æ

# TRANSCRIPTION

7 May 2014

All of cellular life relies on heritable information encoded in DNA, with all cellular functions being carried out by products acquired by transcription of DNA into RNA. In the case of protein-coding loci, these messenger RNAs (mRNAs) must subsequently be translated into strings of amino acids. Gene expression is generally regulated by specific DNA-binding proteins that activate and/or repress transcription in the genes whose regulatory elements they bind. Most genomes encode for hundreds to thousands of such **transcription factors** (hereafter, TFs), each with unique DNA binding-motif requirements called **transcription-factor binding sites** (hereafter, TFBSs). One can certainly imagine simpler mechanisms for using DNA-level information to make proteins (e.g., the use of no RNA intermediates at all), but these are the cards that were dealt to LUCA, and there is now no simple way to erase this legacy of the earliest stages of evolution.

Transcription factors are usually referred to as *trans*-acting regulatory elements, as the genetic loci encoding for them are generally unlinked to those that they regulate (at least in Eukaryotes), whereas TFBSs are generally referred to as *cis*-acting, as they are physically linked to the affected coding regions. This distinction is not always so clear in Prokaryotes, where a TF is sometimes encoded in the same multilocus transcriptional unit (called an **operon**) as its target gene.

A central issue with respect to understanding transcription and its consequences is stochasticity. Genes are generally present just once (haploids) or twice (diploids) within cells, and as will be seen below, specific mRNAs are often present in just a dozen or fewer copies, with proteins typically being an order of magnitude or more abundant. Owing to the small numbers of molecules of individual types relative to the vast space within cells, intermolecular encounters are by no means certain, and as a consequence there can be considerable cell-to-cell variation in gene expression even in a genetically homogeneous population. Thus, before discussing the biology and evolution of transcription, some simple quantitative principles regarding molecules in single cells need to be absorbed.

### Molecular Stochasticity in Single Cells

The fitness of cells ultimately depends on the quality, quantity, and stoichiometric relationships of its underlying functional constituents, our focus here being primarily on proteins. With typically just one (haploids) or two (diploids) genes encoding for each protein within a cell, and stochastic dynamics of transcription and translation

at play, the numbers of individual proteins vary on a cell-to-cell basis, even in a completely homogeneous environment. The many factors governing the probability distributions of numbers of molecules per cell can be subsumed into six rate coefficients: the rates at which an inactive gene enters the transcriptionally active state, and vice versa, $k_{\rm on}$ and $k_{\rm off}$ respectively; the rate at which an active gene transcribes mRNAs, $k_{\rm m}$; the rate at which an mRNA is translated into proteins, $k_{\rm p}$; and the rates of degradation of mRNAs and proteins, $\gamma_{\rm m}$ and $\gamma_{\rm p}$ respectively (Figure 1).

We first consider the numbers of mRNAs found in cells, $n_{\rm m}$. As noted in Figure 2, active cells gain mRNAs by transcription and lose them by degradation, whereas inactive cells (with no TF engaghed with the target TFBS) can only lose mRNAs. Cells can also move back and forth between active and inactive states. The transition rates to smaller numbers of mRNAs increase linearly with increasing $n_{\rm m}$ simply because there are more targets for degradation, and as a consequence, this system converges on an equilibrium distribution regardless of the starting point. As outlined in Details 1, a particularly simple outcome is obtained when a gene is constitutively turned on ($k_{\rm off} = 0$). In this case, $n_{\rm m}$ is Poisson distributed, with both the mean and the variance of the number of mRNAs per cell equal to the ratio of the rates of gain and loss, $k_{\rm m}/\gamma_{\rm m}$. The Poisson distribution is dominated by the zero (mRNA-free) class when the mean is smaller than 1.0, has maximum and equal probability in classes 0 and 1 when the mean is equal to 1.0, and converges on a normal (bell-shaped) distribution with larger mean (Figure 2).

Regulated genes are not continuously expressed, but instead are in the active state only a fraction of the time. Averaging over a sufficiently long period of time, this fractional time is a simple function of the ratio of association and dissociation rates of the TF,

$$P_{\rm on} = k_{\rm on}/(k_{\rm on} + k_{\rm off}). \tag{1}$$

This result arises because under equilibrium conditions, $P_{\rm on}k_{\rm off}$ must equal $(1 - P_{\rm on})k_{\rm on}$. The number of mRNAs per cell is then necessarily more complex, as it involves a mixture of the distributions in active and inactive cells (see Details 1 for the full expression).

Contrary to common belief, gene regulation causes an increase in the variance of mRNA numbers among cells (Figure 2). This occurs because transient switches from active to inactive states result in a heavier weighting towards the categories with small numbers of mRNAs. Indeed, if the rate of switching among active and inactive states is sufficiently slow relative to the rate of degradation of mRNAs, a bimodal distribution can result, with one fraction (the inactive cells) carrying few mRNAs and the remaining active cells having an mRNA-number distribution close to that expected under the constitutive expression model. To take things beyond theory, and understand the likely specific forms of the distribution of mRNA molecules per cell, we next consider the quantitative values for the four key parameters: $k_{\rm m}$, $k_{\rm on}$, $k_{\rm off}$, and $\gamma_{\rm m}$.

Based on observed rates of mRNA chain elongation, and assuming an average transcript length of $\sim 1.0$ kb, an activated *E. coli* gene is capable of producing $\sim 50$ to 150 transcripts/hour (Golding and Cox 2004; Proshkin et al. 2010). These are almost certainly substantial overestimates of the actual rates of transcript production, which also depend on events associated with transcription initiation and termination. Approximate estimates for the yeasts *S. cerevisiae* and *S. pombe*, again largely

based on chain-elongation rates and in this case assuming an average transcript length of 2 kb (Lynch 2007), fall in a lower range of $\sim 10$ to 35/hour (Zenklusen et al. 2008; Sun et al. 2012; Miguel et al. 2013). For mammalian cells grown in the lab, transcription rates across the genome have a roughly log-normal distribution, with a median of 2 to 3 mRNAs/hour and an approximate range of 0.1 to 30/hour (Darzacq et al. 2007; Schwanhäusser et al. 2011; Danko et al. 2013). Vertebrate genes typically contain multiple large introns, which are transcribed prior to removal, and this must contribute substantially to these reduced rates. However, as the latter rates do not account for the time genes spend in the off state, and a substantial fraction of transcription events abort prior to complete elongation ($> 90\%$ in mammals; Darzacq et al. 2007), they must underestimate $k_{\rm m}$.

Transcript degradation rates are often obtained by inhibiting transcription and following the subsequent decline in mRNA numbers. The half life of a molecule, $T_{0.5}$, denotes the time required for an initial concentration to decline by 50% and is related to the degradation rate $\gamma$ by

$$0.5 = e^{\gamma T_{0.5}}. \tag{2}$$

In *E. coli,* $\sim 80\%$ of mRNAs have half-lives between 3 and 8 mins, with a range of 1 to 15 mins and median $\sim 5$ mins (Bernstein et al. 2002; Taniguchi et al. 2010). Estimates of median half lives of mRNAs in *S. cerevisiae,* 22 mins (Wang et al. 2002), and mouse fibroblast cells, 9 hours (Schwanhäusser et al. 2011), are longer. Using the above expression, degradation rates of 0.14, 0.034, 0.012, and 0.0013 per min are implied for half lives of 5 mins, 20 mins, 60 mins, and 9 hours.

Knowledge of the rate at which genes turn on and off transcriptionally provides insight into the dynamics of gene activation/inactivation. For example, the average time between bursts of transcription, which is equivalent to the mean time that a gene remains off, is equal to $1/k_{\rm on}$. Once turned on, a gene remains transciptionally active for an average interval of $1/k_{\rm off}$, so the average number of transcripts produced during a bout of activity is $k_{\rm m}/k_{\rm off}$. Unfortunately, little is known about these rates, although So et al. (2011) estimate $k_{\rm on}$ to average about 0.003/sec in *E. coli*, with $k_{\rm off}$ often being about two to ten-fold lower. Rates of a similar order of magnitude have been observed in mammalian cells (Darzacq et al. 2007). Given that mRNA burst sizes are generally in the range of 1 to 20 (Sanchez and Golding 2013), it follows that $k_{\rm m}$ must typically be on the order of 1 to 20 times larger than $k_{\rm off}$.

The preceding survey implies that the numbers of mRNA molecules associated with individual genes are likely to be quite small. Supposing, for example, that the rate of full mRNA production by an *E.coli* cell in the on state is half the chain elongation rate, i.e., on the order of $k_{\rm m} = 50$ mRNAs per active gene per hour. With a median degradation rate on the order of $\gamma_{\rm m} = 8$/hour, and the gene turned on only a fraction of the time, because the average number of mRNAs per cell is the ratio of the production and elimination rates, $P_{\rm on}k_{\rm m}/\gamma_{\rm m}$ (Details 1), it is clear that genes in this species will commonly be represented by fewer than 10 mRNAs per cell. This rough qualitative prediction is consistent with the observed average number of just 5.0 mRNAs/gene/cell (and a range of 0 to 100) in *E. coli* (Lu et al. 2007; Li and Xie 2011). Even in much larger eukaryotic cells, the numbers of mRNAs per cell can be quite small, with a mean of just 10/gene in *S. cerevisiae* (Lu et al. 2007; Zenklusen et al. 2008), and medians in the vicinity of 20 in mammalian cells (Schwanhäusser

et al. 2011; Marinov et al. 2014). In all cases, there is a broad distribution around the mean, with the variance typically exceeding the mean by several fold (Golding et al. 2005; Raj et al. 2006; Taniguchi et al. 2010), as expected for genes that are not constitutively active (Details 1). Recent work suggests that some modifications of the theory may be necessary for Eukaryotes, where there can be a negative feedback between mRNA synthesis and degradation (Sun et al. 2012; Haimovich et al. 2013).

We now turn to the matter of protein numbers per cell. Because protein production ultimately depends on the presence of mRNAs, the kinds of transcriptional noise noted above (which occur within the life span of a cell) will naturally be transmitted to level of translation. The level of noise propagation would be reduced if the life span of a protein is typically greater than that of its associated mRNA, and that appears to be the case. In *S. cerevisiae,* for example, the majority of proteins outlive their maternal mRNAs, with the ratio of half lives being $\sim 3.0$ (Shahrezaei and Swain 2008). Likewise, in mouse fibroblast cells, the median half life of a protein, approximately two days (with a range of 3 to 500 hours), is $\sim 5\times$ greater than that of mRNAs (Schwanhäusser et al. 2011). The latter study also shows that translation rates per mRNA are roughly $100\times$ greater than transcription rates, with a mode of $\sim 200$ and a range of 1 to $10^4$ proteins/mRNA/hour. Thus, we expect the temporal variation in protein numbers per cell to be dampened in comparison to that for mRNAs (Figure 3).

As a consequence of their greater half lives and higher rates of production, proteins tend to be much more abundant in cells than their cognate mRNAs, with for example, an average ratio of 450 in *E. coli,* 5100 in *S. cerevisiae,* and 2800 in mammalian fibroblasts (Ghaemmaghami et al. 2003; Lu et al. 2007; Schwanhäusser et al. 2011). Bacterial cells have protein copy numbers typically ranging from 10 to 20,000, depending on the species (Ishihama et al. 2008; Malmström et al. 2008; Taniguchi et al. 2010), whereas yeast proteins fall in the range of 100 to $10^6$ copies per cell with a median of $\sim 4000$ (Ghaemmaghami et al. 2003; Newman et al. 2006; Lu et al. 2007), and mammalian proteins range from 10 to $10^7$ copies per cell with a median of 50,000 (Schwanhäusser et al. 2011). Notably, transcription factors tend to be the rarest proteins within cells (Ghaemmaghami et al. 2003; Marinov et al. 2014).

Two key parameters with respect to protein production are the number of mRNAs produced by active genes over the typical life span of a protein,

$$a = k_m/\gamma_p, \tag{3a}$$

and the average number of proteins translated per life span of an mRNA

$$b = k_p/\gamma_m. \tag{3b}$$

Together with the activation and deactivation rates, $k_{\rm on}$ and $k_{\rm off}$, these two parameters define the distribution of protein numbers among cells, as fully developed in Shahrezaei and Swain (2008). The mean number of proteins per cell is a simple extension of the expected value for mRNAs ($\mu_{\rm m}$),

$$\mu_{\rm p} = \frac{P_{\rm on} k_{\rm m} k_{\rm p}}{\gamma_{\rm m}\gamma_{\rm p}} = \frac{\mu_{\rm m} k_{\rm p}}{\gamma_{\rm p}}. \tag{4a}$$

The variance in number of protein molecules is described by

$$\sigma_{\mathrm{p}}^2 = \mu(n_{\mathrm{p}}) \left( 1 + b + \frac{(1 - P_{\mathrm{on}})ab\gamma_{\mathrm{p}}}{\gamma_{\mathrm{p}} + k_{\mathrm{on}} + k_{\mathrm{off}}} \right), \tag{4b}$$

which reduces to $\sigma_{\mathrm{p}}^2 = \mu(n_{\mathrm{p}})(1 + b)$ for a constitutively expressed gene (Thattai and van Oudenaarden 2001). Thus, the dispersion of protein numbers among cells is broader than that expected under a Poisson distribution.

These expressions indicate that there are a wide variety of ways in which cells can control the average numbers of proteins carried by cells, and several observations are suggestive as to how such alterations in protein expression are actually brought about. For example, Schwanhäusser et al. (2011) note a strong correlation between the number of protein molecules per cell and the translation rate, and Wang et al. (2002) find that the decay rates of mRNAs in yeast are coordinated among protein-coding loci whose products interact stoichiometrically.

---

**Details 1. Number of transcripts per cell.** The ultimate manifestation of gene expression, typically some level of protein production, depends on the number of mRNA molecules per cell, which in turn is a function of the rate of production of new transcripts and their subsequent loss by degradative processes. We first consider the situation for a consitutively expressed gene, such that the rate of production of new transcripts is a constant $k_{\mathrm{m}}$, with the rate of decay per transcript being $\gamma_{\mathrm{m}}$ (Figure 4). With constant rates, regardless of the starting conditions, the stochastic distribution of the number of mRNA molecules per cell ($p(n)$) will eventually reach an equilibrium. At this point, the flux rate between the $n = 0$ and $n = 1$ states must be equal in both directions, so that

$$k_{\mathrm{m}} p(0) = \gamma_{\mathrm{m}} p(1), \tag{1.1a}$$

so that

$$p(1) = p(0)(k_{\mathrm{m}}/\gamma_{\mathrm{m}}). \tag{1.1b}$$

Similarly, the flux rates in and out of class $n = 1$ must be equal, so that

$$(k_{\mathrm{m}} + \gamma_{\mathrm{m}})p(1) = k_{\mathrm{m}} p(0) + 2\gamma_{\mathrm{m}} p(2), \tag{1.2a}$$

which after substracting Equation (1.1a) and rearranging implies that

$$p(2) = p(0)(k_{\mathrm{m}}/\gamma_{\mathrm{m}})^2/2. \tag{1.2b}$$

This approach generalizes to

$$p(n) = p(0)(k_{\mathrm{m}}/\gamma_{\mathrm{m}})^n/n!. \tag{1.3}$$

To complete the solution, we require an expression for $p(0)$. Because the sum of all $p(n_{\mathrm{m}})$ is constrained to equal 1.0, it follows that $p(0)$ must be equal to a constant that ensures such equality. By noting that an exponential function can be written as the series expansion,

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \tag{1.4}$$

it follows that

$$p(0) = e^{-k_{\mathrm{m}}/\gamma_{\mathrm{m}}}, \tag{1.5a}$$

and more generally,

$$p(n_{\mathrm{m}}) = (k_{\mathrm{m}}/\gamma_{\mathrm{m}})^n e^{-k_{\mathrm{m}}/\gamma_{\mathrm{m}}}/n!. \tag{1.5b}$$

This is the well-known **Poisson distribution**, which is a function of a single parameter (the mean) and has the unique property of the mean being equal to the variance. In this case, the mean number of transcripts expected per cell is simply the ratio of the rates of production and elimination, $k_{\mathrm{m}}/\gamma_{\mathrm{m}}$.

