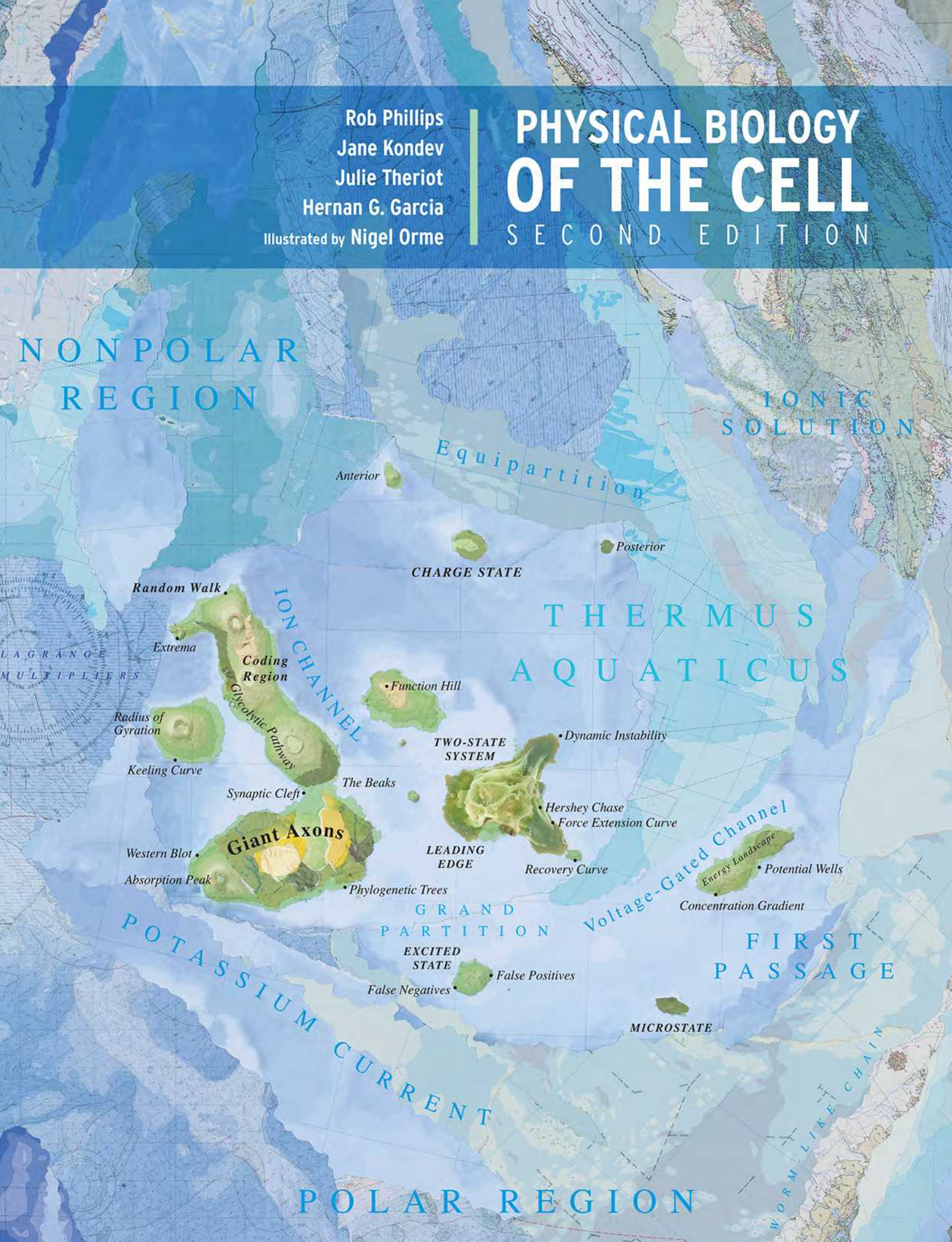
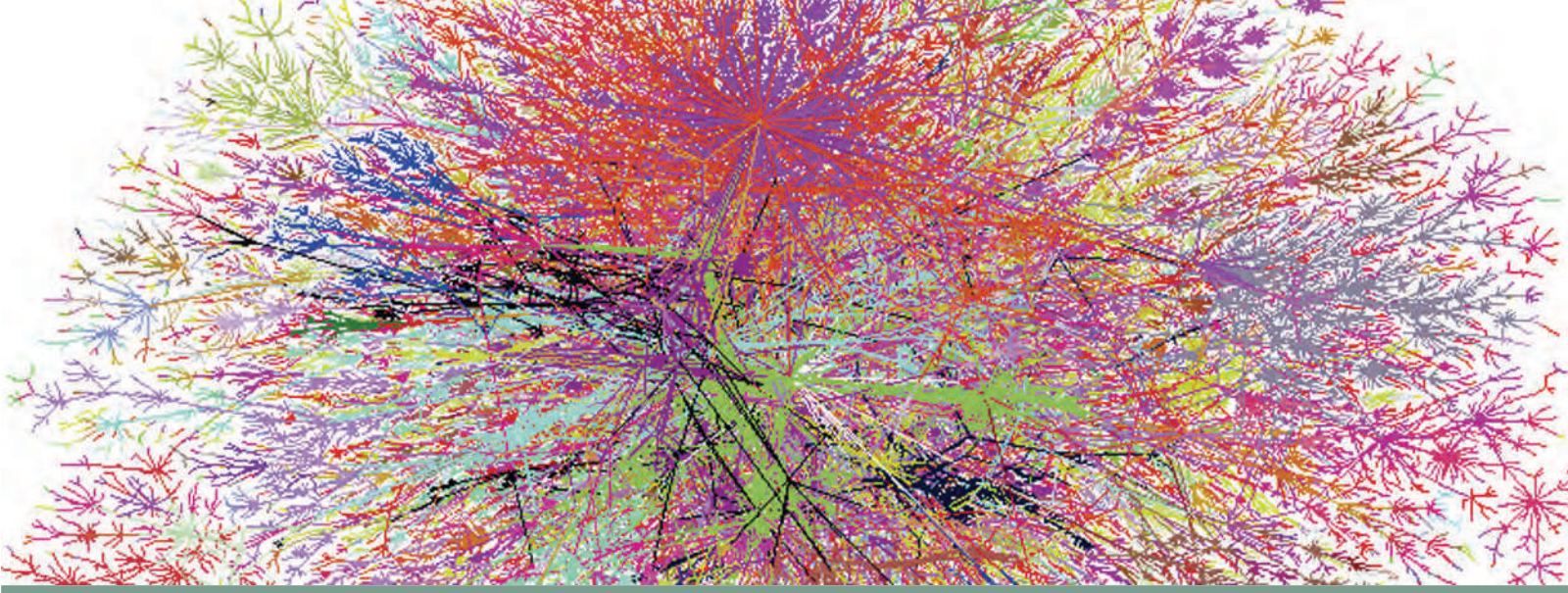

Rob Phillips
Jane Kondev
Julie Theriot
Hernan G. Garcia
Illustrated by Nigel Orme
PHYSICAL BIOLOGY
OF THE CELL
SECOND EDITION
NONPOLAR
REGION
IONIC
SOLUTION
Equipartition
Anterior
Posterior
CHARGE STATE
Random Walk
ION CHANNEL
THERMUS
AQUATICUS
LAGRANGE
MULTIPLIERS
Extrema
Coding
Region
Glycolytic Pathway
Function Hill
Radius of
Gyration
Keeling Curve
TWO-STATE
SYSTEM
Dynamic Instability
The Beaks
Synaptic Cleft
Hershey Chase
Force Extension Curve
Giant Axons
Western Blot
Absorption Peak
LEADING
EDGE
Recovery Curve
Energy Landscape
Potential Wells
Concentration Gradient
Voltage-Gated Channel
Phylogenetic Trees
GRAND
PARTITION
POTASSIUM CURRENT
EXCITED
STATE
False Positives
False Negatives
FIRST
PASSAGE
MICROSTATE
WORM-LIKE CHAIN
POLAR REGION



Organization of Biological Networks

19

Overview: In which statistical mechanics is used to study gene regulation

Specific genes are used only when and where they are needed. For example, we have made much of the classic example of the *lac* operon, which governs the enzymes responsible for lactose digestion. Similar control is exercised over genes in other bacteria, archaea, and eukaryotes. The tools worked out throughout the book leave us poised to consider important quantitative questions about gene regulation such as: how much is a given gene expressed, where in the cell (or the organism) is that gene expressed, and at what time during the cell cycle (or life history) of the organism? The key tools we will use to study these questions are statistical mechanics and rate equations. The statistical mechanical approach will use the probability of promoter occupancy as the key quantity of interest, whereas the rate equation approach will examine the concentrations of protein products over time. These same techniques will also be used to examine signaling with special emphasis on the "decisions" cells make about where to go.

> "The human mind cannot go on forever accumulating facts which remain unconnected and without any mutual bearing and bound together by no law."
>
> **Alfred Russel Wallace**

19.1 Chemical and Informational Organization in the Cell

Many Chemical Reactions in the Cell are Linked in Complex Networks

The reality of the chemical reactions that take place in the cell are a far cry from the relatively sterile and simple kinetic processes described in Chapter 15. In the discussion given there, we showed how to write the time evolution of the concentrations of a set of reactants and products. That theoretical machinery provides an appealing and useful picture for characterizing many of the beautiful *in vitro* experiments

that have powered solution biochemistry. However, biochemistry in living cells has reactants and products linked in a complex set of lineages of biblical proportions where *A* begets *B*, which begets *C*, which in turn begets *D*, and so on, with the added nonanthropomorphic complication that *Z* might just beget *A* again. Indeed, the fact that *Z* can act back on *A* reflects the presence of feedback, which makes the dynamics even richer. Two of the most important classes of reaction that are central to the functioning of cells are those associated with gene regulation and signaling. Indeed, one of the features that most completely distinguishes the chemistry of a cell from that of solution biochemistry is the way in which the reactants are tuned by up- and down-regulation. Similarly, the reactions of the cell are also stimulated by external cues in the form of signaling cascades. In this chapter, we consider regulation and signaling by using a variety of tools developed throughout the book.

Genetic Networks Describe the Linkages Between Different Genes and Their Products

One of the most intriguing reasons why the chemistry of the cell cannot be viewed as a bag of reactants and products is the fact that this chemistry is under the strict control of the genetic machinery of the cell. In particular, if left to its own devices, some particular chemical pathway in the cell might just travel a path to eventual equilibrium. On the other hand, because of both external and internal cues, the machinery of the cell can receive orders via signaling pathways that lead, in turn, to the expression of some gene that results in a new reactant in the original chemical pathway that sends it off in some new direction.

The description of the informational pathways that dictate the cellular concentration profiles in both space and time of the various chemical reactants of interest is founded upon a higher level of abstraction. In particular, there are networks of genes that are linked together in sometimes horrifyingly complex arrays such as that shown in Figure 19.1. This network is an example of a particularly well-characterized genetic network that participates in the embryonic development of sea urchins. One important take-home message concerning this network is that it is a typical network and should leave the reader with a sense of the implied chemical complexity of these systems. In general, genetic networks like that shown in Figure 19.1 make no reference either to the passage of time or to the quantitative distributions of the molecules that mediate these networks. Rather, these networks are an abstraction that shows how genes (and their products) are linked to each other in both space and time. On the other hand, it is important to bear in mind that beneath the surface of these wiring diagrams are actual concentrations of the molecular players of these informational pathways.

Developmental Decisions Are Made by Regulating Genes

Often, genetic networks serve as the basis of the developmental decisions that send a cell or collections of cells down some developmental path. One of the intriguing features of multicellular organisms is that despite the overwhelming cellular diversity, generally, each cell carries the same genetic baggage. However, in general, cells only express a certain fraction of all the available genes. This differentiation is the

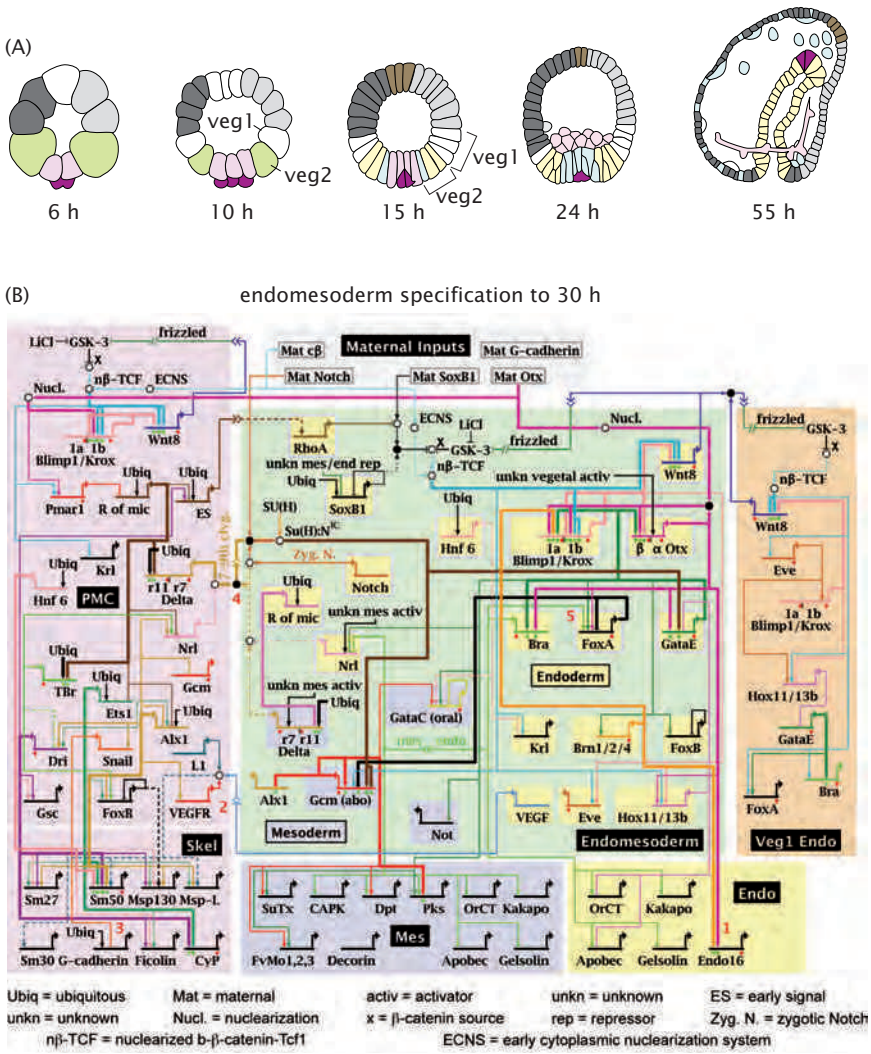


Figure 19.1: Genetic network associated with control of the developmental pathway of the sea urchin embryo. (A) Schematic of stages in the embryonic development of the sea urchin. (B) Genetic network associated with sea urchin development. (Adapted from S. Ben-Tabou de-Leon and E. H. Davidson, *Annu. Rev. Biophys. Biomol. Struct.* 36:191, 2007.)

basis of the development of embryos and the basis of the different structures found in multicellular organisms. The key point is that not all genes are being expressed all the time.

One of the most famous examples of a "developmental decision" is the lambda switch described in Chapter 4 and shown in Figure 4.10 (p. 152). After infecting an *E. coli* bacterium, lambda phage follows one of two developmental pathways. One pathway (the lytic pathway) results in the assembly of new phages and the lysis of the host cell. The second pathway, the lysogenic pathway, involves incorporation of the lambda genome into that of the host cell. Lysogeny can be reversed by damaging the cell with UV light, which triggers lytic replication.

Another compelling example of the role of developmental decisions is that of embryonic development in fruit flies. One of the most celebrated examples is that of the body plan along the long axis of the fly embryo, which is dictated by the distribution of certain proteins along the embryo. Figure 19.2 gives an example of the gradients in four key regulatory proteins that determine the anterior–posterior organization. These proteins determine the pattern of gene expression along the embryo, from which the Eve 2 stripe is the most well-understood example. These ideas were already introduced in Section 2.3.3 (p. 78).

Part of the hard-won wisdom of molecular biology is the recognition that there are many stages in the pathway between DNA and functional protein that can serve as regulatory points. Some of these

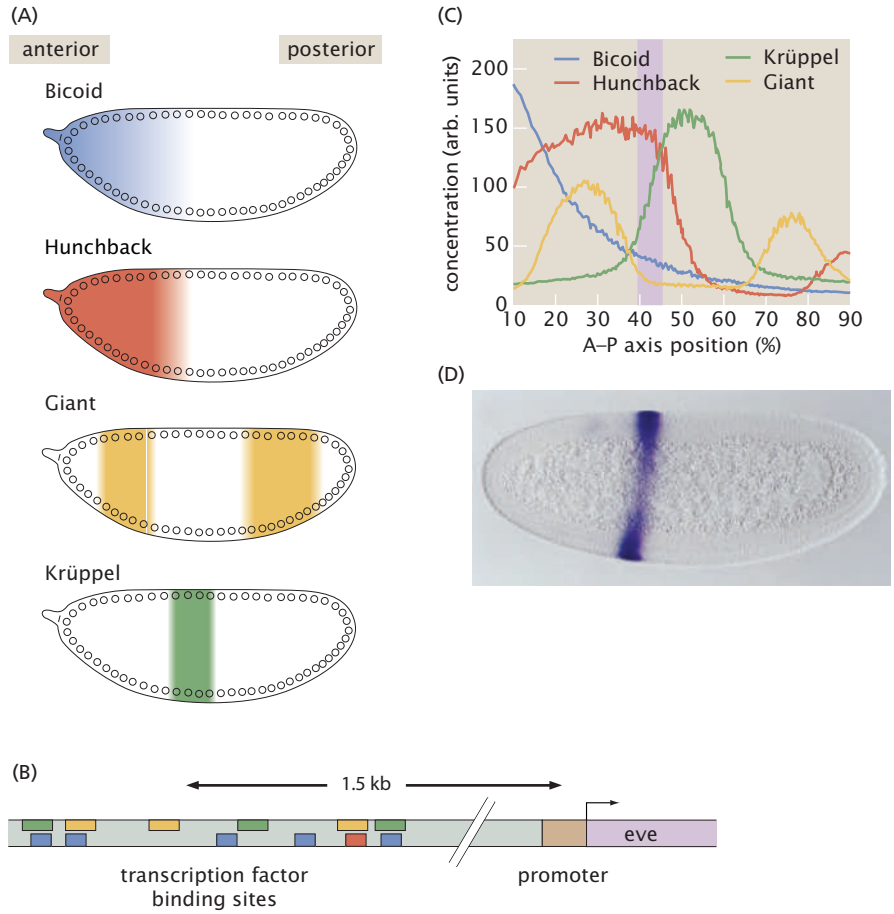


Figure 19.2: Regulatory proteins in the *Drosophila* embryo. The anterior–posterior (A–P) patterning of the fruit fly is dictated by genes that are controlled by spatially varying concentrations of transcription factors. (A) Schematic of the main transcription factors involved in the regulation of stripe 2 of expression of the *even-skipped* gene (eve). (B) Regulatory region of the stripe 2 of the *even-skipped* gene where the binding sites for each transcription factor have been identified. The binding site color on the DNA corresponds to the transcription factor color in (A). (C) Spatial profile of the morphogen gradients measured using immunofluorescence. The purple shaded region corresponds to the striped region shown in (D). (D) Resulting pattern of expression of the regulatory region shown in (B). (B, Adapted from S. Small et al., *EMBO J.* 11:4047, 1992.; C, adapted from E. Myasnikova et al., *Bioinformatics* 17:3, 2001; D, adapted from S. Small et al., *Dev. Biol.* 175:314, 1996.)

different regulatory mechanisms are shown in Figure 6.7 (p. 245). For the purposes of the present discussion, we will focus on one of the most common regulatory mechanisms, namely, transcriptional control, where the key decision that is made is whether or not to produce mRNA.

Gene Expression Is Measured Quantitatively in Terms of How Much, When, and Where

One of our main arguments is that gene expression is a subject that has become increasingly quantitative. In particular, it is now common to measure how much a given gene is expressed, when it is expressed, and where it is expressed. To carry out such measurements, there are a number of useful tools.

Experiments Behind the Facts: Measuring Gene Expression

Quantitative measurement of gene expression can be made at many stages between the decision to start transcription and the emergence of a functional protein product. As noted earlier, such measurements have provided a quantitative window on how much a given gene is expressed, where it is expressed spatially, and when.

One important way to characterize the activity of a gene is by virtue of its protein products. In particular, if the gene product has enzyme activity, that activity can be assayed as a reporter of the extent to which the gene has been expressed

EXPERIMENTS

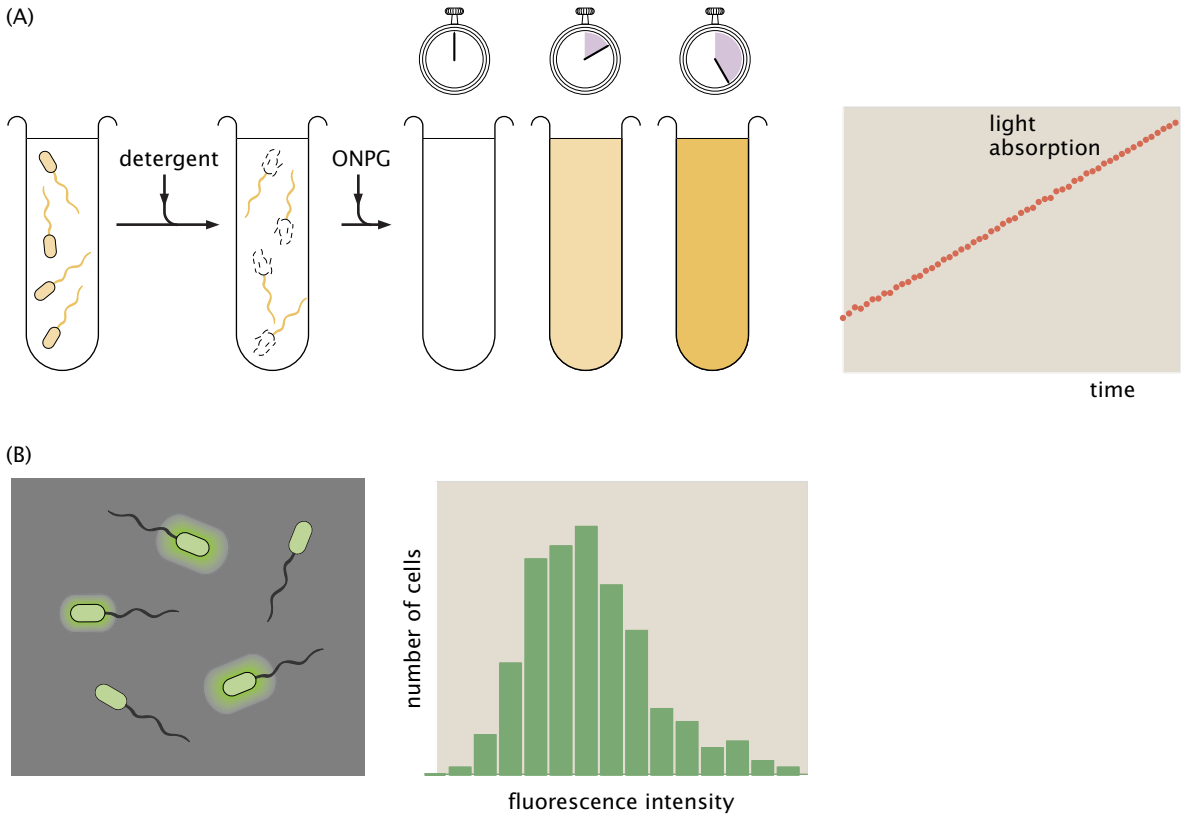


Figure 19.3: Measurement of gene expression. (A) Measurement of gene expression as a result of enzymatic activity. The promoter of interest drives the expression of an enzyme that can cleave a molecule that in the cleaved state is colored. The resulting rate of increase in light absorption is related to the amount of enzyme present in the cells. (B) The promoter of interest drives the expression of a fluorescent protein such as GFP. The amount of fluorescence per cell reports the extent of expression of the gene of interest.

as shown in Figure 19.3(A). Recall that β-galactosidase is the enzymatic product of the *lac* operon, as shown in Figure 4.13 (p. 155), and that the action of this enzyme is to clip lactose molecules. One of the impressive legacies of years of work on this system is a battery of substrates that respond differently to the enzymatic cleavage. One such substrate (ONPG) turns yellow upon cleavage, and measuring the rate at which a solution becomes yellow optically can provide a window on gene expression since it is proportional to the amount of enzyme (over some region of concentrations). By measuring the absorbance at the appropriate wavelengths, one obtains a picture of the amount of active enzyme. Such measurements are typically done on populations of cells. They also require lysing the cells, which means that only end-point assays can be performed with this technique. On the other hand, the sensitivity of this method is superb—to the point where the activity of less than one β-galactosidase molecule per cell can easily be measured. To carry out this kind of assay usually requires routine cloning in which sequences encoding the enzyme are inserted into the genome under the control of the transcription factors of interest.

From a molecular biology perspective, this same strategy of inserting a reporter into the gene of interest can be followed, but with the difference that the "reporter" molecule is a

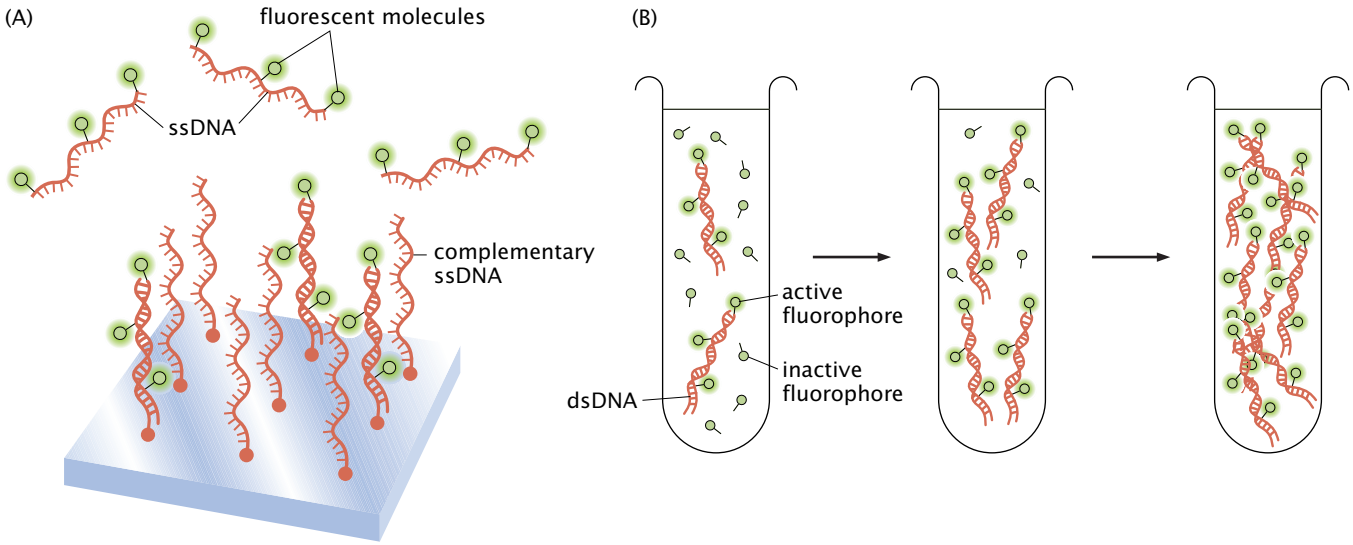


Figure 19.4: Measurement of mRNA concentration. (A) A DNA microarray uses a collection of different molecules on the surface of a slide, each of which has a sequence complementary to the mRNA (or reverse-transcribed ssDNA) associated with the gene of interest. By measuring how much hybridization there is between the sample and the molecules on the surface, one can count the mRNAs. (B) Quantitative PCR uses a template molecule that is produced from the mRNA using reverse transcription. The amount of template determines how many cycles of PCR it will take to reach a critical threshold of amplified DNA using fluorescence as a readout.

fluorescent molecule such as GFP rather than an enzyme. This case is shown in Figure 19.3(B). Relative fluorescence levels of reporters such as GFP are easy to characterize. As shown in Figure 3.3 (p. 93), GFP can be used to track the level of gene expression as a function of time in single living cells. This reporter has its disadvantages, as such fluorescent proteins are subject to photobleaching. Additionally, as we will see in the Computational Exploration on extracting levels of gene expression, the natural constituents of cells have an intrinsic fluorescence, which results in a cellular autofluorescence background that can potentially contaminate the readout from the GFP reporter.

A second scheme for characterizing the extent to which a given gene is expressed is by measuring how much mRNA from the gene of interest is present in the cell. One of the tools of choice for such measurements is the DNA microarray. DNA microarrays are built by labeling a surface with an array of different DNA molecules, each patch of which has small, single-stranded DNA (ssDNA) molecules with the same sequence, as shown in Figure 19.4. These sequences are chosen to be complementary to an entire battery of sequences corresponding to the genes of interest in the experiment. Cells are then broken up and their RNA (or DNA copies made from the RNA) is allowed to flow across the array and hybridize with the molecules on the surface. The various molecules extracted from the cell have been fluorescently labeled, so by looking at the fluorescence intensity at each point on the array, it is possible to read off how much RNA was present.

Another scheme for characterizing the amount of RNA is to use quantitative PCR. Once again, the cell is lysed and the mRNA molecules are turned into DNA using a reverse transcription reaction. Then these molecules are used as templates

in a PCR, and it is seen how many cycles of PCR are needed before the quantity of DNA in the reaction exceeds some threshold. This cycle value is a direct reflection of the number of starting molecules, since starting with lots of template DNA will result in many more molecules at low cycle numbers than will starting with very little material. With quantitative PCR, one can detect mRNA copy numbers as low as 10.

Finally, with the advent of new sequencing technologies that make it possible to generate millions of sequence reads at a reasonable price, it has become commonplace to just sequence the complete mRNA content of cells. By doing so, one can simply count the number of mRNA molecules within the cell corresponding to the various genes of interest, resulting in genome-wide information in one experiment. As with the previous methods, this approach requires the conversion of all cellular mRNA into DNA in order to be sequenced.

As will be described in the remainder of this chapter, a useful surrogate for the actual question of the extent to which a given gene is expressed is to ask whether or not the promoter for the gene of interest is occupied. There are many *in vitro* and *in vivo* methods for finding out whether or not the promoter is bound to polymerase. *Chromatin immunoprecipitation* and *DNA footprinting* are two methods that are sensitive to promoter occupancy. For DNA footprinting, the idea is that the part of DNA where the transcriptional apparatus is bound will react differently when the system is exposed to agents such as restriction enzymes. The most common procedure is to try to digest the DNA using a restriction enzyme. It will not be able to access the DNA over which RNA polymerase is situated, leaving a "footprint" of a longer piece of DNA that can be easily detected. For chromatin immunoprecipitation, DNA is covalently crosslinked to bound proteins using reactive chemicals, and then the DNA is sheared into small fragments. Antibodies specific to polymerase are used to isolate the molecules of polymerase with their associated DNA fragments. Then, the chemical crosslinks are reversed, and the DNA fragments associated with polymerase are sequenced. This same technique can be modified to identify the specific DNA sequences that are associated with any other specific DNA-binding protein of interest, such as a repressor protein. These different methods can also be cleverly combined with the new sequencing technologies in order to perform such assays at the genome-wide scale, as we will see further below.

19.2 Genetic Networks: Doing the Right Thing at the Right Time

In "thermodynamic" models of gene expression, attention is focused on the probability that the promoter is occupied by RNA polymerase. In Section 6.1.2 (p. 244), we showed how the "bare" problem of polymerase molecules interacting with DNA could be solved using simple ideas from statistical mechanics. However, the shortcoming of that approach is that it ignores the existence of molecular gatekeepers

that exercise strict control over the occupancy of promoters. We begin our dissection of gene expression with a consideration of these gatekeepers, which are known as transcription factors.

