Specificity and robustness in transcription control networks

Anirvan M. Sengupta*, Marko Djordjevic†, and Boris I. Shraiman*‡

*Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974; and †Department of Physics, Columbia University, New York, NY 10025

Edited by David R. Nelson, Harvard University, Cambridge, MA, and approved November 8, 2001 (received for review July 26, 2001)

Recognition by transcription factors of the regulatory DNA elements upstream of genes is the fundamental step in controlling gene expression. How does the necessity to provide stability with respect to mutation constrain the organization of transcription control networks? We examine the mutation load of a transcription factor interacting with a set of *n* regulatory response elements as a function of the factor/DNA binding specificity and conclude on theoretical grounds that the optimal specificity decreases with *n*. The predicted correlation between variability of binding sites (for a given transcription factor) and their number is supported by the genomic data for *Escherichia coli*. The analysis of *E. coli* genomic data was carried out using an algorithm suggested by the biophysical model of transcription factor/DNA binding. Complete results of the search for candidate transcription factor binding sites are available at http://www.physics.rockefeller.edu/~boris/public/search_ecoli.

The accumulation of knowledge on control of transcription in simple and complex organisms (1–4) poses many questions regarding its system-level function and organization. What aspects of control network architectures insure their stability with respect to mutation along with their ability to adapt and acquire new function (see ref. 1)? In its turn, could better understanding of the general organization of these networks help to dissect specific systems? Motivated by these questions, we formulate a model of transcription control that captures many of the essential features of the process and can be tested against data. This model is applied to the study of evolutionary stability of bacterial regulons (5), each involving a transcription factor that controls multiple genes by binding to multiple regulatory sequence elements. Evolutionary stability, or stability with respect to mutation, provides a sensible quantitative definition of robustness (1, 6, 7). Factors with highly sequence-specific binding impose severe constraint on regulatory sequences, increasing the probability of failure due to mutation, whereas low specificity of binding increases the probability of spurious interactions. Robustness is maximized by the compromise between these two effects. From this follows a quantitative prediction relating the number of elements in a regulon and the degree of binding site sequence variability. Our model also suggests a method of identifying candidate transcription factor binding sites, which we use in the analysis of the genomic data for *Escherichia coli* (8). Genomic data provides support for the prediction of the theory.

Model of Transcription Factor/DNA Interaction

In modeling the transcription control network, let us concentrate on the flow of information from factors to genes. Active transcription factors bind to the regulatory response element (RE) subsequences associated with genes (2). A given gene may be controlled by multiple repressing or activating factors acting through multiple REs (see Fig. 1). It will suffice for now to assume that all of the controlling factors and REs are nonredundant so that the loss of factor/RE recognition results in a significant detriment of fitness, due to the loss of regulatory linkage.[§]

Binding of a factor to an RE depends on the factor concentration and the binding energy of the pair which together determine the binding probability. The interaction of a transcription factor with a DNA sequence x (of length L) may be approximated (9, 10) by the binding energy $E(x) = \boldsymbol{x}\cdot\boldsymbol{\varepsilon} \equiv \sum_{i=1}^{L} \sum_{\beta=1}^{4} \varepsilon_i^{\beta}\boldsymbol{x}_i^{\beta}$ where ε_i^{β} is the interaction energy with base β at position $i = 1, \ldots, L$ of the DNA string and $\mathrm{x}_{\mathrm{i}}^{\beta} = 1$ if the sequence $\boldsymbol{x}$ contains base β at position i and is 0 otherwise. Thus DNA binding properties of a factor are parametrized by ε_i^{β}.

It is useful to consider the distribution of binding energies $E(x)$ among all possible sequences, which is described by a histogram (or density of states) $\rho(E)$ as shown in Fig. 2. The vast majority of random sequences fall into the approximately Gaussian center of this distribution, whereas the strongest binding, or consensus, sequence defines the leftmost edge (see *Appendix A*). (Note: It is convenient to set the scale of energy by the standard deviation of the binding energy in a random sequence ensemble and to set the average energy—i.e. the energy of nonspecific interaction—as $E = 0$.)

In equilibrium, the probability of any string $\boldsymbol{x}$ to bind a factor is given explicitly by $f(E(\boldsymbol{x})) = [e^{(E(\boldsymbol{x}) - \mu)/k_BT} + 1]^{-1}$, where μ is the chemical potential set by factor concentration. Provided that the characteristic scale of binding energy is large compared to k_BT [e.g., for lac-repressor (11) $E_m/k_BT \sim 10$] one may as a first approximation replace $f(E)$ by a step function; i.e. an "all or nothing" binding condition. The binding condition identifies the subset of all possible sequences (of length L) that at a given physiological concentration of a factor bind to it strongly: sequence, $\boldsymbol{x}$, belongs to such a subset for factor F_i if it satisfies $E_i(\boldsymbol{x}) < \mu_{\mathrm{i}}$. On Fig. 3, sequence subsets binding to different factors are pictured as nonoverlapping discs. Points within a given disk denote response elements controlled by F_i: their number, n_i, is the degree of pleiotropy of the factor. A given factor, present in certain concentration, may be characterized by its "binding specificity," $\sigma \equiv \ln v^{-1}$, defined in terms of the fraction, v, of *random* sequences (of length L) that bind strongly. These sequences lie in the "tail" of energy distribution $\rho(E)$ below μ (see Fig. 2). In Fig. 3 binding specificity is represented by disk area.

In addition to regulatory response elements, DNA regions upstream of genes contain stretches of sequence without direct function. Some of these nonfunctional stretches are truly passive; others, however, may become "active" if through mutation they acquire a binding site for some transcription factor that can then interfere, positively or negatively, with transcription. In Fig. 3, such "potential cis-regulatory sites" are represented as points outside the disks.

