æ

# 21. TRANSCRIPTION

8 January 2023

All life relies on heritable information encoded in DNA, with the majority of cellular functions being derived from products resulting from transcription of DNA into RNA. Some RNAs, such as those in the ribosome, have direct functions, but for protein-coding loci, the first-step messenger RNAs (mRNAs) must be translated subsequently into strings of amino acids. Gene expression is almost always regulated by specific DNA-binding proteins that activate and/or repress the target genes by binding to their regulatory regions. 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 way to erase this legacy of the earliest stages of evolution.

Transcription factors are usually referred to as *trans*-acting, in the sense that the genetic loci encoding for them are generally unlinked to (or at least physically distant from) their regulatory targets. In contrast, TFBSs are generally referred to as *cis*-acting, as they are physically adjacent to the affected coding regions. This distinction can be blurred in prokaryotes, where a TF is sometimes encoded in the same multilocus transcriptional unit (called an operon) as its target gene, and all genes are linked in nonrecombining genomes.

Because transcription factors must bind to the regulatory sites of their target proteins with high affinity relative to off-target sites, gene regulation provides an excellent example of coevolution at the molecular level. A number of questions immediately arise. How long and accurate does a TFBS have to be to ensure high specificity with respect to its cognate TF? Are both the TF and the TFBS free to wander in sequence space, provided an adequate level of joint matching is maintained? What happens when a TF services increasingly large numbers of genes? How do new TF-TFBS interactions arise?

Gene transcription is generally not an autonomous, invariant process, but rather is driven by extra- or intracellular information. The signals range from small inorganic molecules to simple metabolites produced by other genes, which in turn activate (or suppress) other transcription factors, typically by modifying intermediary regulatory proteins (themselves often transcription factors) in functionally significant ways. Central to all of biology, this transfer of environmental information via transcription factors to downstream gene expression is called signal transduction and is the topic of Chapter 22. In addition, although TF proteins and their binding sites provide the dominant mechanism for regulating gene expression, they are by no means the only intervening factor. For example, post-transcriptional regulation

can occur in the form of small complementary RNAs that can bind to transcripts, and post-translational modifications (Chapter 14) can further modify the operational features of gene products. As an overall entrée into the overall field of gene regulation, however, this chapter will focus on processes driven by TF proteins.

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 outlined in Chapter 7, mRNAs of active genes 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 understood.

## Molecular Stochasticity in Single Cells

The fitness of a cell ultimately depends on the quality, quantity, and stoichiometric relationships of its underlying functional constituents. 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 among cells, even in a completely homogeneous external environment and with genetically identical and uniformly aged cells. The many factors governing the probability distributions of numbers of molecules per cell can be subsumed into six coefficients: the rates at which an inactive gene enters the transcriptionally active state and vice versa, $k_{\text{on}}$ and $k_{\text{off}}$ respectively; the rate at which an active gene transcribes mRNAs, $k_{\text{m}}$; the rate at which an mRNA is translated into proteins, $k_{\text{p}}$; and the rates of degradation of mRNAs and proteins, $\delta_{\text{m}}$ and $\delta_{\text{p}}$ respectively (Figure 21.1). We now consider the ways in which these various factors influence the distributions of numbers of transcripts and proteins within different cells. For heuristic purposes, the following assumes a simple system in which gene expression is a function of the binding of a single activating TF, as is the case for many bacterial genes.

**Cellular mRNA abundances.** We first consider the numbers of mRNAs found in cells, $n_{\text{m}}$, as this has cascading effects on protein numbers. As noted in Figure 21.2, active cells gain mRNAs by transcription and lose them by degradation, whereas inactive cells (with no TF engaged with the target TFBS) can only lose mRNAs. Cells can also move back and forth between active and inactive states. The total transition rates to smaller numbers of mRNAs increase linearly with increasing $n_{\text{m}}$ simply because there are more targets for degradation, and as a consequence, such systems converge on a steady-state distribution regardless of the starting point. As outlined in Foundations 21.1, a particularly simple outcome is obtained when a gene is constitutively turned on ($k_{\text{off}} = 0$). In this case, $n_{\text{m}}$ is Poisson distributed, with both the mean and the variance of the number of mRNAs per cell equaling the

ratio of gain and loss rates, $k_{\rm m}/\delta_{\rm m}$. 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 or Gaussian) distribution with larger mean (Figure 21.2).

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, this fractional time is a simple function of the ratio of association and dissociation rates of the TF,

$$P_{\rm on} = k_{\rm on}/(k_{\rm on} + k_{\rm off}). \tag{21.1}$$

This result arises because under steady-state 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 than Poisson, as it involves a mixture of the distributions of $n_m$ in active and inactive cells (see Foundations 21.1 for the full expression).

Contrary to common belief, gene regulation magnifies the variance of mRNA numbers among cells (Figure 21.2). 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. Further insight into the likely specific forms of the distribution of mRNA molecules per cell requires quantitative estimates for the four key parameters: $k_{\rm m}$, $k_{\rm on}$, $k_{\rm off}$, and $\delta_{\rm m}$.

Based on observed rates of mRNA chain elongation (Chapter 20), and assuming an average transcript length of $\sim$ 1 kb, an activated *E. coli* gene is capable of producing as many as 50 to 150 transcripts/hour (Golding and Cox 2004; Proshkin et al. 2010). Approximate estimates for the yeasts *S. cerevisiae* and *S. pombe*, again largely based on chain-elongation rates, and assuming an average transcript length of 2 kb, fall in the range of $\sim$ 10 to 35/hour (Lynch 2007a; 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 contributes 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 estimated 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 $\delta$ by assuming random, exponential decay:

$$0.5 = e^{-\delta T_{0.5}}. \tag{21.2}$$

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* (Wang et al. 2002) and mouse fibroblast cells (Schwanhäusser et al. 2011) are larger, 22 mins and 9 hours respectively. 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.

The rates at which genes turn on and off transcriptionally, $k_{\rm on}$ and $k_{\rm off}$, dictate the dynamics of gene activation/inactivation. For example, the average time between bursts of transcription at a particular locus, which is equivalent to the mean time that a silent gene remains off, is equal to $1/k_{\rm on}$. Once turned on, a gene remains transcriptionally 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 the on and off 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 (per engaged gene) 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 $k_{\rm off}$, which implies $k_{\rm m}$ of the same order of magnitude noted above for this species. A primary determinant of the on-off process in highly expressed genes appears to involve alternating periods of engagement and dissociation of gyrase, a molecule used to relieve positive supercoiling of the DNA that results from the progression of the transcription machinery (Chong et al. 2014), although in more lowly expressed genes, the stochastic engagement with TFs is likely to be involved (below).

The preceding survey provides a mechanistic understanding of why the per-cell numbers of mRNA molecules associated with individual genes are generally quite small (Chapter 7). Suppose, for example, that the rate of full mRNA production by an *E. coli* cell in the on state is $k_{\rm m} \simeq 50$ mRNAs per active gene per hour, with a median degradation rate of $\delta_{\rm m} \simeq 8$/hour, and the gene is turned on only a fraction of the time ($P_{\rm on} < 1$). At equilibrium, the average number of mRNAs per cell equals the ratio of the production and elimination rates, $\mu_m = P_{\rm on}k_{\rm m}/\delta_{\rm m}$ (Foundations 21.1), so the theory predicts that genes in this species should commonly be represented by fewer than 10 mRNAs per cell. This qualitative prediction is consistent with an observed average number of 5 mRNAs/gene/cell (and a range of 0 to 100) in *E. coli* (Lu et al. 2007; Li and Xie 2011).

Modifications of the theory will be necessary for eukaryotes, where there can be a negative feedback between mRNA synthesis and degradation (Sun et al. 2012; Haimovich et al. 2013), and where genes can have much more complex modes of regulation involving multiple transcription factors with higher-order interactions, long-distance enhancer elements, etc., especially in multicellular species (Hafner and Boettiger 2023). Nonetheless, it remains clear that even in large 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.

**Cellular protein abundances.** We now turn to the ultimate manifestation of transcription, the numbers of protein molecules per cell. Because protein production depends on the presence of mRNAs, the kinds of transcriptional noise noted above naturally transmits to the level of translation. However, if a large number of proteins are translated per mRNA, the degree of noise propagation can be reduced, owing to the fact that the life span of a protein is typically greater than that of its associated mRNA. For example, in *S. cerevisiae,* most proteins outlive their maternal mRNAs, with the average ratio of half lives being $\sim 3$ (Shahrezaei and Swain 2008; Martin-Perez and Villén 2017). Likewise, in mouse fibroblast cells, the median half life of a protein, $\sim 2$ 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, the temporal variation in protein numbers per cell is expected to be dampened in comparison to that for mRNAs (Figure 21.3).

As a consequence of their greater half lives and higher rates of production, proteins also tend to be much more abundant in cells than their cognate mRNAs (Chapter 7). For example, the average ratio is 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 per gene (Ishihama et al. 2008; Malmström et al. 2008; Taniguchi et al. 2010). 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 per expressed gene (Schwanhäusser et al. 2011). Notably, transcription factors tend to be the rarest proteins within cells (Ghaemmaghami et al. 2003; Li et al. 2014; Marinov et al. 2014).

Two key determinants of the level of protein production are the number of mRNAs produced by active genes over the typical life span of a protein,

$$a = k_m/\delta_p, \tag{21.3a}$$

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

$$b = k_p/\delta_m, \tag{21.3b}$$

where in both cases the average life span of a molecule is equal to the inverse of the decay rate. Together with the transcriptional activation and deactivation rates, $k_{\text{on}}$ and $k_{\text{off}}$, these parameters define the distribution of protein numbers among cells (Shahrezaei and Swain 2008). The mean number of proteins per cell is a simple extension of the expected value for mRNAs ($\mu_{\text{m}} = P_{\text{on}}k_{\text{m}}/\delta_{\text{m}}$, from Foundations 21.1),

$$\mu_{\text{p}} = \frac{\mu_{\text{m}}k_{\text{p}}}{\delta_{\text{p}}} = P_{\text{on}} \cdot a \cdot b. \tag{21.4a}$$

The variance in number of protein molecules among cells is described by

$$\sigma_{\text{p}}^2 = \mu_{\text{p}}\left(1 + b + \frac{ab(1-P_{\text{on}})\delta_{\text{p}}}{\delta_{\text{p}} + k_{\text{on}} + k_{\text{off}}}\right), \tag{21.4b}$$

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

These composite expressions define the wide variety of ways in which cells might control their steady-state numbers of active proteins, e.g., by adjusting rates of engagement, elongation, and decay. 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. In *E. coli* and *S. cerevisiae,* for proteins that assemble in complexes, the relative rates of protein production associated with each locus are directly proportional to the relative numbers of molecules required in each assembly, suggesting coordinated expression so as to maintain stoichiometric balance (Li et al. 2014). Orthologous genes in closely related yeast species often achieve similar levels of overall expression via compensatory changes in rates of transcription and mRNA degradation (Dori-bachash et al. 2011). Such a pattern is compatible with a bivariate drift barrier in which a particular phenotype can be achieved by interchangeable mechanisms.

**Expression noise and adaptation.** The preceding overview makes clear that phenotypic noise is an inevitable consequence of the structure of biology, resulting from the stochastic features of transcription-factor binding and high rates of mRNA decay. Although muted somewhat, noise created at the level of bursty transcription still has cascading effects at the level of translation. There is essentially no way to completely eliminate such between-cell variation, and based on thermodynamic principles, any process for regulating gene-expression stochasticity must require an energetic investment. Nonetheless, several authors have suggested that expression noise may be promoted by natural selection as a means for coping with a variable environment (Fraser et al. 2004; Tenăse-Nicola and ten Wolde 2008; Wang and Zhang 2011; Levy et al. 2012; Liu et al. 2015; Wolf et al. 2015), operating as a sort of bet-hedging strategy (Chapter 22).

Part of the motivation for this argument is shown in Figure 21.4 – if the expected phenotype (in this case mean expression level) of a particular genotype is far from the optimum dictated by environmental factors, only genotypes producing sufficiently variable progeny will have some hope of gene transmission to the next generation, as the offspring of more noise-suppressed genotypes will all have near-zero fitness. Note, however, that once the mean phenotype evolves to be in accordance with the environmental optimum, the opposite occurs – all members of noise-suppressed genotypes have high fitness, whereas individuals in the tails of the distribution for highly variable genotypes will have near zero fitness. The point has been demonstrated in an experiment in yeast in which both the mean and variance of expression was altered in various alleles for a key gene (Duveau et al. 2018).

There are, however, multiple reasons for skepticism as to whether selection can modulate expression noise on a gene-by-gene basis. First, it is far from clear whether environmental optima exhibit sufficiently large fluctuations to encourage the evolution of an intermediate level of noise, and it is equally unclear whether

environments are typically constant enough to promote noise minimization. Second, any modifier for expression noise would need to be tightly linked to the modulated gene, most likely in the gene body itself, else any benefit associated with the modifier would be quickly disconnected after a few generations of recombination. Third, as outlined in detail at the end of Chapter 9, selection on the level of phenotypic variance production is a second-order effect and, at best, a very weak evolutionary force, as individual genotypes are not filtered on the basis of their own genetic merits but via the nonheritable features of their offspring. Most notably, phenotypic noise actually reduces the response to selection by diminishing the relationship between individual phenotype and the underlying genotype. Related issues on this particular topic are covered by Matsumoto et al. (2015) and Mineta et al. (2015).

Finally, and perhaps most significantly, gene-expression noise is an intrinsic function of average expression, so any selection on the former and vice versa will naturally have cascading effects on both the target locus and other cellular participants. Taking the ratio of the standard deviation in protein number relative to the mean (i.e., the coefficient of variation, or CV), it can be seen from the leading term in Equation 21.4b (which defines the square of the standard deviation), that the CV is roughly inversely proportional to the square root of the mean number of proteins. Lestas et al. (2010) showed more generally that the CV is inversely proportional to the 0.25 to 0.5 power of the number of proteins produced per mRNA. This means that reducing the variation in protein expression by 50% requires roughly a 4- to 16-fold increase in the number of proteins relative to mRNA molecules. Thus, noise suppression comes at the expense of added protein production, whereas noise enhancement will generally require a reduction in protein number, which may compromise basic aspects of cell biology.

A related issue here is that all of the results introduced above were derived under the assumption of haploidy (Foundations 21.1). Diploidy (common in eukaryotes) might further reduce the level of noise depending upon the degree to which expression is coordinated between the two copies of a gene. If allelic expression is completely independent, as some evidence suggests (Sepúlveda et al. 2016; Skinner et al. 2016), then having two genes expressed simultaneously may give a $\sim 15\%$ reduction in the level of noise. On the other hand, with their more complex modes of regulation (often involving multiple transcription factors and/or accessory proteins), eukaryotic genes may commonly exhibit a gradient of transcriptional states (rather than simple on/off) (Corrigan et al. 2016), which will further influence the noise process. These matters seem not to have entered the conversation over the many potential factors that might favor diploidy (Chapter 10).

## The Basic Biology of Transcription

Gene transcription is carried out by multi-subunit DNA-dependent RNA polymerases (Chapter 20), which we will simply call RNA polymerases. However, such complexes are generally nonautonomous in the sense that one to several accessory proteins, including TFs, must be present simultaneously for transcription activation (Jolma et al. 2015; Haberle and Stark 2018; Cramer 2019). The core promoters upon which the transcription machinery assembles typically reside within 100 or so bp of

transcription-initiation sites, whereas enhancer elements containing the TFBSs are generally located further upstream (in multicellular species sometimes up to 100,000 bp away; Hafner and Boettiger 2023). Individual TFs often 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, seven TFs control the expression of $\sim 50\%$ of regulated genes in *E. coli,* whereas $\sim 60$ TFs (about one fifth of the TFs in this species) service single genes (Martinez-Antonio and Collado-Vides 2003).

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

Eukaryotic transcription raises additional issues in that the chromosomes are regularly wrapped around nucleosomes, formed from histones, and often 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 (Chapter 10), help recruit TFs to localized regions (Yan et al. 2013).

