Supplemental Material: Annu. Rev. Biophys. 2019. 48:121-163
https://doi.org/10.1146/annurev-biophys-052118-115525

Figure 1 Theory Meets Figure 2 Experiments in the Study of Gene Expression
Supplementary Information

Rob Phillips[1,2,*],Nathan Belliveau[2], Griffin Chure[2], Hernan Garcia[3], Manuel Razo[2], Clarissa Scholes[4]
1 Dept. of Physics, California Institute of Technology, Pasadena, California, U.S.A
2 Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, U.S.A
3 Department of Molecular & Cell Biology, Department of Physics, Biophysics Graduate Group and Institute for Quantitative Biosciences-QB3, University of California, Berkeley, California, U.S.A
4 Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, U.S.A
∗ E-mail: phillips@pboc.caltech.edu

This Appendix aims to spell out in full detail some of the key technical issues that arise in the attempt to make quantitative theoretical models of transcriptional regulation.

S1.

The theoretical models presented in this work rely on the fundamental assumption that mRNA copy number can act as a proxy for the occupancy of the promoter by RNA polymerase. Only through this assumption are we able to relate experimentally accessible quantities, such as mRNA copy number or number of fluorescent proteins, to the promoter states that are considered theoretically. In this section we explore the validity and reach of this so-called occupancy hypothesis by considering the mathematical relationship between mRNA copy number, m, and the probability of finding RNA polymerase bound to the promoter, p_{bound}.

To make this analysis possible, we consider the simple model of transcription shown in Figure S1. As seen in the figure, we model each step between polymerase binding and mRNA production as a zero-order transition. In this context, the fraction of promoters in the process of initiating transcription, I, is given by

$$\frac{dI}{dt} = r_i p_{bound} - r_e I, \tag{S1}$$

where r_i is the rate of initiation, and r_e is the rate of elongation. As elongation ensues, we will keep track of which base pair the polymerase is located on using the fraction of polymerase molecules occupying base pair j, which we denote by E_j. The fraction of molecules at the first base pair can be obtained by solving

$$\frac{dE_1}{dt} = r_e I - r_e E_1. \tag{S2}$$

Similarly, for base pair $j < N$, where N is the length of the gene being transcribed, we have

$$\frac{dE_j}{dt} = r_e E_{j-1} - r_e E_j. \tag{S3}$$

Finally, the fraction of polymerase molecules at the last base pair is given by

$$\frac{dE_N}{dt} = r_e E_{N-1} - r_t E_N, \tag{S4}$$

where r_t is the rate of termination. Once an mRNA is terminated we assume that it is subject to degradation at a rate γ such that is concentration m is given by

$$\frac{dm}{dt} = r_t E_N - \gamma m. \tag{S5}$$

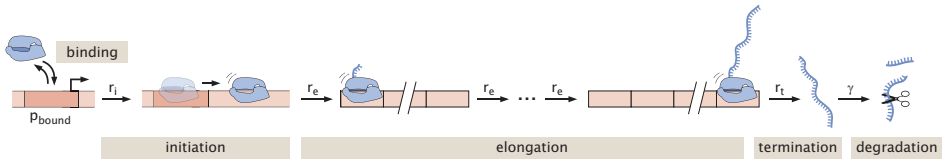


Figure S1. Simple model of mRNA production to probe the occupancy hypothesis. We assume that all steps from RNA polymerase binding to the termination and subsequent degradation of mRNA are described by zero-order kinetics.

By solving the system of equations shown above, we can then relate the magnitude predicted by our models, p_{bound}, to the measurable number of mRNA molecules m.

In order to solve for m using the above equations, we will assume steady-state such that all derivatives are zero. Further, due to the fact that every step in the process shown in Figure S1 is linear in the concentrations of the different molecular species, we can make use of a very convenient property of the system of equations. Specifically, we add up all equations together resulting in

$$0 = r_i p_{bound} - \gamma m \tag{S6}$$

such that

$$m = \frac{r_i}{\gamma} p_{bound}. \tag{S7}$$

This provides us with the first important result. Specifically, under conditions of steady-state and assuming a transcriptional cascade composed of zero-order reactions, we find a simple linear relationship between the mRNA copy number and the occupancy state of the promoter, as determined through p_{bound}.

Under slightly different assumptions, the occupancy hypothesis can also be used to relate p_{bound} to the rate of mRNA production dm/dt as shown in Equation 1. First, we relax the assumption made above that all the processes described by Equations S1 through S5 are in steady-state. Instead we posit that only the processes up until Equation S5 reached this steady-state. To put this in other words, we will set only the derivatives in Equations S1 through S4 to zero. If we, once again, add up the system of equations, we arrive at

$$\frac{dm}{dt} = r_i p_{bound} - \gamma m. \tag{S8}$$

Finally, we consider that mRNA degradation is negligible. This assumption true as long as the rate of initiation is faster than the degradation term such that $r_i \ll \gamma m$. Under this condition, we can neglect the last term on the right-hand side of Equation S8 leading to

$$\frac{dm}{dt} \approx r_i p_{bound} \tag{S9}$$

which is Equation 1 if we identify the rate of transcriptional initiation r_i with the effective rate of mRNA production r used throughout the main text.

S2. Equivalence of thermodynamic and statistical mechanical models of promoter occupancy

We next consider how the the statistical mechanical formulation of expression (Bintu et al. [1]) compares with alternative thermodynamic formulations that use the language of dissociation constants (e.g. Buchler, Gerland, and Hwa [2, 3, 4], and introduced by Shea and Ackers [5, 6]). We begin with the statistical mechanical formulation of the simple repression architecture and calculate the probability of RNA polymerase bound to its target promoter, p_{bound}. We then consider how this formulation

relates to thermodynamic formulations using dissociation constants. In doing so, we are able to show how these dissociation constants implicitly include a factor N_{NS} that was explicitly present in the statistical mechanical formulation and accounts for the reservoir of nonspecific binding sites on the genomic background.