Mutation Load in Transcription Control

Let us now introduce mutations. An accumulation of point mutations in a response element sequence corresponds to a

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

Abbreviation: RE, response element.

‡To whom reprint requests should be addressed at: 1-D236 Bell Labs, 700 Mountain Avenue, Murray Hill, NJ 07974. E-mail: boris@physics.bell-labs.com.

[§] The case of more complex and redundant regulatory interactions can be considered within the same framework.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

www.pnas.org/cgi/doi/10.1073/pnas.022388499

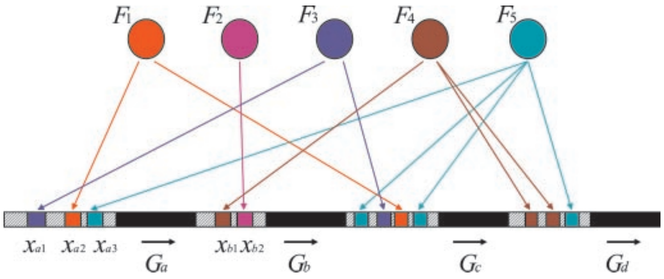

Fig. 1. Schematic model of transcription control. *F*s are active transcription factor proteins, *x*s are response element subsequences upstream of the coding regions of the genes, *G*. Arrows indicate regulatory interactions.

random walk of the RE in sequence space (Fig. 3). A mutation in the DNA binding domain of the transcription factor changes its interaction energy with the sequence, ε. This change would appear in Fig. 3 as a random shift of the disc representing the binding subset in the sequence space. Either process could lead to an RE moving out of the binding subset of sequences, disrupting the interaction with the factor. In our model of a nonredundant network involving only essential genes, such failure is assumed to be "lethal." Another form of regulatory failure involves a mutation of the *potential* regulatory site, causing it to enter the domain of binding with a wrong factor (see Fig. 3). For simplicity, this process is assumed to be lethal as well. To discuss the robustness of any given network, we must be able to calculate the probability (per mutation) of both modes of failure.

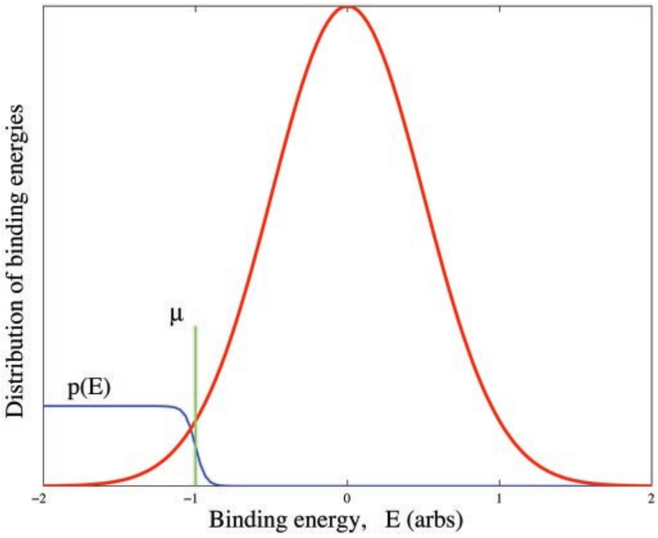

Fig. 2. Typical energy histogram, $\rho(E)$, for a transcription factor interacting with a random DNA subsequences. In equilibrium, strings corresponding to energies below the chemical potential, μ (set by factor concentration), bind the factor with high probability given explicitly by $[e^{(E-\mu)/k_BT} + 1]^{-1}$.

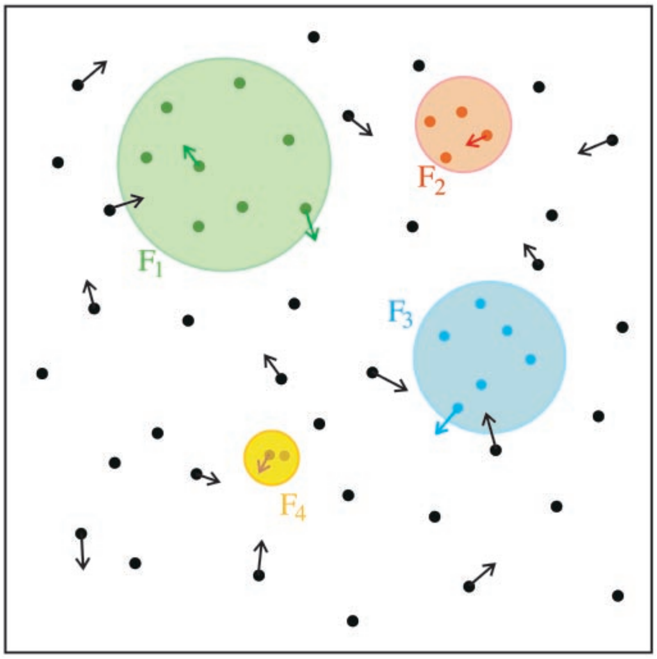

Fig. 3. Response elements (dots) and factor binding subsets (disks) in sequence space. RE located within a disk binds the corresponding factor. Black dots lying outside the discs represent potential cis-regulatory sites, which must not bind transcription factors in order to avoid interference with transcription control. Arrows represent random changes in the sequence of REs (and of potential cis-regulatory sites) due to mutation.