Many dozens of TF families 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 is 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, usually near the lower end, especially in eukaryotes (Luscombe and Thornton 2002).

**A physical model for transcription-factor binding.** The universal mode of transcription, involving the interaction of a specific protein (the TF) with a specific DNA binding site (the TFBS), provides a compelling platform for developing an evolutionary theory of gene expression couched in terms of the biophysics of intermolecular associations (Bintu et al. 2005). However, because many TFs 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). Thus, 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. Empirical data from a variety of sources suggest that the average energetic cost of a single-base mismatch is $\simeq 2$ in Boltzmann units of $k_BT$ (which is $\simeq 0.6$ kcal/mol at most biological temperatures) (Table 21.1). 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 the number of matches with the optimal recognition sequence ($m \leq \ell$). 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; Shultzaberger et al. 2010), although higher-order effects involving the shape of TFBSs can also contribute to the overall binding energy (Yang et al. 2013; Le et al. 2018).

Given a potential TFBS sequence with a particular binding energy for a specific TF, we wish to know the probability of occupancy by the cognate TF, $P_{\rm on}$, as this is a minimal requirement for expression of the associated gene. Recall from Equation 21.1 that $P_{\rm on}$ can be expressed in terms of association/dissociation rates. Here we take a related but more mechanistic approach, treating $P_{\rm on}$ as a function of both the binding site and the features of the intracellular environment that restrict the site's access to cognate TFs. Clearly, $P_{\rm on}$ will increase with the number of TF molecules in the cell ($N_{\rm tf}$), but equally important are the 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 $G$ such nonspecific sites in a haploid cell (with varying degrees of affinity). Letting the excess scaled binding energy of a target TFBS be $2m$, from Foundations 21.2 the probability that a specific target TFBS is occupied by its TF is

$$P_{\rm on} \simeq \frac{1}{1 + Be^{-2m}}, \tag{21.5}$$

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. As $B \to \infty$, $P_{\rm on} \to 0$, whereas as $m \to \infty$, $P_{\rm on} \to 1$.

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 2007a). For the model bacterium *E. 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. Thus, taken together, these observations suggest a range for $B$ on the order of $10^3$ to $10^6$ for prokaryotes. For the yeast *S. 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 $\gg 10^3$.

All of 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. Thought of in a more general way, the composite parameter $B$ can be viewed as a measure of the totality of cellular features working against the binding of a TF to a specific cognate TFBS.

Equation 21.5 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 21.5). 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. This means that a considerable amount of mismatching can be tolerated for a TFBS motif more than a dozen nucleotides in length.

The results in Figure 21.5 highlight two physical constraints on the basic process of transcription regulation by the binding of TFs to DNA. First, because mismatches come in approximately discrete packets (with relative binding energy $\sim 2$/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 (for example, from not all mismatches having exactly the same consequences) will provide some flexibility. Modulation of gene expression might also be accomplished by altering the numbers of TFs residing inside cells (which will influence $B$). However, a secondary consequence of altering the concentration of a TF is that different client genes will also be affected.

**Table 21.1.** Features of the motifs of well-studied transcription factors. Motif sizes are 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 associated with background thermal motion; 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) |

Second, life's transcription mechanism comes at a significant energetic cost, in that to ensure that a particular gene is turned on, a substantial excess number of

TF molecules must be produced to compensate for the unproductive engagements occurring at nonspecific sites. For example, rearrangement of Equation 21.5 shows that for a TFBS motif with $m = 8$ matching bases to achieve a 0.9 probability of being bound to its cognate TF, 10 and 1000 TF molecules are required in cells with genome sizes of $10^7$ and $10^9$ bp, respectively, and $P_{\text{on}} = 0.99$ elevates these numbers to $\sim 110$ and $11,000$. Thus, an unavoidable consequence of biology's mode of gene expression is that far more TFs must be produced than the numbers of genes to be serviced. This cost of living with a system that relies on mRNA production for gene expression necessarily increases in eukaryotic cells with larger genomes.

**Encounter rates between TFs and their binding sites.** The preceding analyses implicitly assume that the distribution of TFs within a cell is typically in a dynamic steady state of bound and unbound molecules. 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 Foundations 21.3, the biophysical properties of cells are such that localization can generally be achieved in a matter of a few seconds or less in bacterial cells.

Despite their passive transport, TFs locate their target sites at rates exceeding the three-dimensional diffusion limit (Riggs et al. 1970), an observation that motivated the facilitated-diffusion model (von Hippel and Berg 1989). Given the minute sizes of individual TFBSs, a newly arisen TF molecule will essentially always first encounter a nonspecific site on a chromosome before locating a proper, more energetically favorable target. 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 if a nondirected one-dimensional diffusion process followed first contact. Finally, it appears that the search for appropriate DNA-binding sites is facilitated by protein-protein interactions within clusters of transcription factors, not just by direct DNA-binding processes (Brodsky et al. 2020). Reviews covering many of the technical issues can be found in Gowers et al. (2005), Halford and Marko (2004), Halford (2009), Kolomeisky (2011), Zhou (2011), Normanno et al. (2012), and Staller (2022).

The extent to which various species alter the spatial configurations of their chromosomal DNA to further assist in the search process remains unclear. However, the spatial issues incurred by the large cells of eukaryotes are of particular interest. The volumes of the nuclei of eukaryotic cells are typically larger than entire prokaryotic cells, and this can result in mean search times of individual TF molecules for a target TFBS of 1 to 200 minutes within the nuclear environment alone (Foundations 21.3). Although the overall search process can be sped up by producing more TF molecules, there is the additional issue of the cytoplasmic cell volume (within which TFs arise by translation), 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 10 to 100× were the genome not concentrated within a nuclear envelope. Thus, 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 (Chapter 15), the challenges of gene expression should be included in this list. The rate of gene expression in large cells might be extremely compromised if the genome were not confined to the restricted space of the nucleus.

### Coevolution of Transcription Factors and Their Binding Sites

To be expressed, essentially every gene in every genome requires interaction with at least one TF. This implies that the TF mode of gene regulation must have been present in LUCA. Nonetheless, 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. 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 closer to linear (Reichmann et al. 2000; van Nimwegen 2003; Aravind et al. 2005; Iyer et al. 2008; Charoensawan et al. 2010).

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, changes in the regulatory vocabulary and the reading machinery have evolved on various time scales. Only 2% of specific DNA-binding domain families are shared across bacteria, archaea, and eukaryotes, and no clear TF orthologs are known across these three superkingdoms (Charoensawan et al. 2010). Dramatic differences appear among the major eukaryotic lineages as well (Reichmann et al. 2000). These observations alone suggest 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, especially among those doing comparative developmental biology in animals, that eukaryotic morphological diversity has been driven by the exploitation of novel TF families and their recruitment to specific sets of genes. However, although there is no question that developmental evolution must involve modifications in gene regulation, it does not follow that the origin of multicellular complexity is an inevitable outcome of transcriptional complexity. As noted above, eukaryotes do not generally invest proportionately more in their TF repertoires at the genomic level than do prokaryotes. Moreover, many of the key TFs deployed in complex development are present in the unicellular relatives of animals and land plants (de Mendoza et al. 2013; Richter et al. 2018).

Another common argument is that most changes in gene regulation are a consequence of alterations in the *cis*-regulatory logic residing upstream of genes rather than a result of modifications in the agents of transcription, 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 assertion is that because individual TFs often service multiple genes, alterations of binding-site specificities of TFs are likely to have large-scale, negative pleiotropic consequences for fitness. Under this view, a change in expression pattern with minimal pleiotropic effects on multiple traits can only be achieved by recruiting, modifying, or eliminating TFBSs on a gene-by-gene basis. However, not all muta-

tions arising in a gene with pleiotropic effects need themselves be pleiotropic, and as discussed further below, considerable evidence suggests that functional changes in TFs often have minimal side consequences (Hsia and McGinnis 2003; Lynch and Wagner 2008; Wagner and Lynch 2008). Moreover, the target size for *trans*-acting mutations can be hundreds of times larger than that for *cis*-acting mutations (Gruber et al. 2012; Metzger et al. 2016), meaning that by sheer numerical dominance, such mutations can be quantitatively important.

**General observations.** Although high-throughput methodologies for genome-wide identification of TFs and their corresponding TFBSs promise to substantially expand our understanding of how such systems diversify (e.g., Berger and Bulyk 2009; Carey et al. 2012; Furey 2012; Ding et al. 2013; Smith et al. 2013; Levo and Segal 2014; Hill et al. 2021), most current insight into the mechanisms of gene-regulatory evolution still derives from observations from the usual key model systems – the bacterium *E. coli,* the yeast *S. cerevisiae,* the fly *D. melanogaster,* mouse, and human. From this limited set of taxa, several generalizations have started to emerge, imposing a need for evolutionary explanation.

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 recognized by one of the two members of a homodimeric TF.

Second, the evolutionary features of TFs appear to depend on the number of host genes serviced. For example, from comparisons of multiple gammaproteobacteria, Rajewsky et al. (2002) found that TFs with larger numbers of target genes are more evolutionarily conserved at the amino-acid sequence level and with respect to the TFBS recognition sequence. Nevertheless, Sengupta et al. (2002) observed a decline in binding-site specificity with increasing numbers of genes serviced by a TF in both *E. coli* and yeast. In principle, the latter condition may evolve so as to minimize the mutational burden on an organism, as a large number of TFBSs increases the overall mutational target size. However, an alternative explanation is that TFs with low specificity are recruited more frequently into various regulatory pathways over evolutionary time.

Third, in eukaryotes, multiple motifs for a particular TF are frequently present in the upstream regions of client genes (e.g., Gotea et al. 2010). Although it is commonly argued that such redundancy is maintained by natural selection, TFBS clustering can also arise naturally by small-scale duplication processes (Lusk and Eisen 2010; Nourmohammad and Lässig 2011). Thus, while the presence of multiple binding sites might help ensure that an adjacent gene will be activated, there is as yet no formal evidence that such configurations are anything more than a simple consequence of physical processes.

**Evolutionary distributions of binding-site motifs.** Transcription factors and their binding sites provide an explicit framework for evolutionary analysis, in that specific DNA-level features can be directly related to fitness (Gerland and Hwa 2002; Berg et al. 2004; Lässig 2007; Stewart et al. 2012; Lynch and Hagner 2015; Tuğrul

et al. 2015). 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 a TFBS with $m$ matching sites is expected to be bound by its cognate TF, e.g.,

$$W(m) = 1 + \alpha P_{\text{on}}(m), \tag{21.6}$$

where $\alpha$ is a scaling factor relating binding probability to fitness, and $P_{\text{on}}(m)$ is defined as Equation 21.5. As $\alpha \to 0$, $W(m) \to 1$, implying neutrality. Equation 21.6 is often referred to as a mesa fitness function, because fitness increases asymptotically from 1 to $(1 + \alpha)$ as the probability of gene activation increases from 0 to 1.

Alternative models relating $m$ to fitness are certainly possible, but once $W(m)$ is defined, and additional information is available on rates of mutational movement between alternative TFBS states, a number of basic issues regarding TFBS evolution can be examined using the methods outlined in Foundations 21.4. For example, it is well known that the genomic set of binding sites associated with a particular TF often exhibit variable motif sequences. Although such variation might partially result from selection for alternative levels of locus-specific gene expression, because of the diminishing-returns nature of the fitness function (Figure 21.5), variation in motif matching is also expected to arise naturally as selection pushes a population towards the drift barrier, where alternative high-$m$ states are selectively equivalent (Berg and von Hippel 1987).

Over evolutionary time, the frequency distribution of the number of matches in the various TFBSs serviced by a particular TF is expected to reach an equilibrium between the mutational forces causing mismatches and the selective forces favoring 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 Foundations 21.4, provided all nucleotides mutate to all others at equal rates, the equilibrium distribution takes on a simple form,

$$\widetilde{P}(m) = C \left[ \binom{\ell}{m} 3^{\ell - m} \right] e^{2N_e W(m)}, \tag{21.7}$$

where $C$ is simply a normalization constant that ensures that the full set of probabilities, $\widetilde{P}(m)$, sum to one.

The equilibrium distribution $\widetilde{P}(m)$ can be viewed as either the long-term average probability of states at a particular TFBS as it wanders through evolutionary time, or as the expected distribution of $m$ for a full set of equivalent TFBSs (for different genes within a particular host genome) at any one point in time. The exponential term in Equation 21.7 is a constant when $W(m)$ is invariant, and so the term within brackets is equivalent to the expected distribution in the absence of selection, $\widetilde{P}_n(m)$. Thus, Equation 21.7 indicates that the evolutionary distribution of binding-site matching is equal to the neutral expectation weighted by an exponential gradient of the fitness surface relative to the power of random genetic drift, $W(m)$ divided by $1/(2N_e)$.

Solution of Equation 21.7 illustrates several general principles (Figure 21.6). First, the equilibrium distribution is completely independent of the mutation rate. The factor of three enters because it is assumed that there are three ways for a matching nucleotide to mutate to a mismatch but only one way for a reversion to

arise. Because the former is simply a multiplicative function of the latter, the actual mutation rate cancels out.

Second, regardless of the set of parameter values, substantial variation in $m$ is almost always expected among sites. Unless the motif size is small (e.g., $\ell = 8$) 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 owing to the plateau of $P_{\text{on}}(m)$ at high $m$. Indeed, with a motif size of 16 bp, essentially no TFBS is expected to be perfect, unless the power of selection is unrealistically high ($N_e\alpha \geq 10^6$). The exact form of $\widetilde{P}(m)$ will vary with different forms of the fitness function, $W(m)$, but provided the upper end of $W(m)$ becomes progressively flatter, the drift barrier combined with the multiplicity of sequences with identical matching levels will encourage substantial motif variation. Thus, the theory provides an explicit nonadaptive explanation for the high level of interspecific divergence in TFBS motifs routinely seen in comparative studies as well as for the substantial variation in motif sequences at the intraspecific level (Zheng et al. 2011; Heinz et al. 2013).

Third, with relatively weak selection pressure ($N_e\alpha \ll 1$), $\widetilde{P}(m)$ is very heavily skewed towards small (but nonzero) numbers of matches (essentially the neutral expectation). This intrinsic weighting towards low numbers of matches is due to both biased mutation pressure and the high multiplicity of configurations leading to the same $m$ with increasing numbers of mismatches.

Fourth, because the neutral distribution is heavily weighted toward low $m$, there can be a sharp "phase transition" as $N_e\alpha$ crosses the threshold value of $\sim 1.0$. Notably, cases can even 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 driven by selection pressure. As the motifs within the different peaks of such distributions will deviate in both length and sequence, this result may help explain the widespread use of secondary TFBS motifs by TFs (noted above).

Fifth, although the preceding results have been derived for a interaction in which the TF is evolutionarily invariant (e.g., due to pleiotropic constraints associated with its use with other genetic substrates), when the TF coevolves with its TFBS, the overall results noted above largely remain except that the average equilibrium degree of matching declines (Lynch and Hagner 2015). This results because mutations in both the TF and the TFBS present a constant stream of changes in each other's selective landscape, in effect preventing strong specialization. A side consequence of this behavior is that when both components of a TF/TFBS system are free to evolve, the underlying recognition motif is free to explore all of sequence space, conditional on the constraint of maintaining an adequate degree of matching at all points of time. This results in the origin of incompatibilities of TF/TFBS pairs in different phylogenetic lineages as the two systems drift apart to the extent that they no longer recognize each other in heterospecific combinations.

Sixth, in the case of a one-to-many scenario in which the TF interacts with multiple TFBSs, there can be a substantial degree of asymmetry in the rates of evolution of the two components (Lynch and Hagner 2015). Owing to its need to satisfy multiple partners, the TF experiences the strongest selective constraints, with the overall rate of evolutionary divergence declining with increasing numbers

of partners. In effect, the master controlling element is no longer free to coevolve with single interacting partners, becoming increasingly constrained to accept only the small subset of mutations that is either effectively neutral for all partners or the even smaller subset with a net overall positive impact. In contrast, the TFBSs themselves continue to evolve in an essentially independent fashion, with distribution and rate features identical to what would be expected in a highly specific system, e.g., Equation 21.7. These results appear to be consistent with the observations, noted above, that TFs with larger numbers of target genes are more evolutionarily conserved at the amino-acid sequence level and evolve lower levels of binding-site specificities.

