulatory region of the regulated gene. Regulation can be quantified by the "response characteristics," i.e., the level of gene expression as a function of the concentrations of (activated) TFs.[†] Although we consider protein concentrations as continuous variables, essential features of the response characteristics can often be represented more compactly by a binary "logic function," which specifies whether a gene is "ON" (expressed) or "OFF" (silent, or expressed at basal level) at different extremes of cellular TF concentrations, e.g., a "low" value of a few molecules per bacterium ($\approx$1 nM) or a "high" value of $\approx$1,000 molecules per bacterium ($\approx 1\ \mu$M). In Fig. 1*a* we show the logic-function representations of six different genetic responses (*g1–g6*) to two TFs, A and B. Some of these responses are commonly encountered in bacterial transcription control, e.g., the response of *g1* represents approximately the regulation of the well known *lac* operon by LacR (A) and CRP (B) in *Escherichia coli* (8). Here, the repressive effect of A is achieved by competitive binding of A and RNAP to the same region of DNA, and the activating effect of B results from the cooperative interaction between B and RNAP when they are both bound to their sites (see Fig. 1*b*).

Can similar schemes involving merely the arrangement of TF-binding sites in the cis-regulatory region be used to implement the other functions listed in Fig. 1*a* as well as more complex ones involving regulation by three or more TFs? Ptashne and Gann (9, 10) postulated that a wide range of regulatory functions might indeed be realizable, simply through the "regulated recruitment" of TFs and the RNAP, without invoking complex (e.g., allosteric) protein–protein interactions. To test this postulate, we formulated a quantitative model of regulated recruitment based on the well characterized bacterial transcription system. Specifically, we endow proteins with only weak "glue-like" contact interaction and explore the possibility of implementing control functions of increasing complexity via the appropriate arrangement of their DNA-binding targets. The model is briefly outlined below, with details provided in *Supporting Text*, which is published as supporting information on the PNAS web site, www.pnas.org (see also ref. 11).

Model. We adopt and generalize the approach of Shea and Ackers (12), describing transcription regulation in bacteria by a thermodynamic treatment. The degree of gene transcription is quantified by the equilibrium binding probability P of the RNAP to its DNA target, the promoter, given the cellular concentrations of all of the TFs. Crucial to our model are two ingredients that we regard as the quantitative formulation of regulated recruitment (9, 10).

1. The binding strength of a TF-binding site on the DNA (an operator) is assumed to be continuously tunable through choice of the binding sequence. In our model, we quantify the binding strength of a site i by an effective dissociation constant K_i, defined as the TF concentration for half-maximal binding of the site in the presence of the genomic background (see *Supporting Text* for details). As shown by experimental studies on exemplary TFs (13) and expected on theoretical grounds for a large class

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: TF, transcription factor; RNAP, RNA polymerase; DNF, disjunctive normal form; CNF, conjunctive normal form.

*To whom correspondence should be addressed. E-mail: hwa@ucsd.edu.

[†]We consider only the concentration of TFs in an activated state, i.e., a state that allows specific DNA binding and affects the expression of the regulated gene. Activation of TFs can be controlled by a number of mechanisms such as phosphorylation and ligand binding.

 www.pnas.org/cgi/doi/10.1073/pnas.0930314100

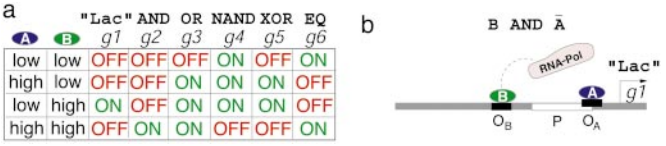

Fig. 1. (*a*) Some possible gene responses (ON or OFF) according to the specific activation patterns of two TFs, A and B, as denoted by their cellular concentrations (high or low). The logical equivalents of these gene responses are listed above each column. (*b*) The cis-regulatory implementation for the response of gene *g1*, as adapted from the *E. coli lac* operon. To achieve the desired effects, the operator sites need to be strong (filled boxes) and the promoter needs to be weak (open box). In this and subsequent cis-regulatory constructs, we use the offset, overlapping boxes to indicate mutual repression and the dashed lines to indicate cooperative interaction. The logic function that this system implements is indicated above the construct, with the overline denoting the ''inverse'' of A, or NOT A.

of bacterial TFs (14), K_i can typically be tuned across and beyond the relevant range of cellular protein concentrations (e.g., $K_i \approx$ 1–10,000 nM) individually for each site i.

2. A weak glue-like interaction between two proteins (TFs and/or RNAP) is assumed possible if the relative placements of the DNA-binding sites allow for direct contact of appropriate regions of the proteins. On the molecular level, weak glue-like interactions can occur, for instance, via contact of hydrophobic patches (15). For a number of well studied proteins (see refs. 10, 12, and 16 and references therein), such interactions fall within the range of ≈1–4 kcal/mol. Here we assume for simplicity the same interaction energy for all protein pairs and choose a conservative value of $E_{\rm int} = -2$ kcal/mol. A repulsive interaction ($E_{\rm int} = +\infty$) between two proteins results if their respective binding sites overlap. No effective interaction ($E_{\rm int} = 0$) is obtained when the binding sites for the two proteins are on opposite sides of the DNA or at an appropriate distance such that they will not bind to their sites and contact each other simultaneously. Quantifying the interaction between two proteins bound to two sites i and j by a cooperativity factor $\omega_{i,j} = e^{-E_{\rm int}/RT}$, where $RT \approx 0.6$ kcal/mol, we see that interaction between each pair of sites can be selected from the values $\omega_{i,j} = \{0, 1, \approx 20\}$ just by arranging the positions of the binding sites in the regulatory region.