Promoter Occupancy Is Dictated by the Presence of Regulatory Proteins Called Transcription Factors

In Figure 6.8 (p. 246) we showed a cartoon of some gene of interest and the promoter and DNA upstream from it. As a first cut at the problem of promoter occupancy, we examined the probability of RNA polymerase binding as a competition between this promoter and nonspecific sites, both of which can be occupied by polymerase molecules. We now expand that discussion to account for the presence of a host of important accessory proteins that can either enhance (activate) or reduce (repress) the probability of promoter occupancy.

As before, we focus primarily on bacteria. What this means concretely is that we will treat RNA polymerase as a single molecule and ask the precise mathematical (but biologically oversimplified) question of whether or not the promoter is occupied by such an RNA polymerase molecule. In the eukaryotic case, this question is less easily posed, since the basal transcription apparatus consists of many parts, all of which need to be present simultaneously in order to start transcription.

19.2.1 The Molecular Implementation of Regulation: Promoters, Activators, and Repressors

Repressor Molecules Are the Proteins That Implement Negative Control

One of the key control mechanisms of genetic networks is negative regulation of transcription. What this means is that the decision to express the gene of interest is made very early on in the set of processes leading from DNA to protein, namely, at the point where RNA is synthesized. If there is little or no mRNA that codes for a given protein, then clearly the ribosomes are in no position to produce the corresponding protein. The molecular implementation of negative control is through protein molecules known as repressors, such as the Lac repressor introduced in Figures 4.13 (on p. 155) and 8.19 (on p. 334). In the case of bacteria, repressors can often be viewed as carrying out a blocking action in the sense that through DNA–protein interactions, they occupy the DNA in a region (called the operator) that overlaps the region where RNA polymerase binds (the promoter). The action of such repressor molecules is illustrated schematically in Figure 19.5. Note that the activity of repressors can, in turn, be regulated by small molecules, or inducers, that can bind and generate a conformational (or allosteric) change that alters the binding probability of the transcription factor for the DNA. Later in this chapter, we give a statistical mechanical interpretation of such cartoons.

It is important to recall that the point of cartoons like that in Figure 19.5 is to convey a conceptual picture and not a detailed molecular rendering of the explicit action of the various molecular participants. On the other hand, the fact that such cartoons can be constructed in the first place is often the result of having digested the significance of hard-won structural determinations from X-ray crystallography. Indeed, sometimes, not only the structures of the bare

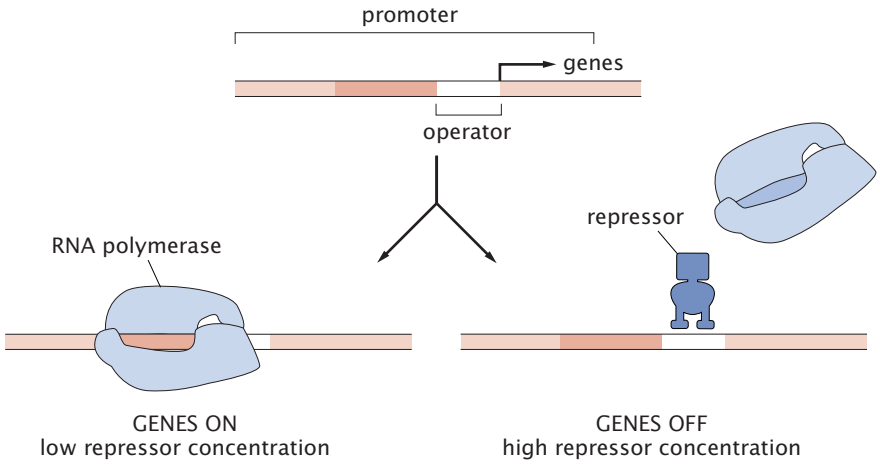


Figure 19.5: The process of repression. Cartoon representation showing the action of repressor molecules in forbidding RNA polymerase from binding to its promoter, or alternatively, if bound, from initiating transcription.

repressors are known, but even the structures of these repressors when complexed with DNA. In fact, there are a variety of structural implementations of repression, some famed examples of which are shown in Figure 19.6.

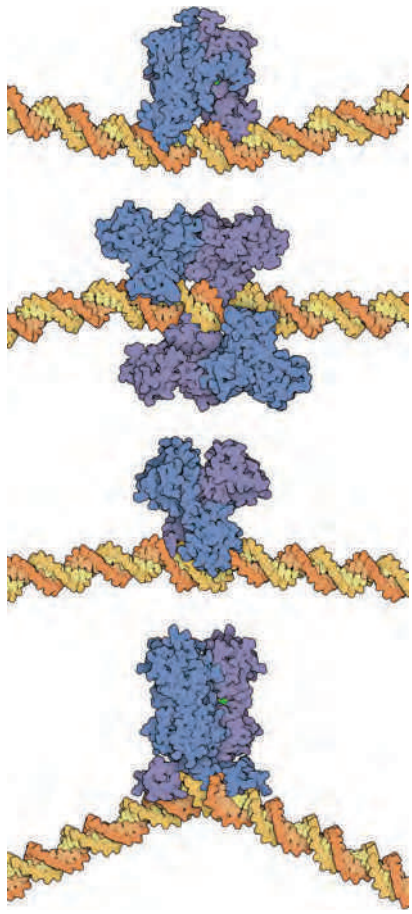


Figure 19.6: Examples of repressor molecules interacting with DNA. From top to bottom, the repressors are TetR (pdb 1QPI), IdeR (pdb 1U8R), FadR (pdb 1HW2), and PurR (pdb 1PNR). The point of the figure is to give an impression of the relative sizes of repressors and their target regions on DNA and to illustrate how these transcription factors deform the DNA double helix in the vicinity of their binding site. These drawings are renditions of actual structures from X-ray crystallography. (Courtesy of D. Goodsell.)

Activators Are the Proteins That Implement Positive Control

A second key mechanism for altering the extent to which a given gene is expressed is known as positive regulation of transcription, or, more provocatively, regulated recruitment. Here too, the idea is that the overall process of protein synthesis of a given gene product is regulated very early on where an accessory molecule enhances the probability of promoter occupancy by RNA polymerase. This mechanism is built around the idea of proteins other than RNA polymerase that bind to DNA and increase the probability that the RNA polymerase itself will bind the promoter. Just as repressors interfere with the ability of RNA polymerase to bind to its promoter, activators bind in the vicinity of the promoter and have adhesive interactions with RNA polymerase itself that enhance the likelihood of RNA polymerase binding. The key point is that the RNA polymerase molecule interacts not only with the DNA to which it is bound, but also through "glue-like" interactions with the activator molecule. A cartoon representation of the process of regulated recruitment (that is, activation) is shown in Figure 19.7.

As with the study of repressors, structural biology has permitted a range of atomic-level insights into the mechanisms of transcriptional activation. Figure 19.8 provides a gallery of some key activators, reveals their sizes relative to the DNA molecule, and illustrates the way in which they distort and occlude the DNA when bound.

Genes Can Be Regulated During Processes Other Than Transcription

Our discussion will focus primarily on transcriptional regulation. On the other hand, as shown in Figure 6.7 (p. 245), there are many points along the route connecting DNA to its protein products where gene expression can be controlled. Two of the most obvious and important ways in which the concentration of active protein is controlled are through the post-translational modifications phosphorylation and protein degradation. In addition, in recent years, a whole host of regulatory RNAs have been discovered that have greatly enriched the study

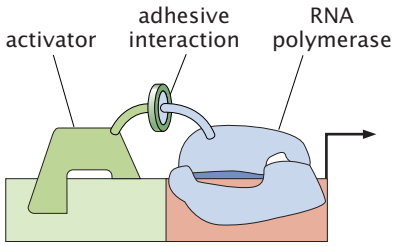


Figure 19.7: The process of activation. Schematic of the way in which activator molecules can recruit the transcription apparatus. Though both the activator and RNA polymerase have their own private interaction energies with the DNA, the enhancement in their occupancies is mediated by the adhesive interaction between them.

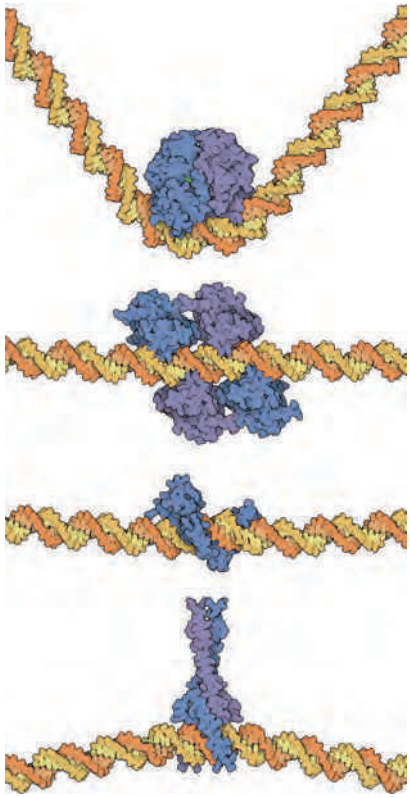


Figure 19.8: Structures of activator molecules. From top to bottom, the activators are CAP (pdb 1CGP), p53 tumor suppressor (pdb 3KMD), zinc finger DNA-binding domain (pdb 2GLI), and leucine zipper DNA-binding domain (pdb 1AN2). (Courtesy of D. Goodsell.)

of regulatory biology. For the moment, we focus on the way in which p_{bound} (the probability that the promoter is occupied by RNA polymerase) can be altered through the action of transcription factors such as repressors and activators.

19.2.2 The Mathematics of Recruitment and Rejection

Recruitment of Proteins Reflects Cooperativity Between Different DNA-Binding Proteins

One of the key general ideas that pervade the description of transcriptional control (and beyond) is the idea of molecular recruitment. In the anthropomorphic terms suggested by the word "recruitment," the idea is that a given molecule that is bound on DNA summons some second molecule to the DNA, where it can then perform its task. For example, we think of RNA polymerase being summoned by some activator molecule such as a transcription factor (and vice versa) and exemplified by the CAP protein in the case of the *lac* operon. Though this colorful language is suggestive and conjures up a useful physical picture, from the perspective of the rules of statistical mechanics, this is nothing more than the well-worn idea of cooperativity cloaked in different verbal clothing.

Activators are proteins that regulate transcription by binding to a specific site on the DNA so as to recruit an RNA polymerase onto a nearby promoter site. It has been suggested that weak, nonspecific binding of the activator protein and the RNA polymerase can greatly enhance the probability of the polymerase binding to DNA, even for the very low concentrations of activator proteins typical of the cellular environment. To assess the feasibility of this strategy, we compute the probability of the polymerase being bound in the presence of an activator protein using a simple model that is depicted in cartoon form in Figure 19.9. The basic point of this cartoon is to show the different allowed states of polymerase and activator molecules and to use this enumeration of states to compute the probability that the promoter will be occupied. Indeed, this is the same "states-and-weights" mentality used throughout the book.

The first step in our analysis of this problem is to write the total partition function. Note that the partition function is obtained by summing over all of the eventualities associated with the activators and polymerase molecules being distributed on the DNA (both nonspecific sites and the promoter). As shown in Figure 19.9, there are four classes of outcomes, namely, both the activator site and promoter unoccupied, just the promoter occupied by polymerase, just the activator binding site occupied by activator, and, finally, both of the specific sites occupied. This is represented mathematically as

$$Z_{\text{tot}}(P, A; N_{\text{NS}}) = \underbrace{Z(P, A; N_{\text{NS}})}_{\text{empty promoter}} + \underbrace{Z(P-1, A; N_{\text{NS}}) e^{-\beta \varepsilon^{\text{S}}_{\text{pd}}}}_{\text{RNAP}} + \underbrace{Z(P, A-1; N_{\text{NS}}) e^{-\beta \varepsilon^{\text{S}}_{\text{ad}}}}_{\text{activator}} + \underbrace{Z(P-1, A-1; N_{\text{NS}}) e^{-\beta(\varepsilon^{\text{S}}_{\text{ad}} + \varepsilon^{\text{S}}_{\text{pd}} + \varepsilon_{\text{pa}})}}_{\text{RNAP + activator}}. \quad (19.1)$$

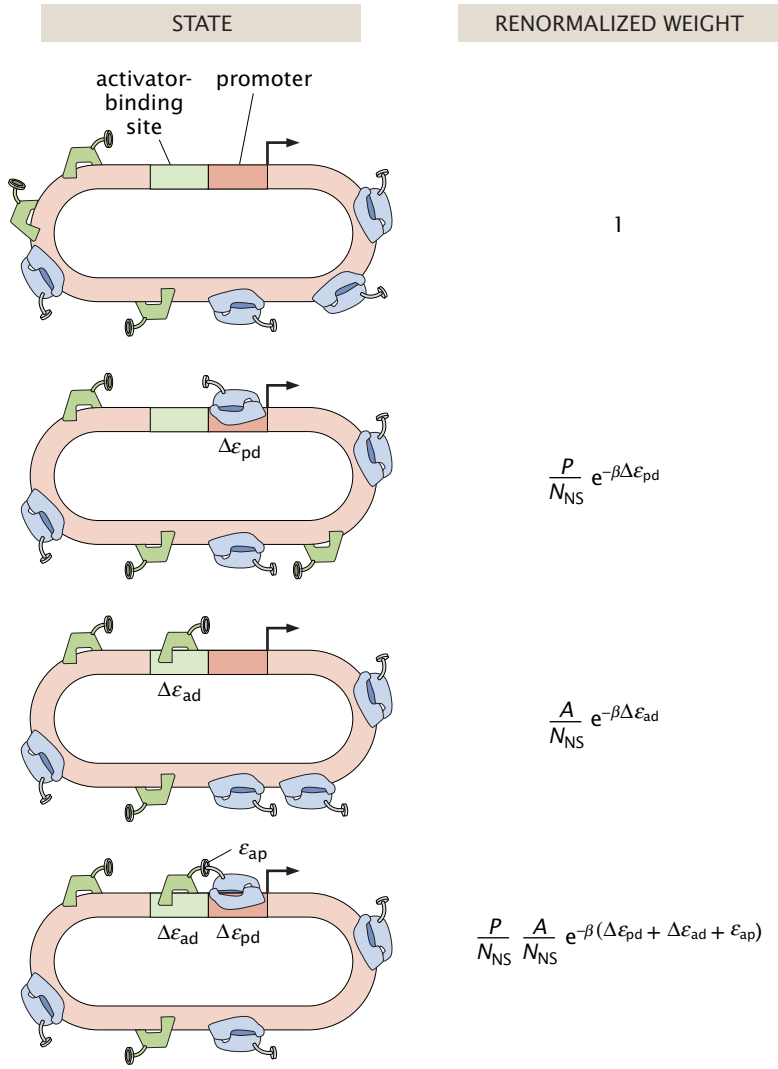


Figure 19.9: Schematic representation of the simple statistical mechanical model of recruitment. The states-and-weights diagram shows the different binding scenarios in the vicinity of the promoter of interest and the corresponding renormalized statistical weights obtained using statistical mechanics. We make the simplifying assumption that the nonspecific binding energy is *constant*. The large circular DNA is a cartoon representation of the bacterial genome.

Note that, notationally, the meaning of $Z(P, A; N_{NS})$ is that it is the partition function for P polymerase molecules and A activator molecules to be bound on the N_{NS} nonspecific sites and is given by

$$Z(P, A; N_{NS}) = \underbrace{\frac{N_{NS}!}{P!A!(N_{NS} - P - A)!}}_{\text{number of arrangements}} \times \underbrace{e^{-\beta P \varepsilon_{pd}^{NS}} e^{-\beta A \varepsilon_{ad}^{NS}}}_{\text{weight of each state}}. \tag{19.2}$$

We have also introduced the notation ε_{pa} to account for the "glue" interaction between the polymerase and activator. Like in Section 6.1.2 (p. 244) for the case of RNA polymerase, we introduce ε_{ad}^{S} and ε_{ad}^{NS} to characterize the binding energy of activator with its specific and nonspecific DNA targets, respectively. Our expression involves a number of terms of the general form

$$\frac{N_{NS}!}{P!A!(N_{NS} - P - A)!} \times e^{-\beta P \varepsilon_{pd}^{NS}} e^{-\beta A \varepsilon_{ad}^{NS}}. \tag{19.3}$$

As we did earlier, we invoke a simplifying strategy that depends upon the fact that $N_{NS} \gg A + P$ and hence there will be almost zero chance of RNA polymerase and the activator finding each other on

the same nonspecific site on the DNA. This permits the approximation $N_{NS}!/(N_{NS}-A-P)! \approx (N_{NS})^{A+P}$ introduced in Section 6.1.2 (see p. 244).

To compute the probability of promoter occupancy, we construct the ratio of all of those outcomes that are favorable (that is, polymerase bound to the promoter) to the total set of outcomes ($Z_{tot}(P, A; N_{NS})$), namely,

$$p_{bound}(P, A; N_{NS}) = \frac{Z(P-1, A; N_{NS})e^{-\beta\varepsilon_{pd}^{S}} + Z(P-1, A-1; N_{NS})e^{-\beta(\varepsilon_{ad}^{S}+\varepsilon_{pd}^{S}+\varepsilon_{pa})}}{Z_{tot}(P, A; N_{NS})}. \quad (19.4)$$

We now propose to simplify this result by dividing both numerator and denominator by the numerator, resulting in

$$p_{bound}(P, A; N_{NS}) = \frac{1}{1 + [N_{NS}/PF_{reg}(A)]e^{\beta\Delta\varepsilon_{pd}}}, \quad (19.5)$$

where we introduce the regulation factor $F_{reg}(A)$, which is given by

$$F_{reg}(A) = \frac{1 + (A/N_{NS})e^{-\beta\Delta\varepsilon_{ad}}e^{-\beta\varepsilon_{ap}}}{1 + (A/N_{NS})e^{-\beta\Delta\varepsilon_{ad}}}, \quad (19.6)$$

and where we have defined $\Delta\varepsilon_{pd} = \varepsilon_{pd}^{S} - \varepsilon_{pd}^{NS}$ and $\Delta\varepsilon_{ad} = \varepsilon_{ad}^{S} - \varepsilon_{ad}^{NS}$. The details of the derivation are left to the problems at the end of the chapter. Note that in the limit that the adhesive interaction between polymerase and activator goes to zero, the regulation factor itself goes to unity. Further, note that for negative values of this adhesive interaction (that is, activator and polymerase like to be near each other), the regulation factor is greater than 1, which is translated into an effective increase in the number of polymerase molecules. The probability of RNA polymerase binding as a function of the number of activators is plotted in Figure 19.10.

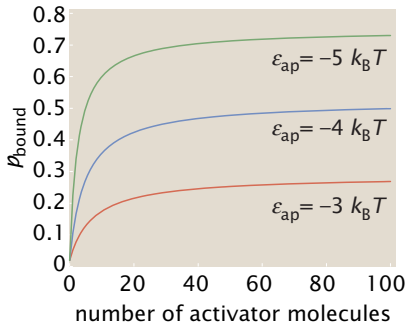


Figure 19.10: Illustration of the recruitment concept. This plot shows the probability of binding when the number of polymerase molecules is $P = 500$ and the binding parameters are $\Delta\varepsilon_{pd} = -5.3\ k_BT$ and $\Delta\varepsilon_{ad} = -13.12\ k_BT$. The three curves correspond to different choices of the adhesive interaction energy between polymerase and the activator.

The Regulation Factor Dictates How the Bare RNA Polymerase Binding Probability Is Altered by Transcription Factors

One of the intriguing claims that we will make is that a simple change in the effective number of RNA polymerase molecules ($P \rightarrow P_{eff}$) will suffice to capture the action of regulatory chaperones such as activators and repressors. This interpretation of the meaning of the regulation factor is shown in Figure 19.11. As a result of the presence of activators, it is as though the number of RNA polymerase molecules has been changed from P to $F_{reg}P$. For the case of activators, the regulation factor is greater than 1 and leads to an effective increase in the number of polymerase molecules. By way of contrast, we will show below that when repressors are present, they result in a regulation factor that is less than 1 and a concomitant decrease in the effective number of polymerase molecules.

In order for our calculations to really carry weight, we need to examine what they have to say about experiments. One of the primary measurables in *in vivo* experiments on regulation is the relative

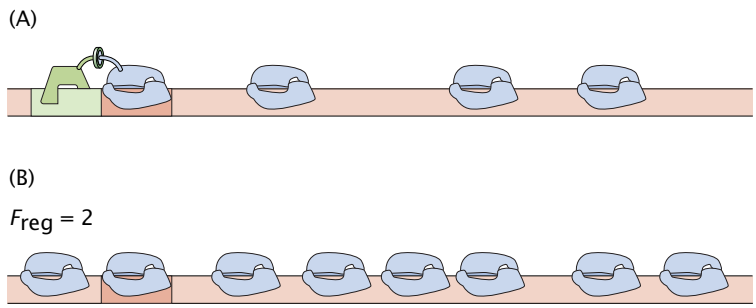


Figure 19.11: Regulation factor and the effective number of polymerase molecules. The presence of activators is equivalent to a problem with just polymerase molecules but a larger number of them. (A) The "bare" problem with activators and polymerase present. (B) The "effective" problem in which the presence of activators is treated as a change in the number of polymerase molecules.

expression for cases in which the transcription factor of interest is present or not. This qualitative notion is made quantitative by introducing the idea of the fold-change in activity, defined in the activation setting as

$$\text{fold-change} = \frac{p_{\text{bound}}(A \neq 0)}{p_{\text{bound}}(A = 0)} = \frac{1 + (N_{\text{NS}}/P)e^{\beta \Delta \varepsilon_{\text{pd}}}}{1 + [N_{\text{NS}}/PF_{\text{reg}}(A)]e^{\beta \Delta \varepsilon_{\text{pd}}}}. \tag{19.7}$$

What this expression reveals is how much more expression there is in the presence of activators relative to the "basal" state in which there is no activation.

As before, an inherent assumption in this analysis is the idea that the relative change in what is measured (for example, protein product, mRNA concentration, or promoter occupancy) is equal to the relative change in p_{bound}. Figure 19.12 illustrates the fold-change in gene expression for the problem of simple activation with a choice of parameters dictated by *in vitro* experiments for a value of $\Delta\varepsilon_{\text{ad}}$ in conjunction with an educated guess for ε_{ap} that results in typical fold-changes in activity reported *in vivo* of about 50. Note that a weak promoter satisfies the condition $(N_{\text{NS}}/P)e^{\beta \Delta \varepsilon_{\text{pd}}} \gg 1$, which implies that the fold-change in activity can be rewritten as

$$\text{fold-change} \approx F_{\text{reg}}(A). \tag{19.8}$$

Here we have also assumed that $(N_{\text{NS}}/PF_{\text{reg}})e^{\beta \Delta \varepsilon_{\text{pd}}} \gg 1$, which means that the promoter is not too strong even in the regulated case. The conclusion is that in the case of a weak promoter the actual details of the promoter, such as its binding energy, factor out of the problem.

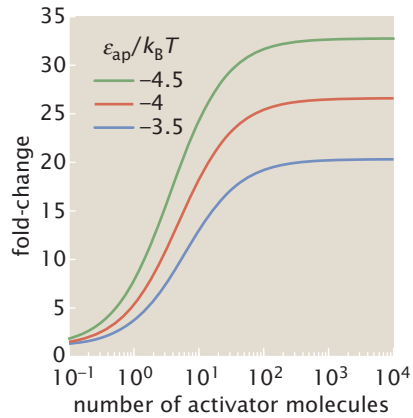


Figure 19.12: Fold-change due to activators. Fold-change in gene expression as a function of the number of activators for different activator–RNA polymerase interaction energies using $P = 500$, $\Delta\varepsilon_{\text{pd}} = -5.3\ k_BT$, and $\Delta\varepsilon_{\text{ad}} = -13.12\ k_BT$ based on *in vitro* measurements.

Activator Bypass Experiments Show That Activators Work by Recruitment

The simple picture of regulated recruitment introduced here is based in part upon a series of classic experiments known as activator bypass experiments. The key idea of such experiments is shown in Figure 19.13. These experiments involve a mix-and-match approach where the DNA-binding domain from one protein is fused with the activator domain of a second protein. A second version of this experiment is based upon direct tethering of the activator and the polymerase. After making the activator bypass constructs, it was found that the gene of interest was still activated. Our ambition here is to consider these experiments more quantitatively and to note that, if viewed from a mathematical perspective, these two classes of experiments lead to different quantitative outcomes that can be used to further test the full range of validity of the notion of regulated recruitment.

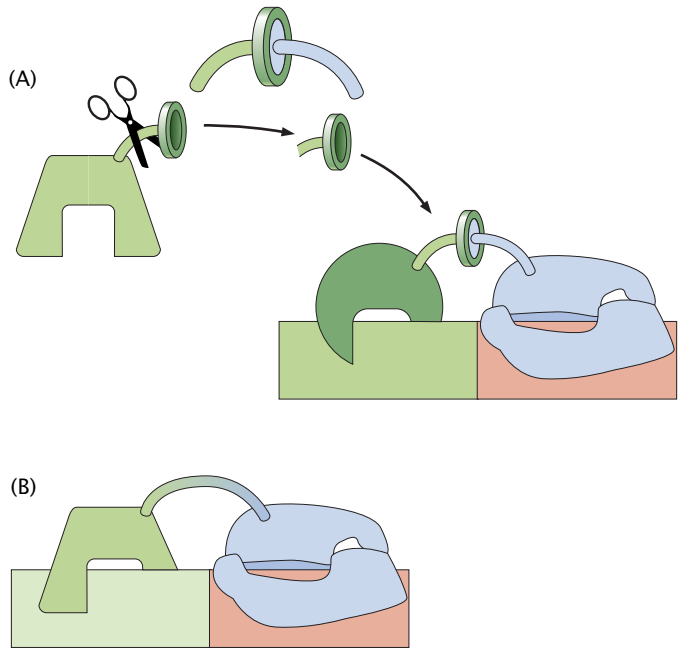


Figure 19.13: Schematic of activator bypass experiments. (A) Activator bypass type 1 in which activation is mediated by proteins with designer DNA-binding regions. (B) Activator bypass type 2 in which the activator is tethered directly to polymerase.

We have already worked out the regulation factor that is associated with activator bypass type 1 experiments. The only change relative to Equation 19.6 is that, by using different proteins, quantities such as $\Delta\varepsilon_{ad}$ and ε_{pa} will have different numerical values, which means that the actual level of activation can be different in this experiment relative to its "wild-type" value. On the other hand, the entire functional form for the regulation factor is different in the case of activator bypass type 2. In this case, there are only two states we really need to consider, namely, polymerase with and without tethered activator bound at the promoter with weights $(P/N_{NS})e^{-\beta(\Delta\varepsilon_{pd}+\Delta\varepsilon_{ad})}$ and 1, respectively. This implies that the probability that polymerase will be bound is

$$p_{bound}(P; N_{NS}) = \frac{1}{1 + (N_{NS}/P)e^{\beta\Delta\varepsilon_{ad}}e^{\beta\Delta\varepsilon_{pd}}}. \tag{19.9}$$

This implies, in turn, that the regulation factor takes the particularly simple form

$$F_{reg} = e^{-\beta\Delta\varepsilon_{ad}}, \tag{19.10}$$

which amounts to the statement that the effective binding energy of polymerase is shifted and nothing more.

Repressor Molecules Reduce the Probability Polymerase Will Bind to the Promoter

The same logic that was introduced above to consider the case of pure activation (that is, recruitment) can be brought to bear on the problem of repression. Once again, we are faced with considering all of the ways of distributing the repressor and RNA polymerase molecules and

it is convenient to introduce the partition function associated with the binding of these molecules to nonspecific sites as

$$Z(P, R : N_{NS}) = \frac{N_{NS}!}{P!R!(N_{NS} - P - R)!} e^{-\beta P \varepsilon_{pd}^{NS}} e^{-\beta R \varepsilon_{rd}^{NS}}, \tag{19.11}$$

which is formally identical to Equation 19.2, but where we have introduced the notation ε_{rd}^{NS} to describe the nonspecific binding of repressor to DNA (ε_{rd}^{S} will be reserved for the specific binding energy of repressor to its operator). In order to write the *total* partition function for all the allowed states, we now need to sum over the states in which the promoter is occupied either by a repressor molecule or by an RNA polymerase molecule. The set of allowed states in this simple model as well as their associated weights are shown in Figure 19.14. Note that in considering this particular model, we do not enter into structural fine points such as whether or not the RNA polymerase can be on its promoter at the same time as the repressor is bound to its operator—the model is intended to be the simplest treatment of the statistical mechanics of the competition between repressors and RNA polymerase.

The total partition function is given by

$$Z_{tot}(P, R; N_{NS}) = \underbrace{Z(P, R; N_{NS})}_{\text{empty promoter}} + \underbrace{Z(P-1, R; N_{NS}) e^{-\beta \varepsilon_{pd}^{S}}}_{\text{RNAP on promoter}} + \underbrace{Z(P, R-1; N_{NS}) e^{-\beta \varepsilon_{rd}^{S}}}_{\text{repressor on promoter}}. \tag{19.12}$$

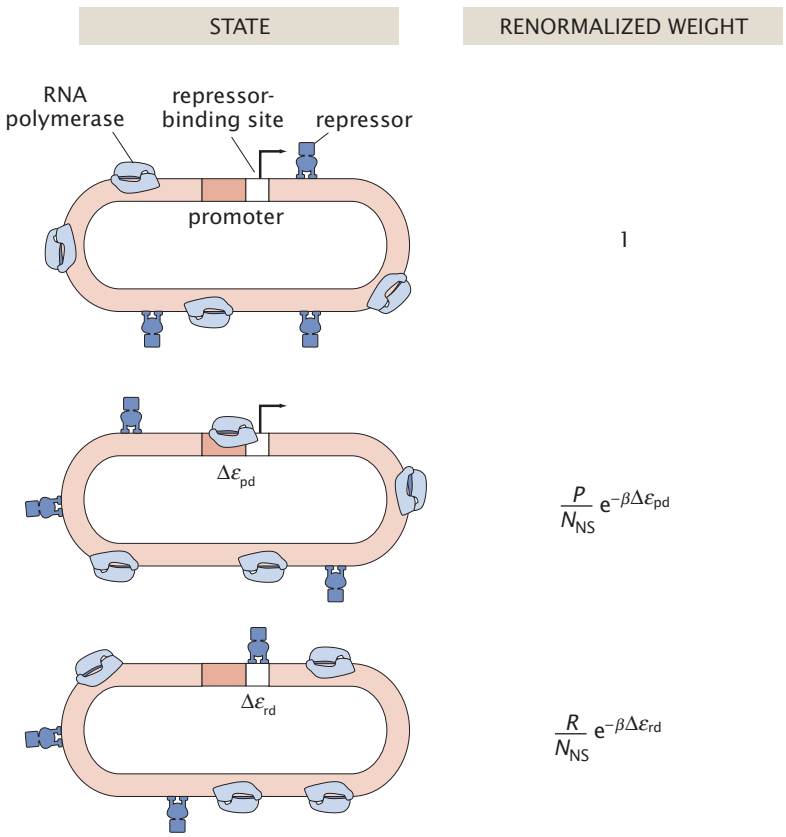


Figure 19.14: States and weights for the case of simple repression. The states of promoter occupancy are empty promoter, RNA polymerase on the promoter, and repressor on the promoter.