Under more complex scenarios of gene regulation, the number of transcripts per cell deviates from the Poisson distribution, and the distributions of $n$ need to be evaluated by more complex methods (Thattai and van Oudenaarden 2001; Phillips et al. 2013). The case of a two-state model in which the gene is turned on with some probability $P_{\mathrm{on}}$ was derived by Peccoud and Ycart (1995) and has the respective mean and variance

$$\mu_{\mathrm{m}} = \frac{P_{\mathrm{on}} k_{\mathrm{m}}}{\gamma_{\mathrm{m}}} \tag{1.6a}$$

$$\sigma^2_{\mathrm{m}} = \mu(n_{\mathrm{m}}) \left(1 + \frac{(1 - P_{\mathrm{on}})k_{\mathrm{m}}}{k_{\mathrm{on}} + k_{\mathrm{off}} + \gamma_{\mathrm{m}}}\right), \tag{1.6b}$$

where $k_{\mathrm{on}}$ and $k_{\mathrm{off}}$ are, respectively, the rates of transition of cells from the off to on states, and vice versa. The complete distribution, worked out by Raj et al. (2006) and Shahrezaei and Swain (2008), is given by

$$p(n_{\mathrm{m}}) = p^*(n_{\mathrm{m}}) = \frac{\Gamma(k'_{\mathrm{on}} + n_{\mathrm{m}})\Gamma(k'_{\mathrm{on}} + k'_{\mathrm{off}})}{\Gamma(k'_{\mathrm{on}} + k'_{\mathrm{off}} + n_{\mathrm{m}})\Gamma(k'_{\mathrm{on}})} {}_1F_1[k'_{\mathrm{off}}, (k'_{\mathrm{on}} + k'_{\mathrm{off}} + n_{\mathrm{m}}); k_{\mathrm{m}}/\gamma_{\mathrm{m}}],$$

where $p^*(n_{\mathrm{m}})$ is the Poisson distribution defined in Equation (1.5b), $k'_{\mathrm{on}} = k_{\mathrm{on}}/\gamma_{\mathrm{m}}$, and $k'_{\mathrm{off}} = k_{\mathrm{off}}/\gamma_{\mathrm{m}}$. $\Gamma(\cdots)$ is the gamma function, and ${}_1F_1[\cdots]$ is the confluent hypergeometric function of the first kind, both of which can be approximated using expressions in Abramowitz and Stegun (1972).

This model is agnostic with respect to the mechanisms turning a gene on and off, although it does assume that the switching events are completely random (i.e., have probabilities that do not depend on the length of stay in a previous state). Under this assumption,

$$P_{\mathrm{on}} = \frac{k_{\mathrm{on}}}{k_{\mathrm{on}} + k_{\mathrm{off}}}. \tag{1.7}$$

This expression does rely on the on/off rates being independent of the state of the DNA, and uninfluenced by the presence of cooperative factors, chromatin remodeling, etc. Alternative models that allow for these higher-order complexities can be found in Phillips et al. (2013) and Hammar et al. (2014).

Below, a more mechanistic description of the determinants of $P_{\mathrm{on}}$, in terms of transcription-factor binding, is provided. The central point to note here is that, owing to the population of cells being heterogeneous with respect to the on and off states, there is a greater dispersion in mRNA number per cell relative to the case of constitutive gene expression (as can be seen from the degree to which the variance of $n_{\mathrm{m}}$ exceeds the mean).

---

## The Basic Biology of Transcription

Gene transcription is carried out by a multi-subunit DNA-dependent RNA polymerase, which we will simply call RNA polymerase. However, RNA polymerase is generally nonautonomous in the sense that one to several accessory TFs must be present simultaneously to activate RNA polymerase. It is common for individual TFs to service multiple genes, which facilitates coregulation of gene expression, but specialized one-to-one relationships between TFs and their client genes are not uncommon. For example, in *E. coli,* seven TFs control the expression of $\sim 50\%$ of regulated genes, whereas $\sim 60$ TFs (about one fifth of the TFs in this species) service single genes (Martinez-Antonio and Collado-Vides 2003). TFBSs usually reside in a window of several hundred base pairs upstream of a gene's translation start site, although many exceptions are known.

Liaisons between TFs and their cognate TFBSs on the DNA are usually governed by hydrogen bonds and van der Waals attractions between the two molecules. However, all TFs also inevitably engage in promiscuous interactions as a consequence of the negatively charged phosphate backbones of the DNA and positively charged residues on the protein. These unavoidable nonspecific interactions impose a substantial challenge for any TF, which must have a high enough affinity to its own specific binding sites to avoid too great of a burden of sequestration in nonfunctional locations.

Eukaryotic transcription raises additional issues in that the chromosomes are heavily populated by structural proteins involved in the packaging of the genome into manageable spatial units (e.g., nucleosomes formed from histones, which are usually further packed into higher-order structures). On the one hand, such structures can reduce the accessibility of a TF to a hidden TFBS, but the occlusion of TFs from nonregulatory DNA can also reduce the time spent on nonproductive searching (Charoensawan et al. 2012; Thurman et al. 2012). In addition, some proteins such as cohesins, which encircle sister chromosomes during cell division, help recruit TFs to localized regions (Yan et al. 2013).

Many dozens of families of TFs are known to exist across the Tree of Life, each structurally reliant on different DNA-binding domains. However, although each TF has maximum affinity for a specific DNA motif, there appears to be no general regulatory code in TFs, i.e., no specific language involving one-to-one recognition matching between the amino-acid sequence of a TF and the nucleotide sequence of its binding site. Typically, 10 to 50 amino-acid residues are involved in contacts with the DNA, whereas TFBS motifs generally consist of 6 to 30 nucleotides (Luscombe and Thornton 2002).

**A physical model for transcription-factor binding.** The universality of the basic mode of transcription – the interaction of a specific protein (the TF) with a specific DNA binding site (the TFBS) – provides a compelling platform for an evolutionary theory of gene expression based on the fundamental biophysics of intermolecular associations (Bintu et al. 2005). However, because many proteins are commonly present in less than a few dozen copies per cell, an understanding of transcription requires insight into the consequences of stochastic aspects of single-cell biology as opposed to measuring just the average features of entire populations. To move forward, we require a probabilistic framework for understanding the likelihood of TFs being bound to their specific TFBS targets in individual cells.

Consider a TF that recognizes an optimal binding motif containing $\ell$ key nucleotide sites. Assuming that each nucleotide within a potential binding site contributes to the total binding energy in an additive fashion, then for a particular sequence denoted by the vector $\mathbf{a}$, the total binding energy can be described as

$$E(\mathbf{a}) = \sum_{i=1}^{\ell} \epsilon_i(a_i), \tag{5}$$

where $\epsilon_i(a_i)$ is the binding energy associated with nucleotide site $i$ if occupied with nucleotide $a_i$, $a_i$ is A, C, G, or T. Letting $a_i^*$ denote the optimal nucleotide for binding at site $i$, $\mathbf{a}^*$ is the sequence yielding maximum binding strength. Empirical data from a variety of sources suggest that the average energetic cost of a mismatch ($a_i \neq a_i^*$) is typically $\simeq 2.0$ in Boltzmann units of $K_BT$ (which is $\simeq 0.6$ kcal/mol at most biological temperatures) (Table 1). Thus, under the assumption that binding strength scales linearly with the degree of correspondence between a TFBS and the optimal binding motif of its TF, the relevant phenotype from the perspective of binding efficiency of a target site can be viewed as approximately $2m$, where $m$ is the number of matches with the optimal recognition sequence. Numerous empirical studies support the additivity assumption as a first-order approximation (von Hippel and Berg 1986; Sarai and Takeda 1989; Takeda et al. 1989; Fields et al. 1997), although higher-order effects involving the shape of TFBSs can also make contributions to the overall binding energy (Yang et al. 2013).

**Table 1.** Features of the motifs of well-studied transcription factors (TFs). Motif size is based on consensus sequences. The estimated costs of mismatches are obtained from binding-strength experiments in which single-base changes were made in motifs. Costs of single-base mismatches are in units of kcal/mol; these average to 1.4 across the full set of studies, or in terms of Boltzmann units ($K_BT \simeq 0.6$ kcal/mol) to 2.3.

| TF | Species | Motif (bp) | Cost of Mismatch Mean (Range) | References |
|---|---|---|---|---|
| CI | Lambda phage | 17 | 1.4 (0.5 – 3.5) | Sarai and Takeda (1989) |
| Cro | Lambda phage | 9 | 1.4 (0.5 – 2.5) | Takeda et al. 1989 |
| Mnt | *Salmonella* phage P22 | 21 | 1.0 (0.3 – 1.6) | Fields et al. (1997); Berggrun and Sauer (2001) |
| CRP | *Escherichia coli* | 22 | 1.7 (0.9 – 2.5) | Gunasekera et al. (1992); Kinney et al. (2010) |
| CRP | *Synechocystis* sp. | 22 | 1.8 (0.7 – 3.0) | Omagari et al. (2004) |
| ArcA | *Shewanella oneidensis* | 15 | 1.3 (0.1 – 3.4) | Schildbach et al. (1999); Wang et al. (2008) |
| Gcn4 | *Saccharomyces cerevisiae* | 11 | 1.0 (0.5 – 1.7) | Nutiu et al. (2011) |
| c-Myb | *Homo sapiens* | 6 | 1.6 (0.6 – 2.8) | Oda et al. (1998) |

Given a potential TFBS sequence with a particular binding energy, we wish to know the probability of occupancy by a cognate TF, $P_{\rm on}$, as this is a minimal requirement for expression of the associated gene. This probability is a function of both the binding site itself and of the features of the intracellular environment that restrict the accessibility of the site to cognate TFs. Clearly, $P_{\rm on}$ will increase with the number of TF molecules accessible to the DNA ($N_{\rm tf}$), but equally important is

the number of ways in which individual TF molecules can become side-tracked by binding to alternative genomic sites. Other genes serviced by the TF (numbering $N_{\rm ot}$, where ot denotes off-target) will compete for the pool of TFs, but nonspecific binding of TFs across the genome can be numerically more important. Letting $G$ denote the total genome size (in bp), because $\ell \ll G$, there are essentially $G - \ell \simeq G$ such nonspecific sites in a haploid cell. From Details 2, letting the excess scaled binding energy of a target TFBS be $2m$, the probability that a specific target TFBS is occupied by its TF is

$$P_{\rm on} \simeq \frac{1}{1 + Be^{-2m}}, \tag{6}$$

where $B = G/(N_{\rm tf} - N_{\rm ot})$ is a measure of the concentration of background (nonspecific) binding sites relative to the number of TF molecules available for the specific target site. More broadly, the composite parameter $B$ can be viewed as a summarization of the totality of cellular issues working against the binding of a TF to a specific cognate TFBS.

A rough idea of the magnitude of $B$ can be inferred by noting that $G$ is generally in the range of $10^6$ to $10^{10}$ bp, with prokaryotes falling at the lower end and multicellular eukaryotes at the higher end of the range (Lynch 2007). For the model bacterium *Escherichia coli,* the numbers of molecules per cell for particular TFs, $N_{\rm tf}$, are often in the range of 100 to 1000, with just a few cases ranging as high as 50,000 (Robison et al. 1998). Somewhat lower numbers are estimated for another bacterium *Leptospira interrogans* (Malmström et al. 2009). In such species, it is unusual for the number of genes serviced by a particular TF to exceed 100, i.e., $N_{\rm ot} < 100$. Taken together, these observations for prokaryotes suggest a range for $B$ on the order of $10^3$ to $10^6$.

For the yeast *Saccharomyces cerevisiae,* proteomic data suggest that the average number of molecules for individual TFs is on the order of 8000 per cell (Ghaemmaghami et al. 2003), so with a genome size of 12 Mb, $B$ should be in the vicinity of $10^3$ to $10^4$. Proteins within mammalian cells appear to be about $10\times$ as numerous as those in yeast (Schwanhäusser et al. 2011), but with a genome size of $\sim 3000$ Mb, $B$ can be expected to be $\gg 10^3$. Note that these estimates of background interference assume that the primary mechanism reducing TF accessibility is nonspecific binding on DNA. If other sources of interference exist (such as promiscuous binding to other proteins), $B$ would be accordingly higher. On the other hand, DNA binding proteins, such as histones in eukaryotes, could reduce $B$ by restricting access of a TF to only a fraction of the genome.

Examination of Equation (6) provides insight into the conditions necessary for a high probability of binding. For example, $m = 0.5\ln(B)$ represents a key pivot point below which background interference results in $P_{\rm on} < 0.5$. If the binding probability is to exceed 0.99, the number of matches must exceed 6 for $B = 10^3$ and 11 for $B = 10^7$ (Figure 4). Thus, unless the level of background interference greatly exceeds $B = 10^7$, there is little to be gained in terms of binding probability for a motif in excess of a dozen bases. Thus, a considerable amount of mismatching can be tolerated for a TFBS motif more than a dozen or two nucleotides in length.

The results in Figure 4 also highlight two physical constraints on the basic process of transcription regulation by the binding of TFs to DNA. First, because mismatches come in discrete packets (with relative binding energy $\sim 2.0$/site), the

opportunities for fine-tuning gene expression by altering the numbers of mismatches in a TFBS may be limited, although variation around this expectation (from, for example, not all mismatches having exactly the same consequences) will provide more flexibility. Thus, if fine-scale modulation of gene expression is to be accomplished, it must largely rely on differences in the numbers of TFs residing inside cells (through the influence on $B$). However, a side effect of altering the concentration of a TF is that different client genes will also be affected.