Given the binding strengths K_i and the cooperativity factors $\omega_{i,j}$ for all the DNA sites, the binding probability P of the RNAP to the promoter can be computed straightforwardly (see refs. 11 and 12 and *Supporting Text*). The task of implementing various regulatory functions is then reduced to arranging the binding sites in the cis-regulatory region such that the interaction parameters K_i and $\omega_{i,j}$ produce the desired P for the various TF concentrations.

Cis-Regulatory Implementations. To illustrate how different regulatory functions can be implemented by using the model described above, let us consider the response of *g2* in Fig. 2*a*, which corresponds to the logic function AND, and the implementation of which is referred to as the AND gate. It can be obtained by choosing weak binding sites for both A and B and placing them adjacent to each other (see Fig. 2*a*) such that each TF alone cannot bind to its site, but when both are present binding occurs with the help of the additional cooperative interaction. This is quantitatively verified by the full response characteristics $P([A],[B])$ plotted across the range of physiological TF concentrations (≈1–1,000 nM). Similarly, one can implement the responses for the genes *g3* and *g4* corresponding to the OR and NAND gates (see Fig. 2 *b* and *c*). The maximal fold change obtained is ≈10 for all three logic gates. (With stronger interaction energy $E_{\rm int}$ or by using multiple binding sites, larger fold changes can be readily obtained for these and more complex logic gates; here we are concerned primarily with obtaining the qualitative behaviors rather than their optimization.) Examples of these control functions can be found in natural and artificially constructed regulatory systems in bacteria (17–19), and the basic molecular mechanisms of their operations are similar to those described above.

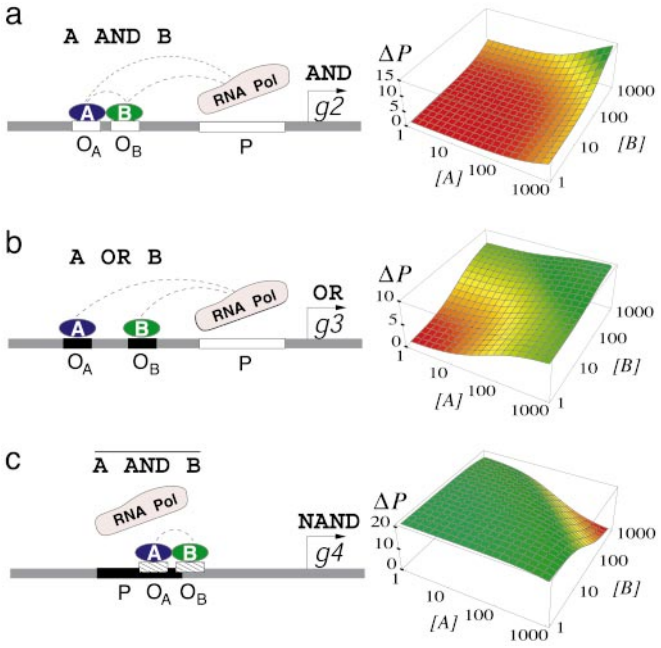

Fig. 2. Cis-regulatory constructs and response characteristics of the AND (*a*), OR (*b*), and NAND (*c*) gates. Filled, hatched, and open boxes denote strong, moderate, and weak binding sites, respectively. Dashed lines indicate cooperative interaction with $\omega_{i,j} = 20$, and overlapping boxes indicate repulsive interaction with $\omega_{i,j} = 0$. Plotted to the right of each construct is the fold change in RNAP-binding probability, $\Delta P \equiv P([A], [B])/P_{\rm min}$ for typical cellular TF concentrations $[A]$ and $[B]$ (in nM). See *Supporting Text* for the actual forms of $P([A], [B])$ and the strengths of the binding sites. Qualitative features of these plots are insensitive to the precise values of the parameters used.

The responses for *g5* and *g6* exemplify an increased level of complexity: The effect of a TF is not always activating or repressing (as is the case for *g1–g4*) but depends on the state of the other TF. For example, protein B activates *g5* in the absence of protein A but represses *g5* in the presence of A, making the gene ON if either one but not both of the TFs are activated; this control is known commonly as the "exclusive-or" (XOR) gate. Analogous to electronic circuit design, *g5* could be achieved via a "gene cascade," e.g., by applying the gene products of *g3* and *g4* on *g2* (see Fig. 3*a*). More simply, the regulatory regions of *g3* and *g4* could be combined into a single region as shown in Fig. 3*b*, which achieves the desired characteristics without any intermediate genes, thereby avoiding many potential problems associated with their expressions (e.g., time delay and stochasticity). The cis-regulatory implementation of the XOR gate is not unique, e.g., an alternative design uses two promoters positioned sequentially in the regulatory region, with one promoter functional only when B is activated and A is not (as in Fig. 1*b*) and *vice versa* for the other (see Fig. 3*c*).