For a single factor/RE link, the binding probability depends only on the interaction energy $E(x) = \boldsymbol{\varepsilon}\cdot\boldsymbol{x}$ which changes randomly with mutations in $\boldsymbol{x}$ and $\boldsymbol{\varepsilon}$. Although this is a discrete process, most qualitative features could be understood in the limit where the energy changes continuously. This approximation is accurate when the binding energy has contribution from many sites, i.e. L is large, and mutation at an individual site results in only a small change in the total energy. Consider a mutating 'population' of single factor/RE links with $n(E, t)$ denoting the number of links with binding energy in the E, $E + dE$ interval. It is possible to derive (*Appendix B*) an equation which governs the time evolution of $n(E, t)$. It has the form of biased diffusion in the energy variable:

$$\partial_t n(E, t) = \partial_E^2 n(E, t) + \partial_E[En(E, t)] \qquad [1]$$

where unit of time is set by the point mutation rate. In addition we impose the boundary condition $n(E, t)|_{E=\mu} = 0$ which implements the assumption that "escape" of the *RE* from the domain of binding represents a lethal failure.[¶] The first term on the right hand side represents "diffusion" in E arising from small random changes of the binding energy, whereas the second term describes the drift toward energies corresponding to the larger number of sequences; i.e., toward higher density of states $\rho(E)$ (see Fig. 2). In the absence of selection (i.e., without the boundary condition) Eq. **1** is solved by $n(E) \sim \exp(-E^2/2)$—the Gaussian distribution that approximates the distribution of E in random sequence ensemble (Fig. 2 and *Appendix A*).

After a long time the distribution of binding energies behaves as $n(E, t) \approx e^{-\kappa_l t} n_\infty(E)$ with κ_l being the smallest eigenvalue of Eq. **1**. Thinking of the population of factor/RE links (or more generally, transcription control networks) as a population of "organisms" one can draw on the ideas from population genetics (12). The asymptotic "death rate," κ_l, determines the minimal rate of replication that would be necessary to maintain a stable population and is therefore identified as *mutation load* (12).

The lowest eigenvalue of Eq. **1** can be computed (*Appendix B*) as a function of binding specificity σ, yielding $\kappa_l(\sigma) \approx \sigma/2$, accurate for $\sigma \gg 1$. This computation makes quantitative the intuitive expectation that more specific interaction is more sensitive to mutation. The same calculation (*Appendix B*) determines $n_\infty(E)$, which gives the distribution of factor/RE binding

[¶] This evolution equation may be readily generalized to include binding-energy-dependent "fitness" term $V(E)n(E,t)$.

energies in a population after many mutations. This distribution exhibits a sharp maximum near the boundary of the domain of binding, $E = \mu$, meaning that most of the response elements are about as far from the consensus sequence (the sequence which binds most strongly) as the binding condition permits. The reason is that there are many more possible sequences (i.e., higher density of states) near the boundary of the domain than at its center. The entropic, or sequence-space volume effects dominate the evolution of response elements!

To calculate the rate with which a factor can acquire a spurious regulatory site, we compute the rate with which a sequence outside the domain of binding, $E > \mu$ would "diffuse" in. This requires solving Eq. **1** with the same boundary condition as before, but in the range $E > \mu$. The probability (per mutation), $\kappa_{sp}(\sigma)$, with which a spurious site enters the domain of binding is given by (*Appendix C*) $\kappa_{sp}(\sigma) \approx (\pi/2)\sigma e^{-\sigma}$ (computed as before in the high specificity limit). This rate is exponentially small, because now most of the sequence space is outside the domain of binding. However, because the number of *potential* cis-regulatory sites, N_{cR}, is large, the total probability for the factor to acquire a spurious regulatory target, $N_{cR}\kappa_{sp}(\sigma)$, may not be negligible. The two modes of failure, κ_{sp} and κ_l have the opposite dependence on binding specificity and their balance will determine the optimal choice for the latter.

Robustness Optimization for a Regulon

Let us now estimate the total rate of mutation induced failure, γ, in a control network consisting of N_f factors with factor i controlling n_i genes (or operons) by the same number of response elements. It is given by the sum of failure rates[‖] for all of the links plus the total "spurious site" acquisition rate

$$\gamma = \sum_{i=1}^{N_f} n_i \kappa_l(\sigma_i) + \sum_{i=1}^{N_f} N_{cR}\ \kappa_{sp}(\sigma_i). \qquad [2]$$

As we already noted, the total rate of failure, γ, or "death rate" per mutation in a population of network "organisms" evolving under a selection constraint, is also known as *mutation load* (12). It sets the minimal rate of replication that would sustain a steady population. An organism (or network architecture) more stable, or robust, with respect to mutation has lower mutation load and has an evolutionary advantage. In fact we can quantitatively define evolutionary robustness as the inverse mutation load, γ^{-1}. (Note that the connection with the lowest eigenvalue of the evolution operator means that this definition is readily generalizable.) Evolutionary interpretation makes it meaningful to minimize the total mutation load γ in Eq. **2** with respect to binding specificities, σ_i. This yields

$$n_i/N_{cR} \approx \pi\sigma_i e^{-\sigma_i}, \qquad [3]$$

which relates the optimal binding specificity of a factor with the number of its regulatory targets—its degree of pleiotropy—n_i. Because $\sigma_i = \ln v_i^{-1}$, Eq. **3** implies that up to logarithmic corrections, n_i is proportional to the volume fraction v_i. Note that a linear scaling would also hold for the number of binding sites within a segment of *random* DNA of length N, $n_i^{\mathrm{rand}} = Nv_i$. The similarity is accidental: in deriving Eq. **3** we considered *nonrandom* response elements, which evolve under selection, and in contrast with the random case, the proportionality constant $n_i/(v_iN_{cR})$ is not equal to 1. The decrease of the optimal specificity with the increasing degree of pleiotropy is forced by the need to reduce the mutation induced failure rate *per RE*, which is achieved by allowing corresponding REs to occupy a larger fraction of sequence space.