This result now provides us with the tools with which to evaluate the probability that the promoter will be occupied by RNA polymerase. This probability is given by the ratio of the favorable outcomes to all of the outcomes. In mathematical terms, that is

$$p_{\text{bound}}(P, R; N_{\text{NS}}) = \frac{Z(P-1, R; N_{\text{NS}})e^{-\beta\varepsilon^{\text{S}}_{\text{pd}}}}{Z(P, R; N_{\text{NS}}) + Z(P-1, R; N_{\text{NS}})e^{-\beta\varepsilon^{\text{S}}_{\text{pd}}} + Z(P, R-1; N_{\text{NS}})e^{-\beta\varepsilon^{\text{S}}_{\text{rd}}}}. \tag{19.13}$$

As argued above, this result can be rewritten in compact form using the regulation factor by dividing top and bottom by $Z(P-1, R; N_{\text{NS}})e^{-\beta\varepsilon^{\text{S}}_{\text{pd}}}$ and by invoking the approximation

$$\frac{N_{\text{NS}}!}{P!R!(N_{\text{NS}}-P-R)!} \approx \frac{N_{\text{NS}}^{\text{P}}}{P!}\frac{N_{\text{NS}}^{\text{R}}}{R!}, \tag{19.14}$$

which amounts to the physical statement that there are so few polymerase and repressor molecules in comparison with the number of available sites, N_{NS}, that each of these molecules can more or less

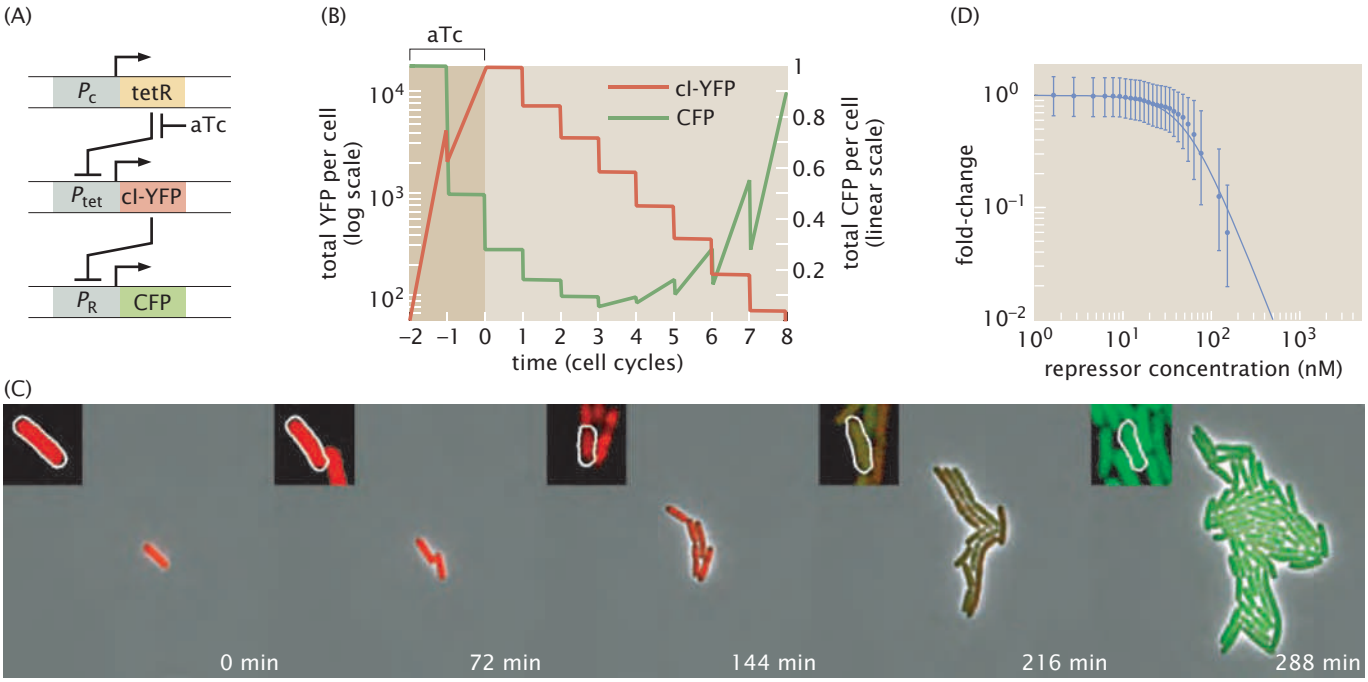


Figure 19.15: Dilution experiment and the measurement of fold-change in repression. (A) Diagram of the circuit. In the absence of the inducer aTc, the repressor TetR shuts down production of the transcription factor cI fused to YFP. This transcription factor, in turn, regulates the expression of the reporter CFP. (B) Schematic of the time course of an experiment. Adding aTc for a short period of time leads to the production of cI-YFP. Upon removal of aTc, no new cI-YFP is produced. As a result, in each new generation, there will be decreasing numbers of cI-YFP per cell, resulting in an ever-higher rate of expression of the downstream CFP gene. This dilution also permits the calibration of YFP fluorescence into absolute numbers of cI-YFP as discussed in the text. (C) Representative snapshots from the time course of an experiment. (D) Fold change (1/repression) as a function of cI repressor concentration measured using the dilution method. (Adapted from N. Rosenfeld et al. *Science* 307: 1962, 2005.)

fully explore those N_{NS} sites. The resulting probability is

$$p_{\text{bound}}(P, R; N_{NS}) = \frac{1}{1 + (N_{NS}/P)e^{\beta(\varepsilon^{S}_{pd} - \varepsilon^{NS}_{pd})}[1 + (R/N_{NS}]e^{-\beta(\varepsilon^{S}_{rd} - \varepsilon^{NS}_{rd})}}. \tag{19.15}$$

This result can be couched in regulation factor language with the observation that the regulation factor itself is given by

$$F_{\text{reg}}(R) = \left(1 + \frac{R}{N_{NS}} e^{-\beta \Delta\varepsilon_{rd}}\right)^{-1}, \tag{19.16}$$

with $\Delta\varepsilon_{rd} = \varepsilon^{S}_{rd} - \varepsilon^{NS}_{rd}$. Note that the regulation factor in the case of repression satisfies the inequality $F_{\text{reg}} < 1$, which can be interpreted as a reduction in the effective number of RNA polymerase molecules. We explore this in more detail in Section 19.2.5 when discussing the particular case of the *lac* operon, though Figure 19.15 gives an example of an extremely elegant measurement of the effect of repression using the beautiful dilution method introduced in the Computational Exploration on p. 46.

COMPUTATIONAL EXPLORATION

Computational Exploration: Extracting Level of Gene Expression from Microscopy Images One way to determine the level of gene expression is to use microscopy images of cells expressing some fluorescent reporter. In this Computational Exploration, the reader is invited to use Matlab to extract the fluorescence intensities from a collection of cells and to use them to determine the fold-change in simple repression.

The logical progression associated with this analysis is introduced schematically in Figure 19.16. Note that we have images of the cells in two different channels. In particular, for each field of view, we have both a phase contrast image and a fluorescence image. Like with the example where we determined the cell cycle time of *E. coli* (p. 100), the first step is to find the cells in an automated fashion using some segmentation scheme. Additionally, we need to choose which one of the two images we want to do the segmentation with. Detecting cells using the fluorescence image is certainly appealing due to the absence of any other fluorescent objects. However, it is clear that for dimmer cells the segmentation might not work as well. As a result, we would risk biasing our segmentation based on the level of expression of the cells, the quantity we are actually interested in measuring! Instead, we choose to segment the phase contrast image, which should, in principle, not be subject to bias resulting from the level of fluorescence within each cell.

Following the procedure outlined in the example on the cell division time in *E. coli* (p. 100), once we have performed the thresholding, we will be left with a mask image with discrete regions that we identify as cells denoted by the different colors in Figure 19.16(C). These ideas are illustrated in the Matlab code associated with this exploration. Once the segmentation

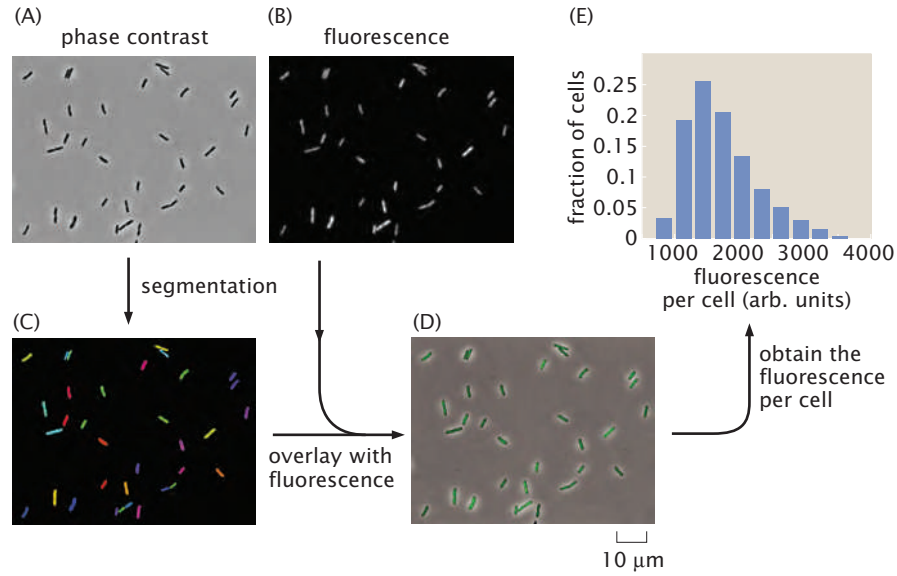


Figure 19.16: Schematic of the image segmentation algorithm to quantify levels of gene expression in bacteria. Two images of bacteria expressing a fluorescent protein are obtained, (A) one in phase contrast and (B) one in fluorescence. The phase contrast image is an imaging scheme that makes it possible to see the bacteria as dark objects. (C) These objects are automatically detected and segmented using computer software that assigns an identity to each segmented bacterium (represented by the different colors). (D) The mask generated by this procedure is applied to the fluorescence image in order to generate an overlay and integrate the fluorescence within the mask of each segmented cell. (E) By repeating this for multiple images and many cells, the distribution of fluorescence per cell can be computed.

process is complete, we can then obtain the fluorescence intensity in each of our cells. To do so, we use the segmented image from the previous step to find the individual cells and then, within each such cell, we ask for the fluorescence intensity of all of the pixels and sum them up. The result is a distribution of fluorescence per cell as shown in Figure 19.16(E). However, there is an extra subtlety that has to be taken into account when obtaining such fluorescence distributions. In particular, because of the intrinsic fluorescence of the cells themselves, there is a spurious contribution to the total fluorescence that we measure, F_{total}, which is given by

$$F_{\text{total}} = F_{\text{reporter}} + F_{\text{cell}}, \tag{19.17}$$

where F_{reporter} is the signal stemming from the fluorescent reporter while F_{cell} is the autofluorescence of the cell. As a result, we need to be able to subtract the cells' average autofluorescence if we want to report only on F_{reporter}. This can be easily done by following the steps outlined in Figure 19.16 and described above, but now for a strain of bacteria that lacks any fluorescent reporter. As a result, we will be able to measure the mean contribution of the cell autofluorescence to the total fluroescence, $\langle F_{\text{cell}} \rangle$, which can then be subtracted from the fluorescence values in the presence of the reporter.

With the fluorescence intensities in hand, we are now prepared to compute the fold-change itself so that we can examine the accord between the model of simple repression presented in Equation 19.16 and the data itself. The logic of this part of the analysis is presented in Figure 19.17. Here the idea is to use our mean fluorescence intensities, corrected for the fluorescence background, for both the regulated and unregulated promoters and then to construct the ratio of these means.

Examples of Matlab code that could be used to perform this Computational Exploration, as well as images of *E. coli* suitable for this analysis, can be found on the book's website.

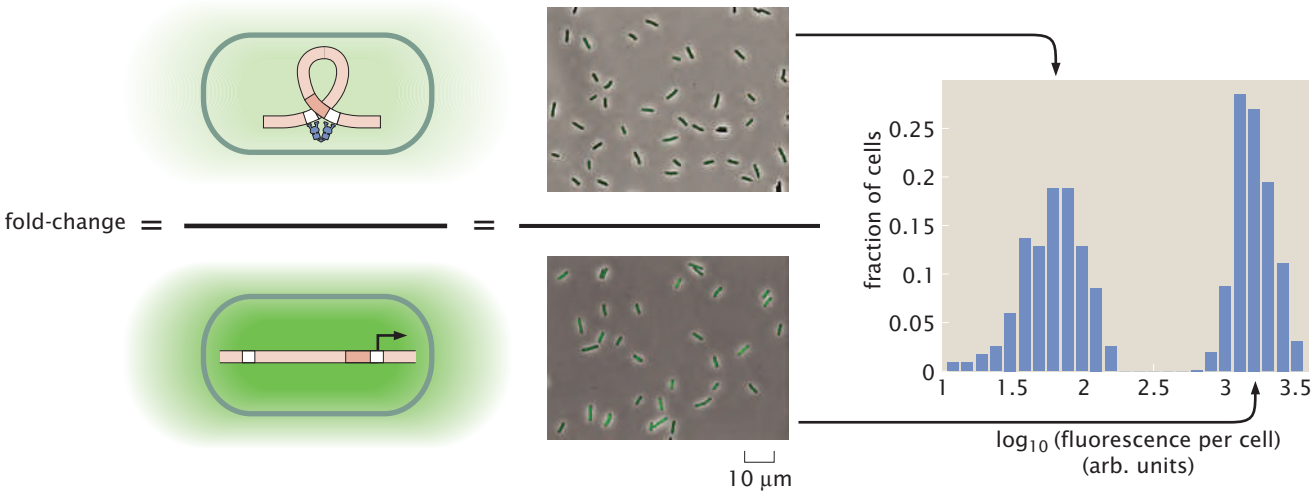


Figure 19.17: Converting image intensities to fold-change. The fold-change in gene expression is defined as the ratio of the levels of gene expression coming from a strain bearing the transcription factor of interest over a strain with a deletion of such transcription factor. For each one of these two strains, the procedure described in Figure 19.16 can be performed, leading to a distribution of fluorescence for each strain. Additionally, the cell autofluorescence is subtracted from each sample by analyzing a strain bearing no fluorescent protein. The means of each distribution can be divided in order to calculate the fold-change in gene expression.

19.2.3 Transcriptional Regulation by the Numbers: Binding Energies and Equilibrium Constants

We have heard it said that "physics isn't worth a damn unless you put in some numbers!" The abstract expressions obtained so far are much more interesting when viewed through the prism of particular measurements. Binding energies quantify the affinity of RNA polymerase or transcription factors for their DNA targets. In particular, RNA polymerase and transcription factors perform molecular recognition as a result of a rank ordering of their preferences for different sequences of nucleotides. Indeed, the sequence associated with a given promoter distinguishes it from some random sequence to which RNA polymerase would bind with a nonspecific binding energy $\varepsilon_{\text{pd}}^{\text{NS}}$. Specific binding energies can also be tuned. For example, even though there might be one very strong consensus promoter, that binding strength can be reduced by introducing mismatches in the sequence. A strong promoter, with a p_{bound} close to 1, will have a strong level of expression. On the other hand, by weakening a given promoter, cells can broaden their dynamic range by introducing a codependency on a battery of transcription factors that effectively tune the range of binding affinities and permit the regulation of promoter occupancy.

Equilibrium Constants Can Be Used To Determine Regulation Factors

In order to compute the regulation factors for the various regulatory scenarios under consideration in this chapter, we need to make estimates for the energy associated with binding protein X to the DNA, both specifically and nonspecifically; protein X can be a repressor or an activator. Binding energies are determined indirectly in experiments that measure the equilibrium constant for binding X to DNA (D). In particular, we consider the reaction

$$\text{X} + \text{D} \rightleftharpoons \text{XD} \tag{19.18}$$

with an equilibrium binding constant

$$K_X^{(\text{bind})} = \frac{[XD]}{[X][D]}. \tag{19.19}$$

Here, $[\cdots]$ denotes concentrations of the various species taking part in the reaction.

When a single X binds to DNA, there is an overall change in the free energy, Δf_{XD}. The more negative this quantity is, the more likely X will be bound to DNA. Similarly, a larger $K_X^{(\text{bind})}$ implies that the bound state is more likely. More precisely, the probability that a particular binding site on the DNA is occupied is equal to the ratio of the number of occupied sites to the total number of sites, as was first introduced in Section 6.4.1 (p. 270). In terms of concentrations, this can be written

$$p_{\text{bound}} = \frac{[XD]}{[D] + [XD]} = \frac{K_X^{(\text{bind})}[X]}{1 + K_X^{(\text{bind})}[X]}, \tag{19.20}$$

where the final expression follows from Equation 19.19. On the other hand, given that there are $[X]V_{\text{cell}}$ copies of protein X in the cell (V_{cell} is the volume of the cell), the probability of a DNA-binding site being occupied is

$$p_{\text{bound}} = \frac{[X]V_{\text{cell}} e^{-\beta \Delta f_{XD}}}{1 + [X]V_{\text{cell}} e^{-\beta \Delta f_{XD}}}. \tag{19.21}$$

Comparison of the two expressions for p_{bound} allows us to relate the microscopic and macroscopic views of binding through the relation

$$\frac{K_X^{(\text{bind})}}{V_{\text{cell}}} = e^{-\beta \Delta f_{XD}}. \tag{19.22}$$

Using this relation, we can compute the binding free energies for RNA polymerase and the various transcription factors in *E. coli*, which provides an alternative description of the same underlying processes. Presently, we use these ideas to tackle the *lac* operon, which features both positive and negative regulation.

19.2.4 A Simple Statistical Mechanical Model of Positive and Negative Regulation

Real regulatory architectures in cells often involve both repression and activation simultaneously. In this case, we consider the five distinct outcomes shown in Figure 19.18 and captured through the total partition function

$$\begin{aligned} Z_{\text{tot}}(P, A, R; N_{\text{NS}}) &= \underbrace{Z(P, A, R; N_{\text{NS}})}_{\text{empty promoter}} + \underbrace{Z(P-1, A, R; N_{\text{NS}}) e^{-\beta \varepsilon_{\text{pd}}^{\text{S}}}}_{\text{RNAP}} \\ &+ \underbrace{Z(P, A-1, R; N_{\text{NS}}) e^{-\beta \varepsilon_{\text{ad}}^{\text{S}}}}_{\text{activator}} + \underbrace{Z(P-1, A-1, R; N_{\text{NS}}) e^{-\beta(\varepsilon_{\text{ad}}^{\text{S}} + \varepsilon_{\text{pd}}^{\text{S}} + \varepsilon_{\text{pa}})}}_{\text{RNAP + activator}} \\ &+ \underbrace{Z(P, A, R-1; N_{\text{NS}}) e^{-\beta \varepsilon_{\text{rd}}^{\text{S}}}}_{\text{repressor}} + \underbrace{Z(P, A-1, R-1; N_{\text{NS}}) e^{-\beta(\varepsilon_{\text{ad}}^{\text{S}} + \varepsilon_{\text{rd}}^{\text{S}})}}_{\text{activator + repressor}}. \end{aligned} \tag{19.23}$$

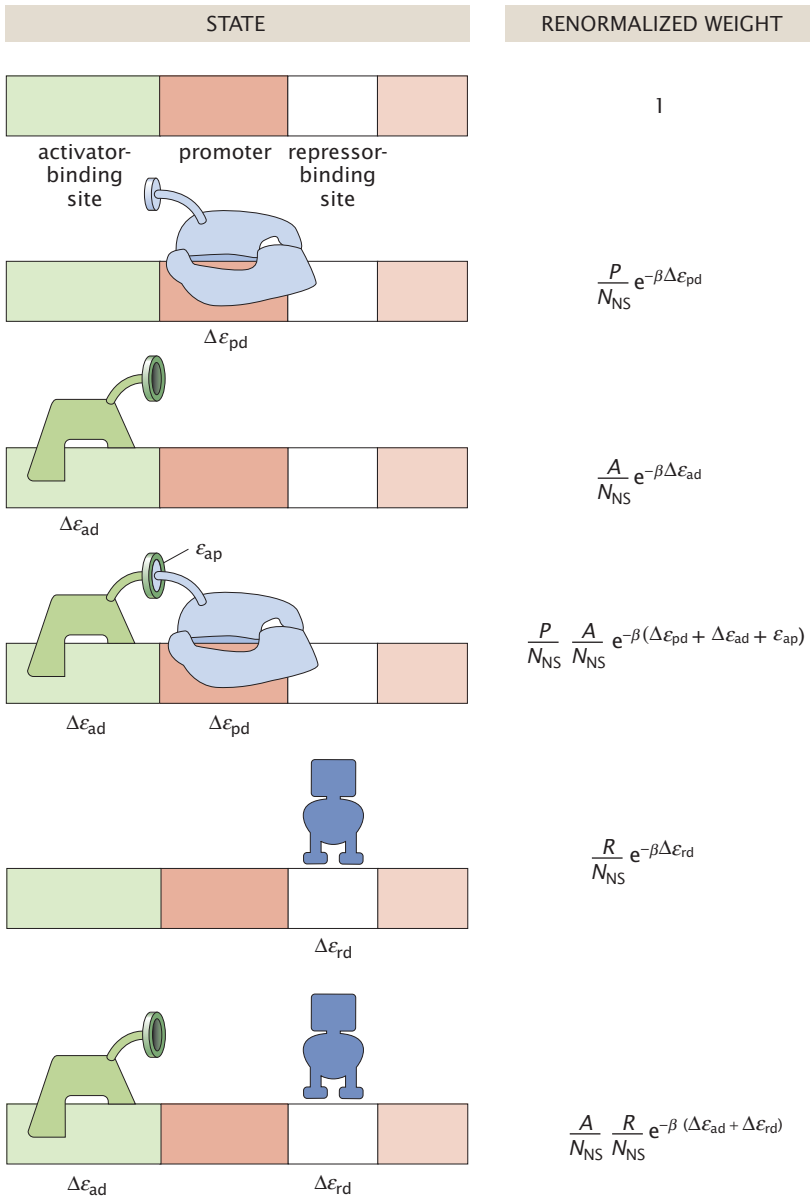


Figure 19.18: Schematic representation of the simple statistical mechanical model of recruitment and repression. States and weights for the case in which activation and simple repression act simultaneously.

Note that the cartoon shows a schematic representation of the different ways that the region in the vicinity of the promoter can be occupied and what the statistical weights are of each such state of occupancy. We can compute the probability of RNA polymerase binding by considering the ratio of favorable outcomes to the total partition function, resulting in

$$p_{\text{bound}}(P, A, R; N_{NS}) = \frac{Z(P-1, A, R; N_{NS}) e^{-\beta\varepsilon_{pd}^{S}} + Z(P-1, A-1, R; N_{NS}) e^{-\beta(\varepsilon_{ad}^{S} + \varepsilon_{pd}^{S} + \varepsilon_{pa})}}{Z_{tot}(P, A, R; N_{NS})}. \tag{19.24}$$

As before, perhaps the simplest way to interpret this result is with reference to the regulation factor, resulting in

$$p_{\text{bound}}(P, A, R; N_{NS}) = \frac{1}{1 + [N_{NS}/PF_{reg}(A, R)] e^{\beta(\varepsilon_{pd}^{S} - \varepsilon_{pd}^{NS})}}, \tag{19.25}$$

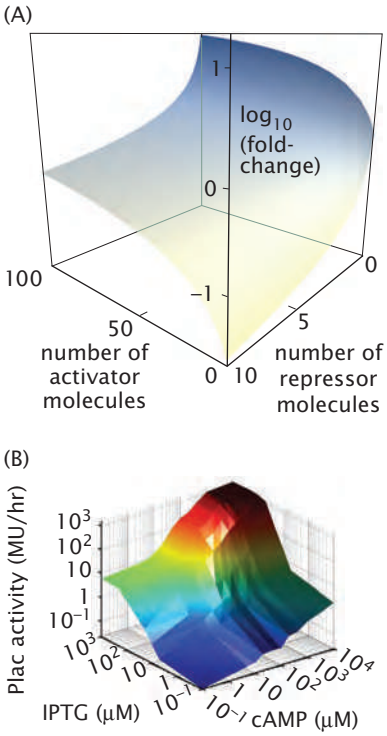


Figure 19.19: Combined regulation by repressor and activator. (A) The fold-change in gene expression as a function of the number of transcription factors shows their combinatorial action. The parameters used are $\Delta\varepsilon_{ad} = -10\ k_BT$, $\varepsilon_{ap} = -3.9\ k_BT$ and $\Delta\varepsilon_{rd} = -16.9\ k_BT$. (B) Activity of the *lac* operon measured in Miller units (MU) per hour as a function of the concentration of IPTG and cAMP, which regulate the binding of Lac repressor and CRP to the DNA, respectively. (B, adapted from T. Kuhlman et al., *Proc. Natl. Acad. Sci. USA* 104:6043, 2007.)

where the regulation factor itself is now a function of both the number of activators, A, and the number of repressors, R. In particular, the regulation factor is given by

$$F_{reg}(A,R) = \frac{1 + (A/N_{NS})e^{-\beta(\Delta\varepsilon_{ad}+\varepsilon_{ap})}}{1 + (A/N_{NS})e^{-\beta\Delta\varepsilon_{ad}} + (R/N_{NS})e^{-\beta\Delta\varepsilon_{rd}} + (A/N_{NS})(R/N_{NS})e^{-\beta(\Delta\varepsilon_{ad}+\Delta\varepsilon_{pd})}}. \tag{19.26}$$

The variation in fold-change in gene expression due to this regulatory architecture in the weak promoter approximation is shown in Figure 19.19(A). The objective of this figure is to illustrate the combinatorial control that can be reached when different transcription factors act in unison. Perhaps nowhere is this interplay of negative and positive regulation better known than in our old friend, the *lac* operon. In fact, Figure 19.19(B) reveals this interplay between activation and repression in the particular context of the *lac* operon. Here, instead of varying the intracellular number of transcription factors, the simpler approach of measuring the activity of the *lac* promoter as a function of the two inducers that control the binding of repressor and activator to DNA (IPTG and cAMP, respectively) is taken.

19.2.5 The *lac* Operon

Both repression and activation are key parts of the equipment of bacteria. Perhaps the most famous example of these effects is provided by the *lac* operon and is shown in Figure 4.15 (p. 158). Indeed, the *lac* operon has served as one of the central workhorses of the entire book, and the present section is the *denouement* of that discussion. In this case, the activator is the catabolite activator protein (CAP), also known as cyclic AMP receptor protein (CRP). In order to be able to recruit RNA polymerase, CAP has to be bound to cyclic AMP (cAMP), a molecule whose concentration goes up when the amount of glucose decreases. The repressor, known as Lac repressor, decreases the level of transcription unless it is bound to allolactose, which is a byproduct of lactose metabolism.

The *lac* Operon Has Features of Both Negative and Positive Regulation

Recall that the *lac* operon oversees the management of the enzymes that are responsible for lactose uptake and digestion. In particular, when *E. coli* cells find themselves simultaneously deprived of glucose and supplied with lactose, the genes of the *lac* operon are turned on so as to take metabolic advantage of the lactose. We have already described the way in which the Lac repressor forbids transcription of the genes associated with lactose digestion by binding on its operator. However, our earlier discussion was a bit too blithe, since we said nothing of what happens in the case where glucose and lactose are simultaneously available. If we were to adopt the picture of negative control described above, then our expectation would be that in this case there should be substantive transcription of the genes of the *lac* operon. However, there is a second element of positive control that completes the story. In particular, in the *absence* of glucose, the activator CAP binds to a site near the promoter (the RNA polymerase-binding

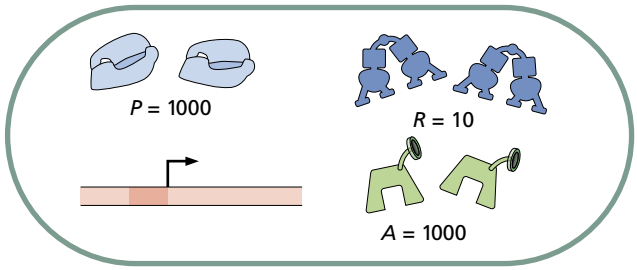


Figure 19.20: Census of the relevant molecular actors in the *lac* operon. The figure shows a rough estimate of the number of polymerase molecules, activators, and repressors associated with the *lac* operon.

site) as shown in Figure 4.15 (p. 158) and "recruits" RNA polymerase to the promoter. The census shown in Figure 19.20 gives a rough impression of the number of copies of some of the key molecules associated with the *lac* operon and illustrates the striking fact that some of the transcription factors exist with as few as 10 copies.

The geometry of the regulatory landscape for the *lac* operon is shown in Figure 19.21. Our discussion of Figure 4.15 (p. 158) was oversimplified in the sense that we ignored the presence of auxiliary binding sites for the Lac repressor that are revealed in Figure 19.21. In particular, there are two other binding sites for the Lac repressor. Specifically, there is a binding site known as O2 located 401 bp downstream from O1 and a second such site known as O3 situated 92 bp upstream. Part of our discussion will center on the subtle ways in which repression takes place in this system. Recall that the repressor itself is a tetramer with two "reading heads" that can each bind to a different operator, looping out the intervening DNA.

One of the most important roles for models like those described here is in providing a conceptual framework for thinking about both *in vivo* and *in vitro* data and in suggesting new experiments. A particularly compelling class of *in vivo* experiments using the *lac* operon measured the repression as a function of the strength and placement of the operator sites that are the targets of Lac repressor. In particular, *E. coli* cells were created that had only one operator for Lac repressor as well as mutants with different spacings between operators (a topic we return to below). The first set of experiments we consider are those in which only one operator was present for Lac repressor binding as shown in Figure 19.22. In these experiments, the repression was measured for cases in which the promoter was repressed by each of the operators O1, O2, and O3 individually. From the standpoint of the models considered here, all that is different from one experiment to the next is the binding energy of repressor for the DNA.

Recall that for a single repressor, the regulation factor is given by Equation 19.16. What is measured in the experiment is the ratio of the level of gene expression in the absence of repressor to that in the presence of repressor. For the purposes of our model, we replace

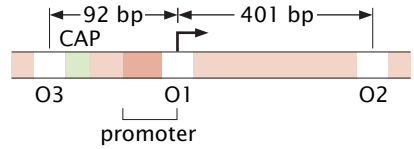
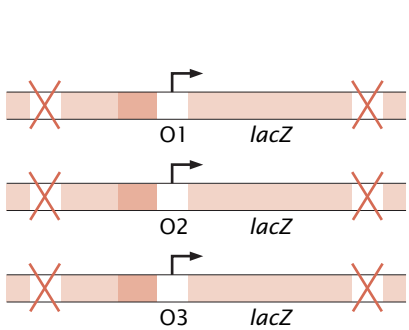


Figure 19.21: Position of the three *lac* operators and the CAP-binding site relative to the promoter. O1 is the main operator, while O2 and O3 are auxiliary binding sites for Lac repressor and are associated with DNA looping.



repression		
50	number of repressors	900
200		4700
21		320
1.3		16

Figure 19.22: Repression in the *lac* operon. The DNA constructs used in these experiments deleted the auxiliary binding sites for repressor and tuned the strength of the main repressor-binding site. Repression, the inverse of the fold-change in gene expression, was measured in each construct for two different concentrations of Lac repressor. (Adapted from S. Oehler et al., *EMBO J.* 13:3348, 1994.)

this definition based upon a measure of protein content (that is, the product of the gene) with a definition based upon examining the probability that the promoter is occupied by RNA polymerase. The implicit assumption here is that the protein content is linearly related to the probability of promoter occupancy. More precisely, we define repression as the ratio of the probability of binding of RNA polymerase to the relevant promoter in the absence of repressor to the probability of such binding in the presence of repressor, namely

$$\text{repression} = \frac{p_{\text{bound}}(R = 0)}{p_{\text{bound}}(R \neq 0)}. \tag{19.27}$$

Concretely, this result depends on the number of repressors, R, and their energy of binding to DNA. If we substitute for p_{bound} using Equation 19.15, we find that the repression can be written as

$$\text{repression}(R) = \frac{1 + (P/N_{\text{NS}})e^{-\beta\Delta\varepsilon_{\text{pd}}} + (R/N_{\text{NS}})e^{-\beta\Delta\varepsilon_{\text{rd}}}}{1 + (P/N_{\text{NS}})e^{-\beta\Delta\varepsilon_{\text{pd}}}}. \tag{19.28}$$

For the case of a weak promoter, this implies in turn that the repression level can be written as

$$\text{repression}(R) = [\text{fold-change}(R)]^{-1} \simeq [F_{\text{reg}}(R)]^{-1} = 1 + \frac{R}{N_{\text{NS}}} e^{-\beta\Delta\varepsilon_{\text{rd}}}. \tag{19.29}$$

One of the interesting opportunities afforded by this expression is the possibility of a direct confrontation with experimental data such as is shown in Figure 19.22.