The above example illustrates a fundamental difference in the style of computation between a gene-regulatory network and an electronic circuit: An electronic circuit features a "deep" architecture with many layers of cascades to take advantage of the vast number of simple but fast nodes. Despite what has been suggested previously (20), we believe a gene-regulatory network cannot afford many stages of cascades because of the slowness and limited number of nodes but can adopt a "broad" architecture integrating complex computations such as the XOR gate into a single node to overcome the slowness. The speed constraint is especially signifi-

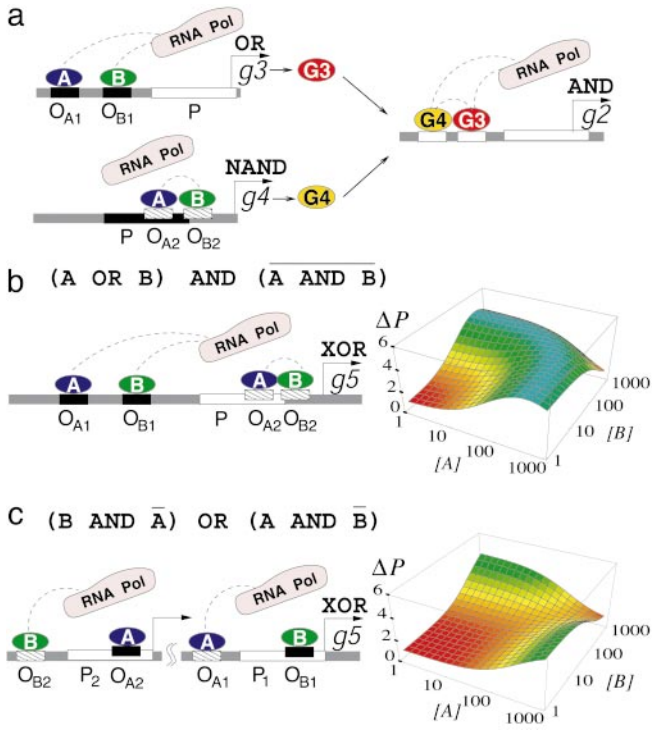

Fig. 3. Various strategies of implementing the XOR function. (*a*) A gene cascade, where the intermediate gene products G3 and G4 themselves are TFs that can interact cooperatively. Alternative cis-regulatory constructs using a single promoter (*b*) or two promoters (*c*) are shown. Notations are the same as those used for Fig. 2, whereas the squiggles in *c* indicate that the two promoters can be at variable distances from one another.

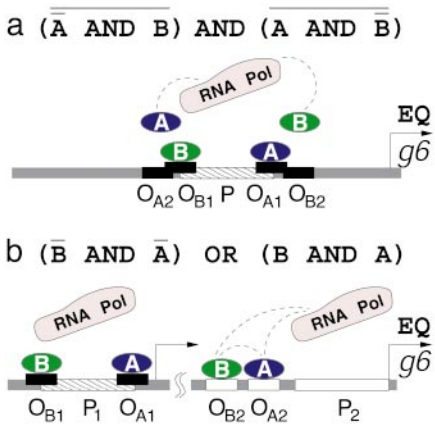

Fig. 4. Cis-regulatory constructs for possible implementations of the EQ gate using a single promoter (*a*) or two promoters (*b*). Notations are the same as those used for Figs. 2 and 3. Both constructs illustrate the problem of promoter overcrowding, a situation that occurs when multiple repressive conditions are needed.

cant in bacteria, where the time scale for gene expression is a significant portion of the generation time. Indeed, the preference for a broad but shallow network architecture has been observed recently in a large-scale analysis of the *E. coli* gene-regulatory circuits (4). Of course, a limited number of gene cascades can be used if speed is not a limiting factor (e.g., in eukaryotes) and may be especially useful in situations such as cell-cycle control (21) and early development (22), where natural temporal orders exist.

Limitations. There are limitations to the control functions one can implement by using only the two ingredients of regulated recruitment formulated thus far. This is illustrated with gene *g6*, the "equivalence" or EQ gate. A strong promoter is required here to turn the gene ON when neither of the TFs are activated, whereas repression is needed under multiple conditions (i.e., when A is activated and B is not, and vice versa). It is difficult to implement both repressive conditions by the direct physical exclusion of RNAP given the small size of the promoter region (see Fig. 4*a*). The situation is improved somewhat in an alternative approach involving two promoters, although multiple repressions are still needed (see Fig. 4*b*). This turns out to be a general problem for the implementation of more complex regulatory functions, which will generically require multiple repression conditions. An effective strategy to overcome promoter overcrowding is repression from a distance. One way to accomplish distal repression is through DNA looping mediated by protein dimerization; see, e.g., the homodimerization of AraC in *E. coli* (23).