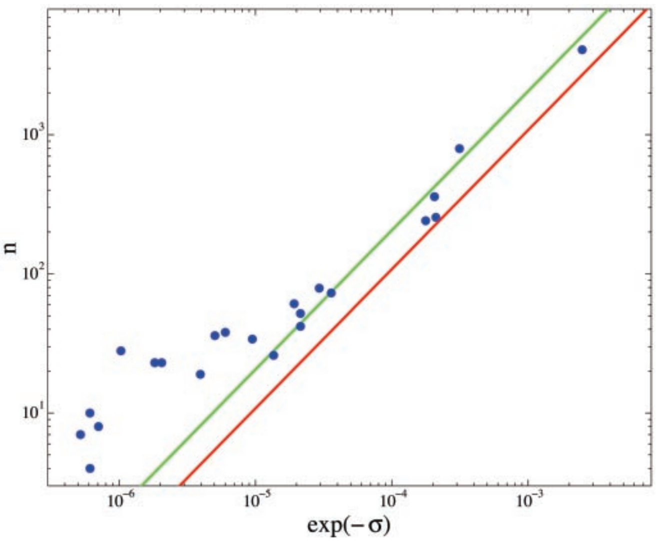

Fig. 4. The number of (candidate) response element sites, n, obtained from the *E. coli* genomic data versus factor binding specificity σ (circles). Note that $\exp(-\sigma)$ is the fraction of random sequences which bind the factor. Red line: expected number of binding sites for the random sequence background (reproducing the base frequency and nearest neighbor correlations of the noncoding segments). Green line: asymptotic fit to the predicted specificity/pleiotropy relation, Eq. **3**.

The predicted, approximately linear, scaling of v_i with n_i, which follows from Eq. **3** arises largely from the exponential dependence of κ_{sp} on σ, which in its turn is due to the fact that potential regulatory sites have an exponentially large fraction of sequence space to "diffuse" in without interfering with any of the transcription factors. We expect this conclusion to persist if the sharp "viability" boundary defined by the "lethal failure" threshold in our model is replaced by a smoother fitness landscape.

Specificity and Pleiotropy for *E. coli* Transcription Factors

The optimal specificity argument, presented above, may be applied in the context of prokaryotic regulons (2, 5). In *E. coli* a single operon is positively or negatively controlled by a small number of transcription factors. A single factor regulates a variable number of operons ranging from one (e.g., *LacI*) to perhaps hundreds, as in the case of cAMP-receptor protein (*Crp*) (5). Thus, factors have different degrees of pleiotropy. From the known functional binding sites one deduces that factors bind with different specificity: e.g., the dimeric target sites of *LacI* are very specific, whereas *Crp* sites are not (an effect due in large degree to a difference in intracellular concentration of the respective factors). A collection of known chromosomal binding sites for a set of 55 transcription factors has been assembled (8). This data set gives us the opportunity to look for an empirical correlation between specificity and degree pleiotropy indicated by Eq. **3**.

We use an algorithm (described in *Appendix D*) to estimate the characteristic parameters ($\boldsymbol{\varepsilon}_j$, μ_j) of each factor (dimer or monomer, as appropriate) from the binding sites in *E. coli* transcription factor database (8). We then search the intergenic regions of *E. coli* genome for sequences, $\boldsymbol{s}$, satisfying $\boldsymbol{\varepsilon}_j{\cdot}\boldsymbol{s} < \mu_{\mathrm{j}}$ condition. Assuming that the number of these *candidate* sites, n_j, is not vastly different from the number of the true binding sites, we use it as an estimate of the degree of pleiotropy. The binding specificity σ_j is computed from the fraction of random background sites satisfying the same binding condition (see supporting information, http://www.physics.rockefeller.edu/~boris/public/search_ecoli). Fig. 4 presents the number of (candidate)

[‖] Assuming for simplicity that different transcription factors do not bind the same REs.

target sites versus the binding specificity (or the sequence volume fraction $v = \exp(-\sigma)$. For comparison, the red line on Fig. 4 gives, as a function of specificity, the number of binding sites, N_{rand}, expected in a stretch of *random* DNA of length ($N \approx 5.4 \times 10^5$) equal to the length of noncoding part of the *E. coli* genome examined in the search. The extent to which n exceeds N_{rand} is a measure of statistical significance of the results of the sequence search: we have set the cutoff at 3 standard deviations (i.e., we have excluded six cases where n is within 3 standard deviations of N_{rand}) so that the counts above the red line plausibly correspond to functional sites.** Green line is a possible asymptotic fit to Eq. **3** (with $N_{cR} \approx N/12$). Note, that although N_{cR}—the number of potential regulatory sites in the promoter regions—is a property of the biological system, it is not known and is determined here only as a fitting parameter. For high specificity/low pleiotropy factors, where very few example binding sites are known, our algorithm overestimates specificity because of overfitting. However, despite the very considerable scatter (note the logarithmic scale of the plot) there is a clear correlation between the two quantities that plausibly follow Eq. **3** in the high pleiotropy regime where our considerations apply. This provides empirical support to the prediction of the robustness optimization argument. Future experimental determination of functional response elements in the *E. coli* will allow more direct and precise determination of both the binding specificity and the degree of pleiotropy of transcription factors.