Regardless of how we arrive at our model of transcriptional regulation, these models are all founded upon an assumption that the observed expression is proportional to the binding probability of RNA polymerase and that an assumption of steady-state is sufficiently valid. Here we begin by outlining the statistical mechanical formulation of the simple repression architecture [7]. We effectively treat the genome as a reservoir containing N_{NS} nonspecific binding sites bound by RNA polymerase and a number of different transcription factors (Figure 10(A)). Due to the high concentration of DNA in the cell it is generally reasonable to assume that most, if not all of the transcription factors in the cell are bound to the genomic DNA [8, 9].

Here we would like to estimate the probability that RNA polymerase is bound to our simple repression promoter, p_{bound}, that is present on the genome. As shown in Figure 10(B), the promoter can either be empty, occupied by RNA polymerase, or occupied by a repressor (in this case, LacI). This probability depends on the difference in free energy associated with each particular state of the system. We will take as a reference state that where all RNA polymerase and LacI proteins are bound nonspecifically to the genomic background. Following this approach, the probability of bound RNA polymerase, p_{bound} can be found to be given by,

$$p_{\text{bound}} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P}}{1 + \frac{R}{N_{NS}} e^{-\beta \Delta \varepsilon_R} + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_P}}, \tag{S10}$$

with $\beta = \frac{1}{k_B T}$, where k_B is the Boltzmann constant and T is the temperature of the system. Here, $\Delta \varepsilon_P$ denotes the difference in binding energy when repressor binds the promoter, relative to nonspecific binding on the genome. $\Delta \varepsilon_P$ similarly denotes the difference in binding energy when RNA polymerase binds the DNA. R and P represent the copy number per cell of repressor and RNA polymerase, respectively. Note that in our formulation, we have assumed that both the repressor and RNA polymerase are unable to bind simultaneously.

Now we can consider the thermodynamic approach that was taken by Buchler, Gerland, and Hwa [3]. In their work, the authors adopted and generalized the approach in the classic work of Shea and Ackers [5, 6] and so we shall begin there. In that classic work, Shea and Ackers developed a statistical mechanical model to describe the bacteriophage lambda switch, enumerating each possible configuration of the regulatory architecture. Following their approach, we will denote $\Delta \acute{G}_P$ as the free energy for binding of RNA polymerase to the promoter, and $\Delta \acute{G}_R$ for binding of LacI to the promoter. In their framework, the probability that RNA polymerase is bound to the promoter, p_{bound}, is then given by

$$p_{\text{bound}} = \frac{[P] e^{-\beta \Delta \acute{G}_P}}{1 + [P] e^{-\beta \Delta \acute{G}_P} + [R] e^{-\beta \Delta \acute{G}_R}}, \tag{S11}$$

where $[P]$ and $[R]$ are the concentrations of unbound RNA polymerase and unbound LacI, respectively. The free energies can be related to corresponding dissociation constants through the standard relationship,

$$\Delta \acute{G}_P = k_B T \ln \frac{K_P}{c_0}, \tag{S12}$$

and

$$\Delta \acute{G}_R = k_B T \ln \frac{K_R}{c_0}, \tag{S13}$$

although note that in each case the argument of the logarithm is normalized by a standard state concentration c_0, normally taken to be 1 M. Here K_P is the dissociation constant for binding by RNA polymerase

to the promoter, and K_R is the dissociation constant for binding of LacI to the promoter. These dissociation constants represent the concentration when each binding site is half-maximally occupied. Using these relationships between energy and dissociation constants in Equation S12 and Equation S13, we can re-write p_{bound} as,

$$p_{\text{bound}} = \frac{\frac{[P]}{K_P}}{1 + \frac{[P]}{K_P} + \frac{[R]}{K_R}}. \tag{S14}$$

This is the thermodynamic representation that would be obtained following the approach of Buchler, Gerland, and Hwa [3]. Here we see that the probability is still determined by considering the set of states available to the promoter, but with the corresponding Boltzmann weight for binding by RNA polymerase defined by $[P]/K_P$, and that of LacI by $[R]/K_R$.

Comparing the statistical mechanical equation of p_{bound} in Equation S10 with the thermodynamic representation in Equation S14 above, we find that

$$K_P = \frac{N_{NS}}{V_{\text{cell}}} e^{-\beta \Delta \varepsilon_P}, \tag{S15}$$

and

$$K_R = \frac{N_{NS}}{V_{\text{cell}}} e^{-\beta \Delta \varepsilon_R}. \tag{S16}$$

Here V_{cell} refers to the volume of the cell and is used to translate between protein copy numbers and concentrations. In the *in vivo* context considered here, the dissociation constants reflect binding by proteins that are otherwise assumed to be bound to the nonspecific genomic background, and will generally differ from what might be obtained from *in vitro* measurements [4]. Hence, we argue that both the statistical mechanical and thermodynamic formulations represent equivalent descriptions. The main distinction is that the statistical mechanical formulation is explicit in describing the nonspecific genomic background through the term N_{NS} and assuming one copy of the promoter.

S3. The equilibrium assumption

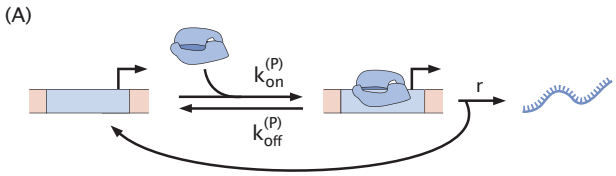
Having established the conditions under which we can connect the probability of finding RNA polymerase bound to the promoter, p_{bound}, with the rate of mRNA production, we now ask whether it is reasonable to use the tools of statistical mechanics to calculate p_{bound}. While we are encouraged by the apparent validity of the theory based on the agreement with experimental data shown throughout the main text, here we will carefully consider the equilibrium assumption that underlies calculating p_{bound} in the context of our minimal parameter set (defined in Figure 13(B)). While it will be shown below that the rates of RNA polymerase binding and unbinding are incompatible with an equilibrium assumption for binding by RNA polymerase, we will find that under the weak-promoter approximation, there exists a regime where it is indeed reasonable to apply a statistical mechanical treatment to calculate p_{bound}.