A simple strategy to implement repression under multiple conditions is to use heterodimers, with two subunits each recognizing a distinct DNA site while associating strongly to each other (quantified by a cooperativity factor $\omega_{i,j} \approx 100$) as shown in Fig. 5*a*. In recent experiments, long-range regulation through heterodimers has been demonstrated *in vivo* by using either two regulatory proteins, each fused with a recognition domain according to the "two-hybrid" approach (24), or a single regulatory protein with two distinct binding domains (25). For our purposes, distal repression can be implemented by overlapping one of the binding sites, say the target of the S subunit, with the promoter. To control the repressive effect solely by the proteins A and B, one can set up a steady background concentration of the heterodimers and make the binding strength of the distal site weak such that the heterodimers only bind to their respective DNA targets when recruited by the appropriate TFs placed adjacent to the distal site. Binding sites for A and B can also be placed overlapping with the distal site to turn off distal repression under desired conditions. A cis-regulatory construct and the corresponding response characteristics of the EQ gate, using the distal repression scheme, is shown in Fig. 5*b* with multiple binding sites for the R subunit used to enforce multiple repression conditions (see *Supporting Text* for details). Alternatively, the EQ gate could be implemented by using a distal activation scheme as shown in Fig. 5*c*, with the target of the S subunit located in close vicinity of the promoter so as to recruit the RNAP.

Complex Transcription Logics

The schemes discussed above with distal activation and repression can be readily extended to describe combinatorial control by multiple TF species. As long as the glue-like contact interaction exists between the TFs and RNAP, one species of TF can be substituted for another by changing the TF-specific DNA-binding sequences in Figs. 1–5. (See below for a discussion on possible adverse effects of promiscuous glue-like interactions.) More complex regulatory functions involving three or more inputs can be implemented by generalizing the constructs of Fig. 5 *b* and *c*. Fig. 6*a* illustrates the general architecture of the regulatory region obtained by using the distal activation scheme. Note that the emerging structure is naturally modular, in the sense that the sequence segment coding for a given logical expression (indicated by brackets) can be moved to different positions in the regulatory region without affecting the regulatory function (6, 22). Because each module recruits RNAP on its own, the regulatory logic function implemented is of the form

$$\mathcal{L} = C_1 \ \mathbf{OR}\ C_2 \ \mathbf{OR} \ldots \mathbf{OR}\ C_M, \qquad [1]$$

where $\mathcal{L}$ indicates the occupation state of the promoter, and C_m is the occupation state of the binding site R_m in the mth module.

Within each module, the recruitment of the R subunit to its target must be accomplished molecularly through contact with TFs

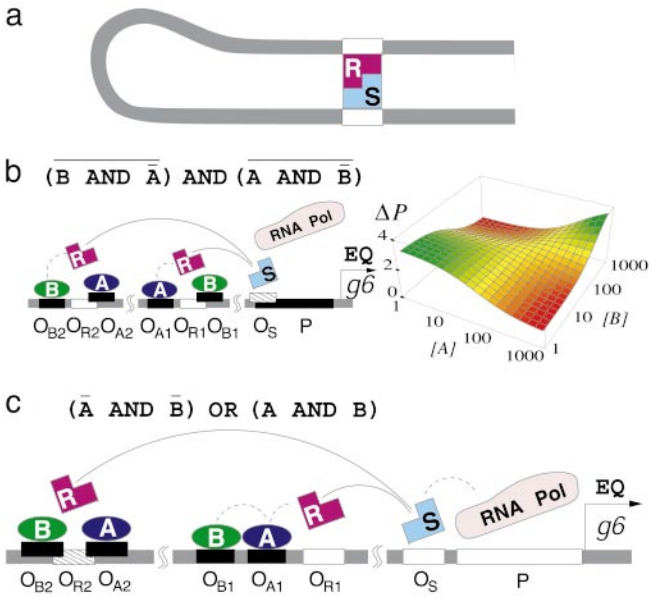

Fig. 5. (*a*) Illustration of distal regulation through ''DNA looping,'' mediated by a heterodimer formed between two subunits, R and S, each recognizing a distinct DNA-binding site. (*b*) The schematic construct and response characteristics of a regulatory region implementing the EQ gate (*g6* of Fig. 1*a*): The operators labeled R1, R2, and S are the targets of the subunits R and S as shown in *a*. The solid lines indicate the relatively strong attraction between the subunits of the heterodimer. (*c*) An alternative implementation of the EQ gate using the distal activation mechanism.

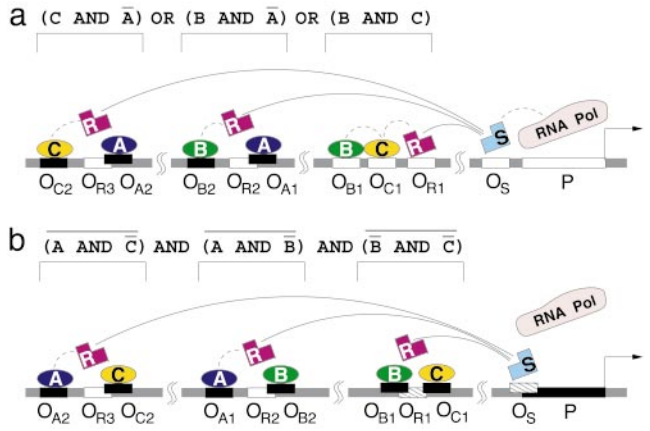

Fig. 6. (*a*) Modular construct of a regulatory function involving three controlling TFs using the distal activation scheme. The operators labeled R1, R2, R3, and S are the targets of the recruited subunits R and the activating subunit S. Each module is bracketed with the corresponding logical syntax written above, and the squiggles indicate that these modules can be at variable distances from one another. (*b*) The same regulatory function using the distal repression scheme.

bound to nearby sites. This implements the logical AND function, leading to the following expression for each "clause" C_m:

$$C_m = b_{m,1}\ \textbf{AND}\ b_{m,2}\ \textbf{AND}\ \ldots\ \textbf{AND}\ b_{m,n(m)}. \qquad [2]$$