Finally, the theory helps clarify why TF motifs are typically so small. Owing to the saturating binding potential embodied in Equation 21.5, even the strongest levels of selection are unlikely to lead to mean binding-motif lengths in excess of 12 bp. Although an overly short TF recognition motif may lead to excessive spurious binding to off-target sites, the challenges here are not too severe. Assuming equal nucleotide usage, the expected number of appearances of any particular sequence of length $\ell$ in a genome containing $G$ bp is $G(1/4)^{\ell}$. Setting this equal to one, and rearranging, we find that less than one random motif is expected to be present by chance on each strand if the motif size exceeds $\ell^* = \ln(G)/\ln(4)$. For genome sizes of 10 to 1000 Mb, $\ell^* = 12$ to 15 bp. Combined, these two points provide a simple explanation for why TFBSs are generally shorter than 15 bp in length.

Although Stewart et al. (2012) have argued that TFBS evolution reflects 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), it is clear that less than maximum matching lengths arise naturally as a consequence of mutation-selection balance. Direct selection for mutational robustness, a second-order effect, need not be invoked. In addition, all of the above results indicate that, without direct empirical validation, the presence of motif variation in genomes, at both the intraspecific and interspecific levels, should not be taken as evidence of adaptive fine-tuning of individual loci.

**Application of the models.** The general model embodied in Equation 21.7 can be used for more than simply predicting the features of TF/TFBS systems. Assuming that the set of TFBSs in a particular species has evolved to its equilibrium distribution of motifs, one can compare the observed distribution of usage to the neutral expectation to estimate the strength of selection on functional binding sites, $2N_e s(m)$, necessary to account for the deviations between the two. This follows by rearranging Equation 21.8 to

$$W(m) = \left(\frac{1}{2N_e}\right) \ln\left(\frac{\widetilde{P}(m)}{\widetilde{P}_n(m)}\right). \tag{21.8}$$

where $\widetilde{P}_n(m)$ is expected distribution under neutrality, which can be obtained from random nucleotide motifs exclusive of known binding sites, i.e., as a fraction of random genomic stretches of length $\ell$ containing $m$ matches to the optimal motif. Note that this sort of application makes no assumptions about the form of the fitness function, and instead relies on the data to infer $W(m)$.

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, with the strength of negative selection declining monotonically with increasing binding affinity (Figure 21.7). An analysis of Abf1 binding sites in yeast yielded similar results (Mustonen et al. 2008), and an analysis of 12 additional transcription factors revealed a positive relationship between fitness and binding energy in each case (Haldane et al. 2014). Thus, all existing analyses appear to support the use of a fitness model that assumes a positive association with binding affinity, as in Equation 21.7. This type of analysis also harbors substantial potential for TFBS discovery in a genome using thermodynamic principles rather than consensus sequence motifs (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 21.4.1, with the entire ensemble of motifs retaining the steady-state 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 among 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 these expectations (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 drift-mutation-selection equilibrium. Due to the multiplicity of binding-site configurations deviating from the optimum, there is room for substantial sequence change via compensatory mutations.

### Evolution of Pathway Architecture

Despite the centrality of TFBSs to gene expression and the common conservation of motifs between distant lineages (Nitta et al. 2015), a diverse set of observations indicates that TFBS locations and motif sequences can vary dramatically among closely related lineages, often with no apparent phenotypic consequences (Borneman et al. 2007; Doniger and Fay 2007; 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) found 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 different 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 regulatory pathways associated with this gene were revealed through several years of study using multiple individuals from a diverse natural population, but this lack of genetic-background control likely influenced the generality of the results. For example, it was subsequently determined that the TFBSs for Endo16

and for 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). Moreover, although the expression patterns of Endo16 appear to be conserved between different sea urchin genera, 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).

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* indicate up to 5% turnover of TFBSs 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 despite the underlying changes in the regulatory regions (Ludwig et al. 2011; Paris et al. 2013).

What remains unclear is whether the observed regulatory-sequence changes in these studies are accompanied by modifications in the DNA-binding domains of the associated TFs. There are, however, clear examples of TF-associated changes in vertebrates. For example, Yokoyama and Pollack (2012) found that a single amino-acid change in the transcription factor SP1, which 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. 2011). 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 changes in TF amino-acid sequence are also known to be associated with changes in binding-site specificities in plants (Sayou et al. 2014).

As in the previous examples with invertebrates, the changes in vertebrate regulatory mechanisms again appear to often occur without noticeable affects on patterns of gene expression. For example, Fisher et al. (2006) found that the control region for the human RET receptor kinase 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. However, in contrast to the situation for messenger RNAs, neither species background is sufficient to drive expression of the ribosomal RNA genes from the other, a phenomenon known as nucleolar dominance (Arnheim 1986).

Taken together, the preceding observations suggest the common existence of evolutionary pathways whereby the underlying mechanisms of gene regulation can be altered with no apparent modification in the outward phenotype. Such regulatory repatterning provides further evidence for evolution at the cellular level by effectively neutral mechanisms, a topic that will be returned to in the final section of the chapter.

We now move on to higher-order issues, in particular with the wide diversity of topological structures of gene-regulation pathways, much of which is unexplained

from an evolutionary perspective. There is much to be considered here, including the numbers and types of steps in regulatory pathways, branching patterns, and the degree to which both pathway topologies and the individual participants remain constant over evolutionary time. Not surprisingly, considerable attention has been given to the idea that regulatory pathways are optimally structured to yield particular performance levels, response times, and stability. However, these conclusions are often reached from the starting assumption of an all-powerful hand of natural selection. It will be argued below that, as with the coevolution of transcription factors and their binding sites, numerous features of pathway-structure evolution are seemingly guided by nonadaptive mechanisms.

**Activators vs. suppressors.** Before considering the higher-order architecture of pathways, it is essential to note that although all of the preceding discussion has been focused on gene activation by TFs, in their simplest form TFs can operate as activators or repressors (Figure 21.8). In the former case, via signal transduction (Chapter 22) the gene for the TF is activated, and the TF then activates the gene of interest. This mechanism of double-positive (++) control ensures that the gene is only active in the face of appropriate demand. However, the same end-result can be obtained with double-negative (−−) control, whereby the transcription factor operates as a repressor of transcription until it is released upon receiving an appropriate signal for gene-usage demand.

In a broad study of the regulation of *E. coli* genes, Savageau (1974, 1977, 1998) found that genes whose products are needed most of the time tend to be subject to ++ regulation, whereas those that are only sporadically needed are generally under −− regulation. To explain this pattern, he proposed a "use it or lose it" hypothesis. The simple basis of this idea is that proteins not carrying out a function are subject to the neutral accumulation of degenerative mutations during such periods of activity. Under this view, an activator TF that is rarely used will be subject to a high rate of pseudogenization, whereas a repressor TF used in this context will only rarely be subject to deleterious-mutation accumulation. In contrast, for a gene whose products are in high and frequent demand, a repressor TF would be unutilized most of the time, and hence subject to degradation. Thus, under this hypothesis, there is a selective premium on the mode of regulation that involves a regulatory protein that is kept at the highest level of utilization. Supplementing this mutation-load argument is the idea that bound transcription factors reduce the likelihood of inadvertent transcription owing to nonspecific binding by other TFs, which amounts to an error-minimization scenario (Shinar et al. 2006).

As pointed out by Gerland and Hwa (2009), the validity of these genetic load arguments depends on the population-genetic environment and the timescale of environmental shifts. They argue that the "use it or lose it" principle is most likely to hold if populations are sufficiently small that conditionally deleterious mutations can drift to high frequency during periods of inactivity. At sufficiently large population sizes, deleterious mutations may rarely have time to rise to high frequencies between bouts of use/nonuse (from Chapter 4, the time to fix a neutral mutation is approximately equal to twice the effective population size). Arguing that all deleterious mutations accumulating for an inactive TF gene will be immediately purged from the population upon demand for the gene product, they invoke a "wear and

tear" principle, whereby the least-used regulator can actually incur a slightly lower long-term average mutation load. The issues are a bit subtle here, but the essential point is that the fitness difference between alternative modes of regulation under this model is less than the mutation rate. This actually makes it highly unlikely that a domain in which the least-used mechanism will be most advantageous will ever be entered, as the mutation rate is weaker than the power of random genetic drift (Chapter 4). Thus, Savageau's hypothesis appears to be quite robust, and is well worth exploring in future studies with other organisms, especially given that it draws support from observations on a high-$N_e$ species, *E. coli.*

**Regulatory rewiring.** As outlined above, there are numerous examples in which the regulatory motifs associated with specific traits vary among species. However, this only touches the surface of the known ways in which regulatory mechanisms change over evolutionary time. Given the large numbers of transcription factors in most cells and their reliance on simple binding sites subject to stochastic mutational turnover, there are many plausible mechanisms for the emergence of novel intracellular transactions by effectively neutral processes (Johnson and Porter 2000; Force et al. 2005; Haag and Molla 2005; Lynch 2007b).

Several well-dissected examples demonstrate the complete rewiring of TF/TFBS associations, mostly in the yeast *S. cerevisiae*, where the study of gene regulation has been especially intense. Such studies strongly support the counterintuitive idea that, even when under strong selection to maintain a stable phenotype, complex regulatory systems are subject to substantial modifications in their underlying structure.

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*. Indeed, 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 occur as well. Some of this rewiring appears to be associated with whole-genome duplication known to have occurred in ancestral *S. cerevisiae* (Wolfe and Shields 1997). For example, an activator and repressor that control ribosomal-protein expression in normal and stress conditions in *S. cerevisiae* are actually subfunctionalized duplicates of an ancestral gene inferred to have had both functions. Moreover, the various TFs involved have associations with novel functions in one or both species, showing expanded/contracted assignments.

Expanding on this theme, 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 regulatory circuitry is completely different. Based on phylogenetic analysis, the ancestral species appears to have had shared (and perhaps redundant) regulatory motifs, with each of the two descendent lineages 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). Even the regulatory mechanisms for the expression of histone proteins, one of the most evolutionary conserved sets of proteins across eukaryotes, are dramatically altered across yeast species, both in terms of the TFs deployed and their binding motifs (Mariño-Ramírez et al. 2006)

The regulatory wiring for the mating-type locus is also dramatically changed

in yeast (Baker et al. 2011, 2012; Sorrells et al. 2015; Britton et al. 2020). Two mating-type cells exist in these species, *a* and $\alpha$ (Chapter 10). In *C. albicans* and other basal yeast lineages, *a*-specific genes are activated by a regulatory protein only present in *a* cells, which keeps the *a*-specific genes off in $\alpha$ cells without any investment in transcription factors. This mode of regulation in $\alpha$ cells diverged in *S. cerevisiae,* where the *a*-specific genes remain constitutively active in *a* cells, as in *C. albicans*, but are kept silent in $\alpha$ cells by a specific repressor protein (i.e., requiring an added investment in gene regulation in such cells). The alterations responsible for these differences again appear to have arisen from an intermediate ancestral state in which two sets of regulation were used simultaneously, and then divergently resolved in descendent lineages.

Many additional examples of regulatory rewiring have been uncovered in comparative analyses of the gene modules of *S. cerevisiae* and *C. albicans* (Tuch et al. 2008; Sarda and Hannenhalli 2015; Nocedal et al. 2017), but these kinds of observations are by no means restricted to yeasts. For example, a number of studies have suggested substantial regulatory rewiring among bacterial 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).

A remarkable feature of all of the above examples of the evolution of different control mechanisms is that they involve 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 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 21.9). 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 TFBS motif in each lineage. Such a condition would lead to relaxed selection on bispecificity, with the TF then being free to lose a complementary motif in each lineage. The net effect of such a scenario would be the continued use of the same TF, but a change in the underlying regulatory language.

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 presence in just a single copy per genome, the recognition motif of this TF differs substantially between mosses and the 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 the two motifs relied upon by other land plants (Sayou et al. 2014). This reciprocal focusing of a bispecific ancestral TF may 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; Jolma et al. 2013; Franco-Zorrilla et al. 2014; Morgunova et al. 2018).

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 again remain the same, although the regulatory network will have become more complex, owing to the specialization of the individual TFs. Analyses in the nematode *C. elegans* (Reece-Hoyes et al. 2013) and the budding yeast *S. cerevisiae* (Pougach et al. 2014) provide considerable support for this model of regulatory rewiring.

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 might become superfluous and subject to loss by mutational degeneration. Under this scenario, a coordinated shift in the entire regulatory mechanism might be achieved by multiple loci, as the initiating event will have been experienced simultaneously by each of the relevant regulatory regions owing to their shared reliance on the first TF.

These kinds of observations have profound implications for how we study biology, the obvious concern being that the molecular details deciphered for the regulatory pathway in one model system need not be relevant to that operating in other species. Yet, almost all molecular, cellular, and developmental biologists eschew intraspecific variation, concentrating instead on typological characterizations of a few model species. Indeed, it has become increasingly common for laboratories in these fields to focus research on just a single strain of a single species, sometimes for decades. The resultant exquisite, painstaking research has led to remarkable advances in our understanding of the details of subcellular mechanisms, but what is the generality of such findings?

Because virtually every complex trait in every species exhibits significant genetic variation (Lynch and Walsh 1998), it is likely that many text-book examples of regulatory pathways derived from single clones or inbred lines are quite unrepresentative of the operational features of related phylogenetic lineages, and some may be positively misleading. Notably, even in the relatively simple bacterium *E. coli,* when a TF is duplicated and one copy is then engineered to have a nonorthologous regulatory region, there are no notable changes in organismal fitness (Isalan et al. 2008), suggesting a high degree of evolutionary flexibility of regulatory systems. Among those with interests in multicellular organisms, these kinds of observations have motivated interest in the process of "developmental system drift," whereby seemingly similar morphological structures in closely related species are achieved by substantially different regulatory mechanisms (Johnson and Porter 2000, 2001, 2007); Weiss and Fullerton 2000; True and Haag 2001; Ruvinsky and Ruvkun 2005; Force et al. 2005; Haag and Molla 2005; Tsong et al. 2006; Lynch 2007b; Pavlicev and Wagner 2012; Sommer 2012; Metzger and Wittkopp 2019).

**Network topology.** As with all other genes, TF expression is often regulated by other TFs, whose control may ultimately be dictated by signal transduction pathways induced by internal or external chemical signals (Chapter 22). Combined with the fact that TFs can operate as either enhancers or repressors, this opens up the possibility of multiple architectures for gene regulatory networks. For example, the joint operation of just two genes can be governed by six different topologies (Figure 21.10a). A common form of network called the feed-forward loop involves just three genes (two TFs, with one regulating the second, and both regulating a

third target gene), but even then still has eight possible topologies not including self-regulatory loops, and expanding to 64 possibilities if the latter are included (Figure 21.10b). Such loops are said to be coherent if the direct effect of the first TF is the same as its indirect effect through the second TF; otherwise, the loop is said to be incoherent.

Many regulatory pathways are much more baroque in form than those just noted (Wilkins 2002, 2005; Lynch 2007b). For example, it is common for linear pathways to consist of a series of genes whose products are essential to the activation/deactivation of the next downstream member, with only the expression of the final component in the series being the ultimate determinant of the phenotype. For example, the product of gene D may be necessary to turn on gene C, whose product is necessary to turn on gene B, whose product finally turns on gene A. Pathways involving only inhibitory steps also exist, and these lead to an alternating series of high and low expression, depending on the state of the first gene in the pathway. For example, gene D may generate a product that inhibits the expression of gene C, whose silence allows gene B to be turned on, which inhibits the expression of gene A. It is often unclear that such complexity has any advantages over simpler two-gene pathways or even self-regulation.