Second, life's transcription mechanism comes with a significant energetic price, in that to ensure that a particular gene is turned on, a substantial excess number of TF molecules must be produced to affect the unproductive engagements occurring at nonspecific sites. For example, rearrangement of Equation (6) shows that for a motif with $m = 8$ matching bases to achieve a 0.9 probability of being bound of a target site, 10 and 1000 TF molecules are required in cells with genome sizes of $10^7$ and $10^9$ bp, respectively, and $P_{\rm on} = 0.99$ elevates these numbers to $\sim 110$ and $11,000$.

---

**Details 2. Occupancy probability for a transcription-factor binding site.** Because gene activation requires that the relevant TFBSs be occupied by their cognate TFs, an understanding of the mechanics of gene expression requires some basic theory for the probability that a particular TFBS is appropriately bound. This, in turn, requires information on the degree to which individual TF molecules are transiently tied to alternative substrates within the cell. Here we consider one particular target TFBS within a genome containing $N_{\rm ot}$ additional off-target but legitimate binding sites for the TF of interest. In addition, we must account for the possibility of binding to random illegitimate sites in the genome. Although such nonspecific binding is expected to be weak on a per-site basis, the number of such sites is enormous, being close to the total number of bases in the genome ($G$) because each nucleotide can serve as an initiation site for binding.

Letting $N_{\rm tf}$ be the number of TF molecules of the type under consideration in the cell, we assume that $N_{\rm ot} \ll N_{\rm tf} \ll G$. The first inequality follows from the fact that a full repertoire of gene expression is extremely unlikely unless the number of TF molecules substantially exceeds the number of genes requiring their services. The second inequality follows from the sheer magnitude of genome sizes (generally, $10^6$ to $10^{10}$ bp).

To compute the probability that a particular TFBS is bound by a cognate TF, we utilize a popular approach from statistical mechanics, evaluating the relative likelihoods of all possible ways in which the $N_{\rm tf}$ TF molecules can be distributed within a cell. Here we assume that all such molecules are situated along the chromosome, either specifically bound to true cognate sites or nonspecifically bound to random genomic regions, although this assumption need not literally be true so long as all off-site sequestration is appropriately accounted for. Ultimately, we require a measure scaling with the total probability that a TF is bound to the site of interest, $Z_{\rm on}$, and another measure scaling with the probability that all $N_{\rm tf}$ TF molecules are engaged elsewhere on the genome, $Z_{\rm off}$. The sum, $(Z_{\rm on} + Z_{\rm off})$, is known as the **partition function**, and the probability that a particular TFBS is occupied is simply

$$P_{\rm on} = \frac{Z_{\rm on}}{Z_{\rm on} + Z_{\rm off}} = \frac{1}{1 + (Z_{\rm off}/Z_{\rm on})}. \tag{1.1}$$

The first step to evaluating the two components of the partition function is to enumerate the full set of relevant configurations of the $N_{\rm tf}$ molecules within the cell,

and weighting each set of states by its multiplicity, i.e., the number of ways in which a particular type of configuration can be distributed over the genome. Consider, for example the situation in which the target TFBS is unoccupied. In this case, all $N_{\rm tf}$ TF molecules might be nonspecifically bound, with none on off-target sites; in this case, there are $G!/(G-N_{\rm tf})!N_{\rm tf}!$ distinct ways in which the TFs can be distributed over the $G$ nonspecific sites (where $x! = x(x-1)(x-2)\cdots 1$ is the factorial product). Alternatively, $N_{\rm tf}-1$ TF molecules might be nonspecifically bound, with one on an off-target site; there would then be $G!/(G-N_{\rm tf}-1)!(N_{\rm tf}-1)!$ distinct ways in which the TFs can be distributed over nonspecific sites, and $N_{\rm ot}$ possible locations for the one off-target TFBS, yielding a total multiplicity of $N_{\rm ot}G!/(G-N_{\rm tf}-1)!(N_{\rm tf}-1)!$. This general enumeration strategy must be extended to the opposite extreme in which all off-target sites are occupied, in each case following the general procedure for determining the distinct number of ways in which $x$ TFs can be distributed over $y$ sites. The same strategy for quantifying multiplicity of configurations applies to the situation in which the target TFBS is occupied, except in this case only $(N_{\rm tf}-1)$ TF molecules are distributed elsewhere.

After enumeration, all of these alternative states must be further weighted by their physical likelihoods dictated by the overall binding energy of each configuration. Here we denote the binding energies of the TF to the target, off-target, and nonspecific sites as $E_{\rm t}$, $E_{\rm ot}$, and $E_{\rm ns}$, respectively. For example, for each configuration in which all TFs reside on nonspecific binding sites, the total weight is $e^{-N_{\rm tf}E_{\rm ns}/(K_{\rm B}T)}$. If one off-target site is occupied along with $(N_{\rm tf}-1)$ nonspecific sites, the weight becomes $e^{-[E_{\rm ot}+(N_{\rm tf}-1)E_{\rm ns})]/(K_{\rm B}T)} = e^{-[(E_{\rm ot}-E_{\rm ns})+N_{\rm tf}E_{\rm ns}]/(K_{\rm B}T)}$. If the target site is occupied, along with one off-target site and $(N_{\rm tf}-2)$ nonspecific sites, the weight becomes $e^{-[E_{\rm t}+E_{\rm ot}+(N_{\rm tf}-2)E_{\rm ns}]/(K_{\rm B}T)} = e^{-[(E_{\rm t}-E_{\rm ns})+E_{\rm ot}-E_{\rm ns})+N_{\rm tf}E_{\rm ns}]/(K_{\rm B}T)}$, etc. In these expressions, $K_{\rm B}T$ is the Boltzmann constant times the temperature (in degrees Kelvin), the standard measure of background thermal energy. With both $K_{\rm B}T$ and the binding energies measured in the same units (usually kcal/mol), the weights are dimensionless. Because the binding energies are negative, with stronger binding being denoted by more negative $E$, the weights increase with the magnitude of binding strength to cognate sites relative to background expectations.

We are now in a position to write down full expressions for each of the two components of the partition function. In each case, this is done by summing over all possible configurations the products of the multiplicity and the energetic weight of each configuration. In the following, we use the abbreviation $\beta = 1/(K_{\rm B}T)$, and let $\Delta E_{\rm t} = E_{\rm t} - E_{\rm ns}$ and $\Delta E_{\rm ot} = E_{\rm ot} - E_{\rm ns}$ denote the differences in binding energies of target and off-target sites relative to background levels. Summing up, some rather complex looking expressions arise,

$$Z_{\rm off} \simeq \frac{G!N_{\rm ot}!e^{-\beta N_{\rm tf}E_{\rm ns}}}{(G-N_{\rm tf})!(N_{\rm tf}-N_{\rm ot})!N_{\rm tf}^{N_{\rm ot}}} \sum_{i=0}^{N_{\rm ot}} \frac{e^{-i\beta N_{\rm tf}\Delta E_{\rm ot}}}{(N_{\rm ot}-i)!i!(G/N_{\rm tf})^i} \tag{1.2a}$$

$$Z_{\rm on} \simeq \frac{G!N_{\rm ot}!e^{-\beta N_{\rm tf}E_{\rm ns}}e^{-\beta\Delta E_{\rm t}}}{(G-N_{\rm tf}+1)!(N_{\rm tf}-N_{\rm ot}-1)!N_{\rm tf}^{N_{\rm ot}}} \sum_{i=0}^{N_{\rm ot}} \frac{e^{-i\beta N_{\rm tf}\Delta E_{\rm ot}}}{(N_{\rm ot}-i)!i!(G/N_{\rm tf})^i} \tag{1.2b}$$

Noting, however, that the summations to the right of Equations (1.2a,b) are identical, and that several of the components on the left are identical or very similar as well, substitution into Equation (1.1) leads to great simplification,

$$P_{\rm on} = \frac{1}{1+[G/(N_{\rm tf}-N_{\rm ot})]e^{\beta\Delta E_{\rm t}}} \tag{1.3}$$

In a succinct fashion, this expression reveals how the magnitude of gene expression is dictated by basic cellular features. First, the probability that a TFBS is occupied depends on the absolute difference in binding strengths between the target and

nonspecific sites; as $E_t$ becomes more negative (implying stronger binding), $P_{on} \rightarrow 1$. Second, the probability of binding at the site declines with increasing concentration of nonspecific sites ($G$) relative to the effective number of transcription factor molecules in excess of the number of cognate binding sites, $(N_{tf} - N_{ot})$. The first effect is a function of the degree of match between the binding motif of the site of interest and the optimal sequence of its cognate TF, whereas the second effect is determined by the size of the genome ($G$), the degree of expression of the TF (the number of molecules in the cell, $N_{tf}$), and the number of additional legitimate sites serviced by the TF ($N_{ot}$).

---

**Encounter rates between TFs and their binding sites.** The preceding analyses implicitly assume that the distribution of TFs within a cell is typically in dynamic equilibrium, which in turn assumes that newly arisen TF molecules rapidly find a location on the genome on the time scale of the life of an individual cell. At first glance, the chances of a TF locating a specific cognate TFBS in a reasonable amount of time would seem to be daunting, but as outlined in Details 2, the biophysical properties of cells are such that localization can typically be achieved in a matter of a few seconds or less.

Despite their passive transport, TFs appear to locate their target sites at rates exceeding the three-dimensional diffusion limit (Riggs et al. 1970), an observation that motivates the **facilitated-diffusion** model (von Hippel and Berg 1989). Given the minute sizes of individual TFBSs, after production a TF molecule will essentially always first encounter a nonspecific site on a chromosome before locating a proper, more energetically favorable target. Although a number of details remain unresolved, the search process involves repeated association-dissociation events involving one-dimensional sliding along DNA molecules interspersed with three-dimensional jumping to new locations. During such episodes of intersegmental transfer, TF molecules are kept in the vicinity of the DNA, thereby avoiding the much larger and unproductive search space of the entire cytoplasm/nucleoplasm. Such three-dimensional wandering also minimizes the redundant interrogation of localized chromosomal space that would occur with a nondirected one-dimensional diffusion process. Reviews covering many of the technical issues can be found in Gowers et al. (2005), Halford and Marko (2004), Halford (2009), Kolomeisky (2011), and Zhou (2011).

The extent to which various species alter the spatial configurations of their chromosomal DNA to assist in this process remains unclear. However, the spatial issues with respect to the large cells of eukaryotes are of interest. Notably, the volumes of the nuclei of eukaryotic cells are typically larger than entire cells of prokaryotes, which results in mean search times of individual TF molecules for a target TFBS on the order of 1 to 200 minutes within the nuclear environment alone (Details 2). Although the overall search process can be sped up by producing more TF molecules, there is the additional issue of cell volume, which is commonly 10 to 100× that of the nucleus. All other things being equal, this would result in an increase in the search time by the same factor were the genome not concentrated within a nuclear envelope. Thus, as will be described in a later chapter, although a number of hypotheses have been proposed for the evolution of the nuclear envelope and its relevance to the expanded sizes of eukaryotic cells, the challenges of gene

expression should be included in this list. Without specific mechanisms of delivery of TFs to their final destinations, the rate of gene expression in large cells would be extremely compromised if the genome were not confined to the restricted space of the nucleus.

---

**Details 3. The biophysics of TFBS localization.** Gene regulation requires that TFs navigate from their point of production through a large cellular volume to their cognate TFBSs. So far as we know, such encounters are established through semi-random diffusive molecular motions, i.e., without the involvement of any directed guidance from specific transport mechanisms. Here we consider the approximate time scale on which encounters are likely to occur, primarily to show that the rapid equilibration assumed in the previous section is indeed likely. We start with a focus on prokaryotic cells, which offer the relative simplicity of a fairly homogeneous cytoplasm. The biophysical basis for the formulae to be used will be described more fully in a later chapter.

Transcription factors have an inherent tendency to bind nonspecifically to DNA. Because the translation of prokaryotic mRNAs is performed in the close vicinity of the chromosome, often co-transcriptionally, it is reasonable to assume that a newly arisen TF is almost immediately bound weakly to a nonspecific site on the DNA. This raises the possibility that a TF could then simply engage in a one-dimensional diffusion process over the chromosome until randomly encountering its cognate TF. The time required for such an encounter can be roughly estimated by noting that the average distance of a particle from its starting point in a one-dimensional diffusion process (and ignoring any boundary conditions) is

$$\overline{d} = \sqrt{2D_1 t}, \tag{3.1}$$

after $t$ time units, with $D_1$ being the one-dimensional diffusion coefficient (with units equal the sqared distance per time). This measure is generally referred to as the **root mean square distance** because in the absence of any directional bias to movement, the particle will end up to the left and right of the starting point with equal probabilities.

A central problem with linear diffusion is its redundancy – with random movement to the right and left, any diffusive event has a 50% probability of returning the molecule to its location two steps earlier. Indeed, for this reason, under one-dimensional diffusion, the *average* location of a molecule always remains at its initial location, with the probability distribution simply broadening, equally to the left and right with time.

Assuming that a TF initially resides at a random location on the genome with respect to its target TFBS, how long would it take to locate a specific target site by one-dimensional diffusion? Because a TFBS can be initiated at any genomic position on either strand of the DNA molecule, with the initial TF position being being half a genome away from the site (with $G$ being the genome size in bp), the TF will have to interrogate an average total of $\sim G$ potential sites to find a specific target. Thus, we require the time solution to Equation (3.1) that yields $\overline{d} = G$. Several studies yield an average estimate of $D_1 \simeq 0.5 \times 10^6$ bp$^2$/sec for a protein moving along the DNA in an *E. coli* cell (Wang et al. 2006; Elf et al. 2007; Marklund et al. 2013). Noting that the *E. coli* genome is $G \simeq 5 \times 10^6$ bp in length, substituting $D_1$ into Equation (2.1) and rearranging, we find that the average time for a single TF molecule to encounter a specific TFBS by one-dimensional diffusion is $\sim 2.5 \times 10^6$ sec (or $\sim 29$ days). With $N_{\rm tf}$

TF molecules searching simultaneously, the average search time would be $1/N_{\rm tf}$ times the single-molecule rate, but even with 1000 molecules per cell (higher than what is seen in this species), the average search time would be $\sim 0.7$ hours. As this is too long to account for the fact that *E. coli* cells are capable of dividing in $< 0.5$ hours, it is clear that linear scanning cannot account for known rates of transcription.