Here the index $i \in \{1, \ldots n(m)\}$ labels the binding sites in the mth module, and the binary "literals" $b_{m,i}$ express the effect (activating/repressing) of a binding site on the occupation of R_m. For an activating site, $b = 1$ (0) if the corresponding TF concentration is high (low), whereas the opposite is true for a repressive site. If we represent the state of the concentration (high/low) of the TF α by a binary variable x_α and its inverse by $\bar{x}_\alpha$, then we have $b_{m,i} \in \{x_{\alpha(m,i)}, \bar{x}_{\alpha(m,i)}\}$ where $\alpha(m,i)$ denotes the identity of the TF (e.g., A, B, C, etc.) targeted by site i in module m.

Eqs. **1** and **2** are a special form of expressing the logic function $\mathcal{L}[x_A, x_B, x_C, \ldots]$, which describes the dependence of the gene activity $\mathcal{L}$ on the TF concentrations. Intuitively, it corresponds to selectively "switching on" rows in a logic table (see Fig. 1*a*) that are OFF by default and corresponds to the so-called disjunctive normal form (DNF) familiar in computer science (26). It is well known that any binary logic function can be expressed in DNF and reduced to a minimal (i.e., the most compact) form. This observation suggests a simple recipe to guide the construction of regulatory regions to implement a wide variety of control functions: Reduce a desired logic function to its minimal DNF and implement each clause using the distal activation scheme as shown in Fig. 6*a*. There are certainly limitations to this scheme: For instance, if a clause contains many repressive conditions, overcrowding of binding sites within a module will limit its implementation.

From the alternative implementations of the EQ gate in Fig. 5 *b* and *c*, we see that it may be possible to reduce the number of repressive conditions within clauses by adopting the distal repression scheme. This scheme is obtained by overlapping the binding site S with the promoter such that each clause C_m can repress the promoter on its own. Consequently, gene expression occurs only if none of the repression clauses are satisfied. The class of logic functions implementable under distal repression are of the form

$$\mathcal{L}' = \overline{C_1}\ \textbf{AND}\ \overline{C_2}\ \textbf{AND}\ \ldots\ \textbf{AND}\ \overline{C_M}, \qquad [3]$$

where the $\overline{C_m}$ are the inverse of the clauses C_m given in Eq. **2**. The generic architecture for the cis-regulatory implementation of logic functions, expressed according to Eqs. **3** and **2**, is shown in Fig. 6*b*. This belongs to the conjunctive normal form (CNF) of logic and corresponds intuitively to selectively "striking out" rows in a logic table that are ON by default. As with the DNF, all logic functions can be reduced to a minimal CNF (26).

Taken together, we see that to implement a given logic function, one can first obtain and compare the minimal CNF and DNF and then choose the one with fewer repressive conditions within clauses. By using two sets of DNA-bending heterodimers, one for distal activation and the other for distal repression, the two schemes could also be combined. Thus, the above theoretical considerations can guide the design of cis-regulatory constructs for a wide variety of complex control functions. However, there may be a practical limit to this approach due to the slow kinetics of assembling very large molecular complexes if there are too many clauses or too many literals within a clause.

Molecular Computing Machine

The transcription machinery can be regarded as a molecular computer, because it is capable of complex logic computations. Specifically, the molecular components (TFs and RNAP) satisfying the two ingredients of regulated recruitment, i.e., continuously tunable protein–DNA-binding strengths and glue-like contact interaction between proteins and further supplemented by distal activation and/or repression mechanisms, constitute a flexible toolkit, a kind of molecular Lego set, that can be assembled in different combinations to perform the desired computations. This machine is a general-purpose computer, because its function can be "programmed" at will through choices and placements of the protein-binding DNA sequences in the regulatory region of any gene. This should be contrasted with an alternative strategy of transcription control based on dedicated, complex (e.g., allosteric) protein–protein interactions: In the latter, complexity of the system is derived from the complexity of proteins, whereas in the former, complexity is derived combinatorially from the composition of the regulatory sequences (the "software") alone without the need of

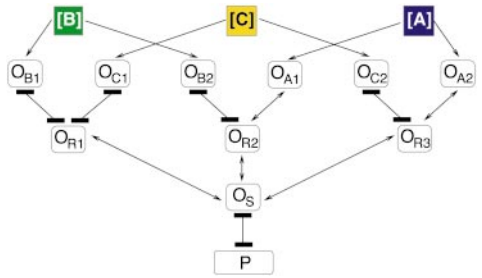

Fig. 7. The construct of Fig. 6*b* maps directly on a well studied model of neural network known as the ''Boltzmann machine'' (see *Supporting Text*). In this mapping, the binding sites are the neurons, the TF concentrations are the inputs, and the promoter is the output neuron. Cooperative/repressive molecular interactions between the TFs play the role of synapses and are denoted with arrows and bars, respectively. Note that the sites R1, R2, R3, and S are not connected to any inputs and are examples of hidden units.

tinkering with the proteins (the "hardware"). A notable advantage of encoding combinatorial control in the regulatory region, as opposed to in the regulatory proteins, is evolvability (10): Unlike the regulatory proteins, each cis-regulatory region controls the expression of a given gene and hence can be programmed with minimal pleiotropic effects.