The mechanisms by which genetic networks become established evolutionarily are far from clear. Many physicists, engineers, computer scientists, and cell and developmental biologists are convinced that biological networks are endowed with features that confer emergent properties' that ostensibly could only be products of natural selection (Gerhart and Kirschner 1997; Shen-Orr et al. 2002; Milo et al. 2002; Barabási and Oltvai 2004; Alon 2006, 2007; Babu et al. 2006; Balaji et al. 2006; Davidson 2006; Tagkopoulos et al. 2008; Burda et al. 2011; Hong et al. 2018; Zitnik et al. 2019). Five popular concepts in biology today – redundancy, robustness, modularity, complexity, and evolvability – invoke a vision of the cell as an electronic circuit, designed by and for adaptation. However, the physical and genetic mechanisms giving give rise to genome architectural features are logically distinct from the adaptive processes utilizing such features as evolutionary resources (Lynch 2007b). Theoretical investigations of network evolution have only rarely examined these matters in the context of well-established evolutionary principles.

Qualitative observations suggest that the complexity of regulatory networks increases from prokaryotes to unicellular eukaryotes to multicellular eukaryotes, with simple autoregulatory loops being more common and multi-component loops less common in microbes (Thieffry et al. 1998; Lee et al. 2002; Wuchty and Almaas 2005; Sellerio et al. 2009). However, it is an open question as to whether complex pathway architectures are a necessary prerequisite for the evolution of complex phenotypes or whether the genome architectures of multicellular species are simply more conducive to the emergence of network connections owing to the elevated power of random genetic drift. The possibility that network-topology evolution is driven by the kinds of nonadaptive processes that generate changes in network participants (noted in the preceding section) clearly merits consideration.

The following arguments illustrate the ease with which commonly observed features of genetic pathways can emerge without any direct selection for such properties. In principle, pathway augmentation may be driven entirely by the nonadaptive processes of duplication, degeneration, and random genetic drift. Consider the series

of events in Figure 21.11. Initially, a single gene A carries out some function in a constitutive fashion, but in a series of steps, it becomes completely reliant on an upstream activation factor B. A scenario like this could unfold in the following way. First, gene A becomes sensitive to activation by gene B, either because gene A has acquired a *cis* modification that causes activation by B, or because some transcription factor B acquires a mutation that causes it to serve as a *trans* activator of A. At this point, gene A has redundant activation pathways, and is therefore vulnerable to loss of one of them. Should a degenerative mutation cause gene A to lose the ability to self-regulate, gene B will have been established as an essential activator. This process can be repeated anew as gene B acquires sensitivity to a further upstream TF and loses the ability to constitutively express.

The probability of establishment of these types of changes is expected to depend on the effective population size ($N_e$). This is because a redundantly regulated allele has a weak mutational advantage equal to the rate of loss of a regulatory site ($u_l$) – one such mutation will result in the nonfunctionalization of either a self-regulated or an upstream-dependent allele, but will leave the function of a redundantly regulated allele unaltered. If $N_e \ll 1/u_l$, such an advantage will be impervious to selection, and the population will evolve to an allelic state that simply depends on the relative rates of gain and loss of regulatory sites ($u_g$ and $u_l$ in Figure 20.10), eventually leading to the establishment of an obligatory pathway. In contrast, if $N_e \gg 1/u_l$, the accumulation of upstream-dependent alleles will be inhibited by their weak mutational burden, as well as by the additional energetic burden (imposed by the expense of an additional pathway component). Although these arguments demonstrate that small population size provides a permissive environment for the emergence of complex genetic networks, without any direct selection for complexity, this does not mean that such alterations cannot occur in very large populations. However, if such changes are to occur in a large $N_e$ context, they must have substantial enough additional advantages to offset the mutational and energetic burden of gene-structural complexity.

As discussed in detail in previous chapters, these simple arguments show how the relative power of the nonadaptive forces of evolution – genetic drift, mutation, and recombination – define the trajectories open to evolutionary exploitation. Although the incorporation of more technical details is needed, previous conclusions on the adaptive basis for the evolution of network topologies that rely on models devoid of population-genetic details should be interpreted with caution (Wagner 2005). Failure to reject a neutral hypothesis is not equivalent to ruling out selection as a governing force. However, the demonstration that the emergence of redundantly regulated genetic pathways is a function of population size and patterns of mutational bias raises doubts about the justification for the search for universal adaptive explanations for the evolution of genetic redundancy and robustness. For similar reasons, the conclusion that convergent evolution of network architectures in distantly related microbes provides compelling evidence for "optimal design" (Conant and Wagner 2003) also appears to be questionable.

As a more explicit example of the issues, one of the most common pathways in bacteria, the coherent feed-forward loop (Figure 21.11) is, in fact, not particularly stable across phylogenetic lineages (Tsoy et al. 2012), and the case has been made that the relative utilization of alternative topologies in bacteria may largely

be a consequence of random patterns of mutational loss and gain of substitutable links (Cordero and Hogeweg 2006; Lynch 2007b; Solé and Valverde 2008; Ruths and Nakhleh 2013). Finally, the ultimate output of a regulatory pathway is dictated not just by its topological form, but by the numerous dynamical properties (e.g., kinetic coefficients) and abundances of its participants that can override any supposed effects of topological structure (Ingram et al. 2006), leading to still more degrees of freedom for pathway rewiring. Although greatly simplified for presentational purposes, the verbal arguments presented here provide the seeds for the development of biologically realistic models for the origins of pathway complexity, which may prove useful in future attempts to infer vs. reject the adaptive significance of such features.

**Summary**

- A defining feature of all gene expression in all organisms is the production of RNA products from DNA templates, activated by various proteins called transcription factors (TFs) and carried out by RNA polymerases.

- Low rates of transcript production combined with high rates of degradation typically result in average steady-state mean numbers of transcripts per cell on the order of 20 or fewer, and bursty transcription generally results in a high level of noise in the number of transcripts per cell.

- As multiple proteins are commonly produced per mRNA and have elevated half-lives relative to the latter, the number of proteins per expressed gene within a cell typically vastly exceeds the number of mRNAs, and the noise level is reduced accordingly.

- Transcription factors link to their binding sites (TFBSs) in ways that can be described by a simple biophysical model, which demonstrates that little is gained in terms of affinity for binding motifs longer than $\sim 12$ nucleotides. The search for target genes is facilitated by the diffusion of TFs over the DNA with periodic jumps from one chromosomal location to another, and the search space in eukaryotes is reduced further by confining the genome to the nucleus.

- Despite these facilitating features, the time for a TF to locate a specific TFBS can be on the order of minutes to hours. Thus, a basic cost of the life's mechanism of gene expression is the necessity of producing substantially more copies of TFs than numbers of genes served in order to ensure a high probability of TFBS binding.

- Although the activation of gene expression by TFs must date to LUCA, there is a

remarkable void of obvious TF orthologs between bacteria, archaea, and eukaryotes. In contrast, within eukaryotes, many of the TFs known to be associated with complex development in animals and land plants are also present in basal unicellular lineages.

- Changes in gene expression evolve through gene-specific modifications of TFBS motifs and/or shifts in the binding affinities and expression patterns of TFs. Nonetheless, many cases are known in which TFBS motifs wander among phylogenetic lineages, while continuity in gene expression is maintained. Such variation is particularly likely when a TF regulates only a few genes, as the TF and its binding sites are relatively susceptible to a coevolutionary dance so long as their mutual compatibility is maintained. In contrast, a TF with a large number of client genes can become frozen in time, as a slight improvement in the binding affinity to one gene may disrupt that for many others.

- TF-systems provide the substrate for the development of a mechanistic evolutionary model directly linking genotype (number of nucleotides in a TFBS matching the optimal TF motif) to phenotype (expression level of the client gene) to fitness. This model predicts the existence of substantial variation of binding-site matching under a wide variety of conditions, especially when population sizes are relatively small.

- Transcription factors can operate as either activators or repressors of expression of client genes. Consistent with a "use it or lose it" hypothesis, the mode of regulation exploited by individual genes is generally the one that keeps the TF at the highest level of utilization, e.g., activation when the client gene is used frequently, and repression when client-gene demands are low.

- A common feature of regulatory-pathway evolution is stasis of performance in the face of substantial regulatory rewiring in different phylogenetic lineages, in some cases to the point of using entirely different TFs to carry out the same tasks. Such cases of regulatory-system drift provide compelling examples of effectively neutral evolution at the subcellular level.

- A multitude of regulatory-pathway topologies exists among genes and organisms, with those in multicellular species being particularly elongated in structure, suggesting a syndrome of overdesign. In addition, the most common topologies in bacteria are often explainable with models invoking random gains and losses of links. Both kinds of observations raise questions about the common assertion that regulatory pathways are optimally designed to minimize expression noise and to maximize robustness and the capacity for future evolvability.

**Foundations 21.1. Numbers of transcripts per cell.** The rate 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 constitutively expressed gene, with a constant rate of production of new transcripts $k_{\mathrm{m}}$, with a rate of decay per transcript $\delta_{\mathrm{m}}$. With constant rates, regardless of the starting conditions, the stochastic probability distribution of the number of mRNA molecules per cell, $p(n_m)$, will eventually reach a steady state. At this point, the flux rate between the $n_m = 0$ and $n_m = 1$ states must be equal in both directions,

$$k_{\mathrm{m}} p(0) = \delta_{\mathrm{m}} p(1), \tag{21.1.1a}$$

so

$$p(1) = p(0)(k_{\mathrm{m}}/\delta_{\mathrm{m}}). \tag{21.1.1b}$$

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

$$(k_{\mathrm{m}} + \delta_{\mathrm{m}})p(1) = k_{\mathrm{m}} p(0) + 2\delta_{\mathrm{m}} p(2). \tag{21.1.2a}$$

After subtracting Equation 21.1.1a and rearranging,

$$p(2) = p(0)(k_{\mathrm{m}}/\delta_{\mathrm{m}})^2/2. \tag{21.1.2b}$$

This approach generalizes to

$$p(n_m) = p(0)(k_{\mathrm{m}}/\delta_{\mathrm{m}})^{n_m}/n_m!, \tag{21.1.3}$$

where $n_m! = n_m(n_m - 1)(n_m - 2)\cdots 1$ is the factorial function.

To complete the solution, we require an expression for $p(0)$. Because the sum of the entire probability distribution, $p(n_m)$, is constrained to equal 1.0, it follows that $p(0)$ must equal a constant that ensures such equality. Noting that an exponential function can be written as the series expansion,

$$e^x = \sum_{n_m=0}^{\infty} \frac{x^{n_m}}{n_m!}, \tag{21.1.4a}$$

which rearranges to

$$1 = e^{-x} \sum_{n_m=0}^{\infty} \frac{x^{n_m}}{n_m!}, \tag{21.1.4b}$$

inspection of Equation 21.1.3, and substituting $x = k_{\mathrm{m}}/\delta_{\mathrm{m}}$, implies

$$p(0) = e^{-k_{\mathrm{m}}/\delta_{\mathrm{m}}}, \tag{21.1.5a}$$

and more generally,

$$p(n_m) = (k_{\mathrm{m}}/\delta_{\mathrm{m}})^{n_m} e^{-k_{\mathrm{m}}/\delta_{\mathrm{m}}}/n_m!. \tag{21.1.5b}$$

This is the well-known Poisson distribution, which is a function of a single parameter (in this case $k_{\mathrm{m}}/\delta_{\mathrm{m}}$), which in turn is equal to both the mean and the variance. Thus, under a model of constitutive gene expression, the mean number of transcripts per cell is simply equal to the ratio of the rates of production and elimination, $k_{\mathrm{m}}/\delta_{\mathrm{m}}$.

Under more complex scenarios of gene regulation, the distribution of the number of transcripts per cell deviates from the Poisson, and needs to be evaluated by more

complex methods (Thattai and van Oudenaarden 2001; Phillips et al. 2012). A solution for the two-state model in which the gene is turned on with some probability $P_{\rm on}$ was derived by Peccoud and Ycart (1995) and has the respective mean and variance

$$\mu_{\rm m} = \frac{P_{\rm on} k_{\rm m}}{\delta_{\rm m}} \tag{21.1.6a}$$

$$\sigma_{\rm m}^2 = \mu_{\rm m} \left( 1 + \frac{(1 - P_{\rm on}) k_{\rm m}}{k_{\rm on} + k_{\rm off} + \delta_{\rm m}} \right), \tag{21.1.6b}$$

where $k_{\rm on}$ and $k_{\rm off}$ 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

$$p(n_m) = p^*(n_m) \cdot \frac{\delta(k'_{\rm on} + n_{\rm m}) \delta(k'_{\rm on} + k'_{\rm off})}{\delta(k'_{\rm on} + k'_{\rm off} + n_{\rm m}) \delta(k'_{\rm on})} \cdot {}_1F_1[k'_{\rm off}, (k'_{\rm on} + k'_{\rm off} + n_{\rm m}); k_{\rm m}/\delta_{\rm m}], \tag{21.1.7}$$

where $p^*(n_m)$ is the Poisson distribution defined in Equation 21.1.5b, $k'_{\rm on} = k_{\rm on}/\delta_{\rm m}$, and $k'_{\rm off} = k_{\rm off}/\delta_{\rm m}$. Here, $\delta(\cdots)$ 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).

Although this model is agnostic with respect to the mechanisms turning a gene on and off, 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,

$$P_{\rm on} = \frac{k_{\rm on}}{k_{\rm on} + k_{\rm off}}. \tag{21.1.8}$$

Alternative models that allow for the on/off rates being dependent of the state of the DNA, and/or influenced by the presence of cooperative factors, chromatin remodeling, etc., can be found in Phillips et al. (2012), Hammar et al. (2014), Corrigan et al. 2016, Sevier al. (2016), and Skinner et al. (2016).

In Foundations 21.2, a more mechanistic description of $P_{\rm on}$ is provided in terms of transcription-factor binding. The central point 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 when $P_{\rm on} < 1$ relative to the case of constitutive gene expression (as can be seen from the degree to which the variance of $n_m$ exceeds the mean).

---

**Foundations 21.2. Occupancy probability for a transcription-factor binding site.** Because gene activation requires that 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, e.g., belonging to other client genes. In addition, we must account for the possibility of erroneous binding to illegitimate sites in the genome. Although such nonspecific binding is expected to be weak on a per-site basis, because each nucleotide site can serve as an initiation site for binding, the number of such sites is enormous, being close to the total number of bases in the genome ($G$).

Letting $N_{\rm tf}$ be the number of cognate TF molecules 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 standard approach from statistical mechanics, evaluating the relative likelihoods of all possible ways in which $N_{\rm tf}$ TF molecules can be distributed within a cell (Bintu et al. 2005; Phillips et al. 2012). Here, we assume that essentially 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 it follows that the probability of a particular TFBS being occupied is simply

$$P_{\rm on} = \frac{Z_{\rm on}}{Z_{\rm on} + Z_{\rm off}} = \frac{1}{1 + (Z_{\rm off}/Z_{\rm on})}. \tag{21.2.1}$$

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, 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; here, 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.

Each of these multiplicities represents the potential for a particular configuration of TF locations within a cell. However, after such enumeration, all of the 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 (Chapter 7). 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 denoted by more negative $E$, the weights increase with the magnitude of binding strength to cognate sites relative to background expectations.

With this substantial amount of bookkeeping in place, 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 $\beta = 1/(K_B T)$, and let $\Delta E_t = E_t - E_{ns}$ and $\Delta E_{ot} = E_{ot} - E_{ns}$ denote the differences in binding energies of target and off-target sites from background levels. Summing up, some rather complex looking expressions arise,

$$Z_{off} \simeq \frac{G! N_{ot}! e^{-\beta N_{tf} E_{ns}}}{(G - N_{tf})!(N_{tf} - N_{ot})! N_{tf}^{N_{ot}}} \sum_{i=0}^{N_{ot}} \frac{e^{-i\beta N_{tf} \Delta E_{ot}}}{(N_{ot} - i)! i! (G/N_{tf})^i} \tag{21.2.2a}$$

$$Z_{on} \simeq \frac{G! N_{ot}! e^{-\beta N_{tf} E_{ns}} e^{-\beta \Delta E_t}}{(G - N_{tf} + 1)!(N_{tf} - N_{ot} - 1)! N_{tf}^{N_{ot}}} \sum_{i=0}^{N_{ot}} \frac{e^{-i\beta N_{tf} \Delta E_{ot}}}{(N_{ot} - i)! i! (G/N_{tf})^i} \tag{21.2.2b}$$

Noting, however, that the summations to the right of Equations 21.2.2a,b are identical, and that several of the components on the left are identical or very similar as well, substitution into Equation 21.2.1 leads to great simplification,

$$P_{on} = \frac{1}{1 + [G/(N_{tf} - N_{ot})] e^{\beta \Delta E_t}}. \tag{21.2.3}$$

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), $P_{on} \to 1$. Second, the probability of binding at the site declines with increasing concentration of nonspecific sites ($G$) relative to the number of transcription factor molecules available to the site, $(N_{tf} - N_{ot})$. 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 ($N_{ot}$).