An alternative way in which the search process might be accomplished is three-dimensional diffusion. In this case, we make use of an expression for the encounter rate per unit concentration,

$$k_e = 4\pi(D_{3n} + D_{3p})(r_n + r_p), \tag{3.2}$$

where $D_{3n}$ and $D_{3p}$ are respectively the diffusion coefficients for the nucleic acid (TFBS) and the protein (TF), and $r_n$ and $r_p$ are their effective radii. This equation assumes that an effective encounter occurs when the centers of the TF and TFBS fall within total distance $r_n + r_p$ of each other. Because of its bulk, it is reasonable to assume that the DNA molecule is effectively immobile relative to the TF, so that $D_{3n} \simeq 0$. Experimental estimates for proteins in *E. coli* suggest that $D_{3p} \simeq 3.5\ \mu\text{m}^2/\text{sec}$ (Elowitz et al. 1999; Elf et al. 2007). Taking an average TFBS motif in this species to be 20 bp in length, and noting that the length of a nucleotide on a DNA molecule is $\simeq 0.34 \times 10^{-3}\,\mu$m, the effective radius of a potential binding site is approximately $r_n = 0.5 \times 20 \times 0.34 \times 10^{-3} = 0.0034\,\mu$m. The effective radii of proteins of the size of a TF are roughly in the range of $r_p = 0.002$ to $0.010\,\mu$m (Wasyl et al. 1971; Erickson et al. 2009), and we will use an average of $0.006\,\mu$m. Substitution of all of these estimates into Equation (3.2) yields an estimated encounter rate of $0.4\mu\text{m}^3/\text{sec}$ per unit concentration.

To obtain an estimate of the actual encounter rate, this specific rate must be multiplied by the products of the concentrations of the TFBS and TF within the cell. The volume of an *E. coli* cell is $\simeq 1\,\mu\text{m}^3$, and so with one TF molecule in search of $10^7$ nonspecific binding sites (summed over both sides of the genome), the rate of encounter with any site on the DNA is $4 \times 10^6$/sec, which implies an average time for a jump between chromosomal locations of just $2.5 \times 10^{-7}$ sec. Elf et al. (2007) estimate that once on the DNA a TF spends $\sim 0.0026$ sec diffusing over $\sim 100$ bp, so essentially all of the search time is spent directly interrogating the DNA. Thus, because approximately $10^5$ 100-bp scans are required to cover the entire genome, the estimated time to locate a site is 260 sec. With $N_{\rm tf}$ molecules in the cell, the search time would be reduced to $260/N_{\rm tf}$. A few prokaryotic species have cell volumes as small as $0.1\,\mu\text{m}^3$ (Shuter et al. 1983), which would reduce the search time further by a factor of ten, and few have volumes exceeding $100\,\mu\text{m}^3$, which would increase the time 100-fold.

How might these results translate to transcription in eukaryotes? First, because eukaryotic TFBSs are about half the length of those of prokaryotes, the encounter rate will be reduced by a factor of 0.5 on the basis of target size. Second, the average rate of diffusion in the nucleoplasm of mammalian cells is on the order of $D_{3p} \simeq 18\,\mu\text{m}^2/\text{sec}$ for proteins (Kühn et al. 2011), which will speed things up by a factor of $18/4 = 4.5$. Third, nuclear volumes in eukaryotic cells are typically larger than the volumes of entire prokaryotic cells, generally in the range of 100 to $10^4\,\mu\text{m}^3$ (Price et al. 1973; Cavalier-Smith 2002). However, the concentration of DNA within nuclei appears to be higher than that within prokaryotic cells – averaging $57 \times 10^6$ bp/$\mu\text{m}^3$ in root-tip cells of land plants (Fujimoto et al. 2005), and $189 \times 10^6$ bp/$\mu\text{m}^3$ in the blood cells of amphibians (Cavalier-Smith 1982), which is $\sim 25\times$ the concentration in an *E. coli* cell. Taken together, these results suggest that, within the nucleus, a TF will encounter DNA at a rate that is on the order of $0.5 \times 4.5 \times 25 = 56$ times faster than the rate calculated above for *E. coli,* which again implies that the time spans required for random jumps among DNA sites are of minor significance in the search process. Estimates for the one-dimensional diffusion parameters do not appear to be available

for eukaryotes. However, assuming they are roughly the same as in *E. coli,* because eukaryotic haploid genome sizes are generally in the range of 10 to 3000 million bp in length, we can anticipate search times on the order of 2 to 600× greater than that for *E. coli*. On the other hand, a substantial fraction of eukaryotic chromosomes are spooled around histones, which will serve to reduce the time needed to search for an exposed TFBS.

Although fairly crude, these estimates clearly indicate that given the architecture of cells, specific motor proteins are not required to guide TFs to their final destinations. All of the above calculations ignore the electrostatic interactions between proteins and nucleic acids, which by increasing the effective radii of interacting particles, would further speed up the localization process (Riggs et al. 1970; Halford 2009). Moreover, initial encounters are expected to be considerably sped up in prokaryotes where the TF is often encoded in a genomic location close to its target genes, and a newly translated TF is ensured a starting point close to its final destination (Kolesov et al. 2007). On the other hand, for eukaryotes, we have ignored the additional problem of a cytoplasmically translated TF finding its way to the nucleus. With eukaryotic cell sizes being on the order of 100 to $10^6$ $\mu$m$^3$ in volume (Price et al. 1973; Cavalier-Smith 1978, 2005; Shuter et al. 1983), this can substantially contribute to the search time.

---

## Evolutionary Considerations

Essentially every gene in every genome requires activation by at least one TF to be expressed, but because many TFs service multiple genes, a fairly small fraction of most genomes is allocated to TF production, typically 1 to 5% of the protein-coding genes within a genome. Among prokaryotic species, the number of TF genes ranges from $\sim 5$ to 500, scaling quadratically with the total number of protein-coding genes, whereas eukaryotic genomes generally encode for at least 100 TFs, with well over 1000 being harbored in multicellular species, and the scaling with total gene number being close to linear (van Nimwegen 2003; Aravind et al. 2005; Charoensawan et al. 2010). As a consequence, Eukaryotes generally invest proportionately less in their TF repertoires at the genomic level than do Prokaryotes. Across the Tree of Life, many dozens of TF families have been identified based on the unique physical structures of their DNA-binding domains. However, despite the fact that the TF mode of gene regulation must date to LUCA, and despite the substantial number of genes in Eubacteria and Archaea with obvious orthologs with Eukaryotes, only 2% of specific DNA-binding domain families are shared across Eubacteria, Archaea, and Eukaryotes, and essentially no obvious TF orthologs are known across the three superkingdoms (Charoensawan et al. 2010). This observation alone suggests a substantial turnover in the specific TFs used in various lineages, a pattern that repeats itself at lower levels of phylogenetic organization, as noted below.

There has been much speculation that eukaryotic morphological diversity has been driven by the exploitation of novel TF families and their recruitment to specific sets of genes. For example, a common refrain among those doing comparative developmental biology in animals is that the origin of novel phenotypes is largely a consequence of alterations in the *cis*-regulatory logic residing upstream of genes

rather than a result of change at the protein level, with some going so far as to claim that *cis*-regulatory modifications are the units of evolutionary change (Carroll et al. 2001; Davidson 2001; Wray 2007). The usual logic underlying this argument is that because individual TFs often service multiple genes, alterations of TF binding-sites specificities are likely to have large-scale, negative pleiotropic consequences for fitness. Under this extreme view, a structural gene can only acquire a change in expression pattern with minimal pleiotropic effects by recruiting or eliminating a pre-existing TFBS.

This idea that the optimal binding sites of TFs are frozen in evolutionary time is inconsistent with considerable evidence that functional changes in TFs often have minimal side consequences (Hsia and McGinnis 2003; Lynch and Wagner 2008; Wagner and Lynch 2008). What seems to have been ignored in previous arguments is the fact that mutations arising in a gene with pleiotropic effects need not themselves be pleiotropic. Further doubts as to whether the expansion of specific TF families played a central role in the evolution of multicellularity and subsequent morphological diversification in animals and land plants are motivated by the observation that many of the key TFs deployed in development are present in the unicellular relatives of both lineages (de Mendoza et al. 2013).

Although the advent of a series of novel methodologies for genome-wide identification of TFs and their corresponding TFBSs promises to substantially expand our understanding of how such systems diversify (e.g., Carey et al. 2012; Furey 2012; Smith et al. 2013), most current insight into the mechanisms of gene-regulatory evolution derives from observations from the usual key model systems – the bacterium *E. coli,* the yeast *S. cerevisiae,* the fly *D. melanogaster,* mouse, and human. Nevertheless, from this limited set of taxa, several clear generalizations have emerged. First, prokaryotes typically harbor substantially longer consensus TFBSs than do eukaryotes (Stewart et al. 2012). Moreover, unlike many eukaryotic TFBSs, prokaryotic binding sites are often palindromic in nature, with each half sequence being 7 to 11 bases in length, and the two halves being recognized by two members of a homodimeric TF.

Second, the molecular evolutionary features of TFs appear to depend on the number of host genes that they service. From comparisons of multiple $\gamma$-Proteobacteria, Rajewsky et al. (2002) found that TFs with larger numbers of target protein-coding genes are more evolutionarily conserved at the amino-acid sequence level, not just at the level of the recognition sequence, but across the entire TF. Sengupta et al. (2002) have observed a decline in binding-site specificity with increasing numbers of genes serviced by a TF in both *E. coli* and yeast, suggesting that this is an evolved mechanism to minimize the mutational burden on an organism. However, it remains unclear whether this genome-wide pattern is actually an outcome of selection, as an alternative explanation for the observed pattern is that TFs with low specificity are recruited more frequently into various regulatory pathways over evolutionary time.

Third, in Eukaryotes, it is common for multiple motifs for a particular TF to appear in the upstream regions of client genes. Although it is commonly argued that such redundancy is maintained by natural selection, using a simple birth-death model with truncation selection, it cah be shown that TFBS clustering can arise naturally by small-scale duplication processes (Lusk and Eisen 2010; Nourmohammad and Lässig 2011). Thus, while it is clear that the presence of multiple binding

sites can help ensure that an adjacent gene will be activated, there is as yet no formal evidence that such configurations anything more than a simple consequence of physical processes.

Finally, despite the centrality of TFBSs to gene expression, a diverse set of observations indicates that TFBS locations and motifs vary dramatically among closely related lineages, often with no apparent phenotypic consequences (Dowell 2010). These sorts of changes are apparently not simply due to random wandering of binding-site sequences, but to functional changes in the TFs themselves. For example, Nakagawa et al. (2013) find that the sequence specificities of members of the forkhead family of TFs have changed over time in the eukaryotic tree, with some evolving bispecificity (i.e., using two motifs), and others subsequently losing the ancestral specificity.

One of the most thoroughly analyzed metazoan promoters is that for the Endo16 gene in the sea urchin *Strongylocentrotus purpuratus,* which is bound by seven different TFs and forms the heart of a complex developmental cascade (Yuh et al. 1998). The developmental pathways associated with this gene were established over a period of several years using multiple individuals derived from a diverse natural population (because inbred laboratory lines have not been established). However, despite the exquisite molecular details of the results, their generality appears questionable given that the TFBSs for Endo16 as well as other regulatory genes in *S. purpuratus* harbor as much (and in some cases more) within-species sequence variation as surrounding, presumably nonfunctional nucleotides (Balhoff and Wray 2005; Garfield et al. 2012). Remarkably, the expression patterns of Endo16 appear to be conserved between species in different genera even though there is virtually no similarity between the regulatory regions (Romano and Wray 2003). Similar kinds of observations have been made on the regulatory regions of developmental genes in different ascidian species (Oda-Ishii et al. 2005). What remains unclear is whether this conservation in gene expression in the face of dramatic regulatory-sequence change is a consequence of alterations in the recognition motifs of the associated TFs or of changes in the specific TFs utilized.

Observations like this are by no means unique to marine invertebrates. Additional examples of apparent stability of gene expression across species with little apparent regulatory-region sequence continuity have been noted in the congeneric nematodes *C. elegans* and *C. briggsae* (Barriére et al. 2011, 2012; Reece-Hoyes et al. 2013). Likewise, multiple studies on developmental genes in *Drosophila* have pointed to up to 5% turnover of TFBs among closely related species (Moses et al. 2006; Crocker et al. 2008; Hare et al. 2008; He et al. 2011), again with conservation of gene-expression patterns being maintained among species despite the underlying changes in the molecular details of regulation (Ludwig et al. 2011; Paris et al. 2013).

Turnover in the regulatory machinery may be even more common in vertebrates. For example, Yokoyama and Pollack (2012) find that a specific single amino-acid change in the transcription factor SP1 that occurred independently in birds and mammals is associated with orchestrated TFBS-motif changes in hundreds of genes in each lineage. Moreover, across the different orders of mammals, which diverged $\sim$ 100 million years ago, at least a third of TFBSs appear not to be shared (Dermitzakis and Clark 2002; Schmidt et al. 2010; Yokoyama et al. 2010). These changes involve alterations in both the TFs utilized in gene expression and the motifs that they bind

to. Substantial changes in the TFs bound to the regulatory regions of orthologous genes have even been observed in closely related mouse species (Stefflova et al. 2013). Small numbers of changes in TF amino-acid sequence are also known to be associated with changes in binding-site specificities in plants (Sayou et al. 2014).

Again, the apparent underlying changes in regulatory mechanisms appear to often occur without noticeable differences in the final patterns of gene expression. For example, focusing on the RET receptor kinase gene, Fisher et al. (2006) found that the control region for the human gene drives expression within zebrafish even though there is no obvious sequence similarity. Wilson et al. (2008) found that when human chromosome 11 is put into mouse cells, the pattern of transcription is very similar to that in human cells. The latter result elegantly shows that even at this substantial level of phylogenetic divergence, transcription is largely dictated by sequences on the DNA and not by epigenetic effects restricted to individual lineages (in this case, nucleosomes and other interacting factors within the mouse nucleus).

Taken together, these observations suggest the existence of evolutionary pathways whereby the underlying mechanisms of gene regulation can be altered while maintaining a constant outward phenotype. The obvious implication of such observations is that the molecular details deciphered for the regulatory pathway in one model system need not be relevant to that in other species. Such regulatory repatterning is an obvious candidate for evolution at the cellular level by effectively neutral mechanisms, the details of which will be further explored below.

**Evolutionary dynamics of binding site motifs.** Transcription factors and their binding sites provide an explicit framework for evolutionary analysis, in that specific genotypic measures can be directly related to fitness (Gerland and Hwa 2002; Berg et al. 2004; Lässig 2007; Stewart et al. 2012). A common approach to understanding the evolution of binding motifs is to consider individual fitness to be a linear function of the fraction of time that the TFBS with $m$ matching sites is expected to be bound by its cognate TF, e.g.,

$$W(m) = 1 + \alpha P_{\text{on}}(m), \tag{7a}$$

where $\alpha$ is a scaling factor relating binding probability to fitness. As $\alpha \to 0$, $W(m) \to 1$, implying neutrality. Equation (7a) is often referred to as a **mesa fitness function**, because fitness increases sigmoidally from 1 to $(1 + \alpha)$ as the probability of gene activation increases from 0 to 1. Using the notation in the preceding section, under this model the fitness of an allele having $m$ matches in its TFBS can be written as

$$W(m) = 1 + \frac{\alpha}{1 + e^{-2m+\ln(B)}}, \tag{7b}$$

with $m = 0$ to $\ell$, where $\ell$ is the length of the optimal binding motif of the TF in bp, and $B$ is a composite measure of the background binding interference.