First, we focus on the model of an unregulated promoter shown in Figure S2(A). Here, the promoter can be unoccupied or occupied by RNA polymerase. The fraction of promoters in each state is denoted by $p_{unbound}$ and p_{bound}, respectively. When RNA polymerase is bound it can also initiate transcription at a rate r. Upon RNA polymerase escape from the promoter, the system is taken back to an unoccupied state. The rate of change in the fraction of occupied promoters is given by

$$\frac{dp_{bound}}{dt} = k_{on}^{(P)} p_{unbound} - k_{off}^{(P)} p_{bound} - r\, p_{bound} \tag{S17}$$

while the rate of mRNA production can be written as

$$\frac{dm}{dt} = r p_{bound} \tag{S18}$$



(B)

$k_{on}^{(P)}$ (1/min)	$k_{off}^{(P)}$ (1/min)	r (mRNA/min)
1	1	0.1
1	0.1	0.1
0.1	1	0.1
0.5	0.5	1

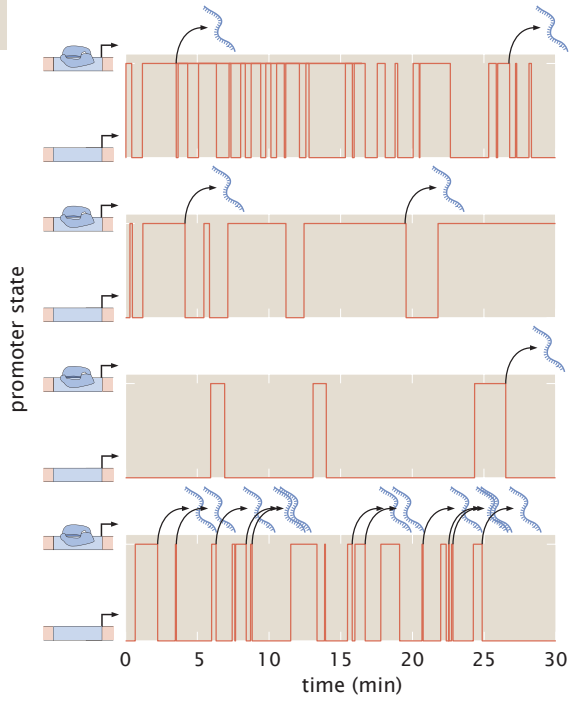


Figure S2. Exploring the equilibrium assumption for the constitutive promoter. (A) Kinetic scheme for a constitutive promoter. (B) Stochastic simulations of promoter state and initiation events for different parameters of the constitutive promoter.

which corresponds to the rate of mRNA production as posited by the occupancy hypothesis.

We next seek to establish under what conditions we can calculate p_{bound} using statistical mechanics. In the equilibrium limit, p_{bound} for this unregulated promoter can be calculated using the states and weights defined in Figure 9(A) such that

$$p_{bound}^{equil} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_p}}{1 + \frac{P}{N_{NS}} e^{-\beta \Delta \varepsilon_p}}. \tag{S19}$$

In Appendix S2, we saw that this same expression can be written in the thermodynamic language as

$$p_{bound}^{equil} = \frac{\frac{[P]}{K_P}}{1 + \frac{[P]}{K_P}}, \tag{S20}$$

where K_P is the dissociation constant between RNA polymerase and the promoter. This expression for p_{bound}^{equil} can be related to the scheme shown in Figure S2(A) by using $[P]/K_P = k_{on}^{(P)}/k_{off}^{(P)}$ such that

$$p_{bound}^{equil} = \frac{k_{on}^{(P)}}{k_{on}^{(P)} + k_{off}^{(P)}}. \tag{S21}$$

In order to calculate p_{bound} without enforcing equilibrium, we invoke steady-state in the fraction of occupied and unoccupied promoters such that Equation S17 can be written as

$$0 = k_{on}^{(P)} p_{unbound} - k_{off}^{(P)} p_{bound} - r\, p_{bound}. \tag{S22}$$

We now make use of the fact that the probabilities are normalized, $p_{bound} + p_{unbound} = 1$ in order to obtain

$$p_{bound} = \frac{k_{on}^{(P)}}{k_{on}^{(P)} + k_{off}^{(P)} + r}. \tag{S23}$$

Clearly, p_{bound} in Equation S23 is not equal to p_{bound}^{equil} in Equation S21. The only way to recover p_{bound}^{equil} is for the rate of initiation r to be much slower that one of the other rates in the system. Namely, we need $r \ll k_{on}^{(P)}$ or $r \ll k_{off}^{(P)}$ such that $k_{on}^{(P)} + k_{off}^{(P)} + r \approx k_{on}^{(P)} + k_{off}^{(P)}$. These different limits are explored in Figure S2(B) through stochastic simulations that calculate the promoter state and initiation events as a function of time. In the first three simulations within Figure S2(B), we show how, when the conditions described above are met, the promoter cycles multiple times between its bound and unbound state before an initiation event ensues. This back-and-forth between the bound and unbound states leads to *quasiequilibrium*. That is, the fact that the transitions between the bound and unbound states are faster than the rate of initiation allows us to invoke separation of time scales such that, at each time point, we can use statistical mechanics to describe the equilibrium between these two states. However, if r is larger than these transition rates, most instances of the promoter being bound lead to an initiation event as shown in the last simulation in the Figure S2(B) and there is no longer a separation of time scales.

Interestingly, the inferred transition rates from Figure 13(B) do not fulfill this condition as $k_{on}^{(P)}, k_{off}^{(P)} < r$. Thus, at least *a priori*, equilibrium cannot be invoked to describe the transcription of an *unregulated lac* promoter. However, the successes of the theory at predicting experiments suggest that, under certain conditions, we are still allowed to invoke the quasi-equilibrium assumption for the *regulated lac* promoter.

We next consider the kinetic scheme for the regulated promoter, shown in Figure S3(A). The reader is reminded that this scheme does not make any assumption about the relative strength of each transition rate or about equilibrium. In this context, we are first interested in asking whether the probability of

finding RNA polymerase bound to the promoter $p(3) = p_{bound}$, which we solved for in Equation 16, is equivalent to the same probability that can be calculated in the equilibrium case, p_{bound}^{equil}, shown in Equation 4.