Coevolution of Factors and Response Elements

So far we have focused on the mutation of response elements. As long as mutations in the DNA-binding domain of the factor produce only small random changes in $\boldsymbol{\varepsilon}$, their effect on factor/RE recognition is still described by Eq. **1** with appropriately adjusted rate of "diffusion" in energy E. For a factor interacting with n REs comprising a regulon, the contribution of factor mutations is small compared to that due to RE mutations and has been neglected in Eq. **2**. Factor mutation however is the limiting step in the evolutionary drift of the consensus sequence of the REs of the regulon. It is possible to estimate how the rate of coevolution of the factor together with its n regulatory targets depends on their number. A calculation of the effective rate of diffusion in sequence space of the consensus RE of a regulon (*Appendix E*) predicts that it should decrease as $1/n$. It is well recognized that more pleiotropic factors are more conserved. Our result however makes a falsifiable quantitative prediction (distinct from the result of ref. 13), which can be checked by comparing orthologous factors between different species of prokaryotes (N. Rajewsi, N. Socci, M. Zapotocki, and E. D. Siggia, unpublished work; ref. 15).

Discussion

In this paper we analyzed the mutation load of the regulon control architecture as a function of specificity of transcription factor binding and found that minimization of the mutation load predicts a correlation between specificity and pleiotropy of the factor. Provided that the sequence dependent contribution to factor/DNA interaction energy is much larger that k_BT, binding specificity can be optimized simply by changing factor concentration.

Our analysis was based on the "all or nothing" fitness model, which associated each violation of the $E < \mu$ binding condition with lethal failure. This model is an abstraction from a more realistic situation where fitness is determined by the ability of a given gene to be switched "on" or "off" by a change in the concentration of the controlling transcription factor. This means that the binding energy of the RE in question must lie in between the μ_{on}, μ_{off} values—the chemical potentials corresponding to the on and off concentrations. Thus, one expects the fitness to be a smooth function of E with a peak at some finite value. Eq. **1** can be extended to this case by effectively replacing each "disk" in Fig. 3 with an annulus. Mutation load on a single RE then still decreases as σ with decreasing specificity, because, as in the analysis above, the mutation load is dominated by the outward drift due to larger number of sequences with lower binding energy. The analysis may be readily extended to include mutation-induced variation in transcription factor concentration levels, by including in the evolution Eq. **1** diffusion in μ. This extension would introduce into the model consideration of robustness with respect to concentration fluctuations. However, the genetic mechanisms controlling transcription factor levels were not presently included in the model.

Another realistic complication arises from the fact that different genes even in the same regulon may be required to turn on at different levels of the transcription factor. For example, operons for metabolism of preferred alternatives of glucose have stronger binding sites for CRP (16) and therefore get turned on at lower concentrations of activated CRP. Hence, REs of different genes comprising the same regulon may have individual constraints on their binding energy. The mutation load on such a factor is still a sum of contributions of regulated genes, but is no longer given simply by $n_i\kappa_l(\sigma_i)$ term as it was in Eq. **2**. The measure of specificity of such a factor would be defined by some weighted average over different thresholds. Yet the scaling of mutation load with (inverse) specificity should still hold because of the phase space considerations—i.e., mutation load on any of the REs being determined by the number of distinct sequences with binding energy E near the corresponding switching threshold.

The key role of the multiplicity of genetic states corresponding to the same phenotype has been emphasized in the "neutral" evolution theory (17) and specifically in the context of RNA folding (6, 18). In the context of transcription control, the neutrality of mutations that preserve the binding energy is plausible to the extent that the regulatory phenotype is determined entirely by the probability of factor/RE binding determined by the binding energy. The multiplicity of response element sequences corresponding to the same binding energy allows the "population" of factor/RE links to spread out over the "equi-fit" (i.e., same binding energy) manifold. This spreading generates an entropic contribution to fitness. Of two network architectures with the same phenotype (and replication rate), the one with larger neutral volume will have the lower mutation load and thus higher rate of growth. Estimating the mutation rate per nucleotide at 10^{-9} per generation in bacteria, we expect the mutation load effects to act on the time scale of 10,000 years. Much work has been done on the effect of finite size of populations on evolution (19). In small populations, random events can dominate over a small selection pressure. The condition for this to happen is that the difference of mutation load is smaller than the inverse of the population size. We assume that over sufficiently long time separate bacterial colonies exchange genetic information, so that the effective size of bacterial population is large enough ($>10^9$ to 10^{10}) for the finite size effects, or "genetic drift" (19), to be ignored. It will be very interesting to investigate the diversity of regulatory binding sites between different strains of *E. coli* and between related bacterial species.

Our quantitative definition of the "effective fitness" or robustness of a transcription factor network by mutation load (and the lowest eigenvalue of the operator describing the evolution of a population of networks) can be extended to complex networks

**The number of candidate binding sites found for *DnaA*, *GlpR*, *Hns*, *MetJ*, *RpoD*, *RpoS*, and *SoxS* is within 3 standard deviations of that expected in the random ensemble. Their significance must be established by further work. We note however that excess over random "background" is not required for the functionality of sites, but is rather an internal check on the bioinformatic approach to motif discovery.

mapping factor activation patterns into patterns of gene expression. Mutations tend to spread the distribution of network parameters (e.g., RE sequences) over the whole domain consistent with the selected phenotype (i.e., factor/gene activity patterns). This domain will in general have very complex structure dependent on the network: different network architectures will have different mutation load (which can be determined numerically) and hence most robust network architecture for a given regulatory task can be identified on a firmly quantitative basis. The term "robustness" has been used with different meanings and is often interpreted as complete insensitivity to parameters, leaving an ambiguity as to which parameters should be considered (or excluded from the consideration) and what the quantitative measure should be. Identifying robustness with (inverse) mutation load solves both difficulties. All and only the parameters subject to mutation enter robustness consideration. Quantitative measure of robustness is provided by comparison of mutation loads for different architectures.