In particular, the data of Figure 19.22 permit us to determine the only unknown in our expression for the repression, namely, the energy parameter $\Delta\varepsilon_{\text{rd}}$. Since the data reflect three different choices of binding strength, we find three different binding energies ($\Delta\varepsilon_{\text{rd}} = -16.9, -14.4$, and $-11.2\ k_{\text{B}}T$ for O1, O2, and O3, respectively). With these energies in hand, we can predict the outcome of repression measurements in which the number of repressors is tuned to other values as shown in Figure 19.23. Note that once the binding-energy difference has been estimated using one data point, it leads to a prediction for the behavior of the system for different numbers of repressor molecules in the cell and will serve as the basis for our analysis of the two-operator case as well.

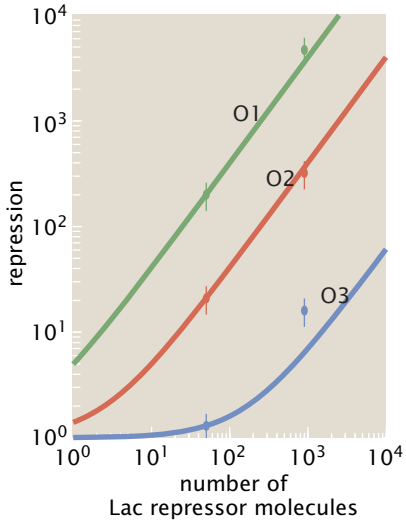


Figure 19.23: Repression model for the *lac* operon. Each curve shows how repression varies as a function of the number of repressor molecules in the cell for constructs with a single main binding site as shown in Figure 19.22. Different curves correspond to different main binding sites (operators) for the Lac repressor. (Data from S. Oehler et al., *EMBO J.* 13:3348, 1994.)

The Free Energy of DNA Looping Affects the Repression of the *lac* Operon

Our discussion of the *lac* operon from the statistical mechanical perspective has thus far ignored one of the more intriguing features of this system, namely, the presence of DNA looping. The behavior of the *lac* operon has been examined in great detail both *in vitro* and *in vivo*. One beautiful set of experiments that is particularly enlightening with reference to the class of models we have described thus far in the chapter examines the repression of the *lac* operon as a function of the spacing between the DNA-binding sites (the operators) for Lac repressor.

The data on repression as a function of interoperator spacing were introduced in Figure 1.11 (p. 19) as an example of the sophisticated quantitative data that exist on biological systems in general, and gene expression in particular. These beautiful experiments and others like

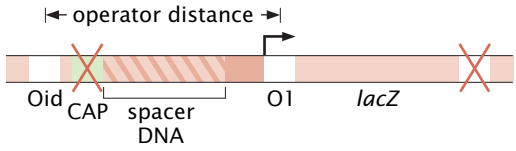


Figure 19.24: Construct used to measure repression in the presence of looping. The binding site for the activator CRP (shown as CAP in the diagram) was deleted, as was the third repressor-binding site. (Adapted from J. Müller et al., *J. Mol. Biol.* 257:21, 1996.)

them show a systematic trend in the promoter activity of the genes in question as a function of the distance between the binding sites for the repressor under consideration. One particularly telling feature of such data is the periodicity that results from the twist degrees of freedom and that reflects the need for particular faces of the DNA to be aligned in order to form a loop.

Figure 19.24 shows the DNA construct that was used to examine the *in vivo* consequences of DNA looping. In this construct, both the binding site for CRP and the operator O2 were deleted, while the promoter was replaced with a stronger promoter. The deletion of the CRP-binding site is intended to remove the question of activation from the problem. Note also that this construct permits the insertion of DNA sequences of arbitrary length between O1 and Oid, where Oid has replaced O3. Oid is a much stronger operator than O3, of approximately the same strength as O1. Finally, the deletion of O2 insures that looping will only occur between the two remaining operators.

In order to confront data like those shown in Figure 1.11 (p. 19), we need to expand our discussion of activators and repressors to include the effect of looping itself. In Figure 19.25, we show a minimal model of the states available to the system when RNA polymerase and Lac repressor are competing for the same region in the vicinity of the promoter. Note that this model permits different repressor molecules to occupy the two operators simultaneously, or a single molecule to occupy both sites and to loop the intervening DNA. We ignore the possibility of activator-binding since the activator-binding site was eliminated as shown in Figure 19.24. Note that this does not unequivocally rule out the possibility of nonspecific CAP binding, which might affect the results as well.

In order to proceed in quantitative terms, as usual, we need to write down the partition function that corresponds to assigning statistical weights to all of the allowed states depicted in Figure 19.25. Using exactly the same logic as in previous sections, the partition function can be written as

$$
\begin{aligned}
Z_{\text{tot}}(P, R; N_{\text{NS}}) = & \underbrace{Z(P, R; N_{\text{NS}})}_{\text{P}^{(0)},\ \text{O}^{(0)}_{\text{main}}\ \text{and}\ \text{O}^{(0)}_{\text{aux}}} + \underbrace{Z(P-1, R; N_{\text{NS}}) e^{-\beta \varepsilon^{\text{S}}_{\text{pd}}}}_{\text{P}^{(1)},\ \text{O}^{(0)}_{\text{main}}\ \text{and}\ \text{O}^{(0)}_{\text{aux}}} \\
& + \underbrace{Z(P-1, R-1; N_{\text{NS}}) e^{-\beta \varepsilon^{\text{S}}_{\text{pd}}} e^{-\beta \varepsilon^{\text{S}}_{\text{rda}}}}_{\text{P}^{(1)},\ \text{O}^{(0)}_{\text{main}}\ \text{and}\ \text{O}^{(1)}_{\text{aux}}} \\
& + \underbrace{Z(P, R-1; N_{\text{NS}}) e^{-\beta \varepsilon^{\text{S}}_{\text{rdm}}}}_{\text{P}^{(0)},\ \text{O}^{(1)}_{\text{main}}\ \text{and}\ \text{O}^{(0)}_{\text{aux}}} + \underbrace{Z(P, R-1; N_{\text{NS}}) e^{-\beta \varepsilon^{\text{S}}_{\text{rda}}}}_{\text{P}^{(0)},\ \text{O}^{(0)}_{\text{main}}\ \text{and}\ \text{O}^{(1)}_{\text{aux}}} \\
& + \underbrace{Z(P, R-2; N_{\text{NS}}) e^{-\beta \varepsilon^{\text{S}}_{\text{rdm}}} e^{-\beta \varepsilon^{\text{S}}_{\text{rda}}}}_{\text{P}^{(0)},\ \text{O}^{(1)}_{\text{main}}\ \text{and}\ \text{O}^{(1)}_{\text{aux}}} \\
& + \underbrace{Z(P, R-1; N_{\text{NS}}) e^{-\beta \varepsilon^{\text{S}}_{\text{rdm}}} e^{-\beta \varepsilon^{\text{S}}_{\text{rda}}} e^{-\beta F_{\text{loop}}}}_{\text{repressor/loop}}, \qquad (19.30)
\end{aligned}
$$

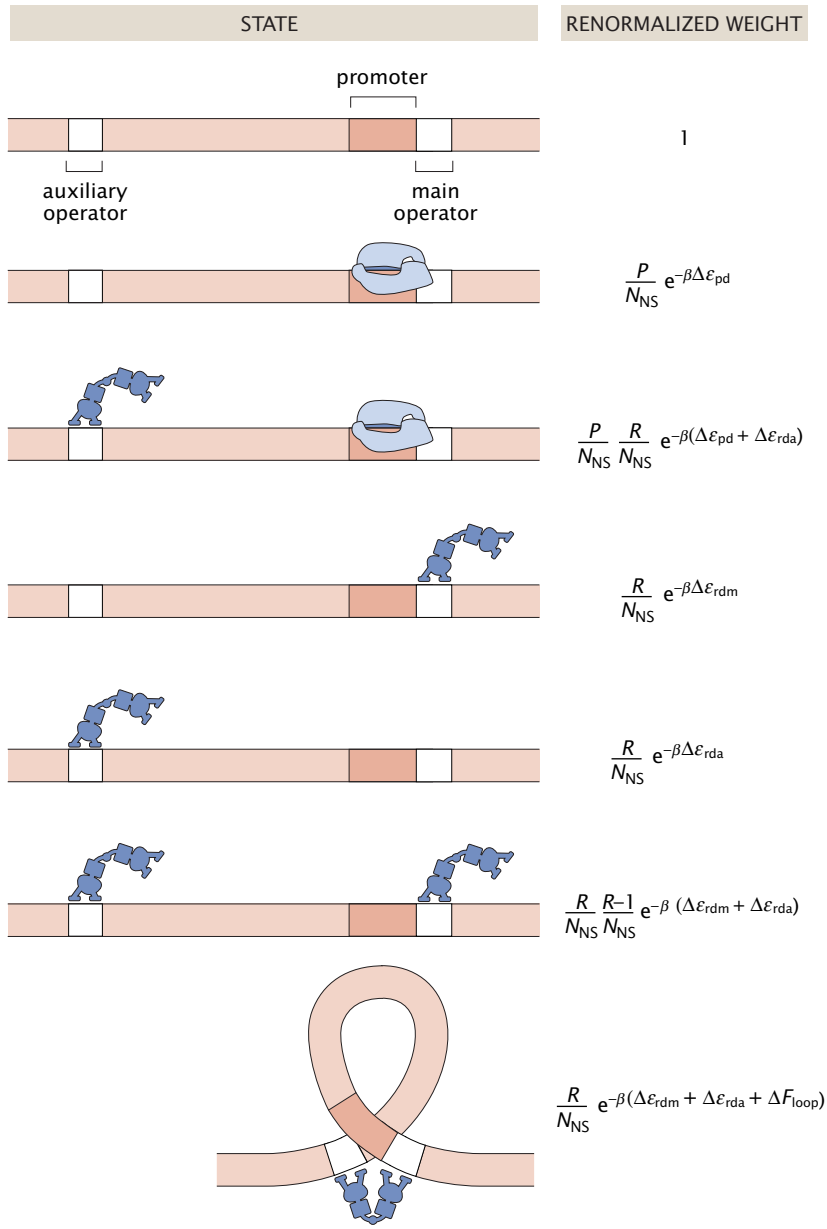


Figure 19.25: Looping states and weights in the *lac* operon. Each state corresponds to a different state of occupancy of the promoter and operators in the operon.

where ε_{rda} is the binding energy of the repressor for the auxiliary operator and ε_{rdm} is the binding energy of the repressor for the main operator. Our notation has clearly become more cumbersome and deserves explanation. First, we introduce $P^{(0)}$, $O^{(0)}_{main}$, and $O^{(0)}_{aux}$ to indicate that the occupancies of the promoter and main and auxiliary operators are zero, respectively. Next, the notation $O^{(1)}_{main}$ indicates that the main operator is occupied. The term with $P^{(0)}$, $O^{(1)}_{main}$, and $O^{(1)}_{aux}$ indicates the states for which there are distinct repressor molecules bound to the two operators and the final term accounts for the looped state.

One of the terms in the expression includes the looping free energy in the form

$$Z(P, R-1; N_{NS}) e^{-\beta \varepsilon^S_{rdm}} e^{-\beta \varepsilon^S_{rda}} e^{-\beta F_{loop}}, \tag{19.31}$$

and the factor $e^{-\beta F_{loop}}$ deserves further comment. Recall that $Z(P, R-1; N_{NS})$ is itself already a sum over all of the possible ways of

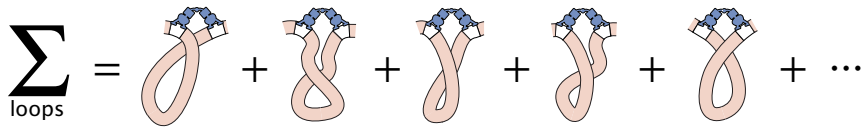


Figure 19.26: Summing over DNA loops. The sum $\sum_{\text{loops}}$ instructs us to sum over all conformations of the DNA loop as indicated schematically here.

distributing the P RNA polymerase molecules and the $R-1$ repressor molecules over the N_{NS} nonspecific binding sites on the DNA, with one of the repressors bound to both operators and looping the intervening DNA. However, for each and every one of these configurations, we have to sum over *all* of the possible geometries of the loop itself. That is, this contribution to the partition function is really of the form

$$Z_{\text{looped}}(P, R-1; N_{\text{NS}}) = \sum_{\text{loops}} Z(P, R-1; N_{\text{NS}}) e^{-\beta \varepsilon^{\text{S}}_{\text{rdm}}} e^{-\beta \varepsilon^{\text{S}}_{\text{rda}}} e^{-\beta \varepsilon_{\text{loop}}}, \tag{19.32}$$

where $\varepsilon_{\text{loop}}$ is the *energy* of a given loop configuration and $\sum_{\text{loops}}$ instructs us to sum over all of the possible loop configurations as schematized in Figure 19.26. Since most of the factors are independent of the looping geometry, we can rewrite this as

$$Z_{\text{looped}}(P, R-1; N_{\text{NS}}) = Z(P, R-1; N_{\text{NS}}) e^{-\beta \varepsilon^{\text{S}}_{\text{rdm}}} e^{-\beta \varepsilon^{\text{S}}_{\text{rda}}} \sum_{\text{loops}} e^{-\beta \varepsilon_{\text{loop}}}, \tag{19.33}$$

where we have pulled all terms out of the sum that do not depend upon the particular choice of looped state. One way to proceed at this point is to appeal to ideas about elasticity to determine $\varepsilon_{\text{loop}}$ and use the random walk as the basis for effecting the sum. On the other hand, a simpler scheme is to replace the sum by $e^{-\beta F_{\text{loop}}}$ and to treat F_{loop} as a phenomenological parameter as we have already done with the various binding energies.

With the partition function in hand, we can compute the probability of RNA polymerase binding by considering the ratio of favorable outcomes to the total partition function, resulting in

$$\begin{aligned} p_{\text{bound}}(P, R; N_{\text{NS}}) = &\frac{P}{N_{\text{NS}}} e^{-\beta \Delta \varepsilon_{\text{pd}}} \left(1 + \frac{R}{N_{\text{NS}}} e^{-\beta \Delta \varepsilon_{\text{rda}}}\right) \\ &\times \left[1 + \frac{P}{N_{\text{NS}}} e^{-\beta \Delta \varepsilon_{\text{pd}}} \left(1 + \frac{R}{N_{\text{NS}}} e^{-\beta \Delta \varepsilon_{\text{rda}}}\right) \right. \\ &+ \frac{R}{N_{\text{NS}}} \left(e^{-\beta \Delta \varepsilon_{\text{rdm}}} + e^{-\beta \Delta \varepsilon_{\text{rda}}}\right) \\ &+ \frac{R(R-1)}{(N_{\text{NS}})^2} e^{-\beta(\Delta \varepsilon_{\text{rdm}} + \Delta \varepsilon_{\text{rda}})} \\ &\left. + \frac{R}{N_{\text{NS}}} e^{-\beta(\Delta \varepsilon_{\text{rdm}} + \Delta \varepsilon_{\text{rda}} + \Delta F_{\text{loop}})} \right]^{-1}, \end{aligned} \tag{19.34}$$

where we have defined $\Delta F_{\text{loop}} = F_{\text{loop}} + \varepsilon^{\text{NS}}_{\text{rd}}$. From this expression, we can obtain the regulation factor

$$\begin{aligned} F_{\text{reg}}(R) = &\left(1 + \frac{R}{N_{\text{NS}}} e^{-\beta \Delta \varepsilon_{\text{rda}}}\right) \left[1 + \frac{R}{N_{\text{NS}}} \left(e^{-\beta \Delta \varepsilon_{\text{rdm}}} + e^{-\beta \Delta \varepsilon_{\text{rda}}}\right) \right. \\ &\left. + \frac{R(R-1)}{(N_{\text{NS}})^2} e^{-\beta(\Delta \varepsilon_{\text{rdm}} + \Delta \varepsilon_{\text{rda}})} + \frac{R}{N_{\text{NS}}} e^{-\beta(\Delta \varepsilon_{\text{rdm}} + \Delta \varepsilon_{\text{rda}} + \Delta F_{\text{loop}})} \right]^{-1}. \end{aligned} \tag{19.35}$$

To make contact with the results of Müller et al. (1996), we now need to write an expression for the repression as a function of the interoperator spacing. Recall that the repression is given by Equation 19.27 and takes the form

$$\begin{aligned}\text{repression}(N_{\text{bp}}) &= (F_{\text{reg}})^{-1} \\ &= \left[1 + \frac{R}{N_{\text{NS}}}(e^{-\beta\Delta\varepsilon_{\text{rdm}}} + e^{-\beta\Delta\varepsilon_{\text{rda}}}) \right. \\ &\quad + \frac{R(R-1)}{(N_{\text{NS}})^2} e^{-\beta(\Delta\varepsilon_{\text{rdm}}+\Delta\varepsilon_{\text{rda}})} \\ &\quad \left. + \frac{R}{N_{\text{NS}}} e^{-\beta(\Delta\varepsilon_{\text{rdm}}+\Delta\varepsilon_{\text{rda}}+\Delta F_{\text{loop}})}\right] \\ &\quad \times \left(1 + \frac{R}{N_{\text{NS}}} e^{-\beta\Delta\varepsilon_{\text{rda}}}\right)^{-1}, \end{aligned} \tag{19.36}$$

where we have written repression(N_{bp}) as a function of the number of base pairs in the loop (N_{bp}) to signal that the looping free energy (and hence the repression) will depend upon the distance between the two operators. We have invoked the approximation that the promoter is weak (that is, $(N_{\text{NS}}/PF_{\text{reg}})e^{\beta\Delta\varepsilon_{\text{pd}}} \gg 1$). In order to examine the significance of our results on looping, we consider the extent to which the model can be used to interpret existing data and to suggest new experiments. Notice that we already know all the parameters in the weights from the previous experiment, with the exception of ΔF_{loop}. We argue that for a given loop size, ΔF_{loop} is a parameter that should be indifferent to which combination of operators is used in these two-operator experiments and, as a result, once ΔF_{loop} is determined, the model is predictive. The results of a simple fit to the looping free energy are shown in Figure 19.27. To obtain these curves, any single data point is used to obtain the looping free energy itself and then the resulting curves are entirely predictive.

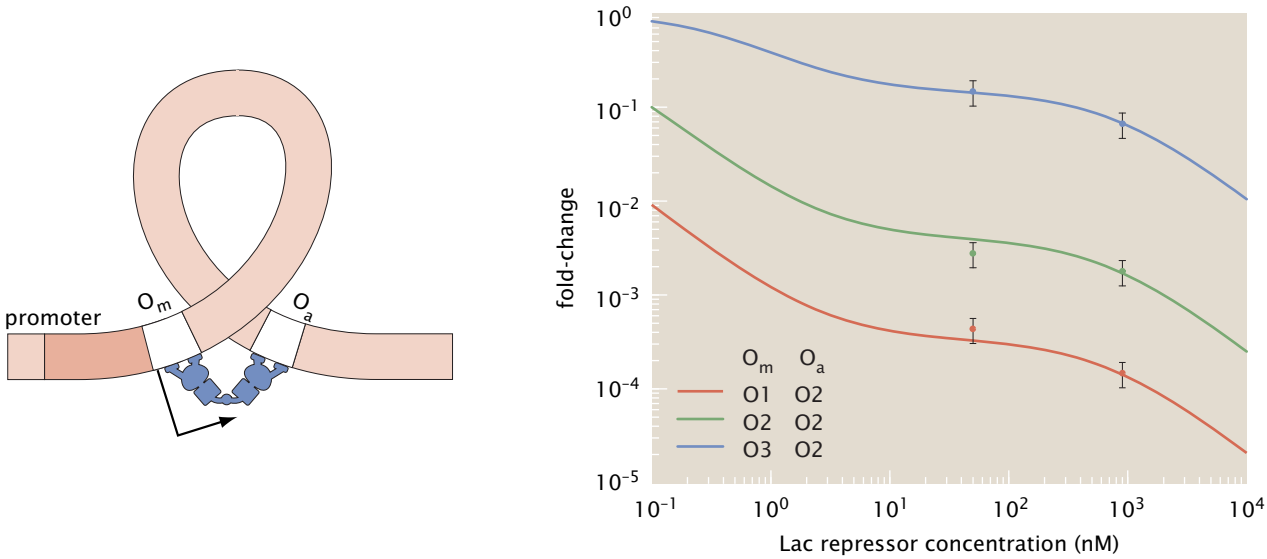


Figure 19.27: Repression and looping. A single fit to ΔF_{loop} giving 9.1 $k_B T$ permits the investigation of multiple configurations of the different operators. (Data from S. Oehler et al., *EMBO J.* 13:3348, 1994.)

Inducers Tune the Level of Regulatory Response

Who regulates the regulators? So far, our story has been built around a set of transcription factors that are themselves above the law. But, in fact, we know that the action of these transcription factors is itself controlled by signaling processes and it is to this subject that we now turn, once again using the *lac* operon as the defining case study to set ideas. The famed Lac repressor is itself controlled by molecules known as inducers, with one of the most important examples being the synthetic inducer IPTG. IPTG binds to Lac repressor, thus reducing its affinity for DNA. In the natural context, the binding of an inducer (allolactose) to Lac repressor provides the feedback that eliminates repression and permits the synthesis of the enzyme that performs the chemical cleavage of lactose.

One of the ways in which the reduction in binding affinity of Lac repressor for DNA has been illustrated is through *in vitro* binding experiments like those schematized in Figure 19.28. This figure shows the *in vitro* occupancy of an operator by Lac repressor as a function of inducer (IPTG) concentration. However, when performing an analogous titration *in vivo* for the wild-type *lac* operon, a much sharper response for the level of gene expression as a function of the IPTG concentration is observed. The "sharpness" of the output signal (gene expression for the *in vivo* measurements or normalized binding probability for the *in vitro* measurements) can be quantified using the thermodynamically inspired Hill function introduced in Section 6.4.3 (p. 273) and given in this case by

$$\text{normalized gene expression} = \frac{\left(\frac{[IPTG]}{K_d}\right)^n}{1 + \left(\frac{[IPTG]}{K_d}\right)^n}. \tag{19.37}$$

Here, $[IPTG]$ is the concentration of IPTG and K_d is an effective dissociation constant. This effective dissociation constant not only reflects the interaction of IPTG with Lac repressor, it also accounts for the change in affinity of Lac repressor to its operator DNA upon binding to the inducer, for the concentration of Lac repressors, and, in the *in vivo* scenario, also for any active pumping of IPTG into the cell. n is the Hill coefficient, a measure of the slope or sensitivity of the output signal with respect to the input concentration of IPTG.

As suggested above, modeling induction of the *lac* operon requires taking into account the passive and active transport by Lac permease (LacY) of IPTG into the cell as well as the binding and DNA looping of Lac repressor described in the previous sections. All these effects conspire to give an *in vivo* Hill coefficient that is significantly different than the *in vitro* counterpart. In Figure 19.28, it is shown how by creating different strains of bacteria bearing a deletion of Lac permease and lacking the auxiliary Lac repressor-binding sites, the *in vivo* and *in vitro* sensitivities can be reconciled.

19.2.6 Other Regulatory Architectures

The *lac* operon is one of the classic case studies in modern biology. Indeed, one of our arguments is that it is precisely the well-characterized biological examples that serve as fertile proving ground for physical biology approaches. However, there is much more to regulatory biology than the *lac* operon! One of the key questions that remains in light of successes with the thermodynamic approach is

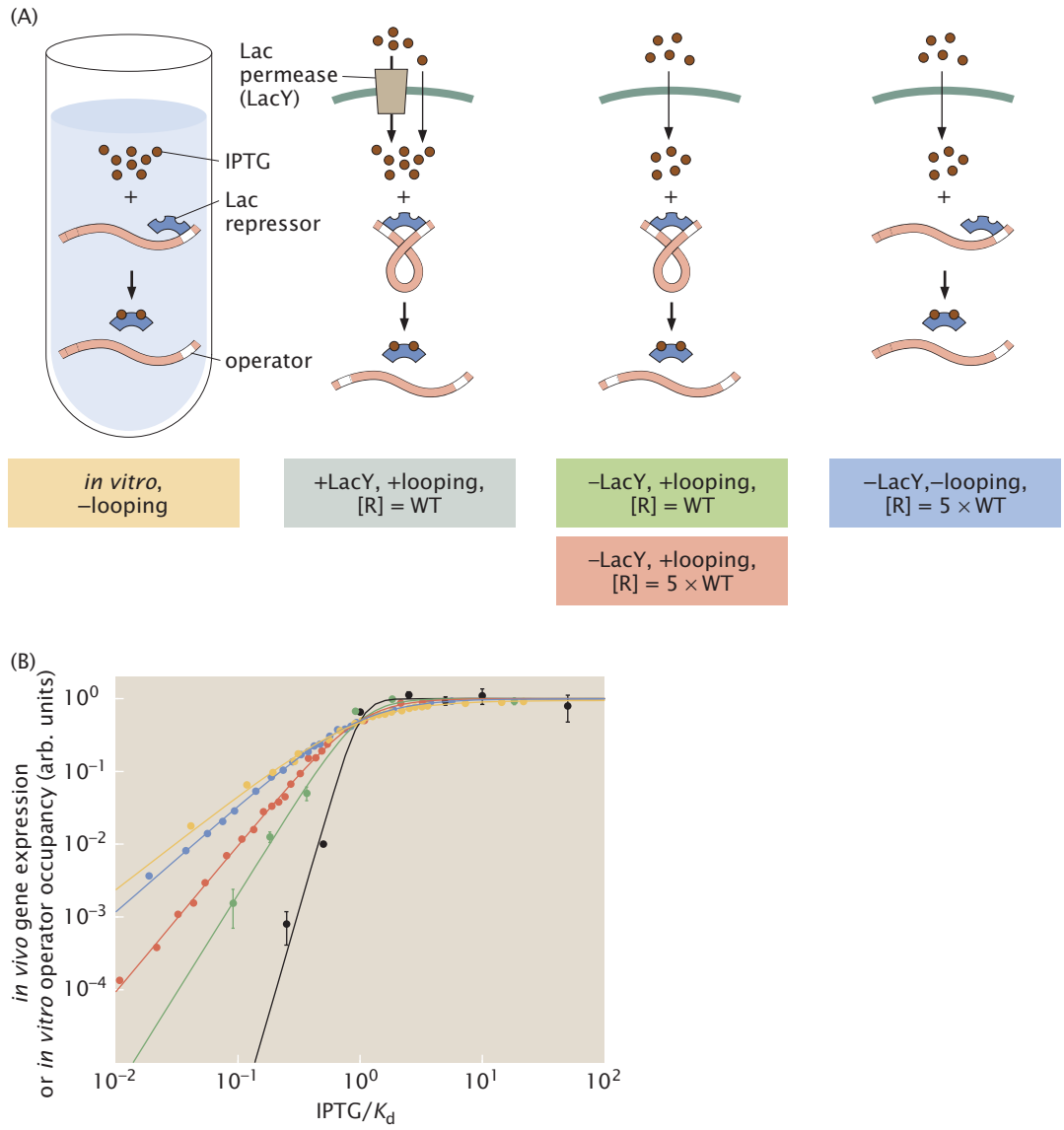


Figure 19.28: Sensitivity of induction in the *lac* operon. (A) The sensitivity can be characterized by measuring the *in vitro* occupancy of an operator as a function of inducer concentration for the simple construct containing only one Lac repressor-binding site. On the other hand, the *in vivo* level of gene expression and its resulting sensitivity can be characterized for the different *lac operon* mutants shown, where Lac permease (which pumps in inducer) or an auxiliary binding site for repressor on DNA were deleted or where the intracellular concentration of repressor, [R], is different from its wild-type (WT) value. (B) The resulting shape of the *in vitro* operator occupancy or *in vivo* level of gene expression are shown as a function of inducer concentration for the series of different mutants shown in (A), where the different curves correspond to the experimental conditions, with the gray box corresponding to the black line. As different elements of the system were deleted, the sensitivity of the induction neared that of the purified *in vitro* system. This sensitivity can be quantified by fitting to a thermodynamically inspired functional form such as the Hill function shown in Equation 19.37. Each curve has been normalized to its corresponding maximum in gene expression (*in vivo* data) or maximum in binding probability (*in vitro* data). (Data adapted from S. Oehler et al., *Nucleic Acids Res.* 34:606, 2006 and T. Kuhlman et al., *Proc. Natl. Acad. Sci. USA* 104:6043, 2007.)

the extent to which the same ideas can be used for other bacterial promoters, and, better yet, in the context of eukaryotic examples.

The Fold-Change for Different Regulatory Motifs Depends Upon Experimentally Accessible Control Parameters

So far, we have shown the cases of simple repression, simple activation, and DNA looping. In each of these cases, our experimental point

of contact has been the fold-change, which tells us how expression changes in the presence of various regulatory motifs. Figure 19.29 summarizes how the thermodynamic models worked out so far in this chapter can be used to predict the input–output response of these regulatory architectures. Conceptually, what we are really after is descriptions of these networks in which we can identify the various regulatory "knobs" such as those shown in Figure 19.30, and tune these knobs both theoretically and in the context of quantitative experiments.

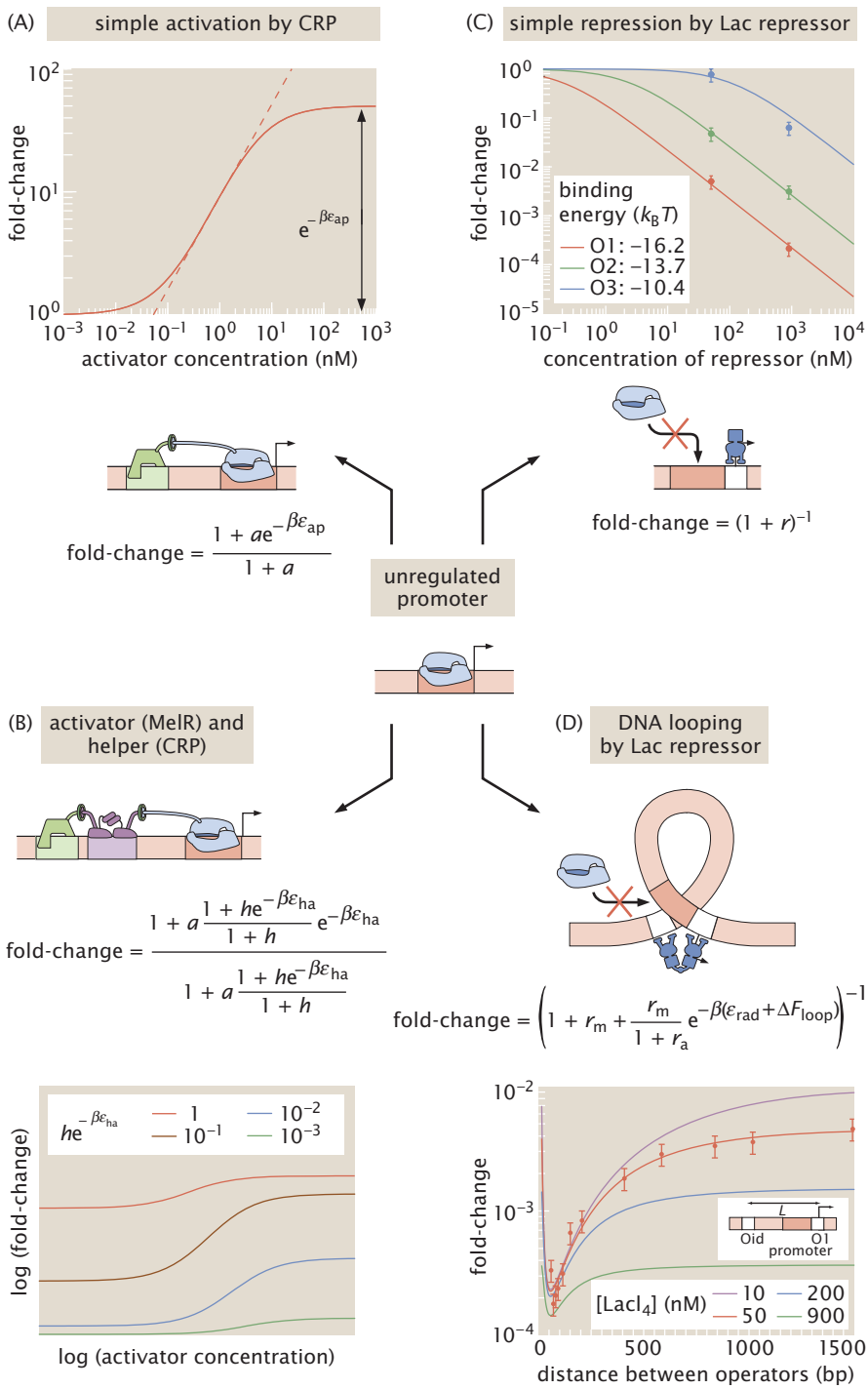


Figure 19.29: Thermodynamic models for diverse regulatory architectures. The thermodynamic models result in a predicted fold-change as a function of parameters such as operator binding strengths and transcription factor copy numbers. The resulting fold-change function serves as a governing equation dictating the regulatory output. The lowercase variables (a, h, r, etc.) correspond to $x = (X/N_{NS})e^{-\beta\Delta\varepsilon_{xd}}$, where X is the intracellular number of the particular transcription factor, $\Delta\varepsilon_{xd}$ its interaction energy with the DNA, and $N_{NS} = 5 \times 10^6$. (A) In simple activation, an activator recruits RNA polymerase to the promoter by interacting with it ($\Delta\varepsilon_{ad} = -13.8\ k_BT$, $\varepsilon_{ap} = 3.9\ k_BT$). (B) A helper molecule can recruit an activator to the promoter, which in turn can recruit RNA polymerase. (C) In simple repression, a repressor binds to a site overlapping the promoter, which results in the exclusion of RNA polymerase from that site. (D) Some repressors can also bind to multiple sites simultaneously, which results in an increase of the level of repression. The looping probability will depend on physical characteristics of the loop, such as its length ($\Delta\varepsilon_{rmd} = -16.7\ k_BT$, $\Delta\varepsilon_{rad} = -18.4\ k_BT$, and $\Delta F_{loop} = A/L + B \ln L - CL - E$, with $A = 140.6\ k_BT \times$ bp, $B = 2.52\ k_BT$, $C = 1.4 \times 10^{-3}\ k_BT$/bp and $E = 19.9\ k_BT$). We assume that one molecule per cell corresponds to 1 nM. (Adapted from L. Bintu et al., *Curr. Opin. Genet. Dev.* 15:125, 2005.)