When combined with information on mutational movements between alternative TFBS states and the influence of random genetic drift on the efficiency of selection for larger $m$, this fitness function can be used to examine a number of basic issues. For example, it is well known that the multiple binding sites within a genome associated with a particular TF exhibit substantial variation in motif sequence. Although such variation might result from selection for alternative levels of locus-specific gene expression, because of the diminishing-returns nature of the fitness function (Figure

4), variation in motif matching is expected to arise naturally as selection pushes a population towards the drift barrier, where alternative levels of high $m$ are selectively equivalent (Berg and von Hippel 1987).

Over evolutionary time, the frequency distribution of the number of matches in the various TFBSs in a genome serviced by a particular TF is expected to reach an equilibrium between the mutational forces causing mismatches and the selective forces encouraging the proliferation of mutant alleles with higher specificity. As always, the efficiency of selection is modulated by the power of genetic drift, which is inversely proportional to the effective population size ($N_e$). From Details 4, provided all nucleotides mutate to all others at equal rates, the equilibrium distribution takes on a simple form, which is independent of the mutation rate,

$$\widetilde{P}(m) = C\left[\binom{\ell}{m} 3^{\ell-m}\right] e^{2N_e W(m)}, \tag{8a}$$

where $C$ is simply a normalization constant that ensures that the full set of probabilities sums to one.

The equilibrium distribution $\widetilde{P}(m)$ can be viewed as either the long-term average probability of states at a particular TFBS through the course of evolutionary time, or as the expected distribution of $m$ for a full set of equivalent TFBSs at any one point in time in a particular host genome. The exponential term in Equation (8a) is a constant when $W(m)$ is invariant, as would the case of the absence of selection, and so the term within brackets is equivalent to the expected distribution in the absence of selection, $\widetilde{P}_n(m)$. Thus, Equation (8a) indicates that the distribution of binding-site matching is equivalent to the neutral expectation weighted by an exponential gradient of the fitness surface relative to the power of random genetic drift, $1/(2N_e)$. Because $W(m) = 1 + s(m)$, where $s(m)$ is the selective advantage relative to a scaled fitness value of 1.0, Equation (8a) can also be written as

$$\widetilde{P}(m) = C\widetilde{P}_n(m)e^{2N_e s(m)}. \tag{8b}$$

Solution of Equation (8b) for relatively small and large motif sizes ($\ell = 8$ and 16) illustrates several general principles (Figure 5). First, regardless of the set of parameter values, substantial variation in $m$ is almost always expected among sites. Unless the motif size is small ($\ell = 8$) and levels of background interference and selection pressures are very high, most motifs are expected to contain mismatches. This behavior arises because the alternative states in the upper range of $m$ are selectively equivalent with respect to each other. Indeed, with a motif size of 16 bp, essentially no TFBS is expected to be perfect, unless the power of selection is extraordinarily high ($N_e\alpha \geq 10^6$). This clearly indicates that the presence of motif variation in genomes need not imply adaptive fine-tuning of individual loci. Second, with relatively weak selection pressure ($N_e\alpha \leq 1$), $\widetilde{P}(m)$ is very heavily skewed towards small numbers of matches (essentially the neutral expectation). This intrinsic weighting towards low numbers of matches is a result of the increasing multiplicity of configurations that lead to the same $m$ with increasing numbers of mismatches. Third, because the neutral distribution is strongly weighted toward low $m$, there can be a strong "phase transition" as $N_e\alpha$ increases from values $< 1.0$ to $> 1.0$. In addition, cases can exist in which $\widetilde{P}(m)$ is bimodal, with a peak to the

left resulting from the high multiplicity of motif configurations driven by mutation and a peak to the right.

These results have interesting implications for how the general features of TFBSs have traditionially been interpreted in an evolutionary sense. For example, Stewart et al. (2012) have argued that TFBS evolution proceeds in the face of an inherent tradeoff between specificity to enhance the stability of gene expression (which increases with matching motif length) and robustness to mutational breakdown (which decreases with increasing length). However, the results just shown demonstrate that less than maximum matching lengths arise naturally as a consequence of mutation-selection balance, without any direct selection for mutational robustness. In addition, while binding-site motif logos (which summarize the incidence of deployment of different nucleotides at each site within a consensus motif) are often used to infer levels of degeneracy within sites (e.g., 50% usage of G and T being taken to imply equal utility of both nucleotides), the multiplicity of partially matching states to any particular motif raises significant questions about the meaningfulness of such interpretations.

Taking into consideration the form of Equation (8b), one can infer the scaled strength of selection, $2N_e s(m)$, from the observed distribution of motifs for a particular TF within a genome. Mustonen and Lässig (2005) performed such an analysis for the cAMP receptor protein (CRP) in *E. coli,* showing that $2N_e s(m)$ for known TFBS sites for this factor is often in the range of 5 to 10 (Figure 6). An analysis of Abf1 binding sites in yeast yielded similar results (Mustonen et al. 2008). This type of analysis harbors substantial potential for TFBS discovery in a genome using thermodynamic principles (Djordjevic et al. 2003; Mustonen and Lässig 2005; Lässig 2007; Mustonen et al. 2008).

As noted above, $\widetilde{P}(m)$ is best described as a quasi-equilibrium, in that each individual motif is expected to wander across the entire distribution over evolutionary time, as described in Equation (4.4),with the entire ensemble of motifs retaining the equilibrium pattern. This general principle leads to a prediction – if the model is correct, and individual motifs are not being kept in their specific states by locus-specific selective pressures, comparison of the differences in binding energies between orthologous sites in different species should yield variances in motif binding consistent with the diffusion model. Observations on the Abf1 transcription factor in four species of *Saccharomyces* are consistent with this model (Mustonen et al. 2008). Thus, consistent with theory, the specific sequences of functional TFBSs appear to be conserved only to the extent that they yield levels of matching consistent with the relevant domain of effectively neutral alternatives. Due to the multiplicity of binding-site configurations deviating from the optimum, there is room for substantial sequence change via compensatory mutations.

The theory to this point ignores the possibility that evolution is not simply restricted to the motifs at TFBS sites, but includes the recognition domain in the TF itself. One of the central issues with respect to TF evolution is the length of the binding motif. High specificity and binding strength of a TF (i.e., a long binding site with low probability of spurious appearance at inappropriate genomic sites) maximizes the probability of binding to appropriate sites, as the probability of a random sequence is $(1/4)^\ell$. For a total genome size of $G$ bp, less than one random motif is expected to be present by chance on each strand if the motif size exceeds

$\ell^* = \ln(G)/\ln(4)$, which for genome sizes of 10 to 1000 Mb translates to $\ell^* = 12$ to 15 bp. However, because most TFs service multiple loci, high binding-site specificity can comprise a substantial mutational hazard, in the sense of imposing a larger cumulative target size for gene-inactivating mutations. These two points provide a simple explanation for why TFBSs are generally shorter than 15 bp in length.

---

**Details 4. The evolutionary dispersion of TFBS matching profiles.** For any TF, there will be a binding-site sequence dictated by the DNA-binding domain that maximizes the binding energy. However, owing to the recurrent introduction of mutations, variation will inevitably arise among the TFBS sequences harbored by different genes. Selection will prevent extreme TFBS degeneration, but there is little to be gained above a high level of binding strength. Thus, we can expect the levels of TF-TFBS matching to wander within the boundaries dictated by these extremes. Such variation will be manifest among the TFBS sequences associated with multiple genes within species as well as among orthologous genes across species. Here we outline a simple model to predict the evolutionary dispersion of such sequences as a function of mutation pressure and the efficiency of selection.

We start with two simple assumptions, the first being that all binding sites with the same number of mismatches ($n$) are equivalent with respect to binding probability, regardless of the position of the mismatches. Second, we assume that each of the four nucleotides mutates to each of the three other states at the same rate $\mu$. Under these conditions, with a TF recogition motif of length $\ell$ nucleotides, there are $\ell$ genotypic classes to consider, each consisting of multiple subclasses with equal expected probabilities under selection-mutation equilibrium. For example, class $n = 1$ consists of $3\ell$ types, as the single mismatch can reside in sites 1 to $\ell$ and there are three mismatching nucleotide types per site. More generally, the multiplicity within each class can be simply determined from the binomial coefficient $3^n\ell!/[(\ell - n)!n!]$. This allows us to reduce a complex problem involving many classes to a more manageable level.

The general **mean-field** approach is to treat a population as generally residing in a near pure state, with a short enough time scale assumed that stochastic changes involve one-step transitions to adjacent states. Denoting the probability that a TFBS resides in class $n$ at time $t$ as $P(n,t)$, where $n = \ell - m$ denotes the number of mismatches ($m$ being the number of matches), the time-dependent behavior of the system is described by

$$\frac{\partial P(n,t)}{t} = N\mu \cdot \{(3(\ell - n + 1)\phi_{n-1,n}P(n-1,t) - [n\phi_{n,n-1} + (3(\ell - n)\phi_{n,n+1}]P(n,t) + (n+1)\phi_{n+1,n}P(n+1,t)\}, \quad (4.1)$$

with the first term being dropped when $n = 0$, and the last being dropped when $n = \ell$. Here we assume a haploid population of $N$ individuals (for a diploid population, $2N$ should be substituted for $N$ throughout). This dynamical equation consists of three terms, the first denoting the influx of probability from the next lower class,with $(\ell - n + 1)$ functional sites mutating to non-matching states at rate $3\mu$ in each gene copy (the 3 accounting for mutation to three alternative nucleotide types), and going on to become fixed in the population with probability $\phi_{n-1,n}$. The second term accounts for the efflux from class $n$ to the next lower and upper classes ($n - 1$ and $n + 1$), again accounting for the number of possible mutations that cause such movement and their probabilities of fixation. The final term describes the influx from the next higher class, which has $n + 1$ mismatches, each back-mutating to a matching state at rate $\mu$.

The fixation probabilities are provided by Kimura's (1962) diffusion equation for newly arisen mutations,

$$\phi_{x,y} = \frac{1 - e^{-2N_e s_{x,y}/N}}{1 - e^{-2N_e s_{x,y}}} \tag{4.2}$$

where $N_e$ is the effective population size, $1/N$ is the initial frequency of a mutation (for a haploid population), and $s_{x,y}$ is the fractional selective advantage of allelic class $y$ over $x$.

Despite its apparent complexity, Equation (4.2) can be solved in a relatively transparent way, which we clarify by starting with the assumption of neutrality, i.e., $s_{x,y} = 0$ for all $(x, y)$. In this case, $\phi_{x,x+1} = \phi_{x,x-1} = 1/N$ for all $x$, and $N$ cancels out in Equation 3.1. The entire array of TFBS states can be represented as a diagram with connecting arrows denoting the flux rates between adjacent classes (Figure 7). Because the upward rate of flux declines and the downward rate of flux increases as $n$ increases, such a system must eventually reach an equilibrium, at which point for each class the total flux from above equals that from below, a condition known as **detailed balance**. For example, denoting the equilibrium solutions with a tilde, detailed balance requires that $3\ell\mu\widetilde{P}(0) = \mu\widetilde{P}(1)$, which tells us that the probability mass in class 1 is $3\ell$ times that in the perfectly matching class 0, i.e., $\widetilde{P}(1)/\widetilde{P}(0) = 3\ell$. More generally, for a linear model of this nature, the full solution for each class can be obtained by simply multiplying all of the coefficients on the arrows pointing up to the class with the product of all of the coefficients pointing down (Lynch 2013), which greatly simplifies to

$$\widetilde{P}(m) = C3^{\ell-m}\binom{\ell}{m}, \tag{4.3}$$

for the equilibrium probability of a site having $m$ matches, where $C = 1/\sum_{i=0}^{\ell} 3^i\binom{\ell}{i}$ is a normalization constant that ensures a total probability mass of 1.0.

There are two notable features of this solution. First, the equilibrium probabilities are completely independent of the mutation rate. Second, the term $3^{\ell-m}\binom{\ell}{m}$ is equivalent to the number of unique ways in which a sequence of length $\ell$ can harbor $n = \ell - m$ mismatches, accounting for both the $\binom{\ell}{m}$ spatial locations of the mismatches and the fact that there are three alternative inappropriate nucleotides for each mismatch.

Extending this approach to include selection is conceptually straight-forward – the coefficient on each arrow in Figure 7 simply needs to be multiplied by the fixation probability between adjacent classes. For example, for the arrows connecting classes 0 and 1, the coefficients become $3\ell\mu\phi_{0,1}$ and $\mu\phi_{1,0}$. The equilibrium probabilities are then again obtained using the rule noted above – multiplying together all of the coefficients leading up to and down to each class. Here, two useful results lead to great simplification: 1) $\phi_{n-1,n}/\phi_{n,n-1} = e^{2N_e s_{n-1,n}}$; and 2) $s_{n-1,n} = W_n - W_{n-1}$, where $W_i$ is the fitness of alleles with $i$ mismatches in their TFBS. Using these equalities, Equation (4.3) generalizes to

$$\widetilde{P}(m) = C3^{\ell-m}\binom{\ell}{m}e^{2N_e W(m)}, \tag{4.4}$$

where $C$ is again a normalization constant (equal to the reciprocal of the sum of the terms to the right of $C$ for all $m$). Thus, with selection, the equilibrium probability distribution is equivalent to a simple modification of the neutral expectation, with each genotypic probability being weighted exponentially by the product of its fitness and the effective population size (which influences the efficiency of selection).

**Regulatory rewiring.** In the preceding section, we encountered numerous examples in which the regulatory machinery associated with specific traits varies among species. We close with several well-dissected and dramatic examples of such rewiring, mostly derived from the yeast *S. cerevisiae*, where the study of gene regulation has been especially intense.

Drawing on earlier work by Tanay et al. (2005) and Hogues et al. (2008), Lavoie et al. (2010) found massive differences in the regulatory machinery associated with the ribosomal protein genes in *S. cerevisiae* and another yeast *Candida albicans*. Nearly every TF used in *S. cerevisiae* is utilized in a different way in the latter species, and shifts in the consensus motifs for orthologous TFs were seen as well. Moreover, the various TFs have associations with other functions, such as telomeres, centromeres, sulfur metabolism, and glycolysis in one or both species, showing still further reassignments. Some of this rewiring appears to be associated with whole-genome duplication. In *S. cerevisiae*, for example, an activator and repressor that control ribosomal-protein expression in normal and stress conditions are actually subfunctionalized duplicates, with the ancestral state inferred to have both functions. Remarkably, numerous other examples of transcriptional rewiring exist for the ribosomal protein genes in ascomycetes.

In another example, Martchenko et al. (2007) found that although *S. cerevisiae* and *C. albicans* have similar patterns of expression for genes associated with galactose metabolism, the underlying circuitry is completely different. Based on phylogenetic analysis, it appears that an intermediate-state species existed in which there were shared (and perhaps redundant) regulatory motifs, with each lineage then going on to divergently utilize just one. Interestingly, the regulatory TF in *S. cerevisiae* (Gal4) is still retained and has similar binding properties in *C. albicans*, but is used in other processes (Askew et al. 2009).