Another way to appreciate the computational power of the transcription machinery is through analogy to a "neural network": As illustrated in Fig. 7, binding sites in a regulatory region can be viewed as "neurons" in a network, with the promoter being the output neuron and the activated TF concentrations being the inputs to the network. The occupancy of a binding site corresponds to the state of a neuron (firing or not), and the binding strength of a site becomes the "firing threshold." Molecular interactions between the proteins play the role of "synapses," which transduce signals between the neurons. This neural network is distinguished by two unique features: synaptic connections are symmetric (because molecular interactions are symmetric), and some neurons in the network are "hidden" (e.g., the binding sites of the heterodimers, which are not linked to the controlling inputs). As shown in *Supporting Text*, such networks are mathematically equivalent to the class of "Boltzmann machines" (27), which are known to be powerful computing machines. Thus, the transcription systems we discuss are molecular realizations of the Boltzmann machine.

A neural network can be "trained" to perform complex tasks by adjusting its synaptic strengths (27). Similarly, the regulatory system we discuss can fine-tune or modify its control function by adjusting molecular interactions through a combination of the programmable protein–DNA and protein–protein interactions. The latter is accomplished in nature by the evolution of DNA sequences in the cis-regulatory region. This architecture of the regulatory system makes it very evolvable (10, 22), because it is straightforward to modify individual DNA-binding sequences (through point substitutions), adjust their positions within a regulatory region (via insertions and deletions), and move/copy them from one regulatory region to another (via duplications and recombination).

Beyond Bacterial Transcription Control

Thus far we have exploited known characteristics of the bacterial transcription system and shown its power for the combinatorial regulation of a single gene. However, most known examples of bacterial transcription control are much simpler than the capabilities described. On the other hand eukaryotes, which rely heavily on complex combinatorial control, use a rather different (and not well characterized) transcription system. Are there crucial limitations in the schemes of combinatorial control we described, which prevent their adoption by bacteria on a genome-wide basis?

Promiscuity of Protein Interaction. A frequent criticism of the regulated recruitment principle is the reliance on rather promiscuous, glue-like interactions between proteins. For example, if all activated TFs in a bacterial cell (or the nucleus of a eukaryote) can interact with each other after contact, then the many possible unintended interactions may overwhelm the required functional interactions, making it impossible for the system to perform any regulatory functions. The frequently observed specificity of TF–TF interactions in bacteria seems to support this criticism. However, a simple estimate shows that unintended interactions are actually not a major concern given the weakness of the glue-like interaction and the limited total activated TF concentration (see *Supporting Text*). For instance, with an interaction energy of $E_{\rm int} = -2$ kcal/mol and typical bacterial genome size of $5{\cdot}10^6$, the adverse effect of promiscuous TF–TF interactions is negligible as long as the total number of activated TF molecules in a cell is below $\approx 10^4$. Thus, at a typical TF concentrations of $\approx$100 molecules per cell, one species of activated TF can interact weakly with $\approx$100 other activated species before unintended interactions become an issue. (Applying a similar estimate to eukaryotes, one finds that one species can roughly interact with 1,000 other species before unintended interactions become significant.)

Although the weak interactions may not be detrimental to the system, there is no reason that they will be maintained over the course of evolution if not needed functionally. Indeed, it has been estimated that the loss of protein–protein interaction is a very rapid evolutionary process (47). The isolated usage of complex combinatorial control in bacteria can thus be responsible for the apparent specificity of TF–TF interactions in bacteria. But as long as weak interactions between protein pairs may be acquired rapidly by evolution (47) when functional demand arises, we may assume a generic promiscuous interaction to study the capabilities of the regulatory system.

Intergenic Cross Talk. A major limitation of the bacterial transcription machinery becomes evident when we attempt to implement the cis-regulatory constructs of Fig. 6 at a genome-wide scale. The problem is that if every gene uses the same heterodimer pair, e.g., the subunits R and S, then they will induce "cross talk" between regulatory regions of different genes. For instance, the recruitment of the R subunit to a site R_m in one gene can cause the recruitment of the S subunit to site S′ of a neighboring gene. This problem is compounded by the fact that there are many more possibilities for the heterodimers to participate in the unintended than the intended distal interactions. Although intragenic interactions generally involve DNA looping over shorter distances than intergenic interactions, the logarithmic dependence of DNA-looping energy on distance implies that intergenic distance must be substantially (e.g., 10 times) greater than intragenic distance before distance can be used as an effective means to prevent cross talk. An alternative way to reduce cross talk is to introduce different heterodimer pairs for different genes; however, this will require many extra genes to code for the heterodimers. The compact bacterial genomes can support neither vast intergenic separations nor a large number of gene-specific heterodimers. Thus, intergenic cross-talk may be a key obstacle for bacteria to adopt complex combinatorial control at a genome-wide scale. However, this does not prevent the implementation of complex control on a few isolated genes spaced far apart along the bacterial chromosome.[‡]

The cross-talk problem is not specific to the use of heterodimers and DNA looping. Rather, it is an unavoidable consequence of the genome-wide use of any distal interaction mechanism, because each regulatory region must be told which gene to regulate. Eukaryotes have developed a number of strategies to cope with the cross-talk problem, e.g., intergenic distances often greatly exceed the size of

[‡]As an example, we note that the NtrC-activated genes (which can interact with the σ^{54} promoters over long distances) are separated by >50,000 bp from each other in *E. coli* (28).

genes in higher eukaryotes, making distance-based controls more feasible, and insulating elements limit the actions of regulatory regions to their designated genes (29).