---

**Foundations 21.3. TFBS localization.** Gene regulation requires that TFs navigate from their point of production by ribosomes to the genomic location of their cognate TFBSs. Such encounters are established through semi-random diffusive molecular motions, i.e., without the involvement of any directed guidance from specific transport mechanisms such as motor proteins. 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 principles underlying the formulae to be used have been described in Chapter 7.

Transcription factors have an inherent tendency to bind nonspecifically to DNA. Thus, 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 genomic site. 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 after $t$ time units the average distance of a particle from its starting point in a one-dimensional diffusion process (and ignoring any boundary conditions) is

$$\overline{d} = \sqrt{2 D_1 t}, \tag{21.3.1}$$

with $D_1$ being the one-dimensional diffusion coefficient (with units equal the squared distance per time).

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 in the preceding step. The *average* location of a molecule always remains at its starting position, with the probability distribution simply broadening, equally to the left and right with time. Because in the absence of any directional bias to movement, the particle will always reside to the left and right of the starting point with equal probabilities, the quantity $\overline{d}$ is generally referred to as the root mean square distance.

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? With the initial random TF position being being half a genome away from the site (with $G$ being the genome size in bp), and the TFBS being potentially on either strand, 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 21.3.1 that yields $\overline{d} = G$. Several studies suggest an average $D_1 \simeq 0.5 \times 10^6$ 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 21.3.1 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 expectation, but even with 1000 TF 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. (This search space could be reduced substantially for a target gene that is contiguous to the location of the TF gene).

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

$$k_e = 4\pi(D_{3\rm n} + D_{3\rm p})(r_{\rm n} + r_{\rm p}), \tag{21.3.2}$$

where $D_{3\rm n}$ and $D_{3\rm p}$ are, respectively, the diffusion coefficients for the nucleic acid (TFBS) and the protein (TF), and $r_{\rm n}$ and $r_{\rm p}$ are their effective radii (Foundations 18.2). This equation assumes that an effective encounter occurs when the centers of the TF and TFBS fall within total distance $r_{\rm n} + r_{\rm p}$ 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 $D_{3\rm n} \simeq 0$. Experimental estimates for proteins in *E. coli* suggest that $D_{3\rm p} \simeq 3.5\ \mu$m$^2$/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 $r_{\rm n} = 0.5 \times 20 \times 0.34 \times 10^{-3} = 0.0034\,\mu$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 these estimates into Equation 21.3.2 yields an estimated encounter rate of $0.4\,\mu$m$^3$/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, and we must also compute the number of times the TF must jump from the DNA to a new location prior to encountering its proper target. The volume of an *E. coli* cell is $\simeq 1\,\mu$m$^3$, and so with one TF molecule in search of $\sim 10^7$ 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. The average time for a jump between chromosomal locations is

the reciprocal of this quantity, $2.5 \times 10^{-7}$ 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, rather than jumping from spot to spot. Thus, because approximately $10^5$ 100-bp scans are required to cover the entire genome, the estimated time to locate a site is $10^5 \times 0.0026 = 260$ 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 $0.1\,\mu{\rm m}^3$, which would reduce the search time further by a factor of ten, and few have volumes exceeding $100\,\mu{\rm m}^3$, 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 $D_{3p} \simeq 18\,\mu{\rm m}^2/{\rm sec}$ 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 $10^4\,\mu{\rm m}^3$ (Chapter 15). However, the concentration of DNA within nuclei appears to be higher than that within prokaryotic cells – averaging $57 \times 10^6$ bp/$\mu{\rm m}^3$ in root-tip cells of land plants (Fujimoto et al. 2005), and $189 \times 10^6$ bp/$\mu{\rm m}^3$ 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, once within the nucleus, a TF will encounter DNA at a rate on the order of $0.5 \times 4.5 \times 25 = 56$ times faster than the rate calculated above for *E. coli.* Estimates for the one-dimensional diffusion parameters do not appear to be available for eukaryotes. However, assuming that 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 $600\times$ 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, ensuring that a newly translated TF has a starting point close to its final destination (Kolesov et al. 2007).

The preceding calculations for eukaryotes ignore the additional problem of a cytoplasmically translated TF finding its way to the nucleus. An estimate of this time can be obtained by referring back to Equation 21.3.2, which defines the expected encounter rate between two diffusing particles. Recalling the range of nuclear volumes cited above and assuming a spherical shape, the radii of nuclei commonly fall in the range of $r_{\rm n} = 3$ to 13 $\mu$m (for nuclear volumes in the range of 100 to $10^4\,\mu{\rm m}^3$, respectively). As this is far larger than the size of proteins, $r_{\rm p}$, the latter can be ignored. From the standpoint of diffusion, it can be assumed that the position of the nucleus is relatively fixed, which implies $D_{3{\rm n}} \simeq 0$, and we again let $D_{3p} \simeq 18\,\mu{\rm m}^2/{\rm sec}$. It then follows that the encounter rate falls in the approximate range of $k_e = 680$ to $2950\,\mu{\rm m}^2/{\rm sec}$. The concentration of a single particle is the reciprocal of the cell volume, and the reciprocal of the product of this and $k_e$ provides an estimate of the mean encounter time. Eukaryotic cell volumes are on the order of 100 to $10^6\,\mu{\rm m}^3$ (Chapter 8), and if we assume that the latter are $\sim 100\times$ the nuclear volume, we obtain time search estimates in the range of 2.5 to 5.5 mins for a TF molecule randomly placed in the cytoplasm. These analyses ignore the additional time to locate and transport through a nuclear pore. Thus, adding in the search time within the nucleus, the total time for an individual eukaryotic TF to locate a specific TFBS is expected to be on

the order of 10 minutes to several hours.

---

**Foundations 21.4. The evolutionary dispersion of TFBS matching profiles.** For any TF, given its specific binding domain, there will also be a specific TFBS sequence on the DNA that maximizes the strength of binding. 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 (Foundations 21.3). 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 different 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 the assumptions that all binding sites with the same number of matches ($m$) are equivalent with respect to binding probability, regardless of the position of the mismatches. In addition, for simplicity 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 recognition 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 $m = \ell - 1$ 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 determined simply from the binomial coefficient $3^n \ell!/[(\ell - m)!m!]$. This reduces a complex problem involving many classes to a more manageable level.

We will further assume a population that usually resides in a near pure state, with a short enough time scale assumed that stochastic changes involve one-step transitions to adjacent states (Figure 21.12). Denoting the probability that a TFBS resides in class $m$ at time $t$ as $P(m, t)$, where $m$ denotes the number of matches, the time-dependent behavior of the system is described by

$$\frac{\partial P(m,t)}{t} = N\mu \cdot \{(3(m+1)\phi_{m+1,m}P(m+1,t) - [(\ell - m)\phi_{m,m+1} + (3m\phi_{m,m-1}]P(m,t) + (\ell - m + 1)\phi_{m-1,m}P(m-1,t)\}. \quad (21.4.1)$$

The front term $N\mu$ denotes the rate of influx of new mutations, whereas all remaining terms denote the probabilities of fixation of various changes conditional on origin by mutation. The first term is dropped when $m = \ell$, and the last is dropped when $m = 0$. 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 higher class, with $(m + 1)$ 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 $\phi_{m+1,m}$. The second term accounts for the efflux from class $m$ to the next upper and lower classes ($m + 1$ and $m - 1$), 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 lower class, which has $\ell - m + 1$ 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,

$$\phi_{x,y} = \frac{1 - e^{-2N_e s_{x,y}/N}}{1 - e^{-2N_e s_{x,y}}} \quad (21.4.2)$$

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 $x$ (Chapter 4).

Despite its apparent complexity, Equation 21.4.1 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 21.4.1. The entire array of TFBS states can be represented as a diagram with connecting arrows denoting the flux rates between adjacent classes (Figure 21.11). Because the rate of flux to matches declines and the rate of flux to mismatches increases as $m$ decreases, such a system must eventually reach an equilibrium, at which point for each class the net flux from above equals that from below. This condition is known as detailed balance.

For example, denoting the equilibrium solutions with a tilde, detailed balance requires that $3\ell\mu\widetilde{P}(\ell) = \mu\widetilde{P}(\ell-1)$, i.e., the flux from class $\ell$ to $(\ell-1)$ matches must equal the reciprocal flux. This tells us that the probability mass in class $(\ell-1)$ is $3\ell$ times that in the perfectly matching class $\ell$, i.e., $\widetilde{P}(\ell-1)/\widetilde{P}(\ell) = 3\ell$. 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). For the case of neutrality, this greatly simplifies to

$$\widetilde{P}(m) = C3^{\ell-m}\binom{\ell}{m}, \tag{21.4.3}$$

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 $m$ matches.

Extending this approach to include selection is conceptually straight-forward. The coefficient on each arrow in Figure 21.11 simply needs to be multiplied by the fixation probability between adjacent classes. For example, for the arrows connecting classes $\ell$ and $(\ell-1)$, the coefficients become $3\ell\mu\phi_{\ell,\ell-1}$ and $\mu\phi_{\ell-1,\ell}$. 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) $\phi_{m-1,m}/\phi_{m,m-1} = e^{2N_e s_{m-1,m}}$; and 2) $s_{m-1,m} = W_m - W_{m-1}$, where $W_m$ is the fitness of alleles with $m$ matches in their TFBS. Using these equalities, Equation 21.4.3 generalizes to

$$\widetilde{P}(m) = C3^{\ell-m}\binom{\ell}{m}e^{2N_e W(m)}, \tag{21.4.4}$$

where $C$ is again a normalization constant (equal to the reciprocal of the sum of the terms to the right of $C$ for all $m$). Equation 21.4.4 shows that with selection the equilibrium probability distribution of alternative binding states is equivalent to a simple modification of the neutral expectation, with each neutral genotypic probability being weighted exponentially by the product of its fitness and the effective population size (which influences the efficiency of selection). Further elaborations of this model can be found in Lynch and Hagner (2015) and Tuğrul et al. (2015).

## Literature Cited

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

Alon, U. 2003. Biological networks: the tinkerer as an engineer. Science 301: 1866-1867.

Alon, U. 2007. Simplicity in biology. Nature 446: 497.

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.

Arnheim, N. 1986. The evolution of transcriptional control signals: coevolution of ribosomal gene promoter sequences and transcription factors, pp. 37-51. In S. Karlin and E. Nevo (eds.) Evolutionary Processes and Theory. Academic Press, Inc., Orlando, FL.

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.

Babu, M. M., S. A. Teichmann, and L. Aravind. 2006. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J. Mol. Biol. 358: 614-633.

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, et al. 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.

Balaji, S., L. M. Iyer, L. Aravind, and M. M. Babu. 2006. Uncovering a hidden distributed architecture behind scale-free transcriptional regulatory networks. J. Mol. Biol. 360: 204-212.

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.

Barabási, A. L., and Z. N. Oltvai. 2004. Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5: 101-113.

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 M. 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.

Berger, M. F., and M. L. Bulyk. 2009. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat. Protoc. 4: 393-411.

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, and 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.

Borneman, A. R., T. A. Gianoulis, Z. D. Zhang, H. Yu, J. Rozowsky, M. R. Seringhaus, L. Y. Wang, M. Gerstein, and M. Snyder. 2007. Divergence of transcription factor binding sites across related yeast species. Science 317: 815-819.

Britton, C. S., T. R. Sorrells, and A. D. Johnson. 2020. Protein-coding changes preceded *cis*-regulatory gains in a newly evolved transcription circuit. Science 367: 96-100.

Brodsky, S., T. Jana, K. Mittelman, M. Chapal, D. K. Kumar, M. Carmi, and N. Barkai. 2020. Intrinsically disordered regions direct transcription factor *in vivo* binding specificity. Mol. Cell 79: 459-471.e4.

Burda, Z., A. Krzywicki, O. C. Martin, and M. Zagorski. 2011. Motifs emerge from function in model gene regulatory networks. Proc. Natl. Acad. Sci. USA 108: 17263-17268.

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. 1982. Skeletal DNA and the evolution of genome size. Annu. Rev. Biophys. Bioeng. 11: 273-302.

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.

Chong, S., C. Chen, H. Ge, and X. S. Xie. 2014. Mechanism of transcriptional bursting in bacteria. Cell 158: 314-326.

Conant, G. C., and A. Wagner. 2003. Convergent evolution of gene circuits. Nat. Genet. 34: 264-266.

Cordero, O. X., and P. Hogeweg. 2006. Feed-forward loop circuits as a side effect of genome evolution. Mol. Biol. Evol. 23: 1931-1936.

Corrigan, A. M., E. Tunnacliffe, D. Cannon, and J. R. Chubb. 2016. A continuum model of transcriptional bursting. eLife 5: e13051.

Cramer, P. 2019. Organization and regulation of gene transcription. Nature 573: 45-54.

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, and 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.

Davidson, E. H. 2006. The Regulatory Genome: Gene Regulatory Networks in Development and Evolution. Academic Press, New York, NY.

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.

Ding, C., D. W. Chan, W. Liu, M. Liu, D. Li, L. Song, C. Li, J. Jin, A. Malovannaya, S. Y. Jung, B. Zhen, Y. Wang, and J. Qin. 2013. Proteome-wide profiling of activated transcription factors with a concatenated tandem array of transcription factor response elements. Proc. Natl. Acad. Sci. USA 110: 6771-6776.

Doniger, S. W., and J. C. Fay. 2007. Frequent gain and loss of functional transcription factor binding sites. PLoS Comput. Biol. 3: e99.

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.

Dori-Bachash, M., E. Shema, and I. Tirosh. 2011. Coupled evolution of transcription and mRNA degradation. PLoS Biol. 9: e1001106.

Duveau, F., A. Hodgins-Davis, B. P. Metzger, B. Yang, S. Tryban, E. A. Walker, T. Lybrook, and P. J. Wittkopp. 2018. Fitness effects of altering gene expression noise in *Saccharomyces cerevisiae*. eLife 7: e37272.

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.

Fraser, H. B., A. E. Hirsh, G. Giaever, J. Kumm, and M. B. Eisen. 2004. Noise minimization in eukaryotic gene expression. PLoS Biol. 2: e137.

Fujimoto, S., M. Ito, S. Matsunaga, and K. Fukui. 2005. An upper limit of the ratio of DNA volume to nuclear volume exists in plants. Genes Genet. Syst. 80: 345-350.

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.

Gerhart, J., and M. Kirschner. 1997. Cells, Embryos and Evolution. Blackwell Science, Malden, MA.

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

Gerland, U., and T. Hwa. 2009. Evolutionary selection between alternative modes of gene regulation. Proc. Natl. Acad. Sci USA 106: 8841-8846.

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.

Gotea, V., A. Visel, J. M. Westlund, M. A. Nobrega, L. A. Pennacchio, and I. Ovcharenko. 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res. 20: 565-577.

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.

Gruber, J. D., K. Vogel, G. Kalay, and P. J. Wittkopp. 2012. Contrasting properties of gene-specific regulatory, coding, and copy number mutations in *Saccharomyces cerevisiae:* frequency, effects, and dominance. PLoS Genet. 8: e1002497.

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: 14713-14720.

Haag, E. S., and M. N. Molla. 2005. Compensatory evolution of interacting gene products through multifunctional intermediates. Evolution 59: 1620-1632.

Haberle, V., and A. Stark. 2018. Eukaryotic core promoters and the functional basis of transcription initiation. Nat. Rev. Mol. Cell Biol. 19: 621-637.