Our present analysis illustrates how modeling in conjunction with genomic data can be used to extract general features of control network organization. Many more of the fundamental principles underlying the organization of genetic networks are yet to be discovered.

Appendix A: Calculating $\rho(E)$ Distribution

Consider the probability of finding a random oligomer that binds to the factor with free energy between E and $E + dE$:

$$\rho(E) = \langle \delta(E - \boldsymbol{x}\cdot\boldsymbol{\varepsilon}) \rangle_{\boldsymbol{x}}, \qquad [4]$$

where $\langle \cdots \rangle_s$ denotes an unrestricted average over the sequences $\boldsymbol{x}$. To calculate $\rho(E)$ let us introduce a Laplace transform

$$\rho(E) = \oint d\beta e^{\beta E} \langle e^{-\beta \boldsymbol{x}\cdot\boldsymbol{\varepsilon}} \rangle_{\boldsymbol{x}} \approx \max_{\beta} e^{\beta E + \ln Z(\beta)}, \qquad [5]$$

where the latter expression is the leading term of the saddle-point approximation to the β integral. We have introduced a "partition function"

$$Z(\beta) = \langle e^{-\beta \boldsymbol{x}\cdot\boldsymbol{\varepsilon}} \rangle_{\boldsymbol{x}} = \prod_{i=1}^{L} \sum_{\alpha} e^{-\beta \varepsilon_{\alpha i}} p_{\alpha}, \qquad [6]$$

where p_α is the frequency of base α. The analogy with the canonical and microcanonical ensembles is evident with the thermodynamic limit implicit in this analogy corresponding to $L \rightarrow \infty$. To evaluate $\rho(E)$ from Eq. **5** one must determine corresponding β from the saddle-point condition

$$E = -\frac{\partial}{\partial \beta} \ln Z(\beta) = \sum_{i} \frac{\Sigma_{\alpha}\, \varepsilon_{\alpha i} e^{-\beta \varepsilon_{\alpha i}} p_{\alpha}}{\Sigma_{\alpha}\, e^{-\beta \varepsilon_{\alpha i}} p_{\alpha}}. \qquad [7]$$

We also have the x ensemble analogue of "entropy," $\ln \rho(E)$, consistent with the definition $\beta = \partial \ln \rho / \partial E$.

Quite generally, for large L, $\rho(E)$ near its peak at $E = 0$ is well approximated by a Gaussian $\rho(E) \sim \exp(-E^2/2\chi^2)$ with $\chi^2 = \partial^2/\partial\beta^2 \ln Z(\beta)|_{\beta=0} = \Sigma_{i=1}^{L} \Sigma_{\alpha}\, p_{\alpha} \varepsilon_{\alpha i}^2$, which provides a measure of sequence specificity of the factor in question. (Note that addition of sequence sites with small ε_i^2 does not contribute much to χ!) Away from the center deviations from Gaussianity appear. In fact, the support of $\rho(E)$ is finite with the bottom of the "band" $E_m = \Sigma_i \min_\alpha \varepsilon_{i\alpha}$ (and the top at $E_M = \Sigma_i \max_\alpha \varepsilon_{i\alpha}$). Note that to simplify notation in the main text we have chosen the energy units so that $\chi^2 = 1$. However, when comparing weakly specific factors (i.e., $\chi/k_B T \approx 1$) the relative magnitude of their χ's becomes important.

For the purpose of establishing the significance of genomic search results it is useful to redo the calculation of $\rho(E)$ in the random ensemble, which reproduces not only the single base statistics of the genome, but includes the correlations between neighboring bases and hence provides an improved representation of genomic "background." Fig. 4 uses specificity σ estimated on the basis of the latter ensemble. The details of this calculation can be found at http://www.physics.rockefeller.edu/~boris/public/search_ecoli.

Appendix B: Derivation of the Evolution Equation

Let $n(E, t)$ denote the number of sites to have binding energy in the E, $E + dE$ interval, which is expressed in terms of the number of occurrences of site $\boldsymbol{x}$: $\eta(\boldsymbol{x}, t)$.

$$n(E, t) = \langle \delta(E - \boldsymbol{x}\cdot\boldsymbol{\varepsilon}) \eta(\boldsymbol{x}, t) \rangle_{\boldsymbol{x}} \qquad [8]$$

As a result of mutations occurring with probability α per unit time the sequence changes by $\Delta \boldsymbol{x} = \boldsymbol{x}' - \boldsymbol{x}$ and we have

$$\frac{d}{dt} \eta(\boldsymbol{x}, t) = \alpha \sum_{\Delta x} [\eta(\boldsymbol{x} - \Delta \boldsymbol{x}, t) - \eta(\boldsymbol{x}, t)] \qquad [9]$$

$$\frac{d}{dt} n(E, t) = \alpha \sum_{\Delta x} \langle \delta(E - \boldsymbol{x}\cdot\boldsymbol{\varepsilon}) [\eta(\boldsymbol{x} - \Delta \boldsymbol{x}, t) - \eta(\boldsymbol{x}, t)] \rangle_{\boldsymbol{x}} \qquad [10]$$

$$= \alpha \frac{\partial}{\partial E} \sum_{\Delta x} \langle \Delta \boldsymbol{x}\cdot\varepsilon \delta(E - \boldsymbol{x}\cdot\boldsymbol{\varepsilon}) \eta(\boldsymbol{x}, t) \rangle_{\boldsymbol{x}} \qquad [11]$$