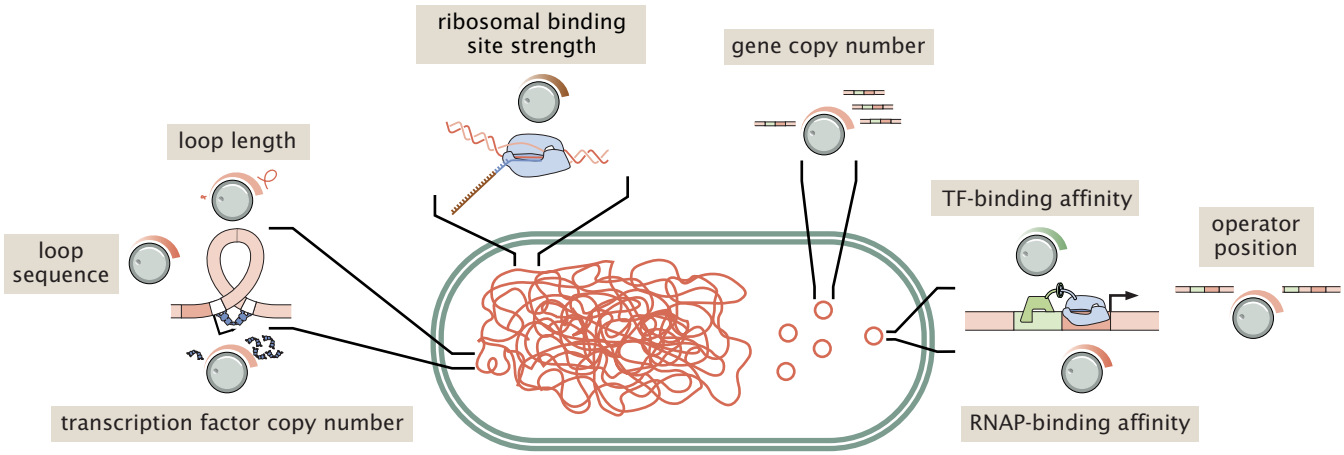


Figure 19.30: Dialing-in transcriptional output. Many parameters can be used to quantitatively tune the level of gene expression, including the number of copies of genes and the transcription factors that control them, as well as the binding strengths of a suite of different proteins that interact with the DNA during gene expression.

Quantitative Analysis of Gene Expression in Eukaryotes Can Also Be Analyzed Using Thermodynamic Models

Our simplified diagrams throughout the book have made it look as though the transcription apparatus is a single light-blue object (the polymerase) that binds to its promoter. However, even in bacteria, the transcription process is much more complicated since the basal transcription apparatus is a complex of multiple factors. In the case of eukaryotic organisms, a huge number of molecular species conspire together to drive transcription. As such, at first cut, one barely dares to use such streamlined "effective" models for which so much of the complexity is blatantly ignored. Nevertheless, the same thermodynamic description applied so far has been used to think about questions such as nucleosome occupancy and how it interferes with transcription and for a host of different eukaryotic regulatory architectures. In this brief discussion, we attempt to whet the reader's appetite by gossiping about several especially renowned eukaryotic examples.

As we will discuss thoroughly in Section 21.3.3 (p. 988), nucleosomes are found to be reliably positioned on eukaryotic genomes. In particular, they can be found in regulatory regions, with the resulting occlusion of transcription factor-binding sites as described in detail in Section 10.4.3 (p. 409). One of the systems where the interplay between nucleosomal occupancy and level of gene expression has been explored quantitatively is the PHO5 promoter in yeast. As shown in Figure 19.31(A), the PHO5 promoter is activated by PHO4, which binds to two sites, one within a nucleosome and the other not. Additionally, there is a TATA binding box at nucleosome −1. This is the binding site for the TATA box binding protein (TBP), which is critical for activation of the gene. Following a logic analogous to that put forth in order to dissect the *lac* operon, one can mutate the regulatory region to move binding sites into nucleosomes or outside of them and change their affinities. Figure 19.31(B) shows that the positioning of these binding sites with respect to those of the nucleosomes matters. In particular, when comparing architectures with sites of identical affinity, but different occlusion by nucleosomes, the features of the input–output function change appreciably. Moreover, Figure 19.31(C)

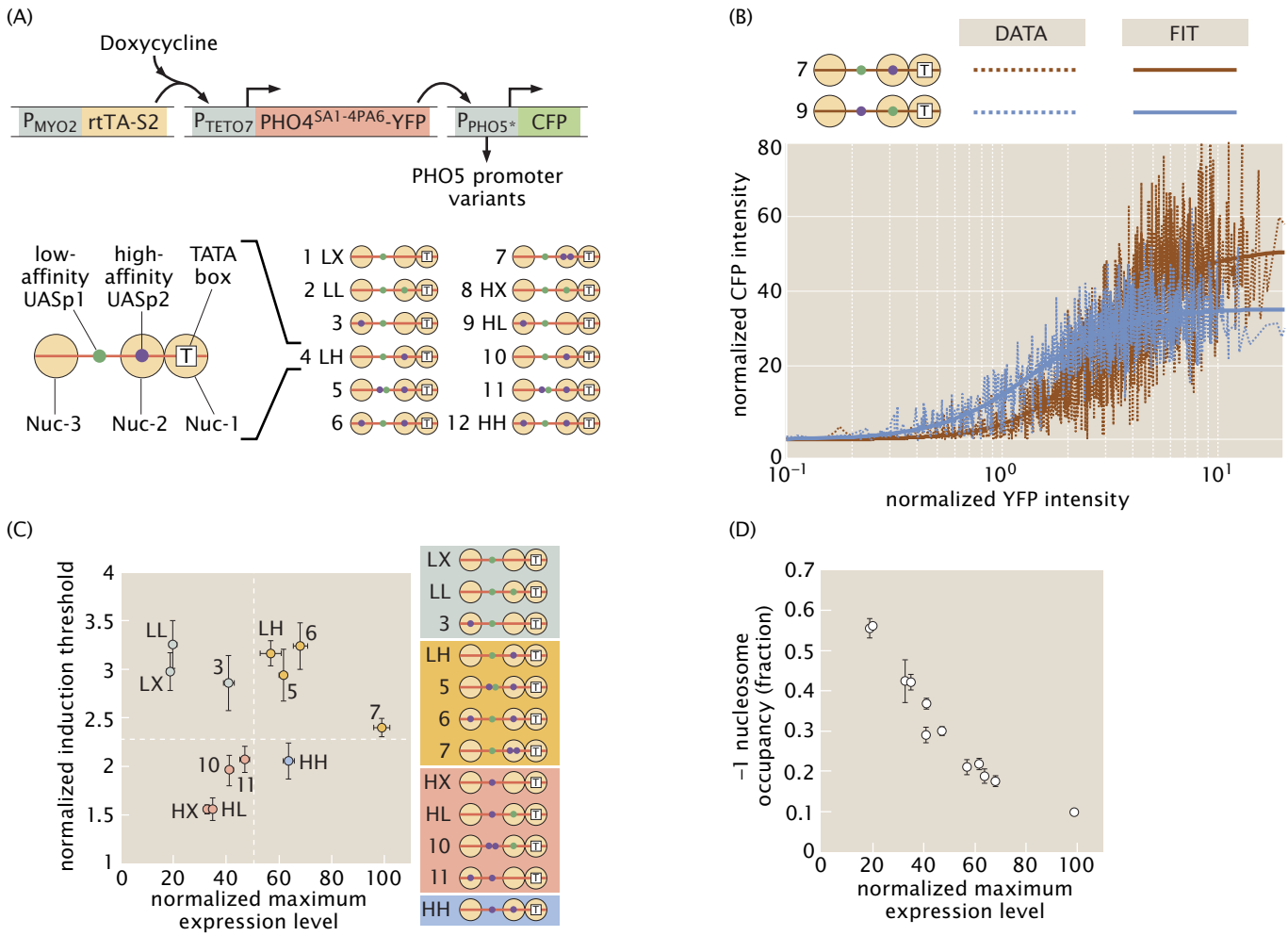


Figure 19.31: Role of nucleosomes in transcriptional regulation. (A) A PHO5 promoter expressing CFP is regulated by PHO4-YFP, with two target binding sites, UASp1 and UASp2. This promoter has a TATA box, which is bound by the TBP activator. The position of nucleosomes in this regulatory region has been mapped such that mutants can be generated where the various promoter features are occluded or not by nucleosomes. (B) Two architectures having the same binding sites but different occlusion geometries by nucleosomes show markedly different input–output functions. (C) Several variants for the PHO5 promoter result in input–output functions with drastically different induction thresholds and maximum levels of expression. (D) The maximum level of expression shows a strong correlation with the occupancy of nucleosome −1, which occludes the TATA box. (Adapted from Kim and O'Shea, *Nat. Struct. Mol. Biol.* 15:1192, 2008.)

shows how subtle changes in the regulatory region can lead to drastic changes in both the threshold and maximum level of expression of the input–output function.

Perhaps the role of nucleosome occupancy is most prominently revealed by Figure 19.31(D). In this case, the maximum level of gene expression shows a strong correlation with the occupancy of nucleosome −1 for each of the different regulatory architectures shown in Figure 19.31(A). Nucleosome −1 overlaps the TATA box, suggesting that one of the roles of PHO4 is to ultimately modulate the occupancy of −1 through the interaction with nucleosome remodeling complexes. This modulation determines, in turn, the absolute level of gene expression of the promoter. It is clear that thermodynamic models can be used to describe the many layers of regulation present in this type of problem. In particular, in Section 10.4.3 (p. 409), we have already shown how thermodynamic thinking can describe the probability of protein accessibility to a binding site that is buried

inside a nucleosome. Further, in Section 21.3.3 (p. 988), we will show how statistical mechanics can lead to simple models that predict the probability landscape of nucleosome occupancy along the genome.

A further challenge in deciphering eukaryotic transcriptional regulation centers on how transcription varies in both space and time during development in multicellular organisms. Perhaps the most well understood such organism is the fruit fly *Drosophila melanogaster*. As shown in Figure 19.2, during the initial stages of development, the fly embryo expresses a battery of transcription factors in a cascade that defines sharper and sharper domains of expression. One of the transcriptional architectures that has been studied in most detail is related to the activation of the transcription factor Hunchback by the transcription factor Bicoid. As shown in Figures 19.2 and 19.32(A), Bicoid is expressed in an exponential profile along the anterior–posterior axis of the developing embryo. Activation by Bicoid is realized by binding to six sites of different strengths that lie upstream from the Hunchback promoter, as seen in Figure 19.32. The resulting pattern of Hunchback expression shown in Figure 19.32(C) presents a domain with a boundary at about 50% of the embryo length. The exquisite

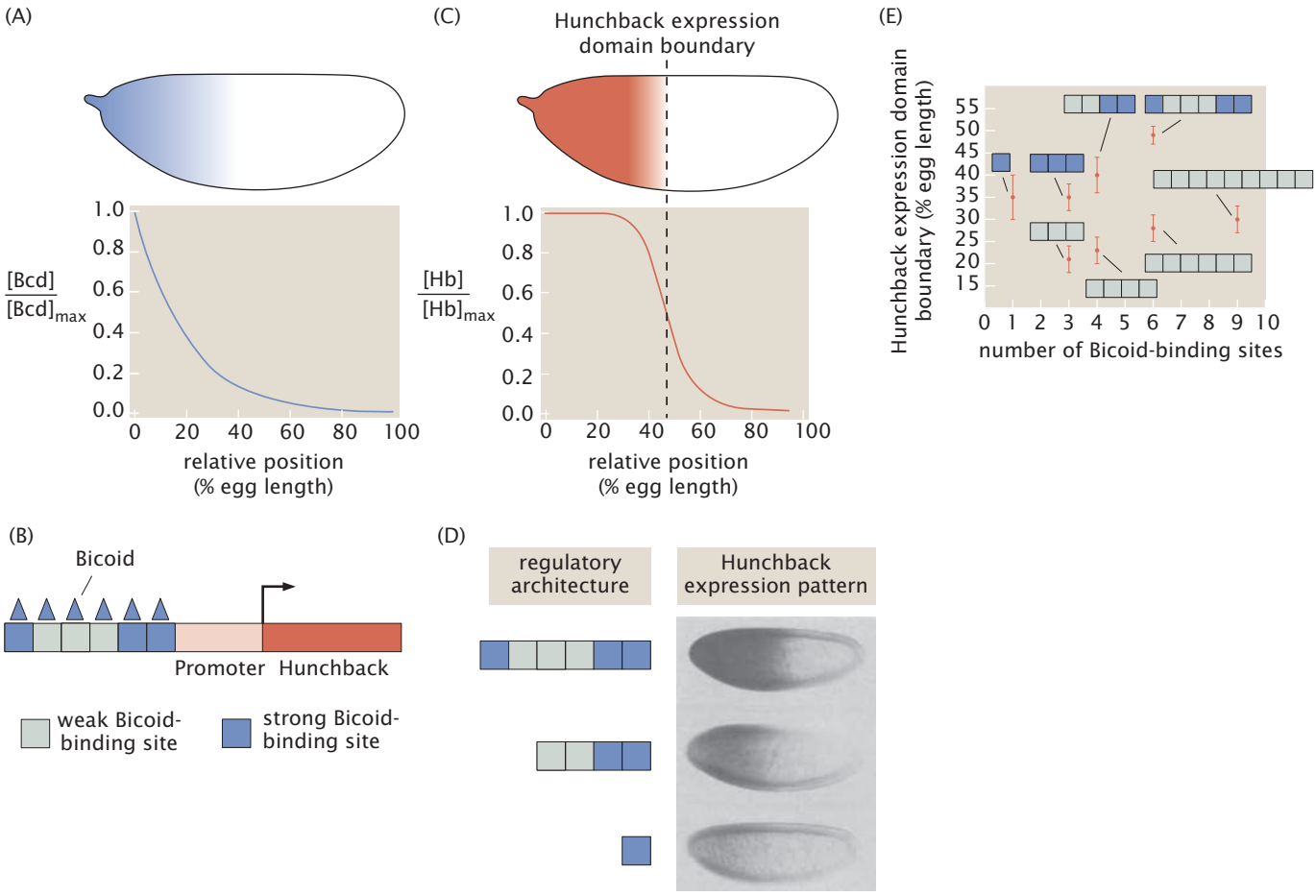


Figure 19.32: Systematic analysis of gene expression in *Drosophila*. (A) The Bicoid transcription factor is expressed in an exponential profile from the anterior to the posterior end of the fly embryo. (B) Bicoid acts as an activator of the Hunchback transcription factor by binding to six binding sites of different strengths located upstream from the Hunchback promoter. (C) The resulting pattern of Bicoid-dependent Hunchback expression domain presents a sharp boundary at about 50% of the embryo length. (D) By creating constructs with different numbers and affinities of binding sites, the boundary of the expression domain can be shifted systematically. (E) Hunchback domain boundary position for several regulatory architectures. (D, E, adapted from W. Driever et al., *Nature* 340:363, 1989.)

genetic control available in fruit flies permits us to query patterns of gene expression for regulatory architectures that have been mutated systematically in much the same fashion as the experiments we have described in the context of the *lac* operon and using the same perturbative philosophy by tuning knobs as shown in Figure 19.30. Figure 19.32(D) shows how different regulatory architectures driving a reporter gene can shift the position of the expression domain. In particular, fewer binding sites corresponds to a shift towards the anterior position, where Bicoid concentrations are higher. This kind of approach was followed systematically for several regulatory architectures, resulting in data like that shown in Figure 19.32(E).

In principle, we can use the same tools presented earlier in the chapter in order to understand the regulatory outcome of these experiments. Given a strength of the array of binding sites there will be a range of Bicoid concentrations over which there will be Hunchback activation. By reducing the number of binding sites or making them weaker, the size of this concentration range is reduced, resulting in a shift of the expression domain. The application of such thermodynamic ideas is spelled out in more detail in our discussion in the next chapter in Section 20.2.3. The work to be described there describes the state of the art in terms of experimental measurements of the dynamics of gene expression in embryonic development in flies and shows how simple thermodynamic models can be used as a basis for discussing these results, though these problems are hugely complex, and simple models like these are akin to highly distorted maps.

19.3 Regulatory Dynamics

19.3.1 The Dynamics of RNA Polymerase and the Promoter

Until now, our treatment of gene regulation has centered on the time-independent output of different regulatory motifs. On the other hand, as is clear from watching the development of any embryo, many of the most beautiful and important questions in regulation center on the orchestration of regulatory decisions over time. Another example that puts questions of the time dependence of gene expression front and center is the study of cells during the cell cycle. As was shown in Chapter 3, entire batteries of genes are expressed at different times during the cell cycle (see Figure 3.23 on p. 119 for a concrete example in the cell cycle of *C. crescentus*). Two of the key dynamical motifs that recur in organisms ranging from bacteria to humans are switches and oscillators. In the case of switches, depending upon some environmental cue, for example, a cell can change the regulatory state associated with particular genes from "off" to "on." Even richer behavior is exhibited by regulatory circuits that give rise to oscillations. So that we can see how switches and oscillators are constructed, we now take up the question of time-dependent gene expression.

The Concentrations of Both RNA and Protein Can Be Described Using Rate Equations

Our conceptual starting point for examining the dynamics of gene expression is the rate equation paradigm introduced in Chapter 15. In particular, we proceed by writing rate equations for the time evolution of the concentrations of various molecular participants in the

regulatory problem of interest. The simplest scenario is to consider a dynamical description that refers only to the time development of the concentrations of the relevant proteins. On the other hand, sometimes it is convenient to characterize the time evolution of the mRNA transcripts as well. In either case, our strategy will be to consider some particular regulatory architecture in which different elements are linked and to write down a dynamical description of their concentrations.

One of the reasons we are forced to go beyond the thermodynamic models favored so far throughout the chapter is the advent of a new generation of experiments aimed at probing regulation. Recent advances on a number of different fronts have now made it possible to query the regulatory response of individual cells. Such experiments make it possible to watch the time evolution of both the mRNAs in individual cells and the proteins they code for.

Examples of the outcome of this recent generation of experiments for the mRNA content of cells are shown in Figure 19.33. One of the immediate insights coming from experiments like those leading to Figure 19.33(B) is that mRNA production is often "bursty." By watching individual cells over time, it becomes evident that there are periods of transcriptional silence occasionally punctuated by bursts of mRNA production. As shown in Figure 19.33(C), these experiments can be used to ask how noisy the gene expression process is, and the calculations in the remainder of the section will confront these questions

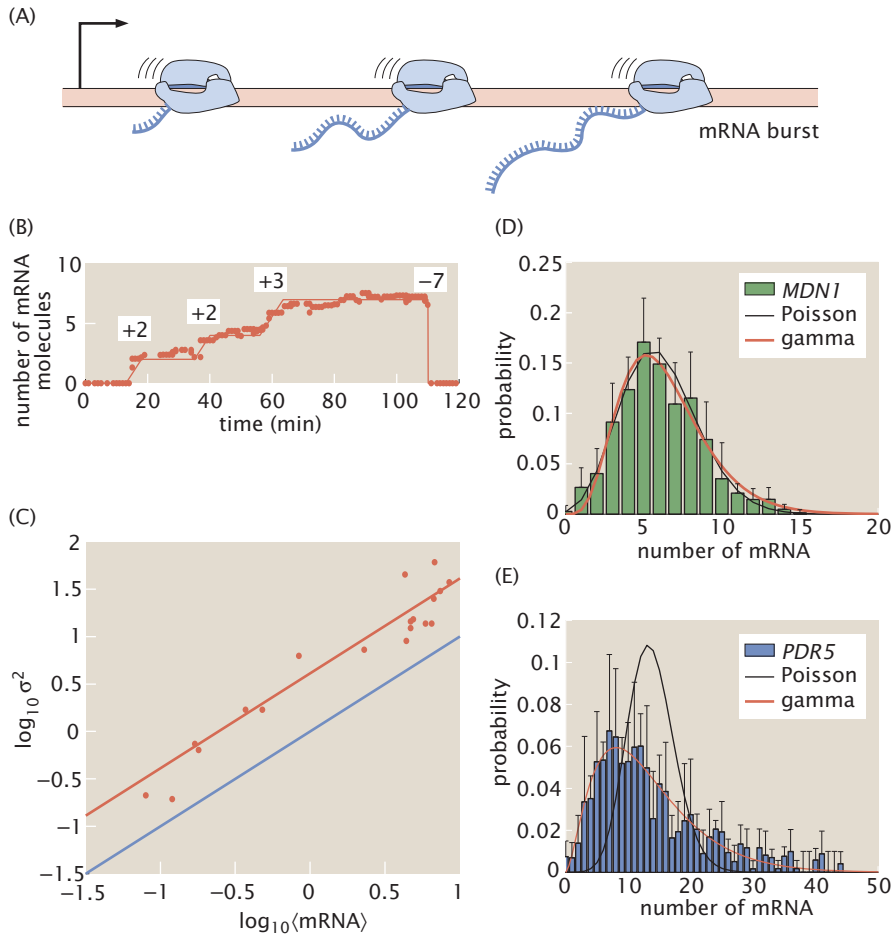


Figure 19.33: Time-dependent dynamics of transcriptional networks. (A) A burst of mRNA production. (B) Time history of the number of mRNA molecules in a given *E. coli* cell, revealing periods of no production punctuated by bursts of production with the size of the burst indicated by the numbers in the white boxes. (C) Noise in *E. coli* gene expression (measured by the variance) as a function of the mean level of gene expression. (D) Distribution of mRNA in yeast for the *MDN1* gene. (E) Distribution of mRNA in yeast for the *PDR5* gene. The distributions in (D) and (E) are fitted to various models considered in the section. (B, C, adapted from I. Golding et al., *Cell* 123:1025, 2005; D, E adapted from D. Zenklusen et al., *Nat. Struct. Mol. Biol.* 15:1263, 2008.)

head on. This noisiness is also revealed by evaluating the entire mRNA distribution over many cells as shown in Figures 19.33(D) and (E).

Before embarking on an analysis of the dynamics of particular regulatory architectures, we return to one of the most elementary questions that can be asked about regulatory dynamics. In particular, our use of statistical mechanics in the previous sections was predicated on the idea that the average amount of transcription from a gene of interest is proportional to the equilibrium occupancy of the promoter by RNA polymerase. This equilibrium assumption is justified when the binding of polymerase to the promoter occurs on a much faster time scale than the time it takes the polymerase to initiate transcription from the bound state. In that case, the polymerase is in rapid pre-equilibrium with the DNA and the amount of transcription is proportional to the fraction of time the polymerase is bound. This is very similar to the arguments put forward in Section 15.2.6 (p. 591), where we examined the conditions under which a chemical reaction can be treated as an equilibrium problem. Interestingly enough, this is not a necessary condition for the equilibrium assumption to hold, as we discuss in the estimate that follows.

Estimate: Dynamics of Transcription by the Numbers

ESTIMATE

The production of mRNA from a typical gene in *E. coli* occurs at a rate of about 10 per minute, while the average lifetime of an mRNA like that for the *lac* operon is a little more than 1 minute (for the distribution of mRNA lifetimes in *E. coli*, see Figure 3.14, p. 110). In steady state, the number of mRNAs created in the cell over any time interval must, on average, balance the number degraded. Since the mRNA molecules are created at 10 per minute, the same number of molecules must be degraded every minute, and we conclude that, on average, there will be 10 mRNA molecules per cell.

If we consider the process of transcription in some detail, it follows a number of biochemical steps. The key steps are RNA polymerase binding to the promoter regulating the gene of interest to form a closed complex with DNA; formation of an open complex in which the two strands of DNA are pulled apart allowing the RNA polymerase to read one of them; and promoter escape, when RNA polymerase begins transcribing the gene. These three steps can be represented by the reaction scheme

$$\mathrm{P} + \mathrm{D} \underset{k_-}{\overset{k_+}{\rightleftharpoons}} \mathrm{PD_c} \xrightarrow{k_{\mathrm{open}}} \mathrm{PD_o} \xrightarrow{k_{\mathrm{escape}}} \text{elongation}, \tag{19.38}$$

where the escape into elongation leaves the promoter in the unbound state, ready to accept a new polymerase. Here, P is free RNA polymerase and D is unbound promoter. $\mathrm{PD_c}$ is the closed complex, while $\mathrm{PD_o}$ is the open complex whose formation is essentially irreversible.

For the binding and unbinding of polymerase to be in rapid pre-equilibrium, the condition $k_\pm \gg k_{\mathrm{open}}$ needs to be satisfied. If RNA polymerase has time to bind and unbind from the promoter multiple times before open complex formation, then we can think of the first step as effectively an equilibrium step characterized by an equilibrium constant $K_{\mathrm{P}} = k_+/k_-$. Indeed, the binding of RNA polymerase to the *lac*UV5 promoter *in vitro* is so fast that the rates $k_\pm$ are not even measured

in typical experiments. Instead, an equilibrium constant of $K_P \approx 200\ \mu M^{-1}$ is measured, while $k_{open} \approx 0.1\ s^{-1}$. This indicates that the rapid pre-equilibrium condition for polymerase binding to this promoter is met.

Once the RNA polymerase has initiated transcription, it elongates at a typical rate of about 50 nucleotides per second, which means that a typical gene of about 1000 nucleotides will be transcribed in about 20 seconds. Given that the average rate of production of mRNA is 10 per minute, this implies that there are about three RNA polymerases per gene at any given time. These numbers are typical for the production of messenger RNAs in *E. coli*, while ribosomal RNA, which is not translated and is one of the key components of ribosomes, is produced at rates of about 1 mRNA per second, almost an order of magnitude faster.

In the case of a regulated promoter such as *lac*UV5, we should also consider the rates at which transcription factors such as the Lac repressor come on and off the regulatory DNA. The diffusion-limited binding rate of Lac repressor to the operator DNA is about $0.003\ s^{-1}\ nM^{-1}$, which corresponds to an on rate of $0.03\ s^{-1}$, assuming 10 repressor molecules in the cell and using our rule of thumb that one molecule in *E. coli* corresponds to a concentration of 1 nM. The dissociation rate from operator DNA varies with the strength of the operator. For Lac repressor, this rate can range from $2\ s^{-1}$ for O3 all the way to $0.002\ s^{-1}$ for Oid. It is interesting to note that these rates are comparable and even slow when compared with the rate of transcription initiation, suggesting that the equilibrium assumption for this promoter might not be valid. However, similar reasoning for the chemical rate equations that are obtained by adding simple repression to the reaction scheme shown in Equation 19.38 leads to the conclusion that equilibrium reasoning is valid in this case as well.

19.3.2 Dynamics of mRNA Distributions

To write the dynamics for the simplest picture of mRNA production, we begin by elucidating the elementary processes that our promoter of interest can undergo in a time step Δt. The dynamics of such a promoter is represented in Figure 19.34, where we see that the

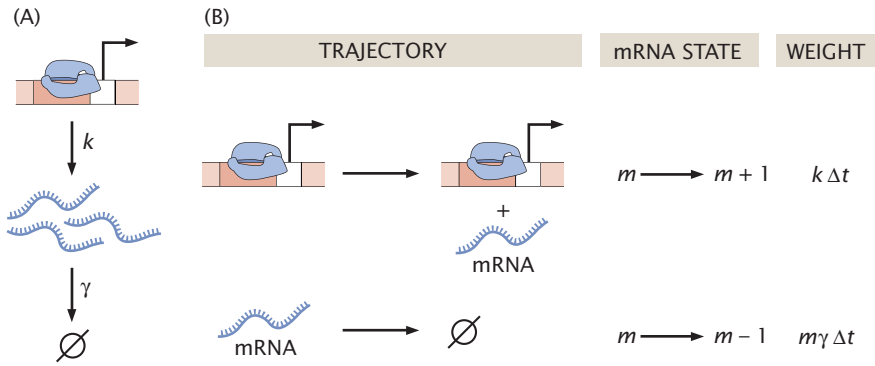


Figure 19.34: Trajectories and weights for the case of an unregulated promoter. (A) Schematic of the processes that can occur for the unregulated promoter, namely, mRNA production with rate k and mRNA decay with rate γ. (B) The individual processes that can occur in a time step of length Δt and their corresponding statistical weights.

mRNA count reflects a competition between the synthesis of new mRNAs with a rate k and their degradation with a rate constant γ. We are interested in determining the dynamics of the probability distribution $p(m,t)$ that tells us the probability of having m mRNA molecules at time t. Using the same "trajectories and weights" strategy adopted earlier and shown in Figure 19.34, we can write the time evolution as

$$\frac{\mathrm{d}p(m,t)}{\mathrm{d}t} = -\underbrace{kp(m,t)}_{m\rightarrow m+1} + \underbrace{kp(m-1,t)}_{m-1\rightarrow m} - \underbrace{\gamma m p(m,t)}_{m\rightarrow m-1} + \underbrace{\gamma(m+1)p(m+1,t)}_{m+1\rightarrow m}. \tag{19.39}$$

The first two terms correspond to the production of new mRNA molecules. Note that one of the terms occurs with a minus sign since the production of another mRNA molecule when we already have m of them leaves the system in a new state with $m+1$ molecules. The last two terms correspond to the situation in which we have m mRNA molecules decaying into $m-1$ molecules and $m+1$ molecules decaying to m molecules, respectively. Care must be taken in the $m=0$ case as the master equation for that state lacks the second term in Equation 19.39, given that the states with a negative number of mRNAs are unphysical. As we will see below, one way to implement this condition is by imposing the condition that $p(m<0,t)=0$.

With this dynamical equation in hand, there are a number of different strategies that can be adopted in order to learn what it implies. One idea is to establish dynamical equations for the moments of the distribution in anticipation of the more challenging situations that arise when we consider a regulated promoter. In those cases, analytic progress aimed at determining the full distribution $p(m,t)$ is very difficult, while the use of the moments such as $\langle m\rangle$ and $\langle m^2\rangle$ is both theoretically tractable and experimentally accessible. In general, we note that the jth moment is given by $\langle m^j\rangle = \sum_{m=0}^{\infty} m^j p(m,t)$. The low-order moments have an intuitive meaning, with the first moment providing the mean and the second moment some measure of the width of the distribution. A second strategy for characterizing our distribution that is also illuminating is to seek the steady-state properties of the distribution. We turn to both of these strategies in the pages that follow.