Hafner, A., and A. Boettiger. 2023. The spatial organization of transcriptional control. Nature Rev. Genet. 24: 53-68.

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.

Haldane, A., M. Manhart, and A. V. Morozov. 2014. Biophysical fitness landscapes for transcription factor binding sites. PLoS Comput Biol. 10: e1003683.

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

Heinz, S., C. E. Romanoski, C. Benner, K. A. Allison, M. U. Kaikkonen, L. D. Orozco, and C. K. Glass. 2013. Effect of natural genetic variation on enhancer selection and function. Nature 503: 487-492.

Hill, M. S., P. Vande Zande, and P. J. Wittkopp. 2021. Molecular and evolutionary processes generating variation in gene expression. Nat. Rev. Genet. 22: 203-215.

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.

Hong, J., N. Brandt, F. Abdul-Rahman, A. Yang, T. Hughes, and D. Gresham. 2018. An incoherent feedforward loop facilitates adaptive tuning of gene expression. eLife 7: e32323.

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

Ingram, P. J., M. P. Stumpf, and J. Stark. 2006. Network motifs: structure does not determine function. BMC Genomics 7: 108.

Isalan, M., C. Lemerle, K. Michalodimitrakis, C. Horn, P. Beltrao, E. Raineri, M. Garriga-Canut, and L. Serrano. 2008. Evolvability and hierarchy in rewired bacterial gene networks. Nature 452: 840-845.

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.

Iyer, L. M., V. Anantharaman, M. Y. Wolf, and L. Aravind. 2008. Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. Int. J. Parasitol. 38: 1-31.

Johnson, N. A., and A. H. Porter. 2000. Rapid speciation via parallel, directional selection on regulatory genetic pathways. J. Theor. Biol. 205: 527-542.

Johnson, N. A., and A. H. Porter. 2001. Toward a new synthesis: population genetics and evolutionary developmental biology. Genetica 112/113: 45-58.

Johnson, N. A., and A. H. Porter. 2007. Evolution of branched regulatory genetic pathways: directional selection on pleiotropic loci accelerates developmental system drift. Genetica 129: 57-70.

Jolma, A., J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, et al. 2013. DNA-binding specificities of human transcription factors. Cell 152: 327-339.

Jolma, A., Y. Yin, K. R. Nitta, K. Dave, A. Popov, M. Taipale, M. Enge, T. Kivioja, E. Morgunova, and J. Taipale. 2015. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature 527: 384-388.

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.

Le, D. D., T. C. Shimko, A. K. Aditham, A. M. Keys, S. A. Longwell, Y. Orenstein, and P. M. Fordyce. 2018. Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. Proc. Natl. Acad. Sci. USA 115: E3702-E3711.

Lee, T. I., N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298: 799-804.

Lestas, I., G. Vinnicombe, and J. Paulsson. 2010. Fundamental limits on the suppression of molecular fluctuations. Nature 467: 174-178.

Levo, M., and E. Segal. 2014. In pursuit of design principles of regulatory sequences. Nat. Rev. Genet. 15: 453-468.

Levy, S. F., N. Ziv, and M. L. Siegal. 2012. Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant. PLoS Biol. 10: e1001325.

Li, G.-W., D. Burkhardt, C. Gross, and J. S. Weissman. 2014. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell 157: 624-635.

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

Liu, J., H. Martin-Yken, F. Bigey, S. Dequin, J. M. Franois, and J. P. Capp. 2015. Natural yeast promoter variants reveal epistasis in the generation of transcriptional-mediated noise and its potential benefit in stressful conditions. Genome Biol. Evol. 7: 969-94.

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

Lynch, M. 2007b. The evolution of genetic networks by nonadaptive processes. Nature Reviews Genetics 8: 803-813.

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

Lynch, M., and K. Hagner. 2015. Evolutionary meandering of intermolecular interactions along

the drift barrier. Proc. Natl. Acad. Sci. USA 112: E30-8

Lynch, M., and J. B. Walsh. 1998. Genetics and Analysis of Quantitative Traits. Sinauer Assocs., Inc., Sunderland, MA.

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.

Mariño-Ramírez, L., I. K. Jordan, and D. Landsman. 2006. Multiple independent evolutionary solutions to core histone gene regulation. Genome Biol. 7: R122.

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. 24: 496-510..

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.

Martin-Perez, M., and J. Villén. 2017. Determinants and regulation of protein turnover in yeast. Cell Syst. 5: 283-294.e5.

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

Matsumoto, T., K. Mineta, N. Osada, and H. Araki. 2015. An individual-based diploid model predicts limited conditions under which stochastic gene expression becomes advantageous. Front. Genet. 6: 336.

Metzger, B. P., F. Duveau, D. C. Yuan, S. Tryban, B. Yang, and P. J. Wittkopp. 2016. Contrasting frequencies and effects of *cis*- and *trans*-regulatory mutations affecting gene expression. Mol. Biol. Evol. 33: 1131-1146.

Metzger, B. P. H., and P. J. Wittkopp. 2019. Compensatory *trans*-regulatory alleles minimizing variation in *TDH3* expression are common within *Saccharomyces cerevisiae*. Evol. Lett. 3: 448-461.

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.

Milo, R., S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. Network motifs: simple building blocks of complex networks. Science 298: 824-827.

Mineta, K., T. Matsumoto, N. Osada, and H. Araki. 2015. Population genetics of non-genetic traits: evolutionary roles of stochasticity in gene expression. Gene 562: 16-21.

Morgunova, E., Y. Yin, P. K. Das, A. Jolma, F. Zhu, A. Popov, Y. Xu, L. Nilsson, and J. Taipale. 2018. Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima. eLife 7: e32963.

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.

Nitta, K. R., A. Jolma, Y. Yin, E. Morgunova, T. Kivioja, J. Akhtar, K. Hens, J. Toivonen, B. Deplancke, E. E. Furlong, and J. Taipale. 2015. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. eLife 4: e04837.

Nocedal, I., E. Mancera, and A. D. Johnson. 2017. Gene regulatory network plasticity predates a switch in function of a conserved transcription regulator. eLife 6: e23250.

Normanno, D., M. Dahan, and X. Darzacq. 2012. Intra-nuclear mobility and target search mechanisms of transcription factors: a single-molecule perspective on gene expression. Biochim. Biophys. Acta 1819: 482-493.

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. 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.

Pavlicev, M., and G. P. Wagner. 2012. A model of developmental evolution: selection, pleiotropy and compensation. Trends Ecol. Evol. 27: 316-322.

Peccoud, J., and B. Ycart. 1995. Markovian modeling of gene-product synthesis. 48: 222-234.

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.

Phillips, R., J. Kondev, J. Theriot, and H. Garcia. 2012. Physical Biology of the Cell, 2nd Ed. Garland Science, New York, NY.

Pougach, K., A. Voet, F. A. Kondrashov, K. Voordeckers, J. F. Christiaens, B. Baying, V. Benes, R. Sakai, J. Aerts, B. Zhu, et al. 2014. Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. Nat. Commun. 5: 4868.

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.

Reece-Hoyes, J. S., C. Pons, A. Diallo, A. Mori, S. Shrestha, S. Kadreppa, J. Nelson, S. Diprima, A. Dricot, B. R. Lajoie, et al. 2013. Extensive rewiring and complex evolutionary dynamics in a *C. elegans* multiparameter transcription factor network. Mol. Cell 51: 116-127.

Richter, D. J., P. Fozouni, M. B. Eisen, and N. King. 2018. Gene family innovation, conservation and loss on the animal stem lineage. eLife 7: e34226.

Riechmann, J. L., J. Heard, G. Martin, L. Reuber, C. Jiang, J. Keddie, L. Adam, O. Pineda, O. J. Ratcliffe, R. R. Samaha, et al. 2000. *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. Science 290: 2105-2110.

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.

Ruths, T., and L. Nakhleh. 2013. Neutral forces acting on intragenomic variability shape the *Escherichia coli* regulatory network topology. Proc. Natl. Acad. Sci. USA 110: 7754-7759.

Ruvinsky, I., and G. Ruvkun. 2003. Functional tests of enhancer conservation between distantly related species. Development 130: 5133-5142.

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.

Sarda, S., and S. Hannenhalli. 2015. High-throughput identification of *cis*-regulatory rewiring events in yeast. Mol. Biol. Evol. 32: 3047-3063.

Savageau, M. A. 1974. Genetic regulatory mechanisms and the ecological niche of *Escherichia coli.* Proc. Natl. Acad. Sci. USA 71: 2453-2455.

Savageau, M. A. 1977. Design of molecular control mechanisms and the demand for gene expression. Proc. Natl. Acad. Sci. USA 74: 5647-5651.

Savageau, M. A. 1998. Demand theory of gene regulation. I. Quantitative development of the theory. Genetics 149: 1665-1676.

Sayou, C., M. Monniaux, M. H. Nanao, E. Moyroud, S. F. Brockington, E. Thévenon, H. Chahtane, N. Warthmann, M. Melkonian, Y. Zhang, et al. 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, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328: 1036-1040.

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.

Sellerio, A. L., B. Bassetti, H. Isambert, and M. Cosentino Lagomarsino. 2009. A comparative evolutionary study of transcription networks. The global role of feedback and hierachical structures. Mol. Biosyst. 5: 170-179.

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.

Sepúlveda, L. A., H. Xu, J. Zhang, M. Wang, and I. Golding. 2016. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. Science 351: 1218-1222.

Sevier, S. A., D. A. Kessler, and H. Levine. 2016. Mechanical bounds to transcriptional noise. Proc. Natl. Acad. Sci. USA 113: 13983-13988.

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

Shen-Orr, S. S., R. Milo, S. Mangan, and U. Alon. 2002. Network motifs in the transcriptional regulation network of *Escherichia coli.* Nature Genet. 31: 64-68.

Shinar, G., E. Dekel, T. Tlusty, and U. Alon. 2006. Rules for biological regulation based on error minimization. Proc. Natl. Acad. Sci. USA 103: 3999-4004.

Shultzaberger, R. K., D. S. Malashock, J. F. Kirsch, and M. B. Eisen. 2010. The fitness landscapes of cis-acting binding sites in different promoter and environmental contexts. PLoS Genet. 6(7):

e1001042.

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.

Skinner, S. O., H. Xu, S. Nagarkar-Jaiswal, P. R. Freire, T. P. Zwaka, and I. Golding. 2016. Single-cell analysis of transcription kinetics across the cell cycle. eLife 5: e12175.

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.

Solé, R. V., and S. Valverde. 2008. Spontaneous emergence of modularity in cellular networks. J. R. Soc. Interface 5: 129-133.

Sommer R. J. 2012. Evolution of regulatory networks: nematode vulva induction as an example of developmental systems drift. Adv. Exp. Med. Biol. 751: 79-91.

Sorrells, T. R., L. N. Booth, B. B. Tuch, and A. D. Johnson. 2015. Intersecting transcription networks constrain gene regulatory evolution. Nature 523: 361-365.

Staller, M. V. 2022. Transcription factors perform a 2-step search of the nucleus. Genetics 222: iyac111.

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

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.

Tagkopoulos, I., Y. C. Liu, and S. Tavazoie. 2008. Predictive behavior within microbial genetic networks. Science 320: 1313-1317.

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.

Tănase-Nicola, S., and P. R. ten Wolde. 2008. Regulatory control and the costs and benefits of biochemical noise. PLoS Comput. Biol. 4: e1000125.

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

Thieffry, D., A. M. Huerta, E. Perez-Rueda, and J. Collado-Vides. 1998. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli.* Bioessays 20: 433-440.

Thurman, R. E.,E. Rynes,R. Humbert,J. Vierstra,M. T. Maurano,E. Haugen,N. C. Sheffield,A. B. Stergachis,H. Wang,B. Vernot,K. Garget al. 2012. The accessible chromatin landscape of the human genome. Nature 489: 75-82.

True, J. R., and E. S. Haag. 2001. Developmental system drift and flexibility in evolutionary trajectories. Evol. Dev. 3: 109-119.

Tsong, A. E., B. B. Tuch, H. Li, and A. D. Johnson. 2006. Evolution of alternative transcriptional circuits with identical logic. Nature 443: 415-420.

Tsoy, O. V., M. A. Pyatnitskiy, M. D. Kazanov, and M. S. Gelfand. 2012. Evolution of transcriptional regulation in closely related bacteria. BMC Evol. Biol. 12: 200.

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

Tuğrul, M., T. Paixão, N. H. Barton, and G. Tkačik. 2015. Dynamics of transcription factor binding site evolution. PLoS Genet. 11: e1005639.

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, A. 2005. Robustness and Evolvability in Living Systems. Princeton Univ. Press, Princeton, NJ.

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, Z., and J. Zhang. 2011. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. Proc. Natl. Acad. Sci. USA 108: E67-E76.

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.

Weiss, K. M., and S. M. Fullerton. 2000. Phenogenetic drift and the evolution of genotype-phenotype relationships. Theor. Popul. Biol. 57: 187-195.

Wilkins, A. S. 2002. The Evolution of Developmental Pathways. Sinauer Assocs., Inc., Sunderland,

MA.

Wilkins, A. S. 2005. Recasting developmental evolution in terms of genetic pathway and network evolution and the implications for comparative biology. Brain Res. Bull. 66: 495-509.

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.

Wolf, L., O. K. Silander, and E. van Nimwegen. 2015. Expression noise facilitates the evolution of gene regulation. eLife 4: e05856.

Wolfe, K. H., and D. C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387: 708-713.

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

Wuchty, S., and E. Almaas. 2005. Evolutionary cores of domain co-occurrence networks. BMC Evol. Biol. 5: 24.

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.

Zheng, W., T. A. Gianoulis, K. J. Karczewski, H. Zhao, and M. Snyder. 2011. Regulatory variation within and between species. Annu. Rev. Genomics Hum. Genet. 12: 327-346.

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.

Zitnik, M., R. Sosič, M. W. Feldman, and J. Leskovec. 2019. Evolution of resilience in protein interactomes across the tree of life. Proc. Natl. Acad. Sci. USA 116: 4426-4433.

**Figure 21.1.** Flow chart for the key determinants of the number of proteins associated with a particular gene within a cell. As transcription factors (blue) 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 produces fully functional mRNAs (orange) at rate $k_{\text{m}}$, which are in turn translated into proteins (purple) at rate $k_{\text{p}}$. Messenger RNAs and proteins are degraded at rates $\delta_{\text{m}}$ and $\delta_{\text{p}}$, respectively (small dashes and spheres).

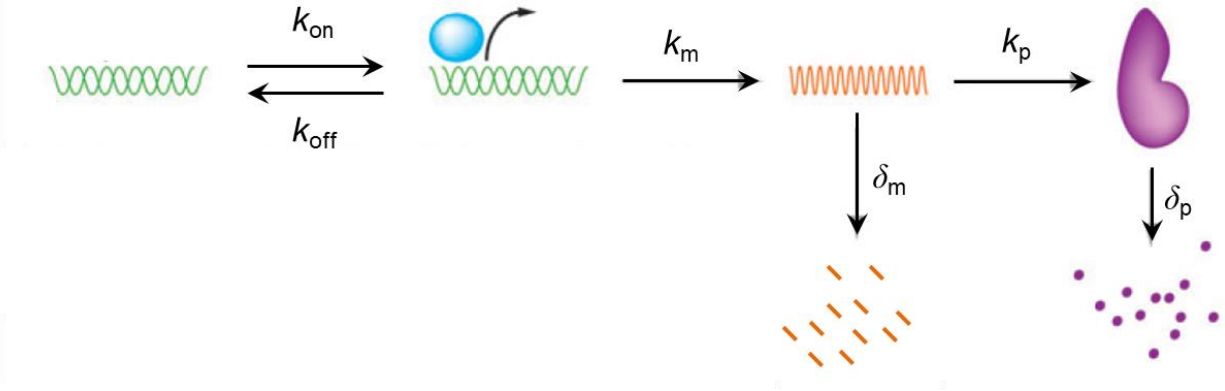
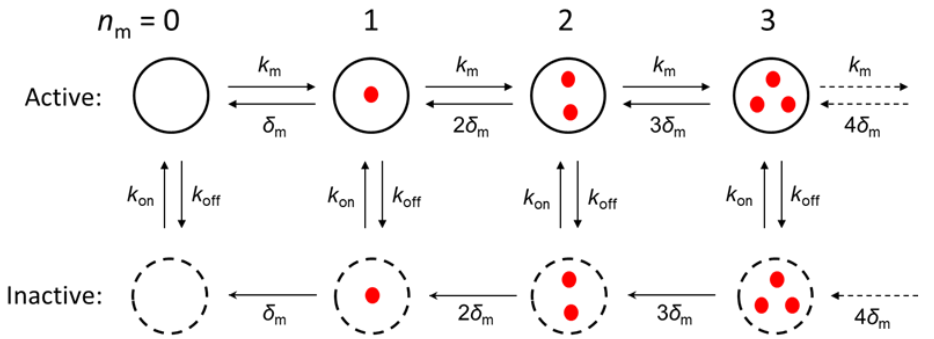


**Figure 21.2.** Determinants of the statistical distribution for the number of mRNAs per cell associated with a particular gene ($n_m$). **Above)** Flow chart for the transitions of numbers of mRNAs within individual cells ($n$). The rates are defined as in Figure 21.1. Circles with solid and dashed lines denote transcribing and nontranscribing cells, respectively, with the red dots denoting the numbers of transcripts in cells of various states. Note that inactive genes can only lose (not gain) mRNAs. **Below)** Probability distributions for the number of mRNAs for a particular gene present in cells, as a function of the ratio of transcription to degradation rates, $k_m/\delta_m$, 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 $k_m/\delta_m$ 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. The functions are obtained with Equation 21.1.7.