To make progress, we rewrite p_{bound}^{equil} in Equation 4 in the language of dissociations constants

$$p_{bound}^{equil} = \frac{\frac{[P]}{K_P}}{1 + \frac{[P]}{K_P} + \frac{[R]}{K_R}}. \tag{S24}$$

Invoking the identities introduced in Section 4.2 such that $k_{on}^{(R)} = k_{+}^{(R)}[R]$ and $k_{on}^{(P)} = k_{+}^{(P)}[P]$, and the definition of the dissociations constant for repressor and RNA polymerase given by $k_{off}^{(R)}/k_{+}^{(R)} = K_R$ and $k_{off}^{(P)}/k_{+}^{(P)} = K_P$, respectively, we obtain

$$p_{bound}^{equil} = \frac{\frac{k_{on}^{(P)}}{k_{off}^{(P)}}}{1 + \frac{k_{on}^{(P)}}{k_{off}^{(P)}} + \frac{k_{on}^{(R)}}{k_{off}^{(R)}}}. \tag{S25}$$

In contrast, p_{bound} from Equation 16, which is absent of any assumption of equilibrium, is given by

$$p_{bound} = \frac{\frac{k_{on}^{(P)}}{k_{off}^{(P)}+r}}{1 + \frac{k_{on}^{(P)}}{k_{off}^{(P)}+r} + \frac{k_{on}^{(R)}}{k_{off}^{(R)}}}. \tag{S26}$$

Again, as with the unregulated promoter, we find that the expression for p_{bound} is not equal to p_{bound}^{equil}. One way to alleviate this discrepancy is through the quasiequilibrium assumption noted above, requiring that the rate of RNA polymerase unbinding is much faster than the rate of initiation, $k_{off}^{(P)} \ll r$. However, Figure 13(B) reveals that $k_{off}^{(P)} \approx r$ and not $k_{off}^{(P)} \ll r$ as demanded above for the quasiequilibrium approximation to apply. Interestingly, at least for the case of simple repression considered here, we will see below that the equilibrium assumption can still be invoked under certain conditions for the calculation of the fold-change in gene expression.

In Equation 17 in the main text, we calculated the fold-change in gene expression corresponding to the kinetic scheme presented in Figure 12 and reproduced in Figure S3(A). This calculation made no assumption regarding equilibrium and resulted in

$$\text{fold-change} = \frac{1 + \frac{k_{on}^{(P)}}{k_{off}^{(P)}+r}}{1 + \frac{k_{on}^{(P)}}{k_{off}^{(P)}+r} + \frac{k_{on}^{(R)}}{k_{off}^{(R)}}}. \tag{S27}$$

Our objective is then to determine under what limits we can reduce this fold-change to its equilibrium counterpart obtained in Equation 7 or in the context of the weak-promoter approximation shown in Equation 8.

As expected from our calculations on the applicability of equilibrium to derive p_{bound}, if we assume that $k_{off}^{(P)} \ll r$, Equation S27 reduces to the fold-change in equilibrium shown in Equation 7. We already saw that this limit is not consistent with the inferred rates. However, instead, consider the limit where $k_{on}^{(P)} \ll k_{off}^{(P)} + r$. In this case, we can neglect the term $\frac{k_{on}^{(P)}}{k_{off}^{(P)}+r}$ in Equation S27 such that the fold-change reduces to

$$\text{fold-change} \approx \frac{1}{1 + \frac{k_{on}^{(R)}}{k_{off}^{(R)}}} = \frac{1}{1 + \frac{[R]}{K_R}}, \tag{S28}$$

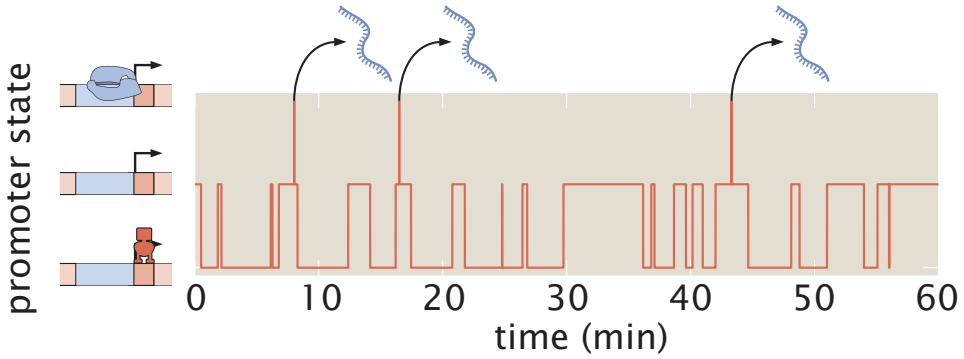


Figure S3. Exploring the equilibrium assumption for simple repression. Stochastic simulations of promoter state and initiation events for the kinetic scheme introduced in Figure 12 for different parameters of the regulated promoter, for the case where $k_{on}^{(P)} \ll k_{off}^{(P)} + r$. Here we observe many more binding and unbinding events by the repressor than by RNA polymerase, characteristic of our statistical mechanical description. The parameters used are $k_{on}^{(P)} = 0.1$ min^{-1}, $k_{off}^{(P)} = 1$ min^{-1}, $k_{on}^{(R)} = 0.5$ min^{-1}, $k_{off}^{(R)} = 0.5$ min^{-1}, and $r = 60$ min^{-1}.

which corresponds to the fold-change in equilibrium under the weak-promoter approximation shown in Equations 8 and 9. In Figure S3(B) we explore this regime using stochastic simulations. The simulation reveals that, in this limit, the promoter mostly transitions between its repressor-occupied state and its empty state. Only rarely will the system transition to the RNA polymerase-bound state and, on these rare occasions, this event almost always leads to the initiation of transcription and the return of the promoter to its empty state. As a result, there is a clear separation of time scales between the process of repressor binding and unbinding and the subsequent steps in the transcriptional cascade. This separation of time scales justifies the applicability of the quasiequilibrium assumptions to calculate the fold-change in gene expression in terms of the probability of repressor binding.