As our first exercise, we calculate the mean of the distribution $p(m,t)$ corresponding to the first moment of the distribution. To obtain the time evolution of the first moment, we multiply both sides of Equation 19.39 by m and sum over all possible values of m. This results in

$$\sum_{m=0}^{\infty} m\frac{\mathrm{d}p(m,t)}{\mathrm{d}t} = \sum_{m=0}^{\infty} m[-kp(m,t) + kp(m-1,t) - \gamma m p(m,t) + \gamma(m+1)p(m+1,t)]. \tag{19.40}$$

Since the derivative is a linear operation, we can rewrite the left-hand side of this equation as

$$\sum_{m=0}^{\infty} m\frac{\mathrm{d}p(m,t)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\sum_{m=0}^{\infty} m p(m,t) = \frac{\mathrm{d}\langle m(t)\rangle}{\mathrm{d}t}, \tag{19.41}$$

which permits us to now rewrite Equation 19.40 as

$$\frac{\mathrm{d}\langle m(t)\rangle}{\mathrm{d}t} = -k\langle m(t)\rangle + k\sum_{m=0}^{\infty} mp(m-1,t) - \gamma\langle m^2(t)\rangle + \gamma\sum_{m=0}^{\infty} m(m+1)p(m+1,t), \tag{19.42}$$

where we have invoked the definition of $\langle m^j\rangle$.

To make further progress, we need to come to terms with the two sums that are left in Equation 19.42. For the first of these sums, we introduce a new variable $m' = m-1$ such that the sum can be rewritten as

$$\sum_{m=0}^{\infty} mp(m-1,t) = \sum_{m'=-1}^{\infty} (m'+1)p(m',t). \tag{19.43}$$

As mentioned above, because it is unphysical to have a negative number of mRNA molecules, we impose $p(-1) = 0$, which allows us to start the summation over m' from 0 rather than from -1. As a result, we have

$$\sum_{m'=-1}^{\infty} (m'+1)p(m',t) = \sum_{m'=0}^{\infty} (m'+1)p(m',t) = \sum_{m'=0}^{\infty} m'p(m',t) + \sum_{m'=0}^{\infty} p(m',t) = \langle m^1(t)\rangle + \langle m^0(t)\rangle. \tag{19.44}$$

Finally, we enforce the fact that the distribution is normalized to 1 at all time points, which means that $\langle m^0(t)\rangle = \sum_{m=0}^{\infty} p(m,t) = 1$. Using a similar strategy, we can now examine the second sum by introducing a variable $m' = m+1$, resulting in

$$\sum_{m=0}^{\infty} m(m+1)p(m+1,t) = \sum_{m'=1}^{\infty} (m'-1)m'p(m',t) = \sum_{m'=0}^{\infty} (m'-1)m'p(m',t), \tag{19.45}$$

where we have used the fact that $(m'-1)m'p(m',t) = 0$ for $m' = 0$ in order to set the starting point of the sum to zero. As a result, we can rewrite the sum as

$$\sum_{m'=0}^{\infty} (m'-1)m'p(m',t) = \sum_{m'=0}^{\infty} m'^2 p(m',t) - m'p(m',t) = \langle m(t)^2\rangle - \langle m(t)\rangle. \tag{19.46}$$

We are now ready to put this all together and rewrite Equation 19.42 as

$$\frac{\mathrm{d}\langle m(t)\rangle}{\mathrm{d}t} = -k\langle m(t)\rangle + k\langle m(t)\rangle + k - \gamma\langle m^2(t)\rangle + \gamma\langle m(t)^2\rangle - \gamma\langle m(t)\rangle, \tag{19.47}$$

which results in

$$\frac{\mathrm{d}\langle m(t)\rangle}{\mathrm{d}t} = k - \gamma\langle m(t)\rangle. \tag{19.48}$$

This important result gives the rate of change of the mean number of mRNAs in a very simple form. One especially important outcome of this analysis is the insight that the steady-state mean level of mRNA expression is given by $\langle m \rangle = k/\gamma$.

Unregulated Promoters Can Be Described By a Poisson Distribution

One key question about these mRNA distributions is the functional form they adopt in steady state. In particular, what is the form of $p(m)$ after all the initial transients have decayed away? In the case of the simple promoter considered here, we make the educated guess that the distribution is of the Poisson form

$$p(m) = \frac{\lambda^m e^{-\lambda}}{m!}, \tag{19.49}$$

where λ is the mean of the distribution. To evaluate this mean, we resort to the definition

$$\langle m \rangle = \sum_{m=0}^{\infty} m \frac{\lambda^m e^{-\lambda}}{m!} = e^{-\lambda} \sum_{m=0}^{\infty} m \frac{\lambda^m}{m!}. \tag{19.50}$$

As we have done throughout the book, this can be evaluated by exploiting the trick of differentiating with respect to a parameter using

$$m\lambda^m = \lambda \frac{d\lambda^m}{d\lambda}. \tag{19.51}$$

In the context of Equation 19.50, this implies

$$\langle m \rangle = e^{-\lambda} \lambda \frac{d}{d\lambda} \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} = e^{-\lambda} \lambda \frac{de^{\lambda}}{d\lambda} = \lambda. \tag{19.52}$$

We can now use these insights to directly substitute the trial solution into Equation 19.39 for the steady-state case, resulting in

$$0 = -k \frac{\lambda^m e^{-\lambda}}{m!} + k \frac{\lambda^{m-1} e^{-\lambda}}{(m-1)!} - \gamma m \frac{\lambda^m e^{-\lambda}}{m!} + \gamma (m+1) \frac{\lambda^{m+1} e^{-\lambda}}{(m+1)!}. \tag{19.53}$$

This can be simplified to the form

$$0 = \frac{1}{m!} (\lambda\gamma - k) + \frac{1}{(m-1)!} \left(\lambda^{-1} k - \gamma \right). \tag{19.54}$$

We see that if we now use the fact that $\lambda = k/\gamma$, the terms in both sets of parentheses are identically zero, confirming that the Poisson distribution is the appropriate steady-state distribution for this simplest dynamical model of transcription. Another interesting feature of this distribution is that its mean equals its variance.

As mentioned earlier, the study of noise in regulatory networks has become one of the central concerns of regulatory biology in recent years. One measure of this noise is provided by the so-called Fano factor, defined as

$$\text{Fano factor} = \frac{\langle m^2 \rangle - \langle m \rangle^2}{\langle m \rangle}. \tag{19.55}$$

Note that the Fano factor measures the relative size of the variance in mRNA number with respect to its mean. Poisson distributions like that considered above have the very special property that the Fano

factor is equal to 1. One of the questions about the distributions like those shown in Figure 19.33 is whether they exhibit the Poisson form. To make further theoretical progress with this question in the context of this simplest of models of transcription, we need to compute the second moment.

We turn to the same trick as before by multiplying both sides of Equation 19.39 by m^2 and summing over all possible values of m. This results in

$$\sum_{m=0}^{\infty} m^2 \frac{\mathrm{d}p(m,t)}{\mathrm{d}t} = \sum_{m=0}^{\infty} m^2[-kp(m,t) + kp(m-1,t) - \gamma mp(m,t) + \gamma(m+1)p(m+1,t)] \tag{19.56}$$

which, using the same logic as before, can immediately be rewritten as

$$\frac{\mathrm{d}\langle m^2(t)\rangle}{\mathrm{d}t} = -k\langle m^2(t)\rangle + k\sum_{m=0}^{\infty} m^2 p(m-1,t) - \gamma\langle m^3(t)\rangle + \gamma\sum_{m=0}^{\infty} m^2(m+1)p(m+1,t). \tag{19.57}$$

Once again, we have to treat the remaining summations in Equation 19.57. For the first of these sums, we make the change of variable $m' = m - 1$, resulting in

$$\sum_{m=0}^{\infty} m^2 p(m-1,t) = \sum_{m'=-1}^{\infty} (m'+1)^2 p(m',t) = \sum_{m'=0}^{\infty} (m'+1)^2 p(m',t), \tag{19.58}$$

where we have changed the starting m' of the sum. We now expand the binomial, resulting in

$$\sum_{m'=0}^{\infty} (m'+1)^2 p(m',t) = \sum_{m'=0}^{\infty} \left(m'^2 + 2m' + 1\right) p(m',t) = \langle m^2(t)\rangle + 2\langle m(t)\rangle + 1. \tag{19.59}$$

For the second sum, we make an analogous change of variable of the form $m' = m + 1$, resulting in

$$\sum_{m=0}^{\infty} m^2(m+1)p(m+1,t) = \sum_{m'=1}^{\infty} (m'-1)^2 m' p(m',t) = \sum_{m'=0}^{\infty} (m'-1)^2 m' p(m',t). \tag{19.60}$$

Once again, we have changed the lower limit of the sum since including the term $m' = 0$ does not change anything. We can now expand the binomial, resulting in

$$\sum_{m'=0}^{\infty} (m'-1)^2 m' p(m',t) = \sum_{m'=0}^{\infty} \left(m'^2 - 2m' + 1\right) m' p(m',t) = \langle m^3(t)\rangle - 2\langle m^2(t)\rangle + \langle m(t)\rangle. \tag{19.61}$$

We can now return to Equation 19.57 by substituting the outcome of our sums, resulting in

$$\frac{\mathrm{d}\langle m^2(t)\rangle}{\mathrm{d}t} = 2k\langle m(t)\rangle + k - 2\gamma\langle m^2(t)\rangle + \gamma\langle m(t)\rangle. \tag{19.62}$$

In steady state, this becomes

$$0 = (2k+\gamma)\langle m\rangle + k - 2\gamma\langle m^2\rangle. \tag{19.63}$$

If we now exploit the result of our calculation for the first moment, namely, $\langle m\rangle = k/\gamma$, this simplifies to

$$\langle m^2\rangle = \langle m\rangle^2 + \langle m\rangle. \tag{19.64}$$

As a result, we now see that the variance σ^2, the quantity that measures the spread of the data around its mean, is given by

$$\sigma^2 = \langle m^2\rangle - \langle m\rangle^2 = \langle m\rangle. \tag{19.65}$$

We see that in this case, our distribution has the special property that the variance is equal to the mean, implying that the Fano factor is equal to 1. This is a very important result because both the variance and the mean can be measured experimentally. As seen in Figure 19.33(C), the resulting variance as a function of the mean for the bacterial promoter measured in that experiment yields a Fano factor closer to 4 shown by the red line, with the blue signifying the Poisson prediction. This suggests that our simple model is wrong, or at least incomplete. In the next section, we try to find a way of resolving this discrepancy.

19.3.3 Dynamics of Regulated Promoters

How can we respond to data like that shown in Figure 19.33 and 19.35? As seen in Figure 19.35, the number of mRNA molecules in a cell as a function of time is not a smoothly increasing function. Rather, the production of mRNA is bursty. One of the insights emerging from our analysis of the thermodynamic models that could account for this burstiness is the fact that there are all sorts of regulatory interventions that can perturb the presumed steady production of mRNA envisaged in the model considered above.

This lesson is driven home simply by the case of simple repression, which we have considered throughout the chapter. The idea illustrated in Figure 19.36 is that the kind of time history shown in Figure 19.35 might result from the promoter switching back and forth between transcriptionally inactive (that is, repressor-bound) and active (that is, polymerase-bound) states.

The thermodynamic models showed us that promoters can exist in many different states: repressed, empty, occupied by RNA polymerase, activated, etc. The one-state model we explored in the previous section corresponds to the simplest situation in which RNA polymerase is always present at the promoter. When transcription starts and that polymerase molecule leaves the promoter, another RNA polymerase will take its place instantaneously. However, this clearly ignores the many regulatory interventions that are possible as a result of the vast array of transcription factors that are present in a cell.

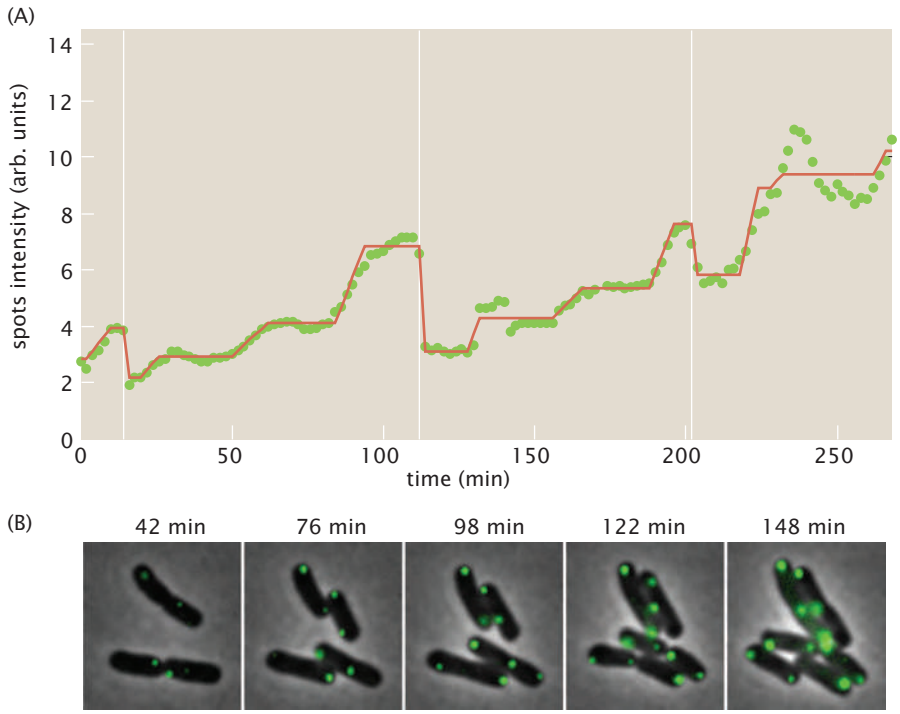


Figure 19.35: Temporal history of mRNA production in *E. coli*. (A) Intensity of spots coming from labeled mRNA molecules as a function of time. Vertical white lines correspond to the cell division process. (B) Microscopy images of cells at various time points in the transcriptional history of the cells. Note that the mRNA molecules in this experiment tend to aggregate. As a result, when counting mRNA molecules, we must consider not only the number of puncta, but also their intensity. (Adapted from I. Golding et al., *Cell* 123:1025, 2005.)

The Two-State Promoter Has a Fano Factor Greater Than One

To increase the realism of our treatment of the kinetics of our promoters, we now include the presence of these transcription factors, which we interpret as conferring different states of promoter activity between which the system can switch back and forth. As a first model, we consider simple repression in which the promoter can switch back

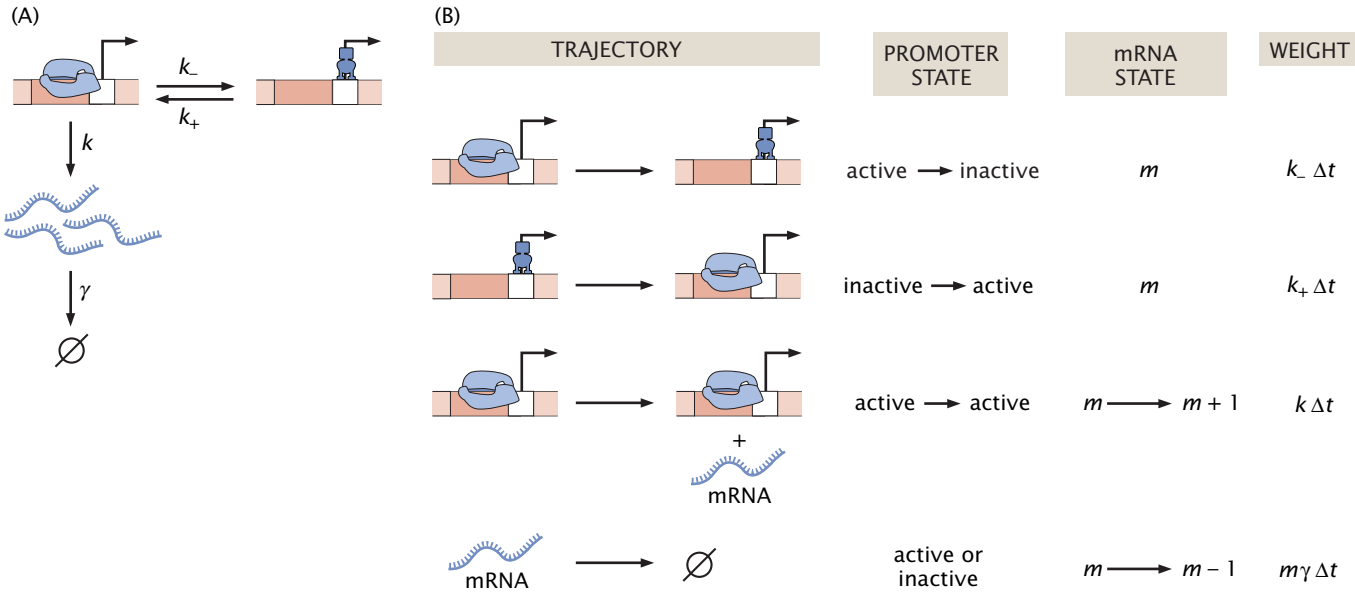


Figure 19.36: Trajectories and weights for the case of simple repression. (A) In an increment of time Δt, the system can suffer several different fates, including switching between the active and inactive states, degradation of an individual mRNA molecule, and production of a mRNA molecule while in the active state. (B) The individual trajectories available to the system in time Δt and their corresponding weights.

and forth between an inactive state (labeled "I") and an active state (labeled "A"). We can describe the kinetics of the promoter using the reactions

$$\mathrm{I} \underset{k_-}{\overset{k^+}{\rightleftharpoons}} \mathrm{A} \overset{k}{\rightharpoonup} \mathrm{mRNA} \overset{\gamma}{\rightharpoonup} \varnothing. \tag{19.66}$$

In the context of this new kinetic model, we now need to keep track of two variables, namely, the state of the promoter (that is, I or A) and the current number of mRNA molecules, m. To do so, we define two different probability distributions corresponding to the two promoter states. $p_\mathrm{I}(m, t)$ describes the probability of having the promoter in the inactive state with m mRNA molecules, whereas $p_\mathrm{A}(m, t)$ describes the probability of finding the promoter in the active state with m mRNA molecules.

As before, our goal is to write equations that describe the time evolution of these probabilities. Intuitively, we see that if we are thinking about how $p_\mathrm{I}(m, t)$ changes over time, there are only a few different processes that can transpire: (i) the promoter can switch from inactive to active, (ii) the promoter can switch from active to inactive, and (iii) an mRNA molecule can decay. This is expressed by the master equation

$$\frac{\mathrm{d}p_\mathrm{I}(m,t)}{\mathrm{d}t} = -\underbrace{k^+ p_\mathrm{I}(m,t)}_{\mathrm{I}\to\mathrm{A}} + \underbrace{k^- p_\mathrm{A}(m,t)}_{\mathrm{A}\to\mathrm{I}} - \underbrace{\gamma m p_\mathrm{I}(m,t)}_{m\to m-1} + \underbrace{\gamma(m+1)p_\mathrm{I}(m+1,t)}_{m+1\to m}. \tag{19.67}$$

The equation describing the time evolution of $p_\mathrm{A}(m, t)$ needs to account for the fact that mRNA molecules can be produced when the promoter is in this state. In this case, we have

$$\begin{aligned}\frac{\mathrm{d}p_\mathrm{A}(m,t)}{\mathrm{d}t} = &\underbrace{k^+ p_\mathrm{I}(m,t)}_{\mathrm{I}\to\mathrm{A}} - \underbrace{k^- p_\mathrm{A}(m,t)}_{\mathrm{A}\to\mathrm{I}} - \underbrace{k p_\mathrm{A}(m,t)}_{m\to m+1} \\ &+ \underbrace{k p_\mathrm{A}(m-1,t)}_{m-1\to m} - \underbrace{\gamma m p_\mathrm{A}(m,t)}_{m\to m-1} + \underbrace{\gamma(m+1)p_\mathrm{A}(m+1,t)}_{m+1\to m}.\end{aligned} \tag{19.68}$$

Our goal is to see what light this model sheds on the Fano factor. To answer this question, we need to evaluate the first two moments of the distribution, namely, $\langle m^1 \rangle$ and $\langle m^2 \rangle$. It is very convenient to define the partial moments $\langle m_\mathrm{I}^j \rangle = \sum_{m=0}^{\infty} m^j p_\mathrm{I}$ and $\langle m_\mathrm{A}^j \rangle = \sum_{m=0}^{\infty} m^j p_\mathrm{A}$ for the inactive and active states, respectively. These partial moments are a mathematical convenience that will allow us to calculate the actual moments of the distribution. In particular, by summing the partial moments of a given order, we have the full moments of the mRNA distribution, namely,

$$\langle m_\mathrm{I}^j \rangle + \langle m_\mathrm{A}^j \rangle = \langle m^j \rangle. \tag{19.69}$$

Solving for the moments with the equations discussed above requires long and cumbersome algebra. As a result, we now resort to a more general matrix scheme that generalizes to any promoter architecture. For this particular case, we begin by defining the vector

$$\mathbf{p}(m,t) = (p_\mathrm{A}(m,t), p_\mathrm{I}(m,t)). \tag{19.70}$$

In addition, we define three matrices $\mathbf{K}$, $\mathbf{R}$, and $\mathbf{\Gamma}$. The matrix $\mathbf{K}$ describes the transitions between the active and inactive states of the

promoter and is given by

$$\mathbf{K} = \begin{pmatrix} -k^- & k^+ \\ k^- & -k^+ \end{pmatrix}. \tag{19.71}$$

The matrix $\mathbf{R}$ describes the production of mRNA,

$$\mathbf{R} = \begin{pmatrix} k & 0 \\ 0 & 0 \end{pmatrix}. \tag{19.72}$$

Finally, the matrix $\mathbf{\Gamma}$ describes the decay of mRNA and is given by

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma & 0 \\ 0 & \gamma \end{pmatrix}. \tag{19.73}$$

Using this notation, the master equations for our two-state promoter can be written as

$$\begin{aligned} \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} &= \mathbf{K}\mathbf{p}(m,t) - \mathbf{R}\mathbf{p}(m,t) + \mathbf{R}\mathbf{p}(m-1,t) \\ &\quad - m\mathbf{\Gamma}\mathbf{p}(m,t) + (m+1)\mathbf{\Gamma}\mathbf{p}(m+1,t). \end{aligned} \tag{19.74}$$

This can be simplified to the form

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = \left(\mathbf{K} - \mathbf{R} - m\mathbf{\Gamma}\right)\mathbf{p}(m,t) + \mathbf{R}\mathbf{p}(m-1,t) + (m+1)\mathbf{\Gamma}\mathbf{p}(m+1,t). \tag{19.75}$$

With these definitions in hand, we can now write matrix equations for the different partial moments. The details are left as an exercise and are discussed in the problems at the end of the chapter, but the general strategy is the same as before: multiply the governing equations by m^i and then sum over all m. Using this strategy, we find that the time evolution of the zeroth moment is given by

$$\frac{\mathrm{d}\langle \mathbf{m}^0 \rangle}{\mathrm{d}t} = \mathbf{K}\langle \mathbf{m}^0(m,t)\rangle. \tag{19.76}$$

For the first moment, similar manipulations result in

$$\frac{\mathbf{d}\langle \mathbf{m}^1 \rangle}{dt} = \left(\mathbf{K} - \mathbf{\Gamma}\right)\langle \mathbf{m}^1(m,t)\rangle + \mathbf{R}\langle \mathbf{m}^0(m,t)\rangle. \tag{19.77}$$

The equation for the time evolution of the second moment then takes the form

$$\frac{\mathrm{d}\langle \mathbf{m}^2 \rangle}{\mathrm{d}t} = (\mathbf{K} - 2\mathbf{\Gamma})\langle \mathbf{m}^2(m,t)\rangle + \left(2\mathbf{R} + \mathbf{\Gamma}\right)\langle \mathbf{m}^1(m,t)\rangle + \mathbf{R}\langle \mathbf{m}^0(m,t)\rangle. \tag{19.78}$$

We interest ourselves in the results for these various moments in steady state. This corresponds to setting the left-hand sides of our dynamical equations to zero and results in the collection of equations

$$1 = \mathbf{u} \cdot \langle \mathbf{m}^0 \rangle, \tag{19.79}$$

$$0 = \mathbf{K}\langle \mathbf{m}^0 \rangle, \tag{19.80}$$

$$0 = \left(\mathbf{K} - \mathbf{\Gamma}\right)\langle \mathbf{m}^1 \rangle + \mathbf{R}\langle \mathbf{m}^0 \rangle \tag{19.81}$$

$$0 = \left(\mathbf{K} - 2\mathbf{\Gamma}\right)\langle \mathbf{m}^2 \rangle + \left(2\mathbf{R} + \mathbf{\Gamma}\right)\langle \mathbf{m}^1 \rangle + \mathbf{R}\langle \mathbf{m}^0 \rangle. \tag{19.82}$$

These equations involve the definition $\mathbf{u} = (1,1)$ as a vector that sums over the components of $\langle \mathbf{m}^j \rangle$. In particular, we have used $\mathbf{u}$ in order to force the normalization of $\langle \mathbf{m}^0 \rangle$.

The idea of the analysis at this point is to bootstrap by successively determining higher-order moments in terms of those we have already determined. If we begin with Equations 19.79 and 19.80, we can determine the partial moments of zeroth order as

$$\begin{aligned}\langle m_{\mathrm{A}}^0\rangle &= \frac{k^+}{k^- + k^+},\\ \langle m_{\mathrm{I}}^0\rangle &= \frac{k^-}{k^- + k^+}.\end{aligned} \tag{19.83}$$

These can be rewritten in terms of the equilibrium constant for the inactive-to-active transition as

$$\begin{aligned}\langle m_{\mathrm{A}}^0\rangle &= \frac{K}{1+K},\\ \langle m_{\mathrm{I}}^0\rangle &= \frac{1}{1+K},\end{aligned} \tag{19.84}$$

where we have defined $K = k^+/k^-$. In fact, what these two partial moments tell us is the probability of finding the system in either the inactive or active states.

With the zeroth moment in hand, we can now turn to Equation 19.81 to determine the first moment. We can rewrite this equation as

$$-\left(\mathbf{K} - \mathbf{\Gamma}\right)^{-1}\mathbf{R}\langle \mathbf{m}^0\rangle = \langle \mathbf{m}^1\rangle. \tag{19.85}$$

After some algebra, this leads in turn to

$$\begin{aligned}\langle m_{\mathrm{A}}^1\rangle &= \frac{k}{\gamma}\frac{k^+ + \gamma}{k^+ + k^- + \gamma}\langle m_{\mathrm{A}}^0\rangle,\\ \langle m_{\mathrm{I}}^1\rangle &= \frac{k}{\gamma}\frac{k^-}{k^+ + k^- + \gamma}\langle m_{\mathrm{A}}^0\rangle.\end{aligned} \tag{19.86}$$

This result tells us that the mean level of mRNA is given by

$$\langle m^1\rangle = \langle m_{\mathrm{A}}^1\rangle + \langle m_{\mathrm{I}}^1\rangle = \frac{k}{\gamma}\langle m_{\mathrm{A}}^0\rangle. \tag{19.87}$$

As expected, the mean level of expression is given by the probability of finding the promoter in the active state times a factor (k/γ) that tells us about the balance of the production and decay of mRNA.

To continue to the point where we can determine the Fano factor, we now need the second moment of the distribution. As a prelude, it is convenient to rewrite Equation 19.82 as

$$-\left(\mathbf{K} - 2\mathbf{\Gamma}\right)^{-1}\left[\left(2\mathbf{R} + \mathbf{\Gamma}\right)\langle \mathbf{m}^1\rangle + \mathbf{R}\langle \mathbf{m}^0\rangle\right] = \langle \mathbf{m}^2\rangle. \tag{19.88}$$

From this equation, we obtain two rather complicated expressions for $\langle m_{\mathrm{A}}^2\rangle$ and $\langle m_{\mathrm{I}}^2\rangle$. However, when we add them together, which is equivalent to evaluating $\mathbf{u}\cdot\langle \mathbf{m}^2\rangle$, we find

$$\langle m^2\rangle = \frac{1}{2\gamma}\left(k\langle m_{\mathrm{A}}^0\rangle + 2k\langle m_{\mathrm{A}}^1\rangle + \gamma\langle m^1\rangle\right). \tag{19.89}$$

Finally, if we use the result from Equation 19.87, this can be rewritten as

$$\langle m^2\rangle = \langle m^1\rangle + \frac{k}{\gamma}\langle m_{\mathrm{A}}^1\rangle. \tag{19.90}$$

Equation 19.90 can be further simplified by appealing to Equation 19.86. The result is given by

$$\langle m^2 \rangle = \langle m^1 \rangle \left(1 + \langle m^1 \rangle \frac{1}{\langle m_A^0 \rangle} \frac{k^+ + \gamma}{k^+ + k^- + \gamma} \right). \tag{19.91}$$

We are now in a position to compute the Fano factor itself, which is given by

$$\frac{\sigma^2}{\langle m^1 \rangle} = \frac{\langle m^2 \rangle - \langle m^1 \rangle^2}{\langle m^1 \rangle} = 1 + \langle m^1 \rangle \left(\frac{1}{\langle m_A^0 \rangle} \frac{k^+ + \gamma}{k^+ + k^- + \gamma} - 1 \right). \tag{19.92}$$

This can be further simplified by using our result for $\langle m_A^0 \rangle$. Making the relevant substitution, we find

$$\frac{\sigma^2}{\langle m^1 \rangle} = 1 + \langle m^1 \rangle \frac{k^-}{k^+} \frac{\gamma}{k^+ + k^- + \gamma}. \tag{19.93}$$

This expression depends only upon the mean level of expression, the switching rates between the on and off states, and the rate of degradation of mRNA. Interestingly, the second term gives us the deviation from a pure Poissonian promoter. To develop intuition for this result, we note that $\gamma/(k^+ + k^- + \gamma)$ is always smaller than one.

To actually compare the results of this analysis with the data revealed in Figure 19.33(C), we have to determine the parameters that appear in our expression for the Fano factor. For the experiments shown in the figure, measurements have shown that these rates are given as follows. First, the mRNA degradation rate is $\gamma = 0.014$ min^{-1}. This degradation rate corresponds to a lifetime of 70 minutes, which is clearly at odds with the average lifetime of mRNA molecules in *E. coli* as shown in Figure 3.14 (p. 110). The reason for this discrepancy is that the experiments performed in Figure 19.33(C) were done using an array of fluorescently tagged mRNA-binding proteins leading to the number of mRNA molecules per cell as a function of time as shown in Figure 19.35. However, the presence of these binding proteins, though useful to detect the production of mRNA molecules, makes the mRNA molecule very stable, such that the only "degradation" is due to dilution upon cell division. Hence, the lifetime of 70 minutes, corresponds to the length of the cell cycle in those experimental conditions. For the case in which the promoter was fully induced (saturating concentration of arabinose), the observed mean number of mRNAs is given by $\langle m^1 \rangle = 10$. In this case, the rates of switching between the on and off states were measured and are given by $k^+ = 0.03$ min^{-1} and $k^- = 0.2$ min^{-1}. If we use these rates, we find that the Fano factor can be evaluated as

$$\begin{aligned} \frac{\sigma^2}{\langle m^1 \rangle} &= 1 + 10 \times \frac{0.2 \text{ min}^{-1}}{0.03 \text{ min}^{-1}} \frac{0.014 \text{ min}^{-1}}{0.03 \text{ min}^{-1} + 0.2 \text{ min}^{-1} + 0.014 \text{ min}^{-1}} \\ &\approx 1 + 3.6 = 4.6. \end{aligned} \tag{19.94}$$

This result is in reasonable accord with the observations reported in Figure 19.33(C). Of course, there are many effects that were not considered in our model. For example, one potential shortcoming of this model is related to the fact that we think about the effects of dilution due to cell division as an exponential process. In reality, this is a discrete process that occurs only at the small time interval corresponding to the separation of the mother cell into the two daughter

cells. The reader is invited to explore the consequences of considering this refined decay mechanism by performing a numerical simulation using the Gillespie algorithm, which we describe in the Computational Exploration below.