The regulatory wiring for the mating-type locus is also dramatically changed in yeast (Baker et al. 2011, 2012). Two mating-type cells exist in these species, a and $\alpha$. In *S. cerevisiae,* the a-specific genes are constitutively on in a cells and repressed in $\alpha$ cells, whereas in *C. albicans*, a-specific genes are activated by a regulatory protein only present in a cells. The alterations responsible for these differences again appeared to arise after an intermediate state was reached in which two sets of regulation were used simultaneously, and then divergently resolved in descendant lineages.

Finally, Tuch et al. (2008) find that Mcm1, a master regulatory factor that influences the expression of 4% of *S. cerevisiae* genes, is dramatically rewired in *Kluyveromyces lactis* and *C. albicans,* both as a consequence of gain and loss of target genes. Only about 20% of target-gene connections are conserved across all three species.

These kinds of observations are not restricted to yeasts. For example, a number of studies have suggested substantial regulatory rewiring among Eubacterial species (Babu et al. 2004; Lozada-Chávez et al. 2006; Price et al. 2007), with the general conclusion being that TFs are much less conserved than their target genes, although detailed examples of closely related species are lacking (see Perez and Groisman 2009a,b). The Eubacterial LexA repressor (to be discussed in the chapter on gene replication) provides a dramatic example of altered mechanisms of gene regulation in the face of conserved function.

In all of the above observations, the evolution of different control mechanisms involves coordinated TFBS changes at multiple target loci. How might multiple genes acquire the same sets of regulatory changes without an intermediate state of massive fitness loss? The simplest routes would appear to require an intermediate phase of redundancy with respect to the TF (Force et al. 2005; Tanay et al. 2005; Tuch et al. 2008) (Figure 8). If, for example, an ancestral TF exhibited bispecificity, i.e., was able to recognize two alternative TFBS motifs, random genetic drift possibly accompanied by alternative mutation pressures might result in the gradual loss of a different motif in each lineage, after which the TF would be free to lose a complementary motif in each lineage. Although such a scenario would result in the continued use of the same TF, the underlying regulatory language will have changed.

An apparent example of such evolutionary divergence is provided by LEAFY, a major regulator of flower development and cell division in land plants. Despite its central role in plant development and presence in just a single copy per genome, the recognition motif of this TF differs substantially between mosses and a clade containing almost all other land plants. However, hornworts, which are basal with respect to the rest of land plants, utilize a third consensus motif while also harboring a capacity to promiscuously recognize both motifs relied upon by the remainder of land plants (Sayou et al. 2014). Thus, the observed phylogenetic diversification of TFBS motifs appears to be a simple consequence of reciprocal focusing of a bispecific ancestral TF. This is likely to be a common mechanism of regulatory rewiring, at least in multicellular species, as roughly half of the TFs in mice and land plants recognize secondary motifs (Badis et al. 2009; Franco-Zorrilla et al. 2014).

Divergence of TFBS motifs can also be achieved by an effectively neutral process of subfunctionalization within a single genome, when an ancestral TF gene with two regulatory motifs becomes duplicated, with the two copies then retaining just single, complementary recognition motifs. In this case, the overall biology of the organism will remain the same, although the regulatory network will have become more complex, owing to the specialization of the TFs.

Finally, the TF used in one particular lineage might fortuitously recruit an unrelated TF through a spurious protein-protein interaction. Although initially neutral, this interaction might then encourage the gradual evolution of local binding sites complementary to the second TF, at which point the first TF would become redundant and subject to loss by drift and/or mutational degeneration. Under this scenario, a coordinated shift in the entire regulatory mechanism can be achieved by multiple loci, as the initiating event was acquired simultaneously by each of the relevant regulatory regions owing to their shared reliance on the first TF.

## Literature Cited

Abramowitz, M., and I. A. Stegun (eds.) 1972. Handbook of Mathematical Functions. Dover Publ., Inc., New York, NY.

Aravind, L., V. Anantharaman, S. Balaji, M. M. Babu, and L. M. Iyer. 2005. The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiol. Rev. 29: 231-262.

Askew, C., A. Sellam, E. Epp, H. Hogues, A. Mullick, A. Nantel, and M. Whiteway. 2009. Transcriptional regulation of carbohydrate metabolism in the human pathogen *Candida albicans.* PLoS Pathog. 5: e1000612.

Babu, M. M., N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann. 2004. Structure and evolution of transcriptional regulatory networks. Curr. Opin. Struct. Biol. 14: 283-291.

Badis, G., M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C. F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk. 2009. Diversity and complexity in DNA recognition by transcription factors. Science 324: 1720-1723.

Baker, C. R., L. N. Booth, T. R. Sorrells, and A. D. Johnson. 2012. Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional network diversification. Cell 151: 80-95.

Baker, C. R., B. B. Tuch, and A. D. Johnson. 2011. Extensive DNA-binding specificity divergence of a conserved transcription regulator. Proc. Natl. Acad. Sci. USA 108: 7493-7498.

Balhoff, J. P., and G. A. Wray. 2005. Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites. Proc. Natl. Acad. Sci. USA 102: 8591-8596.

Barrière, A., K. L. Gordon, and I. Ruvinsky. 2011. Distinct functional constraints partition sequence conservation in a cis-regulatory element. PLoS Genet. 7: e1002095.

Barrière, A., K. L. Gordon, and I. Ruvinsky. 2012. Coevolution within and between regulatory loci can preserve promoter function despite evolutionary rate acceleration. PLoS Genet. 8: e1002961.

Berg, J., S. Willmann, and Lässig. 2004. Adaptive evolution of transcription factor binding sites. BMC Evol. Biol. 4: 42.

Berg, O. G., and P. H. von Hippel. 1987. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J. Mol. Biol. 193: 723-750.

Berggrun, A., and R. T. Sauer. 2001. Contributions of distinct quaternary contacts to cooperative operator binding by Mnt repressor. Proc. Natl. Acad. Sci. USA 98: 2301-2305.

Bernstein, J. A., A. B. Khodursky, P. H. Lin, S. Lin-Chao, S. N. Cohen. 2002. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. Proc. Natl. Acad. Sci. USA 99: 9697-9702.

Bintu, L., N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. 2005. Transcriptional regulation by the numbers: applications. Curr. Opin. Genet. Dev. 15: 125-135.

Carey, M. F., C. L. Peterson, and S. T. Smale. 2012. Experimental strategies for cloning or identifying genes encoding DNA-binding proteins. Cold Spring Harb. Protoc. 2012: 183-192.

Carroll, S. B., J. K. Grenier, and S. D. Weatherbee. 2001. From DNA to Diversity. Blackwell Science, Malden, MA.

Cavalier-Smith, T. 1978. Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J. Cell Sci. 34: 247-278.

Cavalier-Smith, T. 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann. Bot. 95: 147-175.

Charoensawan, V., S. C. Janga, M. L. Bulyk, M. M. Babu, and S. A. Teichmann. 2012. DNA sequence preferences of transcriptional activators correlate more strongly than repressors with nucleosomes. Mol. Cell 47: 183-192.

Charoensawan, V., D. Wilson, and S. A. Teichmann. 2010. Genomic repertoires of DNA-binding transcription factors across the tree of life. Nucleic Acids Res. 38: 7364-7377.

Crocker, J., Y. Tamori, and A. Erives. 2008. Evolution acts on enhancer organization to fine-tune gradient threshold readouts. PLoS Biol. 6: e263.

Danko, C. G., N. Hah, X. Luo, A. L. Martins, L. Core, J. T. Lis, A. Siepel, and W. L. Kraus. 2013. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. Mol. Cell 50: 212-222.

Darzacq, X., Y. Shav-Tal, V. de Turris, Y. Brody, S. M. Shenoy, R. D. Phair, R. H. Singer. 2007. *In vivo* dynamics of RNA polymerase II transcription. Nat. Struct. Mol. Biol. 14: 796-806.

Davidson, E. H. 2001. Genomic Regulatory Systems. Academic Press, San Diego, CA.

de Mendoza, A., A. Sebé-Pedrós, M. S. Šestak, M. Matejcic, G. Torruella, T. Domazet-Loso, and I. Ruiz-Trillo. 2013. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proc. Natl. Acad. Sci. USA 110: E4858-E4866.

Dermitzakis, E. T., and A. G. Clark. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol. 19: 1114-1121.

Dowell, R. D. 2010. Transcription factor binding variation in the evolution of gene regulation. Trends Genet. 26: 468-475.

Djordjevic, M., A. M. Sengupta, and B. I. Shraiman. 2003. A biophysical approach to transcription factor binding site discovery. Genome Res. 13: 2381-2390.

Elf, J., G. W. Li, and X. S. Xie. 2007. Probing transcription factor dynamics at the single-molecule level in a living cell. Science 316: 1191-1194.

Elowitz, M. B., M. G. Surette, P. E. Wolf, J. B. Stock, and S. Leibler. 1999. Protein mobility in the cytoplasm of Escherichia coli. J. Bacteriol. 181: 197-203.

Erickson, H. P. 2009. Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy. Biol. Proceed. Online 11: 32-51.

Fields, D. S., Y. He, A. Y. Al-Uzri, and G. D. Stormo. 1997. Quantitative specificity of the Mnt repressor. J. Mol. Biol. 271: 178-194.

Fisher, S., E. A. Grice, R. M. Vinton, S. L. Bessling, and A. S. McCallion. 2006. Conservation

of RET regulatory function from human to zebrafish without sequence similarity. Science 312: 276-279.

Force, A., W. Cresko, F. B. Pickett, S. Proulx, C. Amemiya, and M. Lynch. 2005. The origin of gene subfunctions and modular gene regulation. Genetics 170: 433-446.

Franco-Zorrilla, J. M., I. López-Vidriero, J. L. Carrasco, M. Godoy, P. Vera, and R. Solano. 2014. DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc. Natl. Acad. Sci. USA 111: 2367-2372.

Furey, T. S. 2012. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat. Rev. Genet. 13: 840-852.

Garfield, D., R. Haygood, W. J. Nielsen, and G. A. Wray. 2012. Population genetics of cis-regulatory sequences that operate during embryonic development in the sea urchin *Strongylocentrotus purpuratus.* Evol. Dev. 14: 152-167.

Gerland, U., and T. Hwa. 2002. On the selection and evolution of regulatory DNA motifs. J. Mol. Evol. 55: 386-400.

Ghaemmaghami, S., W. K. Huh, K. Bower, R. W. Howson, A. Belle, N. Dephoure, E. K. O'Shea, and J. S. Weissman. 2003. Global analysis of protein expression in yeast. Nature 425: 737-741.

Golding, I., amnd E. C. Cox. 2004. RNA dynamics in live *Escherichia coli* cells. Proc. Natl. Acad. Sci. USA 101: 11310-11305.

Golding, I., J. Paulsson, S. M. Zawilski, and E. C. Cox. 2005. Real-time kinetics of gene activity in individual bacteria. Cell 123: 1025-1036.

Gowers, D. M., G. G. Wilson, and S. E. Halford. 2005. Measurement of the contributions of 1D and 3D pathways to the translocation of a protein along DNA. Proc. Natl. Acad. Sci. USA 102: 15883-15888.

Haimovich, G., D. A. Medina, S. Z. Causse, M. Garber, G. Millán-Zambrano, O. Barkai, S. Chávez, J. E. Pérez-Ortín, X. Darzacq, and M. Choder. 2013. Gene expression is circular: factors for mRNA degradation also foster mRNA synthesis. Cell 153: 1000-1011.

Halford, S. E. 2009. An end to 40 years of mistakes in DNA-protein association kinetics? Biochem. Soc. Trans. 37: 343-348.

Halford, S. E., and J. F. Marko. 2004. How do site-specific DNA-binding proteins find their targets? Nucleic Acids Res. 32: 3040-3052.

Hammar, P., M. Walldén, D. Fange, F. Persson, O. Baltekin, G. Ullman, P. Leroy, and J. Elf. 2014. Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation. Nature Genet. 46: 405-408.

Hare, E. E., B. K. Peterson, V. N. Iyer, R. Meier, and M. B. Eisen. 2008. Sepsid even-skipped enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. PLoS Genet. 4: e1000106.

He, B. Z., A. K. Holloway, S. J. Maerkl, and M. Kreitman. 2011. Does positive selection drive transcription factor binding site turnover? A test with *Drosophila* cis-regulatory modules. PLoS Genet. 7: e1002053.

He, Q., A. F. Bardet, B. Patton, J. Purvis, J. Johnston, A. Paulson, M. Gogol, A. Stark, and J. Zeitlinger. 2011. High conservation of transcription factor binding and evidence for combina-

torial regulation across six *Drosophila* species. Nat. Genet. 43: 414-420.

Hogues, H., H. Lavoie, A. Sellam, M. Mangos, T. Roemer, E. Purisima, A. Nantel, and M. Whiteway. 2008. Transcription factor substitution during the evolution of fungal ribosome regulation. Mol. Cell 29: 552-562.

Hsia, C. C., and W. McGinnis. 2003. Evolution of transcription factor function. Curr. Opin. Genet. Dev. 13: 199-206.

Ishihama, Y., T. Schmidt, J. Rappsilber, M. Mann, F. U. Hartl, M. J. Kerner, and D. Frishman. 2008. Protein abundance profiling of the *Escherichia coli* cytosol. BMC Genomics 9: 102.

Gunasekera, A., Y. W. Ebright, and R. H. Ebright. 1992. DNA sequence determinants for binding of the *Escherichia coli* catabolite gene activator protein. J. Biol. Chem. 267: 14713-147120.

Kimura, M. 1962. On the probability of fixation of mutant genes in a population. Genetics 47:713-719.

Kinney, J. B., A. Murugan, C. G. Callan, Jr., and E. C. Cox. 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc. Natl. Acad. Sci. USA 107: 9158-9163.

Kolesov, G., Z. Wunderlich, O. N. Laikova, M. S. Gelfand, and L. A. Mirny. 2007. How gene order is influenced by the biophysics of transcription regulation. Proc. Natl. Acad. Sci. USA 104: 13948-13953.

Kolomeisky, A. B. 2011. Physics of protein-DNA interactions: mechanisms of facilitated target search. Phys. Chem. Chem. Phys. 13: 2088-2095.

Kühn, T., T. O. Ihalainen, J. Hyväluoma, N. Dross, S. F. Willman, J. Langowski, M. Vihinen-Ranta, and J. Timonen. 2011. Protein diffusion in mammalian cell cytoplasm. PLoS One 6: e22962.

Lässig, M. 2007. From biophysics to evolutionary genetics: statistical aspects of gene regulation. BMC Bioinformatics 8 (Suppl. 6): S7.

Lavoie, H., H. Hogues, J. Mallick, A. Sellam, A. Nantel, and M. Whiteway. 2010. Evolutionary tinkering with conserved components of a transcriptional regulatory network. PLoS Biol. 8: e1000329.