$$+ \alpha \frac{1}{2} \frac{\partial^2}{\partial E^2} \sum_{\Delta x} \langle (\Delta \boldsymbol{x}\cdot\boldsymbol{\varepsilon})^2 \delta(E - \boldsymbol{x}\cdot\boldsymbol{\varepsilon}) \eta(\boldsymbol{x}, t) \rangle_{\boldsymbol{x}} \cdots \qquad [12]$$

$$= \alpha \frac{\partial}{\partial E} n(E, t) \sum_{\Delta x} \rho^{-1}(E) \langle (\Delta \boldsymbol{x}\cdot\boldsymbol{\varepsilon}) \delta(E - \boldsymbol{x}\cdot\boldsymbol{\varepsilon}) \rangle_{x} \qquad [13]$$

$$+ \alpha \frac{1}{2} \frac{\partial^2}{\partial E^2} n(E, t) \sum_{\Delta x} \rho^{-1}(E) \langle (\Delta \boldsymbol{x}\cdot\boldsymbol{\varepsilon})^2 \delta(E - \boldsymbol{x}\cdot\boldsymbol{\varepsilon}) \rangle_{x} \cdots, \qquad [14]$$

where the last line was derived by assuming that $\eta(x, t)$ is the same for all x's with $E(x)$ within a narrow shell, E, $E + dE$. Observing that

$$\langle (\Delta \boldsymbol{x}\cdot\boldsymbol{\varepsilon}) \delta(E - \boldsymbol{x}\cdot\boldsymbol{\varepsilon}) \rangle_{x} = \rho(E) \sum_{i} \frac{\sum_{\alpha} p_{\alpha} \sum_{\gamma} p_{\gamma} e^{-\beta \varepsilon_{\alpha i}} (\varepsilon_{\alpha i} - \varepsilon_{\gamma i})}{\sum_{\alpha} p_{\alpha} e^{-\beta \varepsilon_{\alpha i}}}$$

$$= E \rho(E) \qquad [15]$$

and defining the effective "diffusivity" in energy

$$D(E) \equiv \frac{1}{2} \rho^{-1}(E) \langle (\Delta \boldsymbol{x}\cdot\boldsymbol{\varepsilon})^2 \delta(E - \boldsymbol{x}\cdot\boldsymbol{\varepsilon}) \rangle_{x}$$

$$= \frac{1}{2} \sum_{i} \frac{\sum_{\alpha} p_{\alpha} \sum_{\gamma} p_{\gamma} e^{-\beta \varepsilon_{\alpha i}} (\varepsilon_{\alpha i} - \varepsilon_{\gamma i})^2}{\sum_{\alpha} p_{\alpha} e^{-\beta \varepsilon_{\alpha i}}} \qquad [16]$$

we arrive at the evolution equation

$$\frac{d}{dt}n(E,t) = \mathscr{L}n(E,t) = \alpha\left[\frac{\partial}{\partial E}En(E,t) + \frac{\partial^2}{\partial E^2}D(E)n(E,t)\right]. \qquad [17]$$

In the region where $E \sim \chi \sim \sqrt{L}$, $\beta \approx E/\chi^2 \sim 1/\sqrt{L}$, we have $D(E) \approx D(0) = \chi^2$ (the corrections being order of β, which is small when L is large). To simplify notation in the main text we have rescaled energy to eliminate χ and time to absorb α. Thus we have derived Eq. **1**.

Appendix C: Calculating Mutation Load

To calculate the probability of failure we must impose the binding condition $E < \mu$ as the selection constraint. This effect is included by imposing an "absorbing" boundary condition $n(\mu, t) = 0$. In the long time limit, solution of Eq. **1** has the form $n(E, t) \sim e^{-\kappa(\mu)t}n_\infty(E)$, where κ is the smallest magnitude eigenvalue of the evolution operator $\mathscr{L}$ (17). κ is the rate at which sequences move outside the binding region. $n_\infty(E)$ is given in terms of the parabolic cylinder function $U(x, a)$ (20)

$$n_\infty(E) = cst \times U(-E/\chi, -1/2 - \kappa(\mu)/D) \qquad [18]$$

$\kappa(\mu)$ as a function of μ is given by

$$\kappa(\mu) \sim \mu^2/4 \sim \frac{1}{2}\left(\sigma - \frac{1}{2}\ln\sigma + \cdots\right) \text{ for } \mu < 0, |\mu| \gg 1 \qquad [19]$$

$$\sim \sqrt{\frac{\pi}{2}}e^{-\mu^2/2}\frac{\mu}{4} \sim \frac{\pi}{2}\sigma e^{-\sigma} \text{ for } \mu > 0, |\mu| \gg 1 \qquad [20]$$

derived from asymptotics of parabolic cylinder functions. Specificity in terms of μ is given by $\sigma = \ln \nu^{-1} = \mu^2/2^2 + \frac{1}{2}\ln(2\pi\mu^2)$.

Thus, rate of losing a response element from inside the binding region is $\gamma_{loss}(\mu) = \kappa(\mu)$. We can similarly calculate the rate of a string outside the binding region to diffuse in. This process controls the rate of spurious activation (see text). The rate is $\gamma_{sp}(\mu) \approx \kappa(-\mu)$ when μ is not too far from the modal value of energy. One interesting thing about the asymptotic distribution of energy $n_\infty(E)$, for $\mu < 0$, is that most of the weight is concentrated near the boundary, $E = \mu$. Hence most response elements found in the organism would be the "marginally" bound elements, rather different from the sequence that binds most strongly. This is one more example of how the entropy/phase-space effects dominate the evolution of response elements.