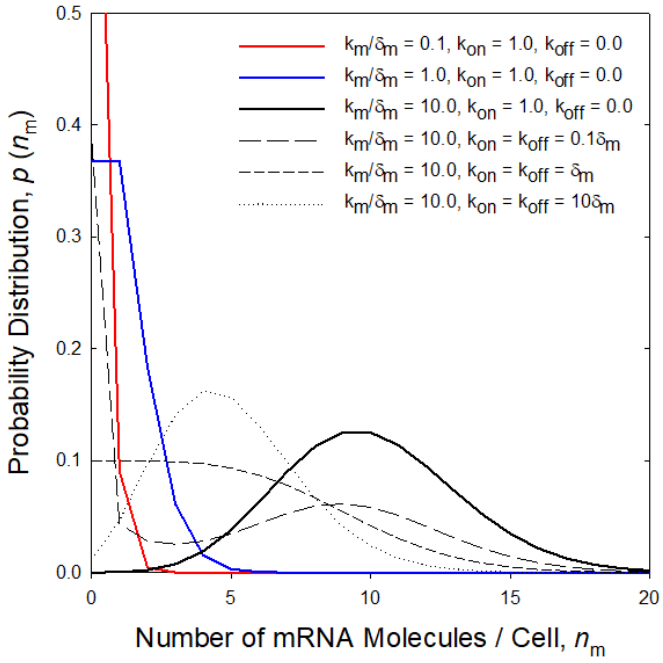



**Figure 21.3.** Idealized view of the temporal variation in gene expression within a cell. The gene is stochastically turned on (blue vertical bars) or off at points depending on the binding of the cognate transcription factors. Messenger RNAs are produced during the on periods, but decline at an exponential 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. However, the fluctuations in protein numbers are damped, owing to their greater longevities than mRNA molecules.

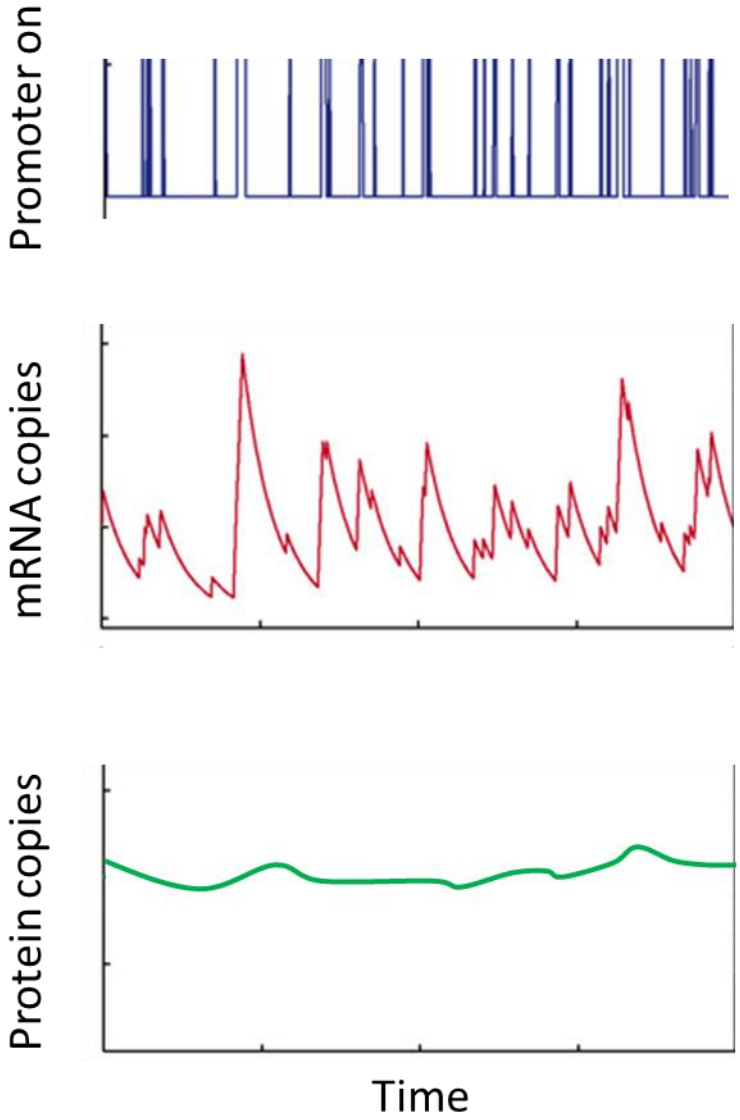


**Figure 21.4.** Phenotypic distributions for two alternative genotypes with high and low levels of expression noise, given by the solid and dashed lines respectively, relative to the fitness function. The red lines illustrate a situation in which two alternative genotypes with the same average expression level deviate far from the optimum phenotype (denoted by the peak of the black curve). The blue lines denote the situation when the genotypes have mean expression levels coinciding with the optimum.

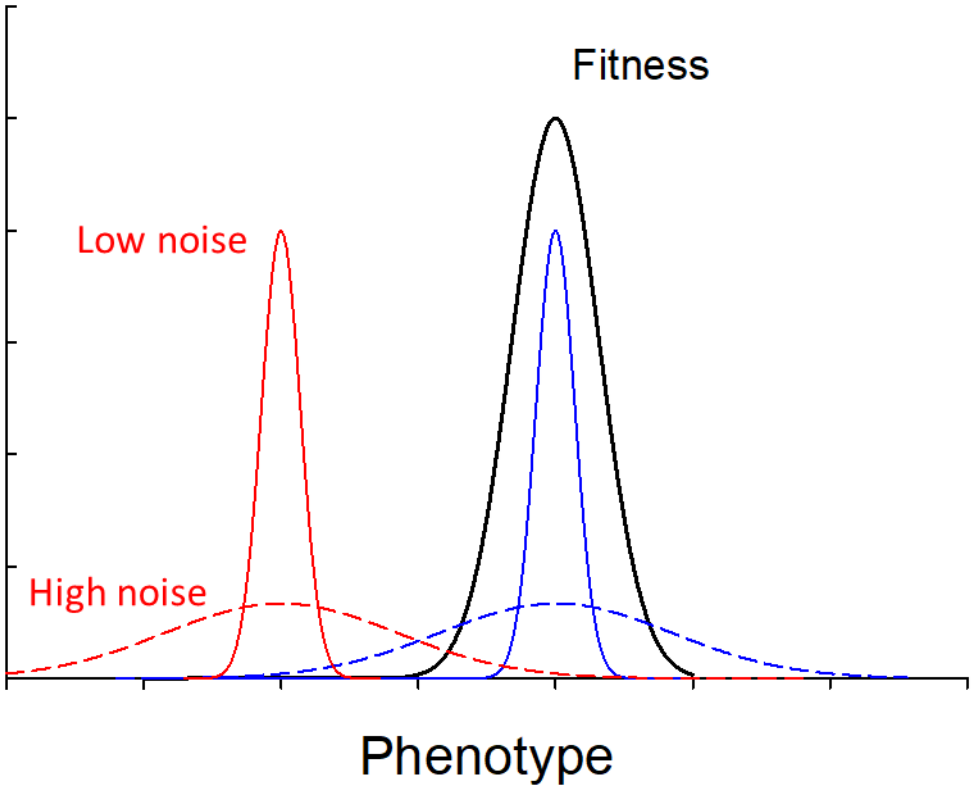


**Figure 21.5.** The probability that a particular transcription-factor binding site (TFBS) is bound by a cognate transcription factor (TF), given the level of background interference ($B$) and the number of nucleotides at the site ($m$) matching the optimal recognition sequence of the TF. The curves, obtained from Equation 21.5, cover the range of biologically plausible values of $B$ (as described in the text).

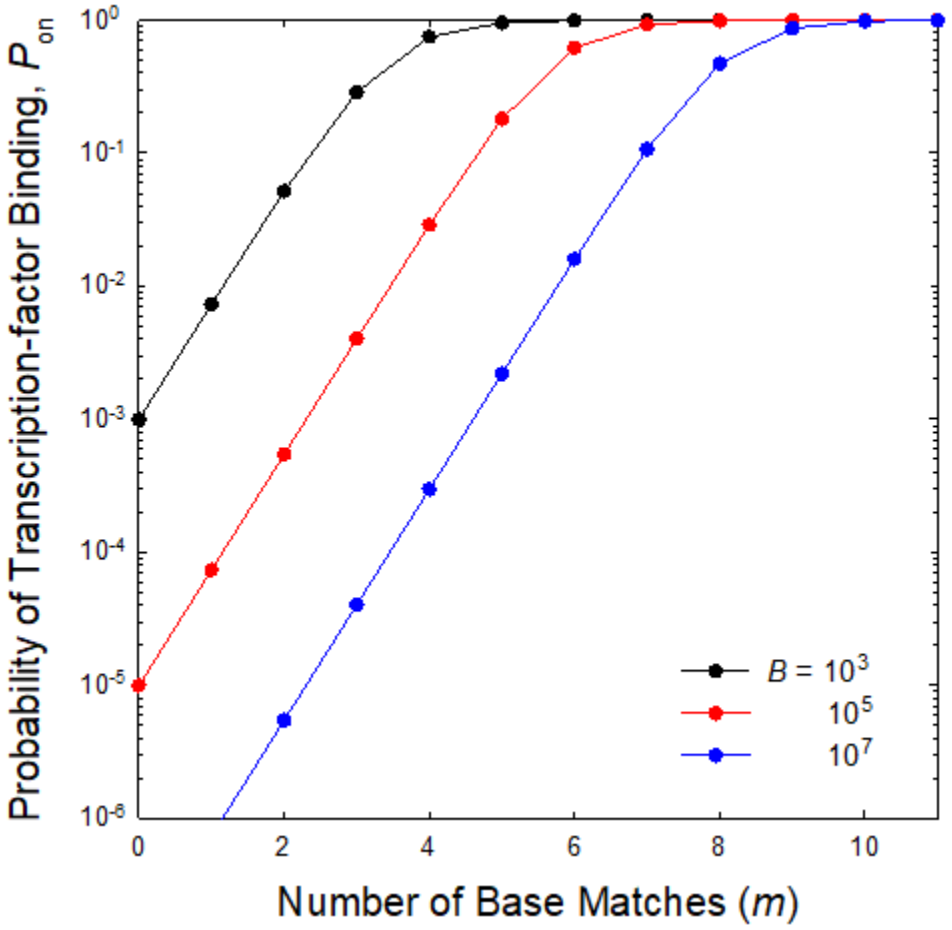


**Figure 21.6.** The expected equilibrium evolutionary distribution of binding-site matches with transcription-factor motifs of lengths $\ell = 8$ and 16. Results are given for various levels of the strength of selection relative to the power of genetic drift ($N_e\alpha$), and two levels of background interference ($B$).

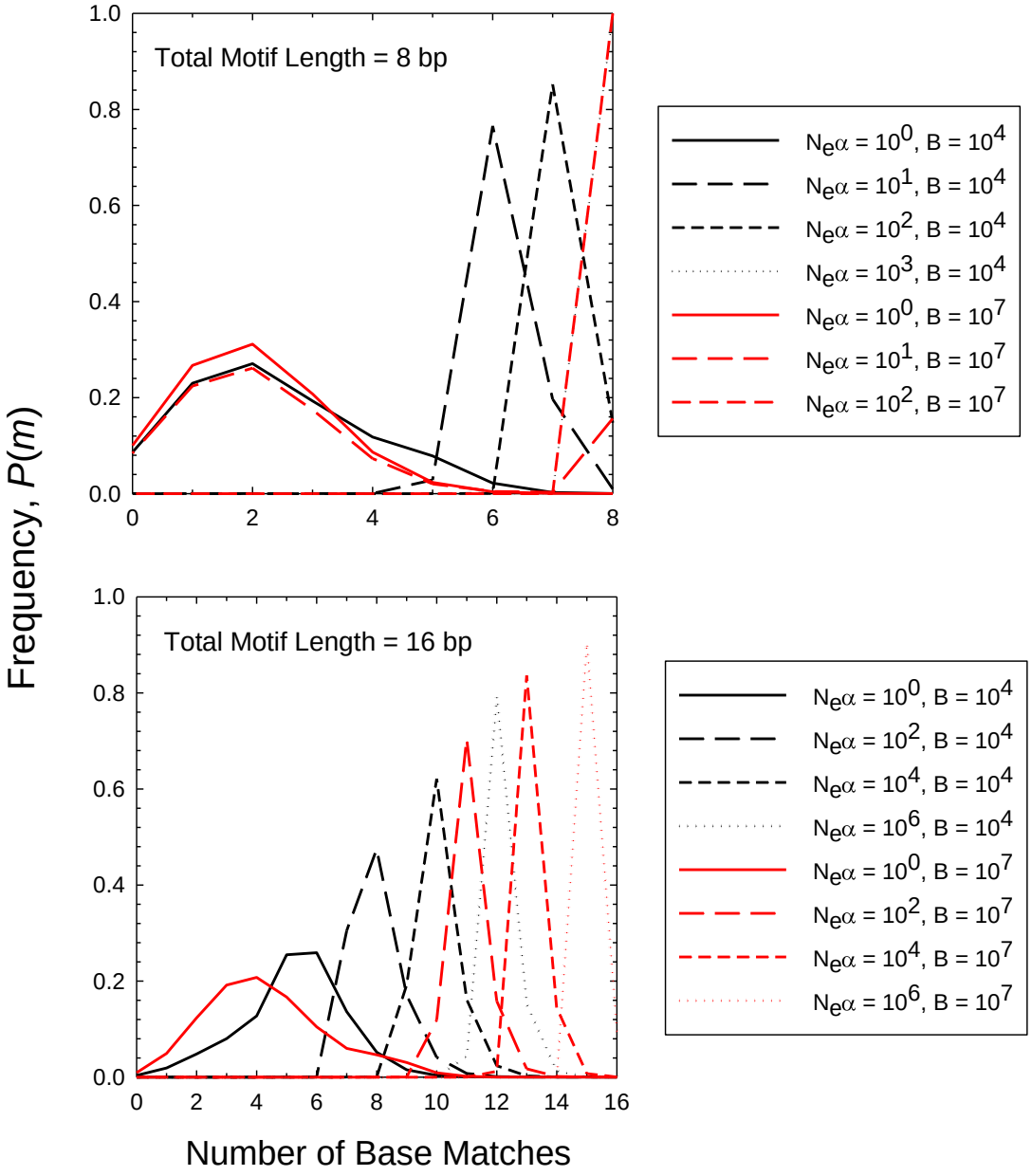


**Figure 21.7.** a) Distribution of binding energies associated with the transcription factor CRP in *E. coli*. Note that contrary to the approach in the text, the binding sites are characterized with respect to energy rather than mismatches, although the two scales are entirely interchangeable. Energies 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 (blue bell-shaped curve 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).