Li, G. W., and X. S. Xie. 2011. Central dogma at the single-molecule level in living cells. Nature 475: 308-315.

Li, Z., J. Yan, C. J. Matheny, T. Corpora, J. Bravo, A. J. Warren, J. H. Bushweller, and N. A. Speck. 2003. Energetic contribution of residues in the Runx1 Runt domain to DNA binding. J. Biol. Chem. 278: 33088-33096.

Lozada-Chávez, I., S. C. Janga, and J. Collado-Vides. 2006. Bacterial regulatory networks are extremely flexible in evolution. Nucleic Acids Res. 34: 3434-3445.

Lu, P., C. Vogel, R. Wang, X. Yao, and E. M. Marcotte. 2007. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nat. Biotechnol. 25: 117-124.

Ludwig, M. Z., R. K. Manu, K. P. White, M. Kreitman. 2011. Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. PLoS Genet. 7: e1002364.

Luscombe, N. M., and J. M. Thornton. 2002. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J. Mol. Biol. 320: 991-1009.

Lusk, R. W., and M. B. Eisen. 2010. Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers. PLoS Genet. 6: e1000829.

Lynch, M. 2007. The Origins of Genome Architecture. Sinauer Assocs., Inc. Sunderland, MA.

Lynch, M. 2013. Evolutionary diversification of the multimeric states of proteins. Proc. Natl. Acad. Sci. USA 110: E2821-E2828.

Lynch, V. J., and G. P. Wagner. 2008. Resurrecting the role of transcription factor change in developmental evolution. Evolution 62: 2131-2154.

Malmström, J., M. Beck, A. Schmidt, V. Lange, E. W. Deutsch, and R. Aebersold. 2009. Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans.* Nature 460: 762-765.

Marinov, G. K., B. A. Williams, K. McCue, G. P. Schroth, J. Gertz, R. M. Myers, and B. J. Wold. 2014. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res.(in press).

Marklund, E. G., A. Mahmutovic, O. G. Berg, P. Hammar, D. van der Spoel, D. Fange, and J. Elf. 2013. Transcription-factor binding and sliding on DNA studied using micro- and macroscopic models. Proc. Natl. Acad. Sci. USA 110: 19796-19801.

Martchenko, M., A. Levitin, H. Hogues, A. Nantel, and M. Whiteway. 2007. Transcriptional rewiring of fungal galactose-metabolism circuitry. Curr. Biol. 17: 1007-1013.

Martínez-Antonio, A., and J. Collado-Vides. 2003. Identifying global regulators in transcriptional regulatory networks in bacteria. Curr. Opin. Microbiol. 6: 482-489.

Miguel, A., F. Montón, T. Li, F. Gómez-Herreros, S. Chávez, P. Alepuz, and J. E. Pérez-Ortín. 2013. External conditions inversely change the RNA polymerase II elongation rate and density in yeast. Biochim. Biophys. Acta 1829: 1248-1255.

Moses, A. M., D. A. Pollard, D. A. Nix, V. N. Iyer, X. Y. Li, M. D. Biggin, and M. B. Eisen. 2006. Large-scale turnover of functional transcription factor binding sites in *Drosophila.* PLoS Comput. Biol. 2: e130.

Mustonen, V., J. Kinney, C. G. Callan, Jr., and M. Lässig. 2008. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. Proc. Natl. Acad. Sci USA 105: 12376-12381.

Mustonen, V., and M. Lässig. 2005. Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. Proc. Natl. Acad. Sci. USA 102: 15936-15941.

Nakagawa, S., S. S. Gisselbrecht, J. M. Rogers, D. L. Hartl, and M. L. Bulyk. 2013. DNA-binding specificity changes in the evolution of forkhead transcription factors. Proc. Natl. Acad. Sci. USA 110: 12349-12354.

Newman, J. R., S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman. 2006. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. Nature 441: 840-846.

Nourmohammad, A., and M. Lässig. 2011. Formation of regulatory modules by local sequence duplication. PLoS Comput. Biol. 7: e1002167.

Nutiu, R., R. C. Friedman, S. Luo, I. Khrebtukova, D. Silva, R. Li, L. Zhang, G. P. Schroth, and C. B. Burge CB. 2011. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nat. Biotechnol. 29: 659-664.

Oda, M., K. Furukawa, K. Ogata, A. Sarai, and H. Nakamura. 1998. Thermodynamics of specific and nonspecific DNA binding by the c-Myb DNA-binding domain. J. Mol. Biol. 276: 571-590.

Oda-Ishii, I., V. Bertrand, I. Matsuo, P. Lemaire, and H. Saiga. 2005. Making very similar embryos with divergent genomes: conservation of regulatory mechanisms of Otx between the ascidians *Halocynthia roretzi* and *Ciona intestinalis*. Development 132: 1663-1674.

Omagari, K., H. Yoshimura, M. Takano, D. Hao, M. Ohmori, A. Sarai, and A. Suyama. 2004. Systematic single base-pair substitution analysis of DNA binding by the cAMP receptor protein in cyanobacterium *Synechocystis* sp. PCC 6803. FEBS Lett. 563: 55-58.

Paris, M., T. Kaplan, X. Y. Li, J. E. Villalta, S. E. Lott, and M. B. Eisen. 2013. Extensive divergence of transcription factor binding in *Drosophila* embryos with highly conserved gene expression. PLoS Genet. 9: e1003748.

Peccoud, J., and B. Ycart. 1995. Markovian modeling of gene-product synthesis. 48: 222234.

Perez, J. C., and E. A. Groisman. 2009a. Evolution of transcriptional regulatory circuits in bacteria. Cell 138: 233-244.

Perez, J. C., and E. A. Groisman. 2009b. Transcription factor function and promoter architecture govern the evolution of bacterial regulons. Proc. Natl. Acad. Sci. USA 106: 4319-4324.

Price, H. J., A. H. Sparrow, and A. F. Nauman. 1973. Correlations between nuclear volume, cell volume and DNA content in meristematic cells of herbaceous angiosperms. Experientia 29: 1028-1029.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2007. Orthologous transcription factors in bacteria have different functions and regulate different genes. PLoS Comput. Biol. 3: 1739-1750.

Proshkin, S., A. R. Rahmouni, A. Mironov, and E. Nudler. 2010. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. Science 328: 504-508.

Raj, A., C. S. Peskin, D. Tranchina, and D. Y. Vargas, and S. Tyagi. 2006. Stochastic mRNA synthesis in mammalian cells. PLoS Biol. 4: e309.

Rajewsky, N., N. D. Socci, M. Zapotocky, and E. D. Siggia. 2002. The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. Genome Res. 12: 298-308.

Rajkumar, A. S., N. Dénervaud, and S. J. Maerkl. 2013. Mapping the fine structure of a eukaryotic promoter input-output function. Nat. Genet. 45: 1207-1215.

Reece-Hoyes, J. S., C. Pons, A. Diallo, A. Mori, S. Shrestha, S. Kadreppa, J. Nelson, S. Diprima, A. Dricot, B. R. Lajoie, P. S. Ribeiro, M. T. Weirauch, D. E. Hill, T. R. Hughes, C. L. Myers, and A. J. Walhout. 2013. Extensive rewiring and complex evolutionary dynamics in a *C. elegans* multiparameter transcription factor network. Mol. Cell 51: 116-127.

Riggs, A. D., S. Bourgeois, and M. Cohn. 1970 The lac repressor-operator interaction. 3. Kinetic studies. J. Mol. Biol. 53: 401-417.

Robison, K., A. M. McGuire, and G. M. Church. 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. J. Mol. Biol. 284: 241-254.

Romano, L. A., and G. A. Wray. 2003. Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. Development 130: 4187-4199.

Sanchez, A., and I. Golding. 2013. Genetic determinants and cellular constraints in noisy gene expression. Science 342: 1188-1193.

Sarai, A., and Y. Takeda. 1989. Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. Proc. Natl. Acad. Sci. USA 86: 6513-6517.

Sayou, C., M. Monniaux, M. H. Nanao, E. Moyroud, S. F. Brockington, E. Thévenon, H. Chahtane, N. Warthmann, M. Melkonian, Y. Zhang, G. K. Wong, D. Weigel, F. Parcy, and R. Dumas. 2014. A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. Science 343: 645-648.

Schildbach, J. F., A. W. Karzai, B. E. Raumann, and R. T. Sauer. 1999. Origins of DNA-binding specificity: role of protein contacts with the DNA backbone. Proc. Natl. Acad. Sci. USA 96: 811-817.

Schmidt, D., M. D. Wilson, B. Ballester, P. C. Schwalie, G. D. Brown, A. Marshall, C. Kutter, S. Watt, C. P. Martinez-Jimenez, S. Mackay, I. Talianidis, P. Flicek, and D. T. Odom. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328: 1036-1040.

Stefflova, K., D. Thybert, M. D. Wilson, I. Streeter, J. Aleksic, P. Karagianni, A. Brazma, D. J. Adams, I. Talianidis, J. C. Marioni, P. Flicek, and D. T. Odom. 2013. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell 154: 530-540.

Schwanhäusser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. 2011. Global quantification of mammalian gene expression control. Nature 473: 337-342.

Sengupta, A. M., M. Djordjevic, and B. I. Shraiman. 2002. Specificity and robustness in transcription control networks. Proc. Natl. Acad. Sci. USA 99: 2072-2077.

Shahrezaei, V., and P. S. Swain. 2008. Analytical distributions for stochastic gene expression. Proc. Natl. Acad. Sci. USA 105: 17256-17261.

Shuter, B. J., J. E. Thomas, W. D. Taylor, and A. M. Zimmerman. 1983. Phenotypic correlates of genomic DNA content in unicellular eukaryotes and other cells. Amer. Natur. 122: 26-44.

Smith, R. P., L. Taher, R. P. Patwardhan, M. J. Kim, F. Inoue, J. Shendure, I. Ovcharenko, and N. Ahituv. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat. Genet. 45: 1021-1028.

So, L. H., A. Ghosh, C. Zong, L. A. Sepúlveda, R. Segev, and I. Golding. 2011. General properties of transcriptional time series in *Escherichia coli.* Nat. Genet. 43: 554-60.

Stewart, A. J., S. Hannenhalli, and J. B. Plotkin. 2012. Why transcription factor binding sites are ten nucleotides long. Genetics 192: 973-985.

Sun, M., B. Schwalb, D. Schulz, N. Pirkl, S. Etzold, L. Larivière, K. C. Maier, M. Seizl, A. Tresch, and P. Cramer. 2012. Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. Genome Res. 22: 1350-1359.

Takeda, Y., A. Sarai, and V. M. Rivera. 1989. Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. Proc.

Natl. Acad. Sci. USA 86: 439-443.

Tanay, A., A. Regev, and R. Shamir. 2005. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc. Natl. Acad. Sci. USA 102: 7203-7208.

Taniguchi, Y., P. J. Choi, G. W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. 2010. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science 329: 533-538.

Thattai, M., and A. van Oudenaarden. 2001. Intrinsic noise in gene regulatory networks. Proc. Natl. Acad. Sci. USA 98: 8614-8619.

Tuch, B. B., H. Li, and A. D. Johnson. 2008. Evolution of eukaryotic transcription circuits. Science 319: 1797-1799.

Thurman, R. E., et al. 2012. The accessible chromatin landscape of the human genome. Nature 489: 75-82.

van Nimwegen, E. 2003. Scaling laws in the functional content of genomes. Trends Genet. 19: 479-484.

von Hippel, P. H., and O. G. Berg. 1986. On the specificity of DNA-protein interactions. Proc. Natl. Acad. Sci. USA 83: 1608-1612.

von Hippel, P. H., and O. G. Berg. 1989. Facilitated target location in biological systems. J. Biol. Chem. 264: 675-678.

Wagner, G. P., and V. J. Lynch. 2008. The gene regulatory logic of transcription factor evolution. Trends Ecol. Evol. 23: 377-385.

Wang, X., H. Gao, Y. Shen, G. M. Weinstock, J. Zhou, and T. Palzkill. 2008. A high-throughput percentage-of-binding strategy to measure binding energies in DNA-protein interactions: application to genome-scale site discovery. Nucleic Acids Res. 36: 4863-4871.

Wang, Y., C. L. Liu, J. D. Storey, R. J. Tibshirani, D. Herschlag, and P. O. Brown. 2002. Precision and functional specificity in mRNA decay. Proc. Natl. Acad. Sci. USA 99: 5860-5865.

Wang, Y. M., R. H. Austin, and E. C. Cox. 2006. Single molecule measurements of repressor protein 1D diffusion on DNA. Phys. Rev. Lett. 97: 048302.

Wasyl, Z., E. Luchter, and W. Bielanski, Jr. 1971. Determination of the effective radius of protein molecules by thin-layer gel filtration. Biochim. Biophys. Acta 243: 11-18.

Wilson, M. D., N. L. Barbosa-Morais, D. Schmidt, C. M. Conboy, L. Vanes, V. L. Tybulewicz, E. M. Fisher, S. Tavaré, and D. T. Odom. 2008. Species-specific transcription in mice carrying human chromosome 21. Science 322: 434-438.

Wray, G. A. 2007. The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. 8: 206-216.

Yan, J., M. Enge, T. Whitington, K. Dave, J. Liu, I. Sur, B. Schmierer, A. Jolma, T. Kivioja, M. Taipale, and J. Taipale. 2013. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. Cell 154: 801-813.

Yang, L., T. Zhou, I. Dror, A. Mathelier, W. W. Wasserman, R. Gordân, and R. Rohs. 2014. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res. 42(Database issue): D148-D155.

Yokoyama, K. D., and D. D. Pollock. 2012. SP transcription factor paralogs and DNA-binding sites coevolve and adaptively converge in mammals and birds. Genome Biol. Evol. 4: 1102-1117.

Yokoyama, K. D., J. L. Thorne, and G. A. Wray. 2011. Coordinated genome-wide modifications within proximal promoter cis-regulatory elements during vertebrate evolution. Genome Biol. Evol. 3: 66-74.

Yuh, C. H., H. Bolouri, and E. H. Davidson. 1998. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science 279: 1896-1902.

Zenklusen, D., D. R. Larson, and R. H. Singer. 2008. Single-RNA counting reveals alternative modes of gene expression in yeast. Nat. Struct. Mol. Biol. 15: 1263-1271.

Zhou, H.-X. 2011. Rapid search for specific sites on DNA through conformational switch of non-specifically bound proteins. Proc. Natl. Acad. Sci. USA 108: 8651-8656.

**Figure 1.** Flow chart for the key determinants of the number of proteins within a cell. As the associated transcription factors join and depart their target binding sites, the associated gene makes transtions to the active and inactive states at respective rates $k_{\text{on}}$ and $k_{\text{off}}$. An actively transcribing gene then produces fully functional mRNAs at rate $k_{\text{m}}$, which are in turn translated into proteins at rate $k_{\text{p}}$. Messenger RNAs and proteins are degraded at rates $\gamma_{\text{m}}$ and $\gamma_{\text{p}}$, respectively.