Given the differences in the molecular mechanisms of gene regulation in prokaryotes and eukaryotes (30), what aspect of our study on combinatorial control could be applicable to eukaryotes? We argue that the qualitative aspects of our study are applicable to eukaryotes regardless of mechanisms, because our main results, e.g., the correspondence of transcription machinery to the Boltzmann machine and the implementation of CNF/DNF, are predicated only on the existence of the two key ingredients of regulated recruitment (i.e., specific protein–DNA interaction and glue-like interaction between nearby proteins) along with the possibility of distal interactions regardless of molecular implementation. Indeed, these ingredients may occur more prevalently in eukaryotes. For example, different TFs within a given class can interact cooperatively when placed adjacently (31), e.g., via contact of hydrophobic patches (15). Moreover, an indirect interaction between two unrelated TFs can also be realized through "collaborative competition" with nucleosomes (16, 32) without actual physical contact. In addition, short-range repression (or "quenching") can be achieved in eukaryotes without the need of overlapping binding sites (33), and distal repression can be accomplished by the recruitment of various chromatin-modification agents (30, 33, 34). Thus, at the qualitative level, the very different eukaryotic transcription system together with the regulated chromatin structure presents a superior molecular platform to implement complex combinatorial control.

Discussion and Outlook

The current knowledge on eukaryotic gene transcription is not sufficient to warrant the construction of quantitative models of transcription regulation (3). Nevertheless, we believe our results are useful in both a qualitative and quantitative way for dissecting the combinatorial transcription control of specific systems. On a qualitative level, the simplest and most natural forms of architecture in complex regulation involving multiple modules are the CNF and DNF. The CNF-like architecture (Fig. 6*b*) requires repression to dominate over activation; it can be accomplished in eukaryotes through the recruitment of repressing complexes such as Tup1 in yeast (34). The DNF-like architecture (Fig. 6*a*) requires activation to dominate over repression and is more natural whenever genes are repressed by default (e.g., through the local chromatin structure). The phenotype exhibited by DNF is "enhancer autonomy," which is observed in *Drosophila* embryonic development. For example, the expression of seven-stripe *even-skipped* is activated by five separate enhancers (22, 35). On a quantitative level, our model as described in *Supporting Text* provides a concrete framework to relate knowledge of cis-regulatory elements to complex gene-expression patterns regardless of molecular mechanisms. This is possible because our model, as a realization of the Boltzmann machine, is sufficiently general to describe a wide range of regulatory control functions. The DNA-binding strengths K_i and the cooperativity factors $\omega_{i,j}$ then constitute meaningful fitting parameters to relate the verified or potential binding sites (the nodes in the Boltzmann machine) to observed gene-expression data. This approach should be particularly useful in cases where a given TF can act both as an activator and a repressor and is hence potentially more powerful than the class of quasilinear models (36, 37) used to correlate gene expression and available regulatory information.

A complementary direction to pursue is the engineering of complex transcription control in bacteria. Although problematic at the genome-wide scale because of intergenic cross talk, the schemes of combinatorial control illustrated in Fig. 6 could be implemented in bacteria for isolated genes, e.g., on plasmids. Designer regulatory sequences could be constructed with our modeling approach as a guide, followed by fine-tuning of interaction parameters (the K_i and $\omega_{i,j}$ values) through directed evolution (38, 39). Such constructs might be used to control gene activities *in vivo* for various bioengineering applications (40, 41). Although many control functions can also be implemented synthetically by a network of genes regulating each other, as demonstrated in several studies (42–45), we believe that the combinatorial cis-regulatory approach is advantageous in a number of ways: Because it does not involve the iterated expression of other genes, combinatorial regulation is fast and useful in instances where timely genetic response is essential. Furthermore, it is less affected by stochastic fluctuations associated with transcription and translation (46) and unintended posttranscriptional, posttranslational, or other cellular controls exerted by the host (45). A few combinatorially regulated genes linked to each other in a network with amplification and feedback, in principle, could perform very complex functions.

We gratefully acknowledge critical comments by A. Danchin, E. H. Davidson, J. Little, W. F. Loomis, A. Murray, J. Reinitz, and J. Widom. This research was supported by National Science Foundation Grants 0211308, 0083704, 0216576, and 0225630. In addition, N.E.B. is supported by a National Science Foundation bioinformatics fellowship, and T.H. is supported by a Burroughs–Wellcome functional genomics award.

1. Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., *et al.* (2002) *Science* **295,** 1669–1679.
2. Hasty, J., McMillen, D., Issacs, F. & Collins, J. J. (2001) *Nat. Rev. Genet.* **2,** 268–279.
3. Gilman, A. & Arkin, A. P. (2002) *Annu. Rev. Genomics Hum. Genet.* **3,** 341–369.
4. Shen-Orr, S., Milo, R., Mangan, S. & Alon, U. (2002) *Nat. Genet.* **31,** 64–68.
5. McAdams, H. H. & Arkin, A. (1998) *Annu. Rev. Biophys. Biomol. Struct.* **27,** 199–224.
6. Davidson, E. H. (2001) *Genomic Regulatory Systems* (Academic, San Diego).
7. Yuh, C.-H., Bolouri, H. & Davidson, E. H. (2001) *Development (Cambridge, U.K.)* **128,** 617–629.
8. Neidhardt, F. C., ed. (1996) *Escherichia coli and Salmonella: Cellular and Molecular Biology* (Am. Soc. Microbiol., Washington, DC).
9. Ptashne, M. & Gann, A. (1997) *Nature* **386,** 569–577.
10. Ptashne, M. & Gann, A. (2002) *Genes and Signals* (Cold Spring Harbor Lab. Press, Plainview, NY).
11. Gerland, U., Buchler, N. E. & Hwa, T. (2003) *Genome Res.*, in press.
12. Shea, M. A. & Ackers, G. K. (1985) *J. Mol. Biol.* **181,** 211–230.
13. Stormo, G. D. & Fields, D. S. (1998) *Trends Biochem. Sci.* **23,** 109–113.
14. Gerland, U., Moroz, J. D. & Hwa, T. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 12015–12020.
15. Wolberger, C. (1999) *Annu. Rev. Biophys. Biomol. Struct.* **28,** 29–56.
16. Polach, K. J. & Widom, J. (1996) *J. Mol. Biol.* **258,** 800–812.
17. Wade, J. T., Belyaeva, T. A., Hyde, E. I. & Busby, S. J. (2001) *EMBO J.* **20,** 7160–7167.
18. Molina-Lopez, J. A. & Santero E. (1999) *Mol. Gen. Genet.* **262,** 291–301.
19. Dmitrova, M., Younes-Cauet, G., Oertel-Buchheit, P., Prote, D., Schnarr, M. & Granger-Schnarr, M. (1998) *Mol. Gen. Genet.* **257,** 205–212.
20. Kauffman, S. A. (1993) *The Origins of Order* (Oxford Univ. Press, New York).
21. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerbert, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., *et al.* (2002) *Science* **298,** 799–804.
22. Gerhardt, J. & Kirschner, M. (1997) *Cells, Embryos, and Evolution* (Blackwell Scientific, Malden, MA).
23. Schleif, R. (2000) *Trends Genet.* **16,** 559–565.
24. Kornacker, M. G., Remsburg, B. & Menzel, R. (1998) *Mol. Microbiol.* **30,** 615–624.
25. Langdon, R. C., Burr, T., Pagan-Westphal, S. & Hochschild, A. (2001) *Mol. Microbiol.* **41,** 885–896.
26. Whitesitt, J. E. (1961) *Boolean Algebra and Its Applications* (Addison–Wesley, Reading, MA).
27. Hertz, J., Krogh, A. & Palmer, R. G. (1991) *Introduction to the Theory of Neural Computation* (Addison–Wesley, Redwood City, CA).
28. Zimmer, D. P., Soupene, E., Lee, H. L., Wendisch, V. F., Khodursky, A. B., Peter, B. J., Bender, R. A. & Kustu, S. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 14674–14679.
29. Bell, A. C., West, A. G. & Felsenfeld, G. (2001) *Science* **291,** 447–450.
30. Struhl, K. (1999) *Cell* **98,** 1–4.
31. Li, R., Pei, H. & Watson, D. K. (2000) *Oncogene* **19,** 6514–6523.
32. Miller, J. A. & Widom, J. (2003) *Mol. Cell. Biol.* **23,** 1623–1632.
33. Gray, S. & Levine, M. (1996) *Curr. Opin. Cell Biol.* **8,** 358–364.
34. Courey, A. J. & Jia, S. T. (2001) *Genes Dev.* **15,** 2786–2796.
35. Arnosti, D. N., Barolo, S., Levine, M. & Small, S. (1996) *Development (Cambridge, U.K.)* **122,** 205–214.
36. Bussemaker, H. J., Li, H. & Siggia, E. D. (2001) *Nat. Genet.* **27,** 167–174.
37. Reinitz, J., Kosman, D., Vanario-Alonso, C. E. & Sharp, D. H. (1998) *Dev. Genet. (Amsterdam)* **23,** 11–27.
38. Stemmer, W. P. C. (1994) *Nature* **370,** 389–391.
39. Yokobayashi, Y., Weiss, R. & Arnold, F. H. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 16587–16591.
40. Hasty, J., McMillen, D. & Collins, J. J. (2002) *Nature* **404,** 224–230.
41. Weiss, R., Homsy, G. E. & Knight, T. F. (1999) in *DIMACS Workshop on Evolution as Computation* (Springer, Princeton), pp. 275–295.
42. Thomas, R. & D'Ari, R. (1990) *Biological Feedback* (CRC, Boca Raton, FL).
43. Elowitz, M. B. & Leibler, S. (2000) *Nature* **403,** 335–338.
44. Gardner, T. S., Cantor, C. R & Collins, J. J. (2000) *Nature* **403,** 339–342.
45. Guet, C. C., Elowitz, M. B., Hsing, W. H. & Leibler, S. (2002) *Science* **296,** 1466–1470.
46. Elowitz, M. B., Levine, A. J. Siggia, E. D. & Swain, P. S. (2002) *Science* **297,** 1183–1186.
47. Wagner, A. (2003) *Proc. R. Soc. London Ser. B* **270**, 457–466.