As seen in Figure 13(B), our estimates for $k_{on}^{(R)}$, $k_{off}^{(R)}$ and r suggest that we are in this regime where the fold-change in gene expression can be calculated using the tools of statistical mechanics despite the fact that the probability of RNA polymerase binding to the promoter cannot be obtained using such equilibrium considerations. Thus, by considering fold-change instead of p_{bound} directly, we are able to ignore the potentially non-equilibrium behavior of RNA polymerase.

S4. The nonspecific genomic background

A simplifying assumption often made in thermodynamic models of transcription is the idea that the binding of transcription factors to nonspecific sites is characterized by a single binding energy as shown in Figure S4(A). In this case, the partition function for putting P polymerases on the nonspecific background is

$$Z_{NS}(P, N_{NS}) = \frac{N_{NS}^P}{P!} e^{-\beta P \varepsilon_{NS}}. \tag{S29}$$

Of course, this is a convenient simplifying assumption that is pedagogically helpful, but raises the question of whether it masks some important effect. In fact, as we show in the remainder of this section, even when the nonspecific background is characterized by a distribution of energies, ultimately, it can be represented by an equation of the form Equation S29, but with the energy ε_{NS} replaced by an *effective energy*.

To get a feeling for how the effective energy arises, we begin with a toy model of the nonspecific background as shown in Figure S4(B). In this case, the P polymerases are distributed between the $N_{NS}/2$ sites available with binding energy $\varepsilon_1 = \bar{\varepsilon} + \Delta$ and the $N_{NS}/2$ sites available with binding energy $\varepsilon_2 = \bar{\varepsilon} - \Delta$ such that $\bar{\varepsilon}$ is the mean non-specific binding energy. To compute the partition function, we

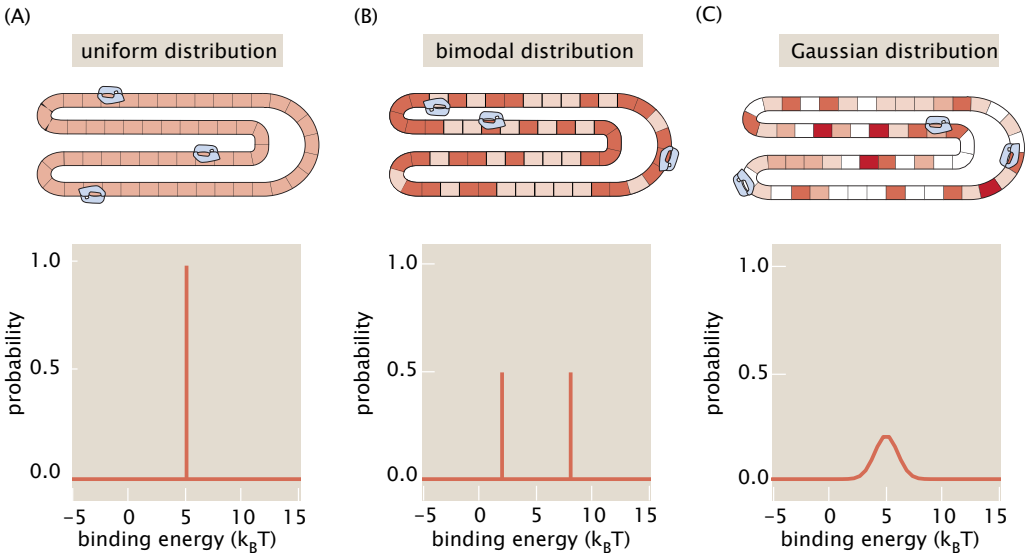


Figure S4. Increasingly sophisticated models of the nonspecific background. (A) Uniform background. (B) Two-state model of the nonspecific background. (C) Nonspecific binding energies characterized by a Gaussian distribution.

need to sum over *all* the ways of distributing the P polymerases over the two nonspecific reservoirs. We imagine that the number bound on reservoir 1 is i and the number bound on reservoir 2 is $P - i$, and then sum over all i ranging from $i = 0$ all the way to $i = P$, resulting in

$$Z_{NS} = \sum_{i=0}^{P} g_1(i) g_2(P-i) e^{-\beta[i\varepsilon_1 + (P-i)\varepsilon_2]}, \tag{S30}$$

where $g_1(i)$ is the number of ways of distributing i polymerases over the $N_{NS}/2$ sites of reservoir 1 and $g_2(P-i)$ is the number of ways of distributing $P-i$ polymerases over the $N_{NS}/2$ sites of reservoir 2. Because $i << N_{NS}/2$, we can write $g_1(i)$ as

$$g_1(i) \approx \frac{(\frac{N_{NS}}{2})^i}{i!} \tag{S31}$$

and similarly write $g_2(P-i)$ as

$$g_2(P-i) \approx \frac{(\frac{N_{NS}}{2})^{P-i}}{(P-i)!}. \tag{S32}$$

In light of these results, we can now rewrite the partition function for nonspecific binding as

$$Z_{NS} = \sum_{i=0}^{P} \frac{(\frac{N_{NS}}{2})^P}{i!(P-i)!} e^{-\beta[i\varepsilon_1 + (P-i)\varepsilon_2]} \tag{S33}$$

which can be rewritten as

$$Z_{NS} = \frac{(\frac{N_{NS}}{2})^P}{P!} e^{-\beta P \varepsilon_2} \sum_{i=0}^{P} \frac{P!}{i!(P-i)!} e^{-\beta i(\varepsilon_1 - \varepsilon_2)}, \tag{S34}$$

where we have multiplied the previous expression by $P!/P! = 1$ in anticipation of beating our formula into the form of a binomial. Indeed, our sum is now of the form of a binomial allowing us to use

$$\sum_{i=0}^{P} \frac{P!}{i!(P-i)!} x^P = (1+x)^P. \tag{S35}$$

As a result, we can write our partition function in the form

$$Z_{NS} = \frac{N_{NS}^P}{P!} \frac{1}{2^P} (e^{-\beta\varepsilon_2}(1 + e^{-\beta(\varepsilon_1 - \varepsilon_2)}))^P. \tag{S36}$$

This should be compared with

$$Z_{NS} = \frac{N_{NS}^P}{P!} e^{-\beta P \varepsilon_{NS}} \tag{S37}$$

which is the result for the partition function for the most simple model in which the nonspecific background is assumed to be uniform.