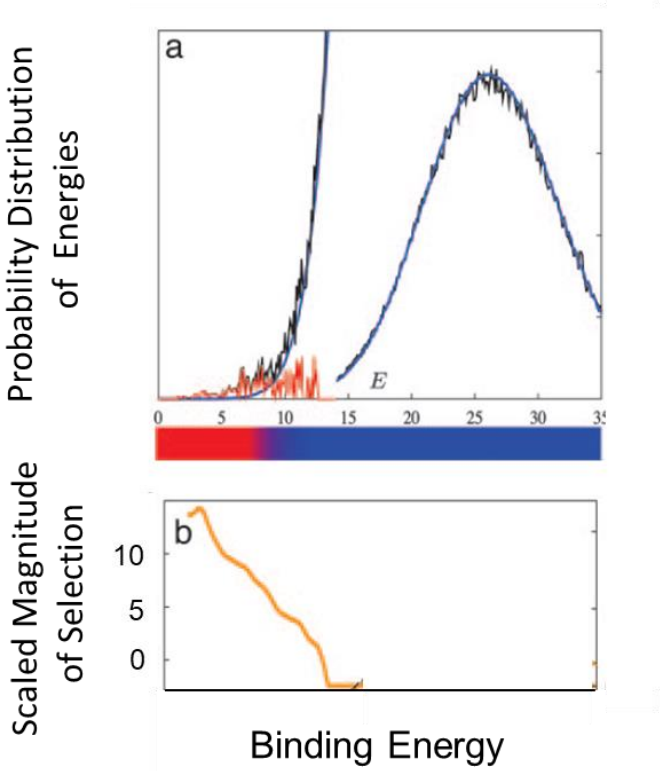


**Figure 21.8.** Gene-expression regulation by an activator vs. repressor transcription factor.

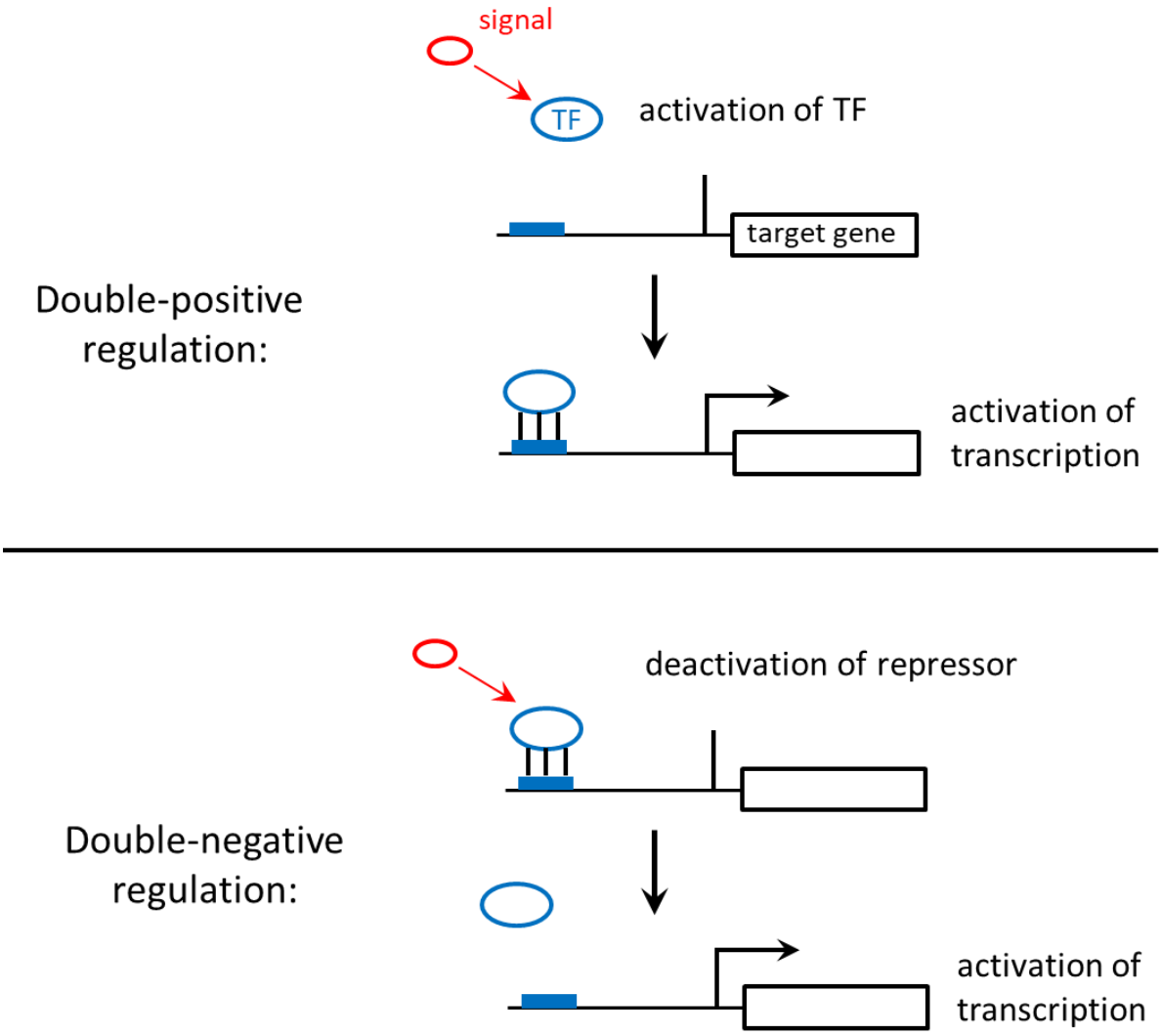


**Figure 21.9.** Two ways in which regulatory mechanisms may diverge between species in an effectively neutral fashion. **Above**) An ancestral transcription factor (purple) is capable of binding to two different, but functionally equivalent and hence redundant, sites (blue and red rectangles). Following isolation by speciation, the TF in each descendent taxon then loses an alternative TFBS, hence becoming more specialized. **Below**) A pre-existing TF (blue) fortuitously acquires binding affinity to another protein (red) that may or may not have DNA-binding activity. This opens up the opportunity for the emergence of a TFBS site with affinity to the recruited protein, resulting in redundant gene regulation, and allowing for the eventual loss of the ancestral regulatory mechanism (not shown).

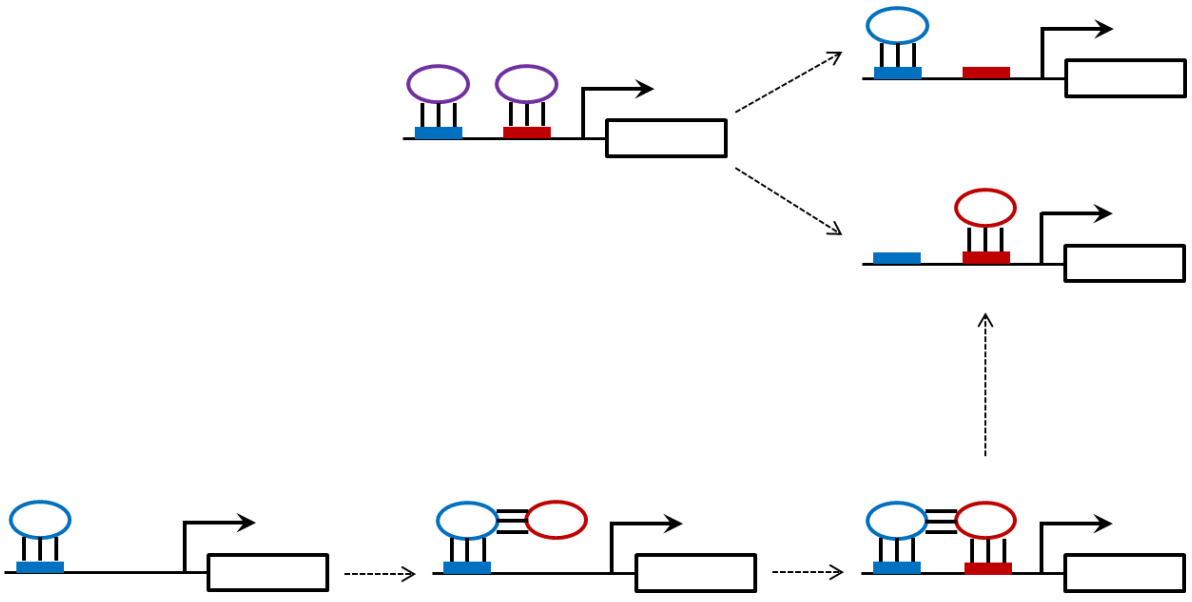
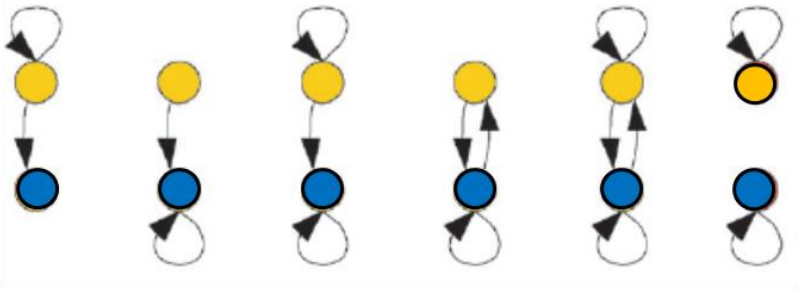
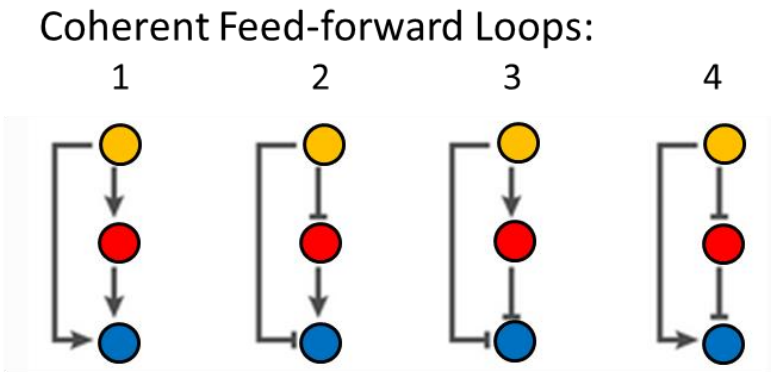


**Figure 21.10.** Examples of possible topologies for simple gene networks. Arrows denote activation (including self-activation for loops), and blunt ends denote repression. **Above**) Two-gene model, with yellow and blue denoting alternative genes, one or both of which might be a transcription factor. Note that examples of repression are not included here, which would further magnify the number of possible topologies. **Below**) Three-gene model, with yellow and red denoting transcription factors, and blue denoting the regulated gene. Examples of self-regulation are not included here, which again would further increase the number of topologies.





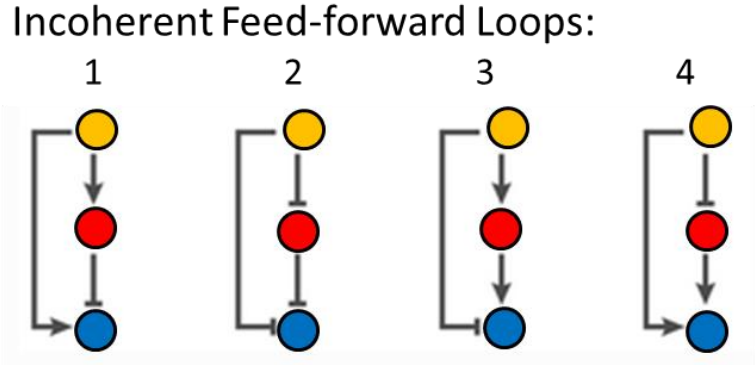



**Figure 21.11.** Flow diagram for the series of evolutionary events leading to pathway expansions and contractions. Only activating interactions are shown, and for simplicity, gain and loss rates ($u_g$ and $u_l$) are assumed to be the same for all links.

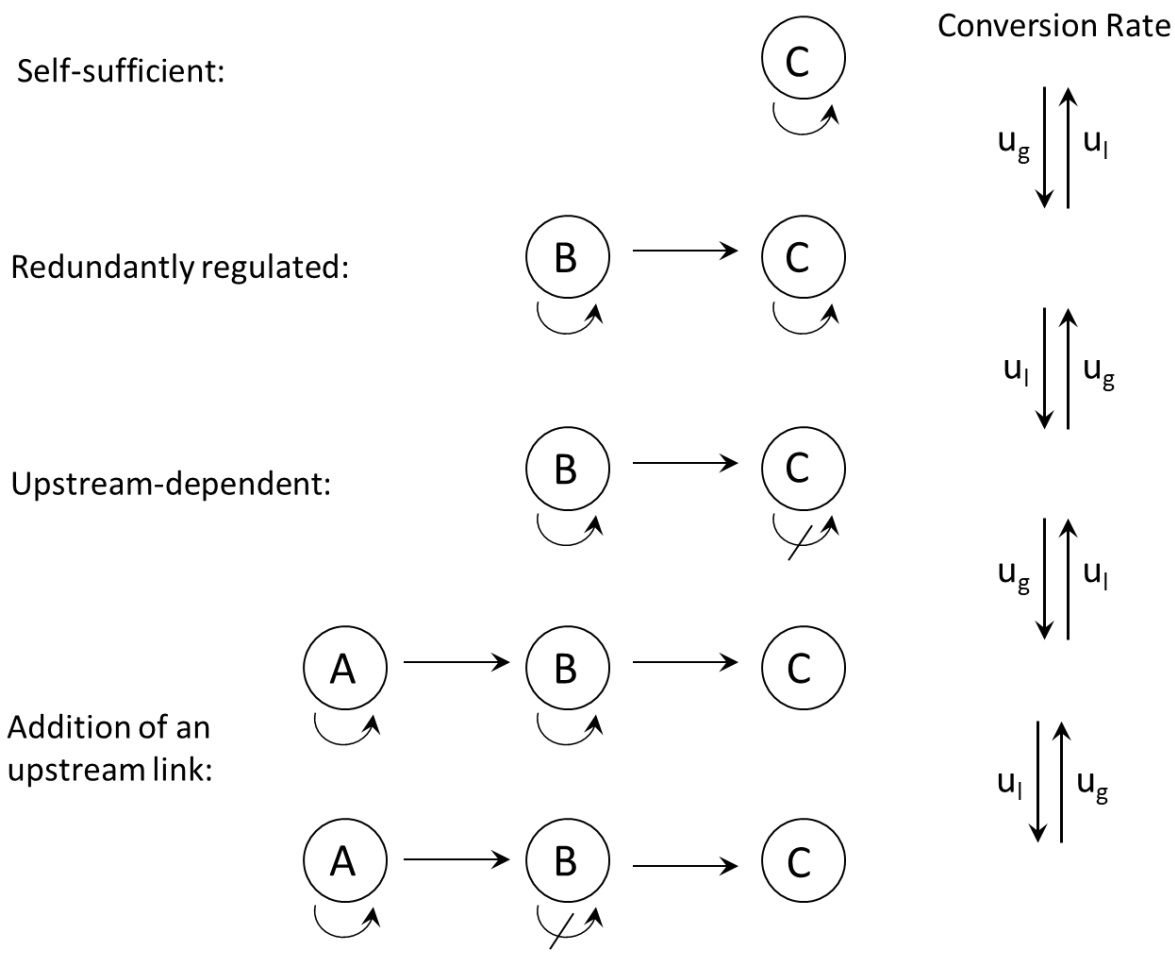


**Figure 21.12.** Flow diagram for the alternative states of a binding site of length $\ell = 5$, with the dot diagrams below simply illustrating one specific type within each category of numbers of mismatches ($m$, denoted by black balls). The transition rates are given on the arrows for the case of 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.