Different Regulatory Architectures Have Different Fano Factors

Our use of the master equation approach advocated in this section has focused exclusively on unregulated promoters and the simple repression motif. However, just as with the thermodynamic models, different choices of regulatory architecture lead to different noise profiles (and Fano factors). Figure 19.37 shows the results of the same kind of analysis we have performed in this section to other regulatory architectures.

COMPUTATIONAL EXPLORATION

Computational Exploration: The Gillespie Algorithm and Stochastic Models of Gene Regulation Earlier in this section, we solved the simple case of an unregulated promoter using a master equation approach. By taking into account the different trajectories available to the system and their corresponding weights as shown in Figure 19.34, we were able to calculate both the evolution of the mean mRNA level and the higher moments of the mRNA distribution. Though such master equations are conceptually simple, they can quickly become intractable from an analytic perspective. An alternative approach is to directly integrate the equations numerically. Such a strategy was presented in the computational exploration in Section 3.1.3 (p. 99), where we showed how we can numerically integrate the logistic equation. As discussed there, when performing such numerical integrations, it is key to choose a time step Δt such that this time scale is shorter than all of the intrinsic time scales characterizing the system.

The stochastic nature of chemical reactions is typically revealed in cases where the number of molecular species is small such that random fluctuations are non-negligible. In cases like this, many integration steps Δt can go by without any reaction actually occurring. This has the annoying side effect that much of the computational time is spent effectively doing nothing "interesting." To put this in specific terms, we consider the simple reaction shown in Figure 19.38(A), where a species A can be converted to a species B and vice versa. For example, if we have only one molecule of species A and no molecules of species B, much time can go by without the reaction A $\rightarrow$ B occurring.

One of the strengths of the Gillespie algorithm for solving stochastic differential equations lies in its economy of computation. Instead of following the trajectory of a system by integrating with a Δt that is constant, it provides a strategy to adapt the Δt at each time step of the integration by randomly determining the time to the next possible reaction from a probability distribution, thus avoiding unwanted time steps in which no reaction occurs. The idea behind this algorithm is to construct a particular realization of the stochastic dynamics of the system. The concept is explored schematically for our simple reaction in Figure 19.38(A). To figure out what happens during each step in the simulation, we draw random numbers

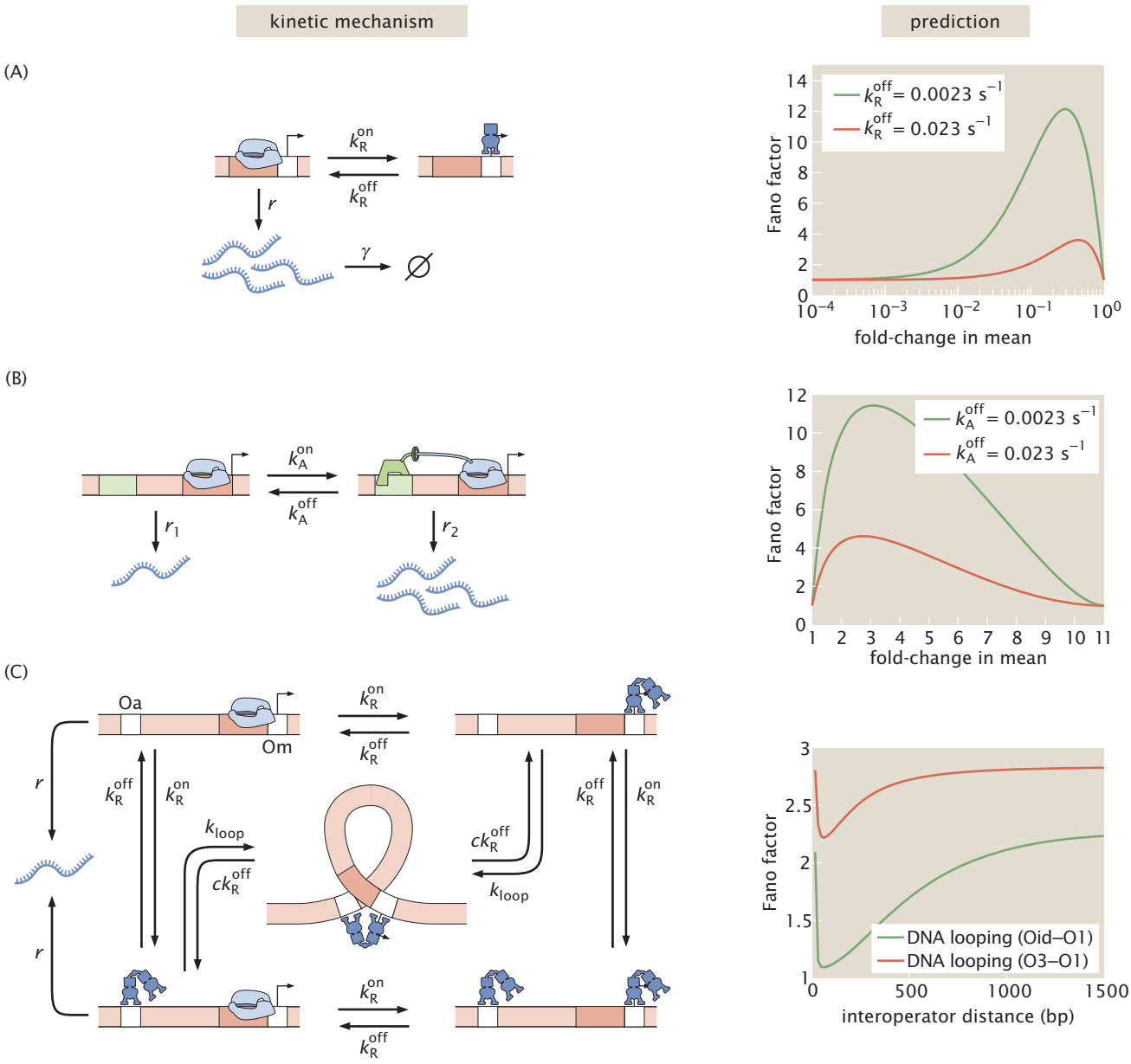


Figure 19.37: Stochastic models of transcriptional regulation for several regulatory architectures. The Fano factor is shown as a function of the fold-change in gene expression for (A) simple repression, (B) simple activation, and (C) repression by DNA looping. The kinetic models assumed for each architecture are shown on the left and their corresponding predictions are shown on the right. For all three figures, we use $r = 0.33$ mRNA s^{-1} and $\gamma = 0.011\ s^{-1}$. The specific parameters for each figure are (B) $r_1 = r$ and $r_2/r_1 = 11$ and (C) $c = 1$, $k_{loop} = [J]k_R^0$, with $[J] = (\ln M)e^{-\beta\Delta F_{loop}}$ the same as given in Figure 19.29 and $k_R^0 = 2.7 \times 10^{-3}\ s^{-1}\ nM^{-1}$, $k_R^{off}(Oid) = 1/(7\ min)$, $k_R^{off}(O1) = 1/(2.4\ min)$ and $k_R^{off}(O3) = 1/(0.47\ s)$. The fold-change in mean gene expression is obtained by varying k_R^{on} and k_A^{on} and is given by (A) $\left(1 + k_R^{on}/k_R^{off}\right)^{-1}$, (B) $\left[\left(k_A^{on}/k_A^{off}\right) r_1/r_2 + 1\right] / \left(k_A^{on}/k_A^{off} + 1\right)$, and (C) $\left[ck_R^{off}(k_R^{off} + k_R^{on})\right] / \left[k_{loop}k_R^{on} + c(k_R^{off} + k_R^{on})^2\right]$. The connection between fold-change in mean gene expression and Fano factor is explored in the problems at the end of the chapter. (Adapted from A. Sanchez et al., *PLoS Comput. Biol.* 7:e1001100, 2011.)

from specific distributions that are detailed below. One of the key insights of this algorithm is that we need to draw a random number twice per step in the reaction. First, we need to draw a random number in order to determine how much time Δt we need to wait until the next reaction takes place. The second drawing of a random number is analogous to a coin flip. We flip a coin (which is unfair) in order to determine which

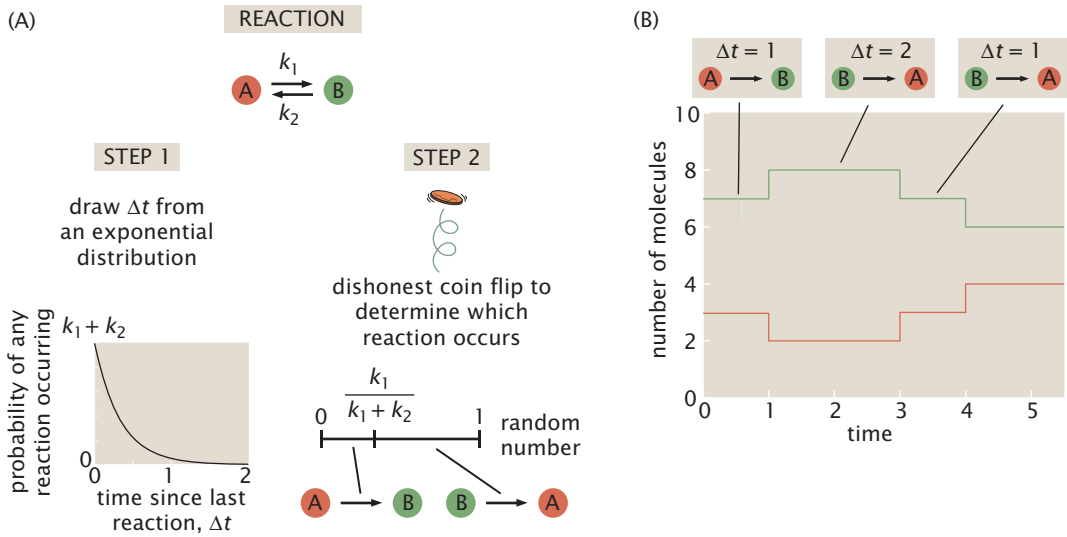


Figure 19.38: Concept of the Gillespie algorithm. (A) Schematic of the Gillespie algorithm for a simple chemical reaction. Here, two random decisions are made. The first step corresponds to drawing a random number from the exponential distribution given by Equation 19.101 in order to determine the time to the next reaction. The second step corresponds to a coin flip that determines which one of the reactions will occur. The bias of the flip is based on the magnitude of the rates corresponding to each possible reaction. (B) Example of a trajectory for the reaction shown in (A).

one of all the possible reactions actually occurred. An example of a realization of the approach for the reaction shown in Figure 19.38(A) is presented in Figure 19.38(B). Here we see how drawing a random number at each step leads to different values for the time interval until the next reaction, Δt, and for which reaction occurred at that time point (either species A to B or species B to A).

How can these ideas be applied to regulatory dynamics? In the unregulated promoter case shown in Figure 19.34(A), we have two possible reactions. First, an mRNA can be produced with a probability k per unit time. Second, an mRNA can decay with a probability γ per unit time and per mRNA molecule. Let's denote the mRNA production reaction as "1" and the mRNA decay reaction as "2" such that their generic rates per unit time will be k_i. This means that $k_1 = k$ and $k_2 = m(t)\gamma$, where $m(t)$ is the number of mRNA molecules at time t. For a time step Δt, we want to determine $P(i, \Delta t)\,\mathrm{d}t$, the probability of reaction i taking place in the time interval $(\Delta t, \Delta t + \mathrm{d}t)$. We construct this probability distribution by first noting that we need to impose that no reaction has already occurred between time points t and $t + \Delta t$. The probability of no reaction before Δt is written as $P_0(\Delta t)$, and hence the probability of the ith reaction occurring between Δt and $\Delta t + \mathrm{d}t$ is given by

$$P(i, \Delta t)\,\mathrm{d}t = P_0(\Delta t) k_i\,\mathrm{d}t, \tag{19.95}$$

where the term $k_i\,\mathrm{d}t$ corresponds to the probability of reaction i occurring in a time step $\mathrm{d}t$. But what is $P_0(\Delta t)$ (that is, what is the probability of no reaction occurring up until the time point Δt)? In order to calculate this, we write an expression for $P_0(\Delta t + \mathrm{d}t)$, the probability of no reaction occurring until the time point $\Delta t + \mathrm{d}t$. This is just the probability that no reactions took place in time Δt multiplied by the probability that

no reactions take place in time $\mathrm{d}t$, namely,

$$P_0(\Delta t + \mathrm{d}t) = P_0(\Delta t)\left(1 - \sum_i k_i\,\mathrm{d}t\right), \tag{19.96}$$

where the index i sums over all possible reactions. If we Taylor-expand $P_0(\Delta t + \mathrm{d}t)$ around Δt, we get

$$P_0(\Delta t + \mathrm{d}t) \approx P_0(\Delta t) + \frac{\mathrm{d}P_0(\Delta t)}{\mathrm{d}\Delta t}\,\mathrm{d}t. \tag{19.97}$$

Comparing the terms in Equations 19.96 and 19.97 leads to the differential equation

$$\frac{\mathrm{d}P_0(\Delta t)}{\mathrm{d}\Delta t} = -P_0(\Delta t)\sum_i k_i. \tag{19.98}$$

This equation can be solved to yield

$$P_0(\Delta t) = \mathrm{e}^{-\sum_i k_i \Delta t} = \mathrm{e}^{-k_0 \Delta t}, \tag{19.99}$$

where we have used the initial condition $P_0(\Delta t = 0) = 1$, stating that at the beginning of our interval no reaction could have occurred yet. We have also defined $k_0 = \sum_i k_i$. As a result, we find

$$P(i, \Delta t)\,\mathrm{d}t = \mathrm{e}^{-k_0 \Delta t} k_i\,\mathrm{d}t. \tag{19.100}$$

In order to make progress, we notice again that the probability distribution shown in Equation 19.100 can be thought of as the product of two probabilities. First, we determine what is the probability of any reaction occurring between time points Δt and $\Delta t + \mathrm{d}t$. In order to do this, we sum the distribution $P(i, \Delta t)\,\mathrm{d}t$ over all possible reactions i,

$$P(\Delta t)\,\mathrm{d}t = \sum_i P(i, \Delta t)\,\mathrm{d}t = \mathrm{e}^{-k_0 \Delta t} k_0\,\mathrm{d}t. \tag{19.101}$$

In Figure 19.38(A), we show this distribution schematically as Step 1 of our algorithm. Here, we see that our algorithm picks the time to the next reaction in a random manner. In particular, the time Δt needs to be picked from the exponential distribution shown in Equation 19.101. To that end, we use a random number x that is uniformly distributed in the interval $(0, 1)$, which is the output of a random number generator in Matlab. From this number, we compute the time interval $\Delta t = (1/k_0)\ln(1/x)$. We can easily convince ourselves that Δt obtained this way is drawn from the exponential distribution shown in Equation 19.101. Namely, if $Q(x) = 1$ is the probability distribution for the variable x, then the probability distribution for Δt satisfies the equation

$$P(\Delta t)\,\mathrm{d}t = Q(x)\,\mathrm{d}x \tag{19.102}$$

which simply states that the probability of finding x in the interval $(x, x + \mathrm{d}x)$ is the same as the probability of finding Δt in the interval that $(x, x + \mathrm{d}x)$ is mapped to by the function $\Delta t(x) = (1/k_0)\ln(1/x)$. (Note that this relationship is very general and very useful when switching from one random variable

to another.) Substituting $Q(x) = 1$ and $x = e^{-k_0 \Delta t}$ into Equation 19.102, and noting that the probability distribution is necessarily positive, leads to $P(\Delta t) = k_0 e^{-k_0 \Delta t}$, as required.

Now that we know when the next reaction will occur, we can ask which one of all the possible reactions will take place. We calculate this probability by integrating Equation 19.100 over time,

$$P(i) = \int_0^{+\infty} P(i, \Delta t)\, dt = \frac{k_i}{k_0}. \tag{19.103}$$

An alternative way to obtain this result is by going back to the definition of k_i as a probability per unit time that reaction i will occur. Given a time interval Δt, the probability of reaction i taking place is then proportional to $k_i \Delta t$ such that

$$P(i) = \frac{k_i \Delta t}{\sum_j k_j \Delta t} = \frac{k_i}{k_0}. \tag{19.104}$$

We see that the probability of reaction i taking place is just given by the the relation between its rate, k_i, and the sum of the rates of all possible transitions in the system, k_0. This is shown schematically as Step 2 in Figure 19.38(A). The bias of our dishonest coin flip is then given by the magnitude of the rates corresponding to the different reactions. An example of how this bias is calculated from a random number picked in the interval [0,1] is shown in the figure.

We are now ready to implement this stochastic algorithm in order to solve for the dynamics of the unregulated promoter from Figure 19.34. For each iteration of the algorithm, we carry out the following steps:

1. Given the current number of mRNA molecules $m(t)$, calculate $k_1 = k$, which does not change as a function of mRNA number, and $k_2 = \gamma m(t)$, which does depend on time through $m(t)$.
2. Calculate a random number x uniformly distributed between 0 and 1 (this is the output of a random number generator). From it, compute the time interval to the next reaction, $\Delta t = (1/k_0) \ln(1/x)$, where $k_0 = k_1 + k_2$. Advance the clock by the time interval Δt.
3. Calculate a random number between 0 and 1. If the number is between 0 and k_1/k_0, then increase the mRNA number by one. If the number is between k_1/k_0 and 1, then decrease the mRNA number by one. Notice that we invoke a number in the interval $[0, 1]$ in an unbiased way. However, we split this interval into $[0, k_1/k_0]$ and $[k_1/k_0, 1]$. The result is that the coin flip is biased based on the values of the different rates.
4. Repeat.

In Figures 19.39(A) and (B), we show the results of the stochastic simulation for an initial condition in which no mRNA molecules are present in the system. In Figure 19.39(A), the time step to the next reaction, Δt, was drawn from the exponential distribution shown in Equation 19.101. In contrast, in Figure 19.39(B), the average time step stemming from that

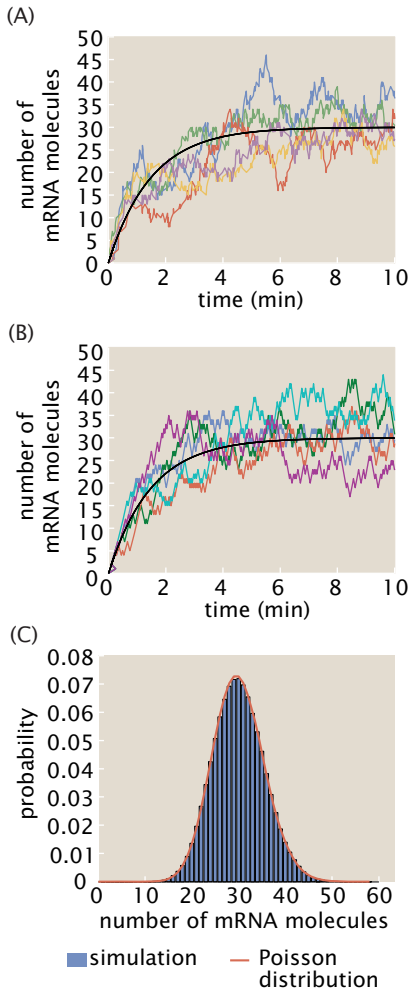


Figure 19.39: Numerical simulation of stochastic effects in the central dogma. (A) Various mRNA trajectories for the unregulated promoter shown in Figure 19.34 using mRNA = 0 as a starting condition. The deterministic solution for the mean number of mRNA molecules per unit time is included for comparison. For these simulations, the time step to the next reaction, Δt, was obtained from the distribution shown in Equation 19.101. (B) mRNA trajectories calculated using the average time step stemming from the exponential distribution shown in Equation 19.101. (C) Steady-state mRNA distribution obtained by the Gillespie algorithm and comparison with the exact solution given by a Poisson distribution. An mRNA transcription rate of 20 min^{-1} and a decay rate of 0.67 min^{-1} $molecule^{-1}$ were used.

distribution was used to carry out the simulations. This is the suggested implementation of the Gillespie algorithm in this Computational Exploration. The reader is invited to explore the statistics of time-stepping for both approaches in the problems at the end of the chapter. Finally, notice how running the algorithm several times leads to different trajectories and how they all compare with the deterministic solution for the mean number of mRNAs given by

$$\langle m(t) \rangle = \left(1 - e^{-\gamma t}\right) \frac{k}{\gamma}. \tag{19.105}$$

In the context of the discussion regarding Figure 19.34, we noticed that in steady state the probability distribution of mRNA molecules can be described by a Poisson distribution with mean k/γ. This can be verified using the Gillespie algorithm as shown in Figure 19.39(B). Here we have run a simulation at steady-state by setting the initial value of the simulation to the deterministic steady state value, $m(t = 0) = k/\gamma$. From the resulting distribution, we determine the number of mRNA molecules as a function of time and generate the corresponding histogram. It is seen that the simulation is in excellent quantitative agreement with the Poisson distribution. This simple example gives a sense of how the simulations can be used as an alternative to the analytic treatment of these systems. The reader is invited to generate graphs like those shown in the figure for him- or herself.

19.3.4 Dynamics of Protein Translation

In the same way that the process of transcription is subject to variability, so are the other steps of the central dogma. For example, even though there might be a well-defined mean number of proteins arising from the translation of a single mRNA, the translation process is subject to variability. To see how this plays out, we start by considering a single mRNA in a cell of interest and compute the probability of n translation events taking place in the lifetime of this mRNA. Later, using this result, we will compute the steady-state protein distribution from a simple model of stochastic transcription and translation, thereby taking into account both key processes of the central dogma. In the simple kinetic model to be exploited here, over each interval of time Δt, two different processes can occur, namely, (i) the mRNA can be translated into a protein and (ii) the mRNA can decay. In Figure 19.40, we show these trajectories with their corresponding weights.

We do the bookkeeping on the mRNA state using the variable m, which keeps track of the number of mRNA molecules. This variable can adopt the values 1 or 0. Initially, we will have one mRNA, such

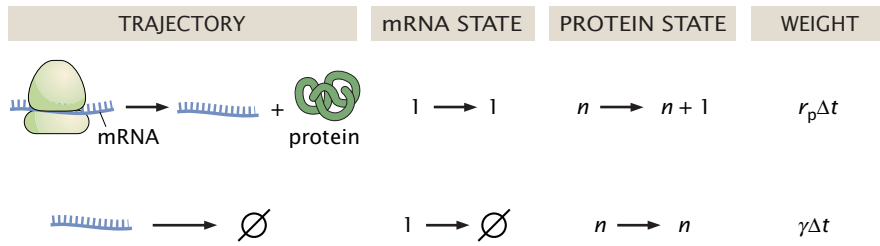


Figure 19.40: Trajectories and weights for translation of a single mRNA molecule. During each interval of time Δt, the mRNA can either be transcribed or decay.

that $m = 1$. However, for long enough times, we know that $m = 0$, as the mRNA will eventually decay. The finite lifetime of mRNA molecules can be appreciated at the genome-wide level in both *E. coli* and yeast as shown in Figure 3.14 (p. 110). We also keep track of the number of proteins through the variable n. Our goal is to calculate the probability distribution $p(n, m; t)\,\mathrm{d}t$, namely the probability that during the time interval $(t, t + \mathrm{d}t)$, the number of mRNAs is m and the number of proteins is n. The master equation describing the time evolution of the state with one mRNA molecule is given by

$$\frac{\partial p(n, m = 1; t)}{\partial t} = -\gamma\, p(n, m = 1; t) + r_\mathrm{p}\left[p(n - 1, m = 1; t) - p(n, m = 1; t)\right]. \tag{19.106}$$

Like in the master equation for mRNA production from an unregulated promoter given in Equation 19.39, we need to be mindful of the $m = 0$ and $n = 0$ cases. For example, since $n < 0$ is unphysical, in the case where there are no proteins, $n = 0$, we need to drop the term in square brackets in Equation 19.106. This can be implemented by imposing $p(n < 0, m; t) = 0$. The state with zero mRNA molecules evolves according to the prescription

$$\frac{\partial p(n, m = 0; t)}{\partial t} = \gamma\, p(n, m = 1; t). \tag{19.107}$$

Solving these coupled differential equations is not straightforward. However, we are concerned with the function $P(n)$ defined as the probability that over the lifetime of the mRNA molecule, n proteins will have been produced. As we will see, the equation for this quantity is more tractable. The probability that the number of proteins synthesized during the lifetime of the mRNA is equal to n is given by $p(n, m = 0, t)$ at very long times, much longer than the decay time $1/\gamma$. In other words, $P(n) = p(n, m = 0, t)$ in the limit $t \to \infty$. We can compute this quantity from Equation 19.107 by integrating both sides from 0 to ∞ and noting that the number of proteins at $t = 0$ is zero (we start the clock when the mRNA is synthesized and no proteins have yet been produced). This results in

$$P(n) = \gamma \int_0^{+\infty} p(n, m = 1; t)\,\mathrm{d}t. \tag{19.108}$$

To make further progress with this, we must compute $p(n, m = 1; t)$ itself. To that end, we integrate both sides of Equation 19.106, resulting in

$$\begin{aligned}\int_0^{+\infty} \frac{\partial p(n, m = 1; t)}{\partial t}\,\mathrm{d}t = &-\int_0^{+\infty} \gamma\, p(n, m = 1; t)\,\mathrm{d}t \\ &+ \int_0^{+\infty} r_\mathrm{p}\left[p(n - 1, m = 1; t) - p(n, m = 1; t)\right]\mathrm{d}t.\end{aligned} \tag{19.109}$$

Using the definition of $P(n)$ given above, we can write this as

$$p(n, m = 1; t \to +\infty) - p(n, 1; t = 0) = -P(n) + \frac{r_\mathrm{p}}{\gamma}\left[P(n - 1) - P(n)\right]. \tag{19.110}$$

This can be further simplified because we know that if we wait long enough, the mRNA will have decayed. As a result, we have $p(n, m = 1; t \to +\infty) = 0$. Further, we can use the initial condition that at time zero there is a single mRNA molecule and no corresponding protein

resulting in the condition $p(n, m = 1; t = 0) = \delta_{n0}$. We are left with the equation

$$-\delta_{n0} = -P(n)\left(\frac{r_p}{\gamma} + 1\right) + \frac{r_p}{\gamma}P(n-1). \tag{19.111}$$

To solve this equation, we make the *ansatz* that $P(n) = A\lambda^n$. We start with the $n = 0$ case,

$$-1 = -P(0)\left(\frac{r_p}{\gamma} + 1\right), \tag{19.112}$$

which results in $A = \gamma/(r_p + \gamma)$. This can now be used in turn for the case when $n > 0$ for which we find $\lambda = r_p/(r_p + \gamma)$, resulting in the distribution

$$P(n) = \frac{\gamma}{r_p + \gamma}\left(\frac{r_p}{r_p + \gamma}\right)^n. \tag{19.113}$$

An interesting quantity to calculate is the mean number of proteins produced per mRNA. This is also called the "burst size," since the idea is that translation and mRNA decay occur on a time scale much faster than any protein decay. From the standpoint of proteins, the result of having one mRNA is a burst of protein production over a small amount of time. The mean can be shown to be

$$\langle n \rangle = \frac{r_p}{\gamma} = b, \tag{19.114}$$

where we have defined b as the burst size. Figure 19.41(B) shows the results from an experiment in which the number of β-galactosidase proteins produced per mRNA was measured in *E. coli*. The curve corresponds to a fit to the distribution calculated above with a mean burst size of five proteins per mRNA, revealing quite reasonable agreement with the distribution.

In the calculation given above, we calculated the probability of obtaining n proteins as the product of translation of a single mRNA molecule. However, the total protein number within a cell is the result of the translation of multiple mRNA molecules. How do we determine the protein distribution in this more complicated case? The most straightforward approach is to write a master equation describing the evolution of both the number of proteins and the number of mRNA molecules within the cell. However, there are approximations that can be made that will simplify that task. For example, we can assume that the lifetime of an mRNA molecule is much shorter than the lifetime of the resulting proteins, as is clearly true for the case of *E. coli* as shown in Figures 3.14 and 3.15 (pp. 110 and 110). As a result, we will see no significant accumulation of mRNA over the life of a protein. Instead, every time an mRNA is transcribed, it will lead to a "burst" of protein production. The trajectories and weights corresponding to this model are shown in Figure 19.42. Note that we use our calculated $P(n)$ to account for the variability in protein production from a single mRNA.

Of course, one way to calculate the actual protein distribution is to solve the master equation stemming from the model shown in Figure 19.42. We define $P_{tot}(n, t)$ as the probability of having n proteins at time t. This equation is then

$$\frac{\partial P_{tot}(n,t)}{\partial t} = -n\gamma_p P_{tot}(n,t) + (n+1)\gamma_p P_{tot}(n+1,t) + \sum_{j=1}^{n} rP(j)P_{tot}(n-j,t) - \sum_{j=1}^{+\infty} rP(j)p(n,t). \tag{19.115}$$

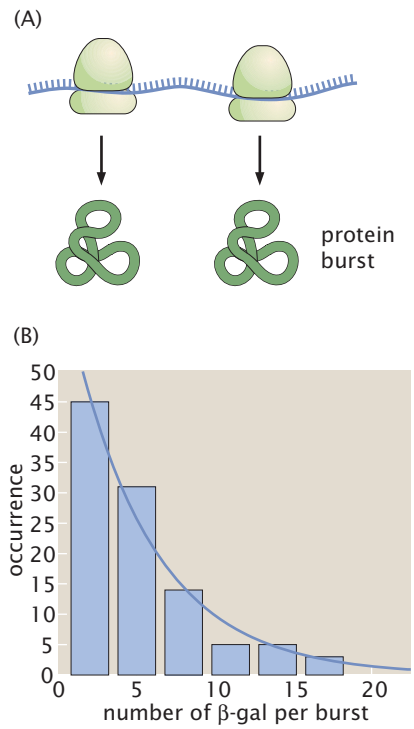


Figure 19.41: Time-dependent dynamics of protein production. (A) Bursts in translation where a single mRNA molecule can give rise to multiple proteins before it decays. (B) Distribution of burst sizes for the production of β-galactosidase in *E. coli* and fit to the probability distribution calculated in Equation 19.113. (B, adapted from Cai et al., *Nature* 440:358, 2006.)

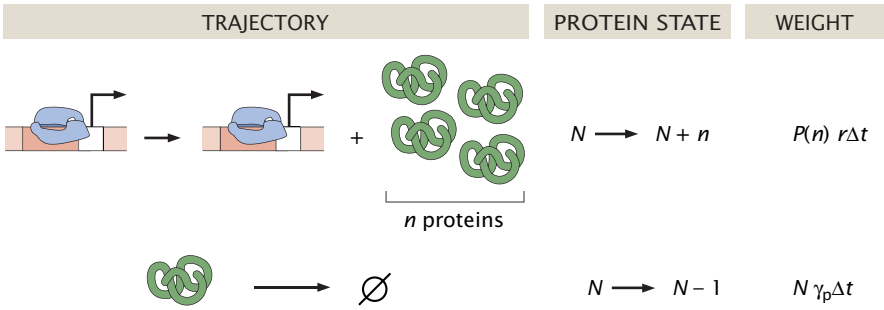


Figure 19.42: Trajectories and weights for a simple model of transcriptional and translational bursts. This model describes the production of mRNA through transcription and the subsequent protein production through the translation of mRNA in a burst before the mRNA decays. The size of the burst is given by $nP(n)$. The main assumption of this model is that the lifetime of an mRNA molecule is much shorter than the protein lifetime.

The last two terms in this equation are related to translation. In the first of these terms we account for all the ways we can go from having $n-j$ proteins to having n proteins from the translation of a single mRNA. This is the reason the probability $P(j)$ shows up in this term. The last term accounts for all the possible ways of leaving the state with n proteins due to the translation of more proteins.