We now want to see whether our expression given in eqn. S36 is equivalent to the single reservoir model. By equating eqn. S36 and eqn. S37 and taking the log of both sides we have

$$\varepsilon_{NS} = k_B T \ \ln 2 + \varepsilon_2 - k_B T \ \ln \ (1 + e^{-\beta(\varepsilon_1 - \varepsilon_2)}) \tag{S38}$$

We can simplify this by noting that the term involving the logarithm can be simplified as

$$\ln(1 + e^{-\beta(\varepsilon_1 - \varepsilon_2)}) = \ln(1 + e^{-2\beta\Delta}) \approx \ln(1 + 1 - 2\beta\Delta) \approx \ln \ 2 + \ln \ (1 - \beta\Delta), \tag{S39}$$

where we have used the fact that $\varepsilon_1 - \varepsilon_2 = 2\Delta$. Given that $\beta\Delta << 1$ (i.e. the energy difference between the two states is small), we can use the Taylor series $\ln \ (1 - x) \approx -x$ with the result that

$$\varepsilon_{NS} = \bar{\varepsilon} \tag{S40}$$

This result shows us that in the toy model of the nonspecific background of Figure S4(B), the two nonspecific backgrounds are equivalent to a single reservoir with an energy given by the mean of the energies of the two reservoirs, establishing that in this pedagogically motivated model we can use a single energy to describe the nonspecific background. Now let's move to the case of realistic distribution of nonspecific energies.

Figure 17 shows the distribution of nonspecific binding energies obtained by taking the energy matrix describing the binding of LacI and applying it to all sites across the *E. coli* genome (also see Figure S4(C) for a comparison with the other models considered thus far). Other examples of the distribution of nonspecific binding energies have been considered as well with similar outcome [2, 10]. As a result, we can write the number of binding sites with energy between E and $E + dE$ as

$$n(E) = \frac{N_{NS}}{\sqrt{2\pi\sigma^2}} e^{-(E-\bar{\varepsilon})^2/2\sigma^2}, \tag{S41}$$

where $\bar{\varepsilon}$ is the mean of the distribution of nonspecific binding energies and σ provides a measure of the width of that distribution.

To compute the partition function for the binding of a polymerase, for example, to this nonuniform genomic background, we need to sum over all the microscopic states available to the polymerase. Symbolically, the quantity we need to evaluate is

$$Z_{NS} = \sum_E n(E) e^{-\beta E}. \tag{S42}$$

In fact, since we are assuming a continuous distribution of energies, this really is an integral of the form

$$Z_{NS} = \int_{\infty}^{\infty} e^{-\beta E} \frac{N_{NS}}{\sqrt{2\pi\sigma^2}} e^{-(E-\bar{\varepsilon})^2/2\sigma^2} dE. \tag{S43}$$

This result can be rewritten as

$$Z_{NS} = \frac{N_{NS}}{\sqrt{2\pi\sigma^2}} e^{-\bar{\varepsilon}^2/2\sigma^2} \int_{\infty}^{\infty} e^{-\frac{1}{2\sigma^2}[E^2 - 2E\bar{\varepsilon} + 2\sigma^2\beta E]} dE. \tag{S44}$$

By completing the square, this integral results in

$$Z_{NS} = N_{NS} e^{-\beta\bar{\varepsilon}} e^{\beta^2\sigma^2/2} \tag{S45}$$

which should be compared with the result we would get if we assumed a homogeneous nonspecific background with only a single binding energy ε_{NS} resulting in the form

$$Z_{NS} = N_{NS} e^{-\beta\varepsilon_{NS}} \tag{S46}$$

By equating these two expressions, we find that we can treat the nonuniform background as though it were a homogenous genomic background with effective binding energy

$$\varepsilon_{eff} = \bar{\varepsilon} - \frac{\beta\sigma^2}{2}. \tag{S47}$$

The result above considered a single polymerase or repressor molecule bound to the nonuniform nonspecific background. What happens in the case where we have P polymerases bound nonspecifically? Because each of those polymerases binds independently of the others (because the number of polymerases is of order $10^3 - 10^4$ and the genome size is greater than 10^6 we don't need to worry about polymerases interfering with each other), the total partition function for all of these polymerases bound to the nonspecific background is given by

$$Z_{NS}(P, N_{NS}) = \frac{(\int_{\infty}^{\infty} e^{-\beta E} \frac{N_{NS}}{\sqrt{2\pi\sigma^2}} e^{-(E-\bar{\varepsilon})^2/2\sigma^2} dE)^P}{P!} = \frac{N_{NS}^P e^{-\beta P \varepsilon_{eff}}}{P!}, \tag{S48}$$

where once again $\varepsilon_{eff} = \bar{\varepsilon} - \frac{\beta\sigma^2}{2}$ and this result shows that if the distribution of binding energies is Gaussian, then we can treat the nonspecific background as being equivalent to a uniform nonspecific background with energy ε_{eff}. The point of all of this analysis was simply to examine the validity of the convenient simplifying assumption of some thermodynamic models of treating the nonspecific background as uniform. As shown elsewhere [2, 10], this approximation is quite reasonable.

S5. Accounting for the effect of nonspecific promoter occupancy

So far our statistical mechanical treatment of the simple repression architecture has treated the RNA polymerase and LacI proteins as isolated from the pool of other transcription factors that are also littered across the genomic DNA. In Figure 6 we plot the abundance of DNA-binding proteins per cell across a number of growth conditions using the proteomic study from Schmidt et al. [11]. These values include nucleoid-associated proteins that also bind the genomic DNA. For growth in M9 minimal media with 0.5% glucose, we find that there are about 3×10^5 DNA-binding proteins per cell and we can use this to make a simple estimate of genomic occupancy by these proteins. Let us assume that each transcription factor binds the DNA as a dimer (this will vary with the transcription factor species) and occupies a DNA

length of 15 bp (this varies from 7 bp to 38 bp in *E. coli* for transcription factors listed on RegulonDB; [12]). For growth in 0.5% glucose, we find that about 2.3 Mbp or about half the genome is occupied (15 bp$\times 3 \times 10^5$DNA-binding proteins$\times 1/2$ dimers per protein).