Appendix D: Empirical Determination of ε, μ Parameters

Given a set of known binding sites (8) for the jth factor, we find parameters $\boldsymbol{\varepsilon}^{(j)}$ and μ_j such that all the known binding site sequences for this factor (the "example sequences") $\boldsymbol{s}^{(k)}$ satisfy $\boldsymbol{s}^{(k)}\cdot\boldsymbol{\varepsilon}^{(j)} \leq \mu_j < 0$ with the maximal possible μ_j^2 (and $\boldsymbol{\varepsilon}^{(j)}\cdot\boldsymbol{\varepsilon}^{(j)} = 1$ constraint). Because specificity increases for large negative thresholds, μ, maximizing μ^2 insures that our model of the factor [i.e., $\boldsymbol{\varepsilon}^{(j)}$ and μ_j] has the highest specificity consistent with the data. This procedure minimizes the probability for a random sequence to satisfy the binding condition and therefore minimizes the number of "false-positive" sites in a genomic search for possible binding sites.

Restated in terms of a scaled variable $\tilde{\boldsymbol{\varepsilon}}^{(j)} \equiv \boldsymbol{\varepsilon}^{(j)}/|\mu|$, the problem becomes that of minimization of a quadratic form $\tilde{\boldsymbol{\varepsilon}}^{(j)}\cdot\tilde{\boldsymbol{\varepsilon}}^{(j)} = 1/\mu^2$ subject to a set of linear constraints $\tilde{\boldsymbol{\varepsilon}}^{(j)}\cdot\boldsymbol{s}^{(k)} < -1$. This problem is solved by "quadratic programming" (14). Our string search algorithm is different from the widely used weight matrix (9, 10) procedure. Its key advantage is the parsimony in the number of the candidate binding sites it generates for the low specificity factors where the weight-matrix approach (8) produces too many genomic "hits." The detailed results of the search are available electronically at http://www.physics.rockefeller.edu/~boris/public/search_ecoli.

Appendix E: Estimate of Coevolution Rate

Consider a transcription factor controlling n binding sites. How fast can its consensus sequence and the whole domain of binding drift through sequence space? Suppose a mutation of the factor causes a change in its DNA interaction so that $\varepsilon \rightarrow \varepsilon + \delta\varepsilon$ with probability $P(\delta\varepsilon)$. Such a mutation may cause each of the n RE links to fail with probability $1 - e^{-q(\delta\varepsilon)}$, so that the survival probability per mutation is given by $p = \int d(\delta\varepsilon)P(\delta\varepsilon)e^{-nq(\delta\varepsilon)}$. The effective diffusivity in ε is given by the variance of the "nonlethal" $\delta\varepsilon$ shift per mutation: $D_\varepsilon = p^{-1} \int d(\delta\varepsilon)P(\delta\varepsilon)e^{-nq(\delta\varepsilon)}(\delta\varepsilon)^2$. Assuming that mutation induces only small shifts in ε, we can approximate $q(\delta\varepsilon) \approx cnst \times (\delta\varepsilon)^2$ and for $n \gg 1$ we arrive at $D_\varepsilon \sim n^{-1}$ scaling. This provides a falsifiable quantitative prediction for the rate of evolutionary drift as a function of the degree of pleiotropy.

1. Gerhart, J. & M. Kirschner. (1997) *Cells, Embryos and Evolution* (Blackwell Scientific, Oxford).
2. Lewin, B. (1997) *Genes VI* (Oxford Univ. Press, Oxford), p. 812.
3. Ptashne, M. (1992) *A Genetic Switch* (Blackwell Scientific, Oxford), 2nd Ed.
4. Yuh, CH, Bolouri, H., Davidson, E. H. (1998) *Science* **279,** 1896–1902.
5. Neidhardt, F. C., ed. (1996) *E. coli and Salmonella: Cellular and Molecular Biology* (Am. Soc. Microbiol., Washington, DC).
6. Barkai, N. & Liebler, S. (1997) *Nature (London)* **387,** 913–917.
7. van Nimwegen, E., Crutchfield, J. P. & Huynen, M. (1997) *Proc. Natl. Acad. Sci. USA* **96,** 9716–9720.
8. Robison, K., McGuire, A. M. & Church, G. M. (1998) *J. Mol. Biol.* **284,** 241–254.
9. von Hippel, P. H. (1979) in *Biological Regulation and Development*, Goldberger, R. F., ed. (Plenum, New York), Vol. 1, pp. 279–347.
10. Stormo, G. D. & Fields, D. S. (1998) *Trends Biochem. Sci.* **23,** 109–113.
11. Muller-Hill, B. (1996) *The lac Operon* (de Gruyter, Berlin).
12. Smith, J. M. (1998) *Evolutionary Genetics* (Oxford Univ. Press, Oxford).
13. Waxman, D. & Peck, J. R. (1998) *Science* **279,** 1210–1213.
14. Fletcher, R. (1987) *Practical Methods of Optimization* (Wiley, New York).
15. McCue, L. et al. (2001) *Nucleic Acids Res.* **29,** 774–782.
16. Berg, O. G. & von Hippel, P. H. (1988) *J. Mol. Biol.* **200,** 709–723.
17. Kimura, M. (1994) *Molecular Evolution and the Neutral Theory* (Univ. of Chicago Press, Chicago).
18. Fontana, W. & Schuster, P. (1998) *Science* **280,** 1451–1455.
19. Ohta, T. (1992) *Annu. Rev. Ecol. Syst.* **23,** 263–286.
20. Abramowitz, M. & Stegun, I. A., eds. (1964) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards Applied Mathematical Series (U.S. Government Printing Office, Washington, DC), Vol. 55.