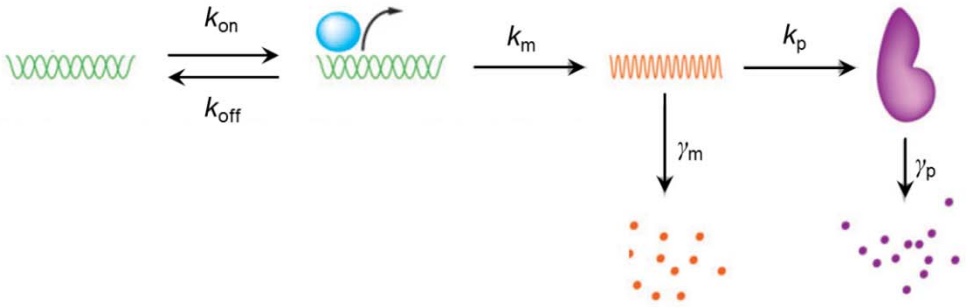
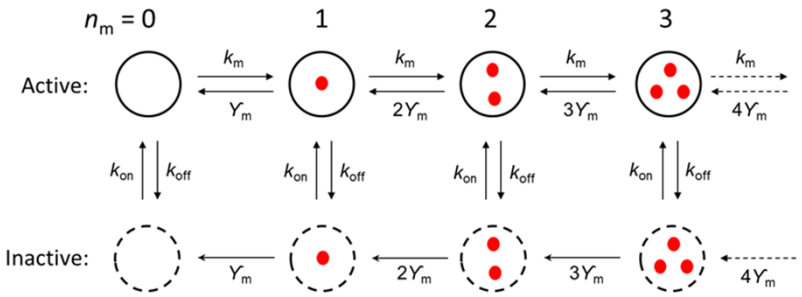


**Figure 2. Top)** Flow chart for the transitions of numbers of mRNAs within individual cells ($N_m$). The rates are defined as in Figure 1. Open and closed circles denote transcribing and nontranscribing cells, respectively, whereas the red dots denote the numbers of transcripts in cells of various states. Note that inactive genes can only lose (not gain) mRNAs. **Bottom)** Probability distributions for the number of mRNAs from a particular gene present in cells, as a function of the ratio of transcription to degradation rates, $k_m/\gamma_m$, and the rates of transition of cells from the off to on states and vice versa, $k_{on}$ and $k_{off}$ respectively. Solid lines are Poisson distributions, with mean $k_m/\gamma_m$ expected for genes that are constitutively on, whereas the black dashed and dotted lines represent situations in which the transition rates to on and off states are equal, with the mode of the distribution shifting to the right with increasing rates of switching.



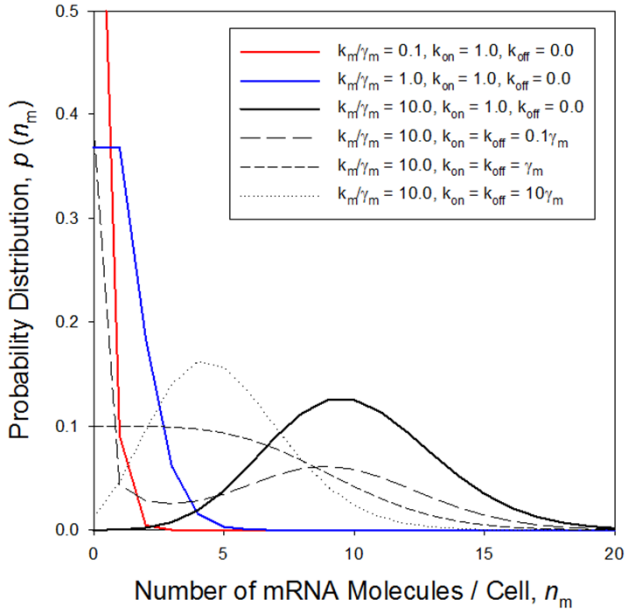



**Figure 3.** Idealized view of the temporal variation in gene expression within a cell. The gene is stochastically turned on (blue vertical bars) or off at points depending on the binding of the cognate transcription factors. Messenger RNAs are produced during the on periods, but they they decline at an exponential rate during off periods. Protein numbers also vary within the cell, rising during periods of mRNA abundance, but then declining via degradation during periods of mRNA rarity. The fluctuations in protein numbers are damped, owing to their greater longevities than mRNA molecules.

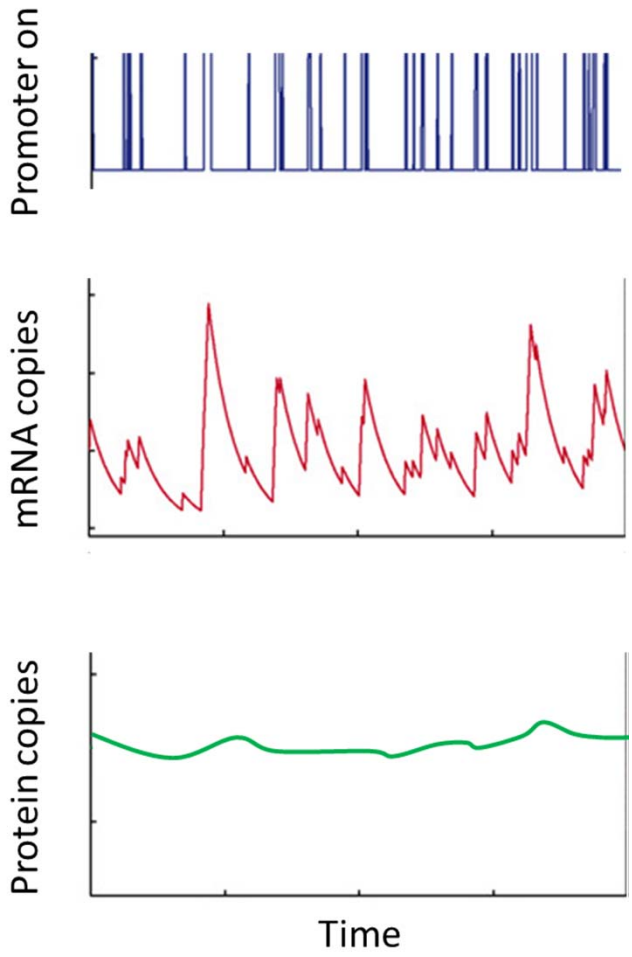


**Figure 4.** The probability that a particular transcription-factor binding site (TFBS) is bound by a cognate transcription factor (TF), given the level of background interference ($B$) and the number of nucleotides at the site ($m$) matching the optimal recognition sequence of the TF. The curves, obtained from Equation (6), cover the range of biologically plausible values of $B$ (as described in the text).

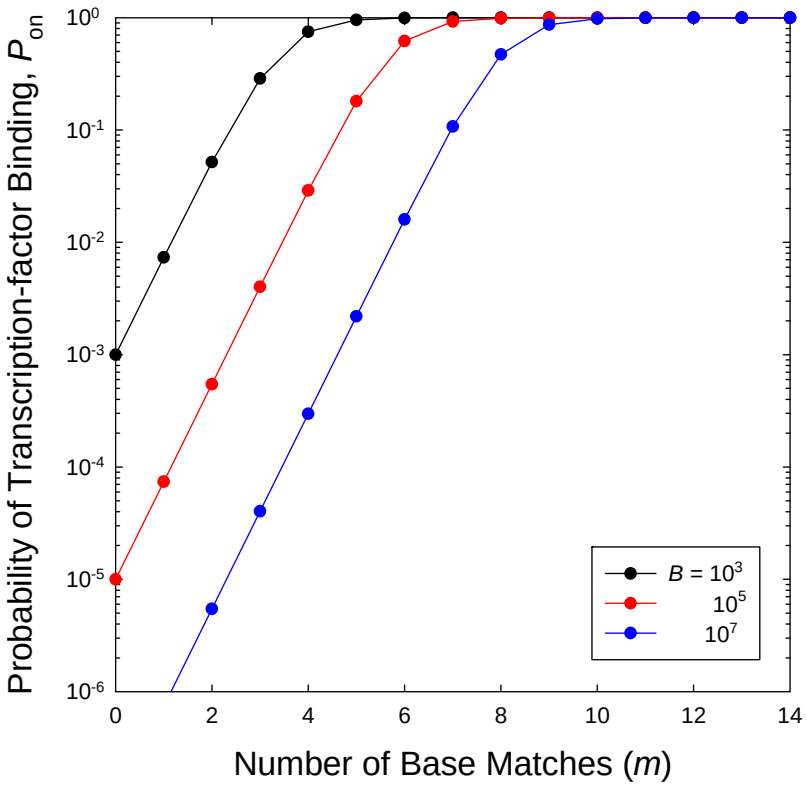


**Figure 5.** The expected equilibrium evolutionary distribution of binding-site matches with transcription-factor motifs of lengths $\ell = 8$ and 16. Results are given for various levels of the strength of selection relative to the power of genetic drift ($N_e\alpha$), and two levels of background interference ($B$).

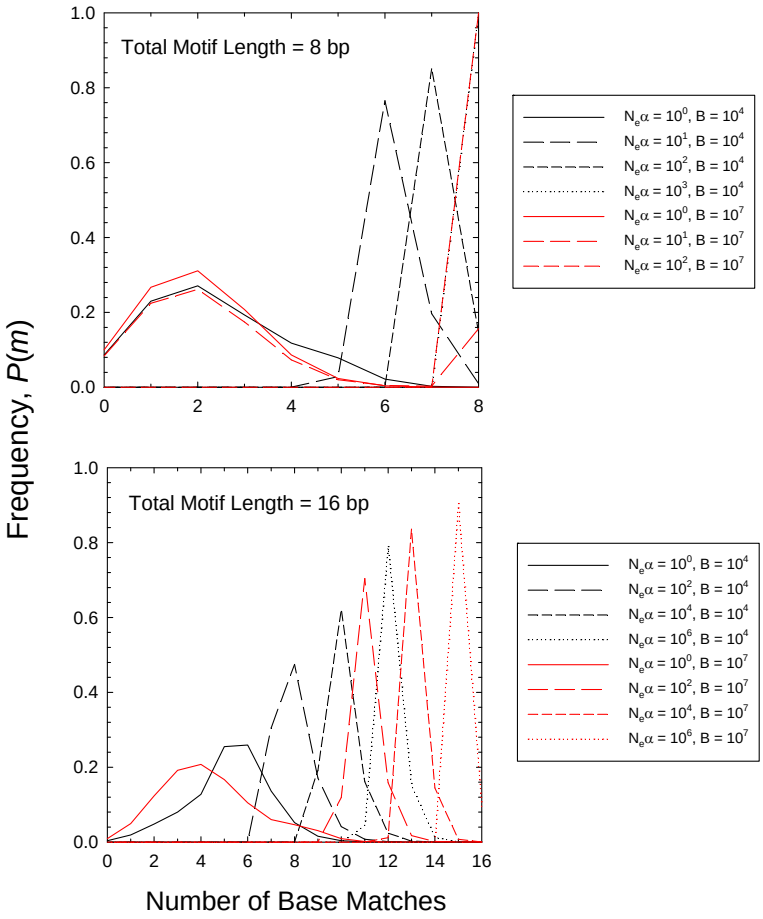


**Figure 6.** **a)** Distribution of binding energies associated with the transcription factor CRP in *E. coli*. Note that contrary to the approach in the text, the binding sites are characterized with respect to energy rather than mismatches, although the two scales are entirely interchangeable. Energies are computed using sliding windows of 22-bp (the length of the consensus TFBS for CRP) sequences across the entire *E. coli* genome. The energy scale is set such that $E = 0$ denotes the strongest possible binding site, with all other (more weakly binding) motif sequences simply being measured as the deviation from this value (and appearing further towards the right). The rapidly rising left curve is the tail of the remainder of the energy distribution (to the right) multiplied by 30 to enhance visualization. The solid lines illustrate the expected distribution based on the full set of possible 22-bp sequences under a random (model using the known distribution of nucleotide types in the *E. coli* genome; these fit very well in the right portion of the distribution, which represents non-specific binding sites. The red line is the excess of motifs in the left tail from this neutral expectation. Motifs in the red region are viewed as true binding sites, whereas all others denote the background resulting from nonspecific binding. **b)** As discussed in the text, for TFBS motifs deemed to be functional, the logarithm of the ratio of observed abundance relative to that expected under neutrality (the red line), $\widetilde{P}/\widetilde{P}_n$, provides an estimate of $2N_e s$, which is equivalent to the selective advantage of each site relative to the power of drift. From Mustonen and Lässig (2005).

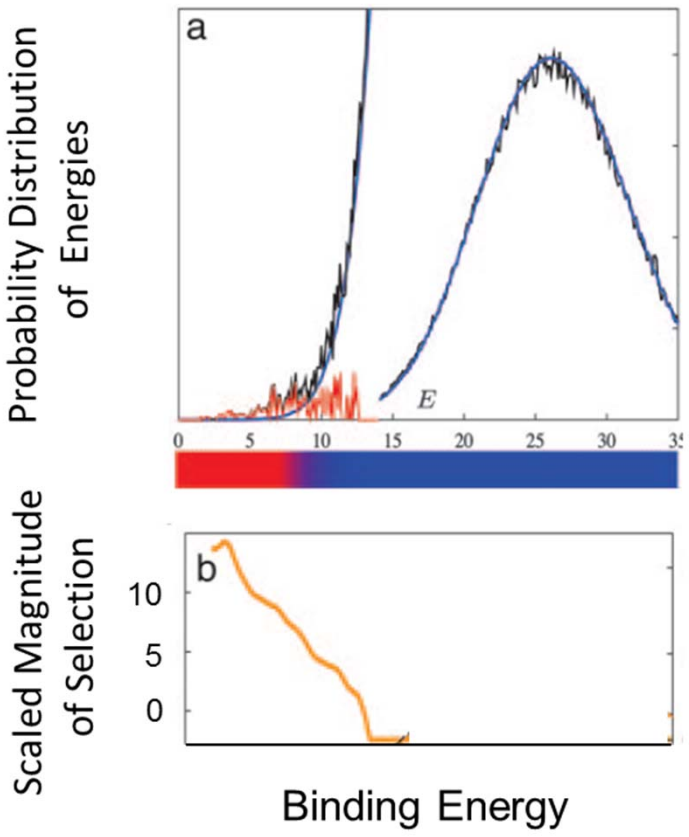


**Figure 7.** Flow diagram for the alternative states of a binding site of length $\ell = 5$, with the dot diagrams below simply illustrating one specific type within each category of numbers of mismatches ($n$). The transition rates are given on the arrows for the case of neutrality, where the probability of fixation is equal to the mutation rate per site, $3\mu$ in the case of single-site losses (arrows to the right) because each appropriate nucleotide can mutate to three others, and $\mu$ in the case of site improvement (arrows to the left) because each mismatch can only mutate to the appropriate state in one way.