Given the high occupancy of DNA-binding proteins on the genomic DNA estimated above, there might be some expectation that, in contrast to our current model of simple repression, the occupancy of the genome by these other DNA-binding proteins cannot be ignored. Here we consider the effect of their occupancy by adding an explicit set of states to represent the case where these additional DNA-binding proteins can occupy the roughly 60 bp promoter region of our simple repression architecture. For simplicity we assume that these proteins only bind nonspecifically, ignoring any potential sequence-specific effects. In Figure S5(A) we show the states and weights of the simple repression promoter, where we have included this additional set of states. We could have extended this further, either by treating each additional DNA-binding protein species separately, or by being more careful about our specification of these additional states. However, the point of this exercise is to see what effect the pool of nonspecifically bound DNA-proteins might have on our model. We can calculate p_{bound} which, if we invoke the weak promoter approximation $\left(\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P} \ll 1\right)$, is given by

$$p_{\text{bound}} = \frac{\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}}{1 + L \cdot \frac{C_{NS}}{N_{NS}} + \frac{R}{N_{NS}} e^{-\beta \Delta\varepsilon_R}}. \tag{S49}$$

L represents the number of ways other DNA-binding proteins may bind the promoter nonspecifically, and for simplicity is taken as the length of the promoter region ($L \approx 60$ bp). C_{NS} represents the copy number of all other DNA-binding proteins bound to the genome that we noted earlier. Fold-change, which is the ratio of $p_{\text{bound}}(R \geq 0)$ to $p_{\text{bound}}(R = 0)$, will then be given by

$$\text{fold-change} = \frac{1 + L \cdot \frac{C_{NS}}{N_{NS}}}{\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}} \cdot \frac{\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}}{1 + L \cdot \frac{C_{NS}}{N_{NS}} + \frac{R}{N_{NS}} e^{-\beta \Delta\varepsilon_R}}. \tag{S50}$$

Here, the RNA polymerase components $\frac{P}{N_{NS}} e^{-\beta \Delta\varepsilon_P}$ cancel out and upon some rearrangement, we find that

$$\text{fold-change} = \frac{1}{1 + \frac{R}{N_{NS}} e^{-\beta \Delta\varepsilon_R} (1 + L \cdot \frac{C_{NS}}{N_{NS}})^{-1}}. \tag{S51}$$

Using $C_{NS} \approx 1.5 \times 10^5$, which is based on our estimate of the total DNA-binding protein copy number found above for growth in glucose (bound as dimers), we calculate a value of $L \cdot \frac{C_{NS}}{N_{NS}} \approx 2$. Importantly, we find that this additional term in our fold-change equation does not depend on the key parameters of our simple repression architecture, namely the repressor copy number or repressor binding energy, and we can arrive back to our original form of fold-change by a defining $N'_{NS} = N_{NS} \times (1 + L \cdot \frac{C_{ns}}{N_{NS}})$.

The estimates so far were based on assuming that cells grow in 0.5% glucose at a particular doubling rate. In different media, the growth rate will change leading also to a modulation in the total number of transcription factors: faster growing cells have a larger protein complement than their slower-growing counterparts. However, faster growing cells also have more copies of the genome as a means to keep up with the fast replication pace. Figure S5(B) shows that these two effects cancel each other out. Specifically, variations in the number of transcription factors as a result of changes in growth rate are counteracted by the corresponding change in the average genome copy number per cell such that the number of nonspecific binding proteins per base pair remains approximately constant throughout a wide range of growth conditions. As a result, the small effect of considering all nonspecifically bound transcription factors remains unaltered regardless of growth rate.

References

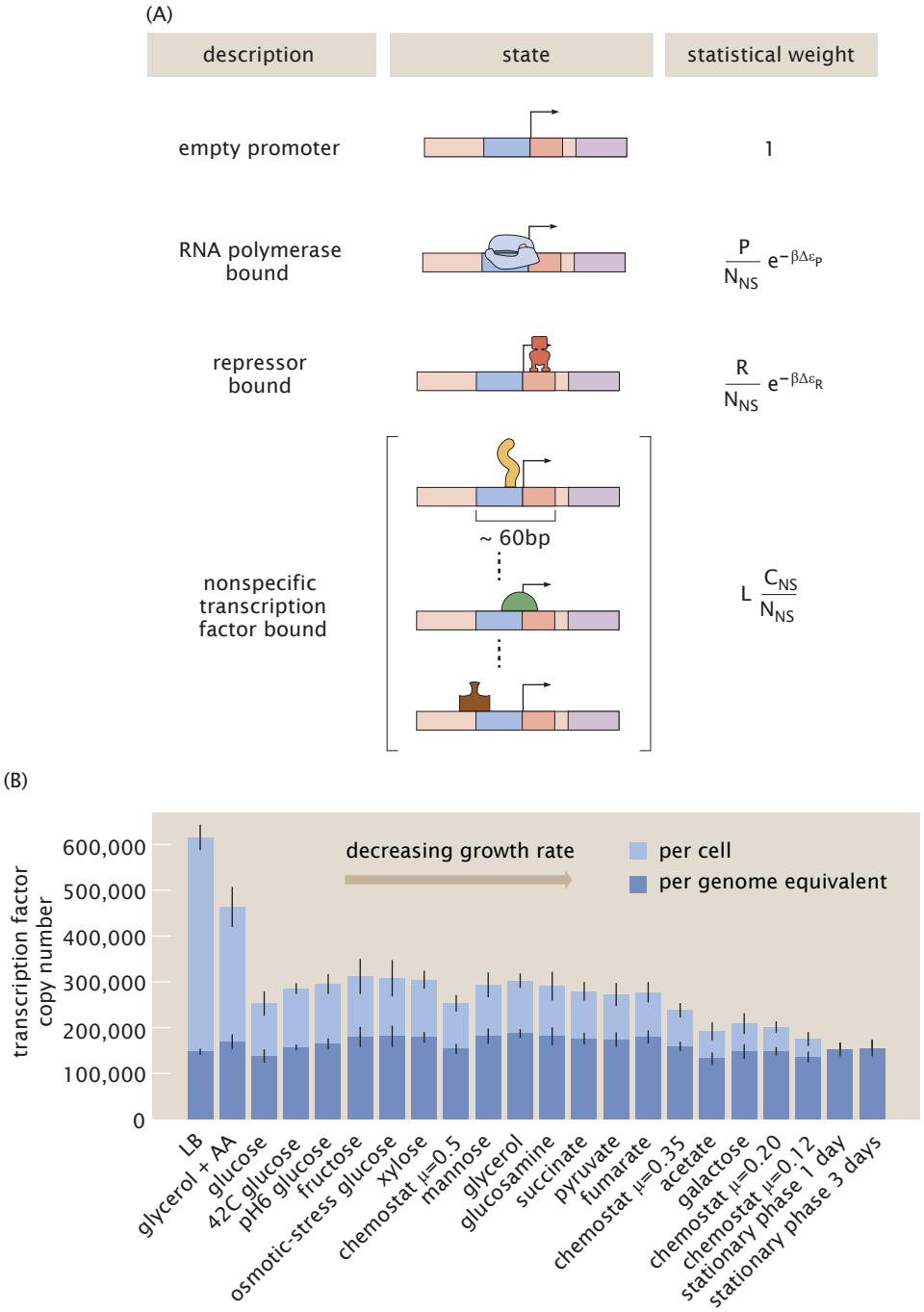


Figure S5. A crowded chromosome. (A) States and Weights for simple repression with a pool of nonspecific DNA binding proteins. RNA polymerase (light blue), a repressor, and other nonspecific DNA binding proteins compete for binding to a promoter. The R repressors and P RNA polymerase bind with energies $\Delta\varepsilon_R$ and $\Delta\varepsilon_P$, respectively. In addition, there are C_{NS} DNA binding proteins per cell that can bind the promoter of length $L \approx 60$ bp. These proteins bind nonspecifically and therefore only contribute an entropic term. N_{NS} represents the number of nonspecific binding sites on the genome. (B) Measured protein copy numbers are shown for DNA binding proteins in *E. coli* across 22 growth conditions. Protein copy numbers per cell were determined by Schmidt et al. [11] with proteins identified based on their annotation in EcoCyc. Error bars are propagated from the reported standard deviations. Protein copy numbers per genome equivalent were calculated by estimating the total genomic content as a function of growth rate using Cooper and Helmstetter's model of *E. coli* chromosomal replication [13, 14, 15].

1. L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, R. Phillips, Transcriptional regulation by the numbers: models, Curr Opin Genet Dev 15 (2005) 116–24.

2. U. Gerland, J. D. Moroz, T. Hwa, Physical constraints and functional characteristics of transcription factor-DNA interaction, Proc Natl Acad Sci U S A 99 (2002) 12015–20.

3. N. E. Buchler, U. Gerland, T. Hwa, On schemes of combinatorial transcription logic, Proc Natl Acad Sci U S A 100 (2003) 5136–41.

4. T. Kuhlman, Z. Zhang, M. H. S. Jr., T. Hwa, Combinatorial transcriptional control of the lactose operon of *escherichia coli*, Proc Natl Acad Sci U S A 104 (2007) 6043–8.

5. G. K. Ackers, A. D. Johnson, M. A. Shea, Quantitative model for gene regulation by lambda phage repressor, Proc Natl Acad Sci U S A 79 (1982) 1129–33.

6. M. A. Shea, G. K. Ackers, The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation, J Mol Biol 181 (1985) 211–30.

7. H. G. Garcia, R. Phillips, Quantitative dissection of the simple repression input-output function, Proc Natl Acad Sci U S A 108 (2011) 12173–8.

8. P. L. deHaseth, C. A. Gross, R. R. Burgess, M. T. J. Record, Measurement of binding constants for protein-DNA interactions by DNA-cellulose chromatography, Biochemistry 16 (1977) 4777–83.

9. Y. Kao-Huang, A. Revzin, A. P. Butler, P. O'Conner, D. W. Noble, P. H. von Hippel, Nonspecific dna binding of genome-regulating proteins as a biological control mechanism: measurement of dna-bound *Escherichia coli lac* repressor *in vivo*, Proc Natl Acad Sci U S A 74 (1977) 4228–32.

10. A. M. Sengupta, M. Djordjevic, B. I. Shraiman, Specificity and robustness in transcription control networks, Proc Natl Acad Sci U S A 99 (2002) 2072–7.

11. A. Schmidt, K. Kochanowski, S. Vedelaar, E. Ahrne, B. Volkmer, L. Callipo, K. Knoops, M. Bauer, R. Aebersold, M. Heinemann, The quantitative and condition-dependent *Escherichia coli* proteome, Nature Biotech 34 (2016) 104–111.

12. S. Gama-Castro, H. Salgado, A. Santos-Zavaleta, D. Ledezma-Tejeida, L. Muñiz-Rascado, J. S. García-Sotelo, K. Alquicira-Hernández, I. Martínez-Flores, L. Pannier, J. A. Castro-Mondragón, A. Medina-Rivera, H. Solano-Lira, C. Bonavides-Martínez, E. Pérez-Rueda, S. Alquicira-Hernández, L. Porrón-Sotelo, A. López-Fuentes, A. Hernández-Koutoucheva, V. D. Moral-Chávez, F. Rinaldi, J. Collado-Vides, RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond, Nucleic Acids Res 44 (2016) D133–D143.

13. S. Cooper, C. E. Helmstetter, Chromosome replication and the division cycle of Escherichia coli Br, J Mol Biol 31 (1968) 519–540.

14. P. P. Dennis, H. Bremer, Modulation of Chemical Composition and Other Parameters of the Cell at Different Exponential Growth Rates, EcoSal Plus 3 (2008).

15. T. E. Kuhlman, E. C. Cox, Gene location and DNA density determine transcription factor distributions in *Escherichia coli*, Mol Syst Biol 8 (2012) 610.