We will solve this equation by making an educated guess for the distribution. We start by rewriting the distribution calculated in Equation 19.113 in terms of the burst size $b = r_p/\gamma$,

$$P(n) = \frac{b^n}{(1+b)^{n+1}}. \tag{19.116}$$

Now, let's say we have two mRNA molecules. We also assume that translation of one mRNA molecule occurs in a completely independent fashion from the second mRNA molecule. Under these conditions, the probability of producing N proteins from the translation of both such mRNA molecules is given by the product of the probability that the first molecule will produce n proteins and that the second will produce $N-n$ proteins, namely,

$$P_2(N) = \sum_{n=0}^{N} P(n)P(N-n). \tag{19.117}$$

Here we have simply used the fact that if one mRNA molecule produces n proteins, the other needs to produce $N-n$ if we want to have a total of N proteins translated. The operation carried out in Equation 19.117 is defined as the convolution of two functions. In particular, here we calculated the convolution of the function $P(N)$ with itself. In the Math Behind the Models at the end of this section, we describe the properties of convolutions in detail.

For three mRNA molecules, we can separate the problem into one and two mRNA molecules, namely,

$$P_3(N) = \sum_{n=0}^{N} P(n)P_2(N-n). \tag{19.118}$$

However, we already have an expression for $P_2(N)$ in terms of $P(N)$, which leads us to

$$P_3(N) = \sum_{n=0}^{N} P(n) \sum_{n'=0}^{N-n} P(n')P(n-n') = \sum_{n=0}^{N} \sum_{n'=0}^{N-n} P(n)P(n')P(N-n-n'). \tag{19.119}$$

What we have just shown is that if we want to calculate the probability distribution for the total number of proteins produced by m mRNA

molecules, then we need to calculate the convolution of the individual probability distributions.

In order to make progress, we will assume that the number of proteins is large enough such that all the sums in the previous equations can be replaced by integrals. For example, for the case of two mRNA molecules, we have

$$P_2(N) = \int_0^N P(n)P(N-n)\,\mathrm{d}n. \tag{19.120}$$

This is again a convolution of $P(n)$ with itself just like in Equation 19.117, but now in integral form. Such integrals are much easier to solve in either Fourier or Laplace space. A careful description of Laplace transforms and how to use them in order to solve convolutions is presented in the Math Behind the Models below. After some algebra, we find the probability of having produced N proteins out of m mRNA molecules as

$$P_m(n) = \left(\frac{b}{1+b}\right)^n \left(\frac{1}{1+b}\right)^m \frac{n^{m-1}}{\Gamma(m)}, \tag{19.121}$$

where $\Gamma(m) = (m-1)!$ for m an integer. This distribution is called the negative binomial distribution (in the limit of large n and $b \gg 1$). Its continuous version is the more popular gamma distribution and can be obtained from Equation 19.121 by taking the limit of large n,

$$P_m(n) \to \frac{n^{m-1}\mathrm{e}^{-n/b}}{b^m\Gamma(m)}. \tag{19.122}$$

Does this negative binomial distribution solve the master equation shown in Equation 19.115? In order to do that, we need to plug our $P_m(n)$ into the master equation. We leave it as a problem to show that m, the mean number of mRNA molecules that a protein sees in its lifetime, is given by $m = r_\mathrm{p}/\gamma$. If a protein is very long-lived, then γ_p will be determined by the cell cycle, since decay will take place due to dilution by division. In that case, we can interpret r_p/γ as the number of mRNA molecules produced per cell cycle. Since each of these molecules leads to a burst of protein production, the inverse of this magnitude is often called the *burst frequency*. Together with the burst size $b = r_\mathrm{p}/\gamma$ the burst frequency fully determines the gamma distribution from Equation 19.122. One interesting thing is that, assuming that the proposed model in Figure 19.42 is correct, one can obtain dynamical information about the transcription and translation process from just looking at steady-state distributions. This concept is shown in Figure 19.43, where we present a strategy to perform such dynamical measurements leading to the values for the burst size and burst frequency. These values are to be compared with those obtained from fitting our gamma distribution to the experimental steady-state protein distribution. As we can see, there is reasonable qualitative agreement between the two techniques, suggesting that at least in an effective way it is valid to think of protein bursts in gene regulation.

The Math Behind the Models: Laplace Transforms and Convolutions Like its cousin the Fourier transform, which we have made use of repeatedly throughout the book, the Laplace transform is a very useful tool in physics for solving linear differential equations, such as those that appear in studies of mechanical and electrical phenomena. The basic premise here, like with other transforms, is to replace the sought function

MATH

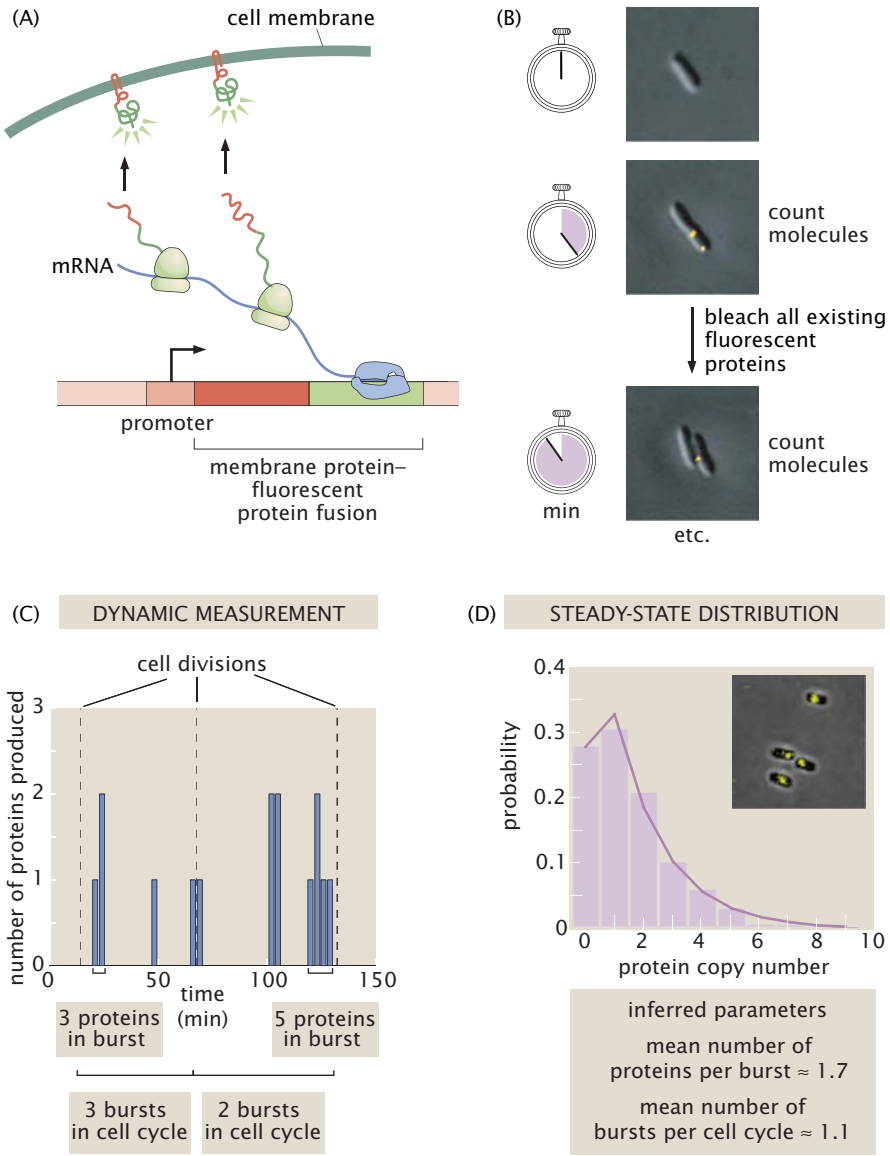


Figure 19.43: Protein bursting in *E. coli*. (A) A membrane protein can be fused to the fluorescent reporter YFP. The translated protein is localized to the membrane, where it can then be quantified. (B) If the level of expression is low, a snapshot of a cell can be taken in order to count proteins stuck to the membrane. The fluorophores can then be photobleached, making it possible to count the number of proteins produced in the next time step, which leads to a protein production rate. Using this method, the effective burst frequency and burst size can be measured. (C) An example of such a measurement of expression dynamics is shown. Here, each bar corresponds to the number of proteins observed on the membrane at that instant in time previous to photobleaching them in order to move to the next time point as shown in (B). Bursts of protein expression over several division cycles can be discerned as clusters of bars, which are associated with translation events off a single mRNA. From many such traces, the mean number of proteins per burst and the mean number of bursts per cell cycle can be calculated. (D) By fitting the steady-state distribution to the discrete gamma distribution from Equation 19.122, the burst frequency and burst size can also be estimated, yielding results comparable to the direct dynamical measurements. Inset: Representative field of view of the cells used to obtain the protein distribution. Note that the results from (C) and (D) correspond to different proteins and cannot be compared directly in a quantitative fashion. (B, C, adapted from J. Yu et al., *Science* 311:1600, 2006; D, adapted from Y. Taniguchi et al., *Science* 329:533, 2010.)

$f(t)$ with its transform $\tilde{f}(s)$, thereby turning the differential equation for $f(t)$ into a much simpler algebraic equation for its transform. The algebraic equation can then be solved for $\tilde{f}(s)$ which in turn can be used to obtain the original $f(t)$ by means of an inverse transform.

The Laplace transform of the function $f(t)$ is given by

$$\tilde{f}(s) = \int_0^{+\infty} f(t) \mathrm{e}^{-st} \, \mathrm{d}t. \tag{19.123}$$

The inverse transform is a more complicated matter, as it involves doing an integral of $f(s)$ in the complex plane, where s is assumed to be a complex number. In practice, one often uses tables of transforms and inverse transforms to solve the problem at hand.

The Laplace transform is particularly useful for solving convolution integrals, such as those that come up in the context of mRNA translation dynamics. The convolution integral of

functions $f(t)$ and $g(t)$ is given by

$$C(t) = \int_0^t f(t')g(t-t')\,\mathrm{d}t' \tag{19.124}$$

and its Laplace transform is related to the Laplace transforms of $f(t)$ and $g(t)$ by the simple relation

$$\tilde{C}(s) = \tilde{f}(s)\tilde{g}(s). \tag{19.125}$$

This can be demonstrated by taking the Laplace transform, as defined by Equation 19.123, of both sides of Equation 19.124, and then making a change of variables $u = t'$ and $v = t - t'$ in the two-dimensional integral that appears on the right-hand side. We leave the mathematical details as an exercise for the interested reader.

In the main text, we have derived the distribution of the number of proteins $P(n)$ obtained by repeated translation of a single mRNA molecule over its lifetime. In order to obtain the equivalent distribution $P_m(n)$ when there are m mRNA molecules in the cell, we make use of the convolution integral. For example, for the case $m = 2$, the sought distribution can be obtained by writing the probability of obtaining n proteins as the product of the probability of making n' proteins by translating the first mRNA and the probability of getting $n - n'$ proteins from the second mRNA molecule. The fact that we can write the probability as a product assumes that translation events from the two mRNAs are independent of each other. Summing over all n' between zero and n then takes into account all the possible ways of ending up with n proteins from two mRNA molecules. If we replace the sum with an integral, we obtain

$$P_2(n) = \int_0^n P(n')P(n-n')\,\mathrm{d}n', \tag{19.126}$$

which is nothing but the convolution of $P(n)$ with itself. We can now repeat the same divide-and-conquer approach for $m = 3$, and the distribution $P_3(n)$ will be given by the convolution integral of $P_2(n)$ and $P(n)$. Further iterations will then yield the sought distribution $P_m(n)$ as the convolution integral of $P_{m-1}(n)$ and $P(n)$.

The laborious procedure of calculating repeated convolution integrals is replaced by a simple multiplication using Laplace transforms. Namely, it follows from Equation 19.125 that $\tilde{P}_2(s) = \tilde{P}(s)\tilde{P}(s)$, and repeated use of Equation 19.125 then yields a simple formula for the Laplace transform of $P_m(n)$,

$$\tilde{P}_m(s) = \left[\tilde{P}(s)\right]^m. \tag{19.127}$$

Therefore, if we want to compute the probability of getting n proteins out of m mRNA molecules, we simply need to calculate the inverse Laplace transform of $\tilde{P}_m(s)$ obtained from this equation.

So, let's start then by calculating the Laplace transform of $P(n)$,

$$\tilde{P}(s) = \int_0^{+\infty} \frac{b^n}{(1+b)^{n+1}} \mathrm{e}^{-sn}\,\mathrm{d}n. \tag{19.128}$$

We can write this as

$$\tilde{P}(s) = \frac{1}{1+b}\int_0^{+\infty}\left(\frac{b}{1+b}e^{-s}\right)^n dn = \frac{1}{1+b}\left.\frac{\left(\frac{b}{1+b}\right)^n e^{-sn}}{\ln\left(\frac{b}{1+b}\right)-s}\right|_0^{+\infty}, \tag{19.129}$$

which leads to

$$\tilde{P}(s) = -\left\{(1+b)\left[\ln\left(\frac{b}{1+b}\right)-s\right]\right\}^{-1}. \tag{19.130}$$

For m mRNA molecules, we need to calculate the inverse Laplace transform of $[\tilde{P}(s)]^m$. This inverse transform has an analytical form, but its calculation is cumbersome. We choose to just quote the result,

$$P_m(n) = \left(\frac{b}{1+b}\right)^n\left(\frac{1}{1+b}\right)^m\frac{n^{m-1}}{\Gamma(m)}, \tag{19.131}$$

where $\Gamma(m) = (m-1)!$ for m an integer. This distribution is called the negative binomial distribution (in the limit of large n and $b \gg 1$).

19.3.5 Genetic Switches: Natural and Synthetic

Switches are an important part of the genetic repertoire of all organisms. To explore the behavior of these switches more carefully, a synthetic version of such a switch was constructed in *E. coli* that had the convenient property that the gene product of the switch is a fluorescent reporter protein such that flipping of the switch can be read out by observing the fluorescent state of the cells. Data from this synthetic switch are shown in Figure 19.44.

The switch described above was constructed by using two repressor proteins whose transcription is mutually regulated as shown in Figure 19.45. This simple design allows us to see one of the most widespread regulatory features, namely, feedback. In particular, the protein that is the output from the first gene serves as a repressor

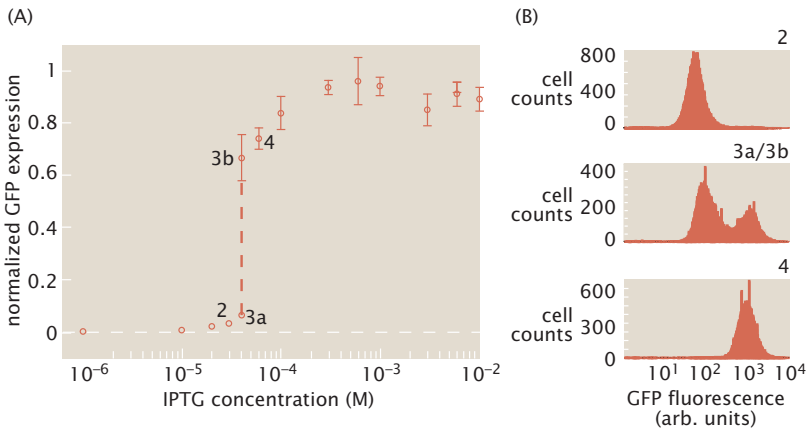


Figure 19.44: Data illustrating the flipping of the genetic switch in *E. coli* cells. (A) Average fluorescence of a population of *E. coli* cells harboring the genetic switch as a function of the concentration of an inducer molecule that flips the switch. In this case, IPTG (a lactose analog) is the inducer, which upon binding to Lac repressor produces an allosteric change that reduces its binding affinity. (B) Flow cytometry data showing the single-cell fluorescence distribution for different inducer concentrations. The labels correspond to points in the curve shown in (A). Bistability is revealed through the fact that there are two populations of cells at the same inducer concentration. (Adapted from T. S. Gardner et al., *Nature* 403:339, 2000.)

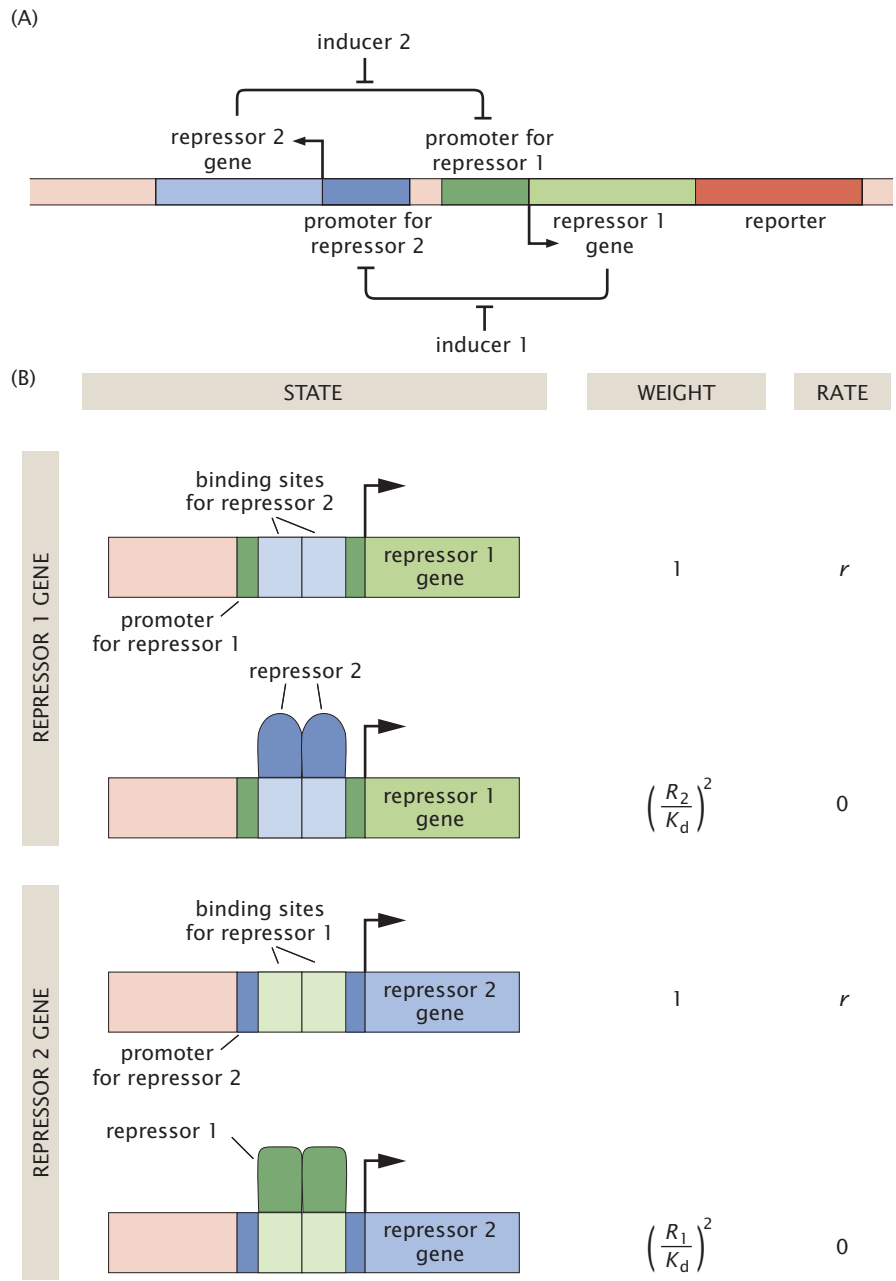


Figure 19.45: Regulatory architecture for a genetic switch. (A) There are two promoters that are under the transcriptional control of the gene product of the partner promoter. (B) States and weights for the two coupled genes making a genetic switch. For the case shown here, the Hill coefficient is $n = 2$ because the repressors bind as dimers. The more general case is considered in the text.

for the second gene. Conversely, the protein that is the output from the second gene serves as a repressor of the first gene. The reader is strongly urged to explore a genetic switch with an even simpler architecture in the problems at the end of the chapter.

We denote the concentrations of the two protein species by c_1 and c_2. We are interested in writing equations for $\mathrm{d}c_1/\mathrm{d}t$ and $\mathrm{d}c_2/\mathrm{d}t$. We consider two classes of processes that can alter the concentrations of these proteins. First, the proteins can be degraded over time. The change in concentration resulting from degradation can be written as $\mathrm{d}c_1/\mathrm{d}t = -\gamma c_1$. Second, protein 2 can bind onto the promoter for protein 1 and repress its production and vice versa. To capture this effect, we introduce a term of the form $r(1 - p_{\mathrm{bound}})$, where r is the basal rate of production and p_{bound} is the probability that the repressor of interest will be bound. When $p_{\mathrm{bound}} = 1$, there is no protein production and when $p_{\mathrm{bound}} = 0$, the rate of protein production takes its basal rate.

Recall from Chapter 6 that for binding described by a Hill function, we have

$$p_{\text{bound}}(c_1) = \frac{K_b c_1^n}{1 + K_b c_1^n}, \tag{19.132}$$

where K_b is the binding constant for the repressor. This implies in turn that the protein production rate for protein 2 is

$$r(1 - p_{\text{bound}}) = \frac{r}{1 + K_b c_1^n}. \tag{19.133}$$

A Hill function (see Section 6.4.3 on p. 273) rather than our statistical mechanical treatment has been used to model p_{bound} so that our treatment is consonant with the original literature. The reader will have the chance to explore the behavior of this circuit using p_{bound} as it has been considered throughout the book in the problems at the end of the chapter. Notice that our treatment of the binding constant here is slightly different than that favored in Section 6.4.3 and Figure 19.45, also for the purposes of consistency with the original literature.

Using the conceptual framework introduced above, the chemical rate equations for the genetic switch are

$$\begin{aligned} \frac{dc_1}{dt} &= -\gamma c_1 + \frac{r}{1 + K_b c_2^n}, \\ \frac{dc_2}{dt} &= -\gamma c_2 + \frac{r}{1 + K_b c_1^n}. \end{aligned} \tag{19.134}$$

The first terms on the right-hand sides of both equations correspond to protein degradation, and for simplicity we assume that the degradation rate (characterized by the parameter γ) of both proteins is the same. For proteins that are stable over time scales longer than the cell cycle (as is the case in the repressors used in this circuit), the dilution rate is determined by the cell doubling time and the subsequent dilution of the protein between the two daughter cells. Therefore, under these conditions, the effective protein degradation rate is the same and is set by the cell division time. The second terms on the right-hand sides of both equations characterize the rate of protein production. As introduced above, the *basal* rate of production is captured in the parameter r. However, this rate is reduced when the repressor is bound to the promoter of interest, as shown above. For simplicity, we assume that the basal production rates and the binding constants that characterize the affinity of the repressors for their binding site are the same for both genes. For a realistic circuit, these assumptions are not necessarily true, but will suffice here to describe the basic operation of the circuit. Another conceptual simplification implicit in these rate equations is the idea that the binding of the repressors is characterized by a Hill function with Hill coefficient n.

From a mathematical perspective, we wonder whether equations like Equations 19.134 yield switch-like solutions. Our assertion is that there are two regions in the space of parameters, one with a single stable solution corresponding to equal concentrations of the two species (decidedly not a switch) and another, more interesting regime, where we find two stable solutions distinguished by having one of the protein concentrations much larger than the other. For values of the parameters where the stable solutions are of this variety, the genetic network exhibits switch-like behavior.

In order to simplify the mathematical analysis of the circuit we resort to a dimensionless form for Equation 19.134. This is achieved by measuring c_1 and c_2 in units of $K_b^{-1/n}$ and time in units of γ^{-1}. This reduces the circuit equations to

$$\begin{aligned} \frac{du}{dt} &= -u + \frac{\alpha}{1+v^n}, \\ \frac{dv}{dt} &= -v + \frac{\alpha}{1+u^n}, \end{aligned} \tag{19.135}$$

where the parameters $\alpha = rK_b^{1/n}/\gamma$ and the Hill coefficient n are the only remaining dimensionless parameters. We have introduced the notation u for the dimensionless concentration of c_1 and v for the dimensionless concentration of c_2. At this point, our goal is to find the steady-state solutions of Equation 19.135 and analyze their stability for different values of α and the Hill coefficient.

To find the steady-state solutions to the rate equations, we set the time derivatives to zero. Since the equations are symmetric with respect to u and v, we immediately conclude that

$$u^* = v^* = \frac{\alpha}{1+v^{*n}} \tag{19.136}$$

is always a solution. Clearly, this result does *not* exhibit the properties of a switch, since the concentrations of both proteins in this case are the same. Are there other solutions that exhibit switching behavior? The equations that determine the steady-state u^* and v^* are of the form $x = f(f(x))$, where $f(x) = \alpha/(1+x^n)$. To see this, solve the first equation for u and substitute that result into the second equation. Since the function f is monotonically decreasing (that is, larger values of x imply $f(x)$ is smaller) the composition $f \circ f$ will be a monotonically *increasing* function, like the function x itself. Therefore, there is the possibility that the two curves x and $f(f(x))$ intersect at more than one point, leading to multiple steady states. The detailed stability analysis is performed in the Math Behind the Models below.

To make these considerations explicit, we consider the case when the Hill coefficient n equals 2, which lends itself to analytic treatment. The steady-state equation for the repressor concentration u^* is

$$u^* = \frac{\alpha}{1+\left(\frac{\alpha}{1+u^{*2}}\right)^2}, \tag{19.137}$$

and the same equation holds for v^*. A little bit of algebra transforms the above equation to a much simpler form given by a product of two polynomials

$$(u^{*2} - \alpha u^* + 1)(u^{*3} + u^* - \alpha) = 0. \tag{19.138}$$

The steady-state solutions to the rate equations for the genetic switch, Equations 19.135, are therefore zeroes of the two polynomials appearing in the above equations.

The cubic polynomial has one real zero, which can be seen from Figure 19.46(A), where we plot the polynomial for different values of α. A mathematically rigorous way to show this is to note that

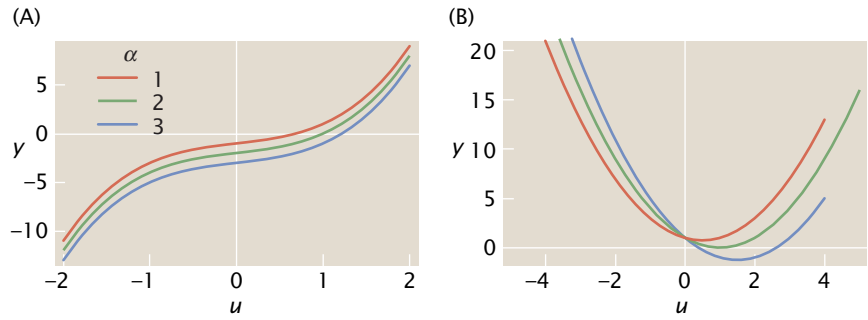


Figure 19.46: Steady-state solutions for protein concentrations in the genetic switch. (A) The function $y = u^3 + u - \alpha$ plotted for various values of α. The solution u^* corresponds to the point at which the curve crosses the u-axis. (B) The function $y = u^2 - \alpha u + 1$ plotted for various values of α. Depending upon the choice of α, there can be 0, 1, or 2 crossings of the u-axis.

the first derivative of this polynomial, $3u^{*2} + 1$, is always positive, which implies that the function is strictly increasing and can therefore intercept the u^*-axis at most once. The equilibrium state that corresponds to the zero of the cubic polynomial has equal concentrations of the two repressor species, since the equation $u^{*3} + u^* - \alpha = 0$ can be rewritten as $u^* = \alpha/(1 + u^{*2})$, and the right-hand side of this equation is v^*.

The quadratic polynomial in Equation 19.138 can have one, two, or no zeroes, depending on the value of α, as observed in Figure 19.46(B). For $\alpha < 2$, there are no zeroes; for $\alpha > 2$, the polynomial has two zeroes; while for $\alpha_c = 2$, the critical value of α, it has one zero at $u^* = 1$. For the two-solution case, the two steady-state values of u^* and v^* correspond to the two different ways of assigning the two roots to each of the dimensionless repressor concentrations. Namely, for a given u^*, the corresponding value of v^* can be calculated using $v^* = \alpha/(1 + u^{*2})$. For these values of u^* and v^*, the equality $u^* + v^* = \alpha$ is satisfied, assuming u^* is one of the zeroes of the quadratic polynomial in Equation 19.138.

In light of the general analysis done above, we see that for $\alpha < 2$ the genetic switch exhibits only one stable equilibrium state with $u^* = v^*$, while for $\alpha > 2$ it has two stable states and one unstable state. In the latter case, the unstable state is the one in which the concentrations of the two repressors are equal, while stable equilibrium states have either repressor u or repressor v in excess.

The dynamical behavior of a system of rate equations like those given in Equations 19.135 can be examined in a different way graphically using the idea of a phase portrait (the mathematics is explained in the Tricks behind the Math at the end of the section). The idea is that we can think of du/dt and dv/dt as the two components of a velocity vector and we can plot the velocity field at every point (u, v). The steady-state solutions will correspond to those points in the phase portrait where the vectors are zero. The solutions represented by those points are stable if for any small excursion away from that point, all the velocity vectors point towards the solution point. An example of this idea for several choices of α is shown in Figure 19.47. The phase portrait provides a convenient graphical representation of the dynamics of the genetic switch. Namely, for a given initial condition u_0, v_0, in order to see how the concentrations will evolve with time, all one has to do is follow the flow depicted by the arrows in the phase portrait. We therefore conclude that the stable steady states of the rate equations are associated with positions in the u–v plane where the phase flow converges from all directions, while diminishing in size, while unsteady states have at least one direction along which the flow is diverging.

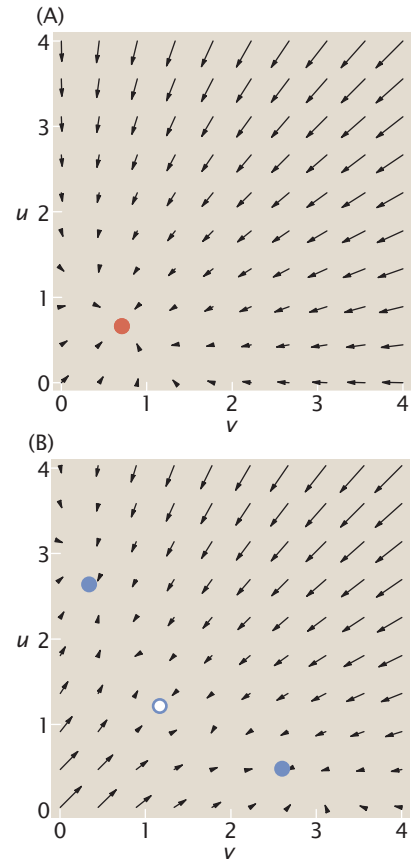


Figure 19.47: Graphical representation of the dynamics of the genetic switch. The phase portraits of the genetic switch for (A) $\alpha = 1$ and (B) $\alpha = 3$. Stable equilibria are represented by filled circles, while the unfilled circle corresponds to an unstable state.

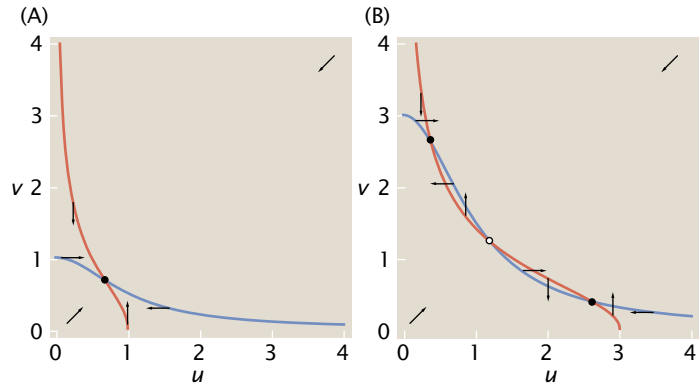


Figure 19.48: Graphical determination of the phase portraits for the genetic switch. Qualitative features of the phase portrait shown in Figure 19.47 can be constructed using the nullclines of the dynamical system described by Equations 19.135. The direction of the phase flow is first determined for large and small values of the dimensionless concentrations of the two repressors, and then the flow in the rest of phase space is determined by continuity. In particular, the direction of the u- or v- component of the flow can only change sign at the nullclines, which are the sets of points along which the rate of change of either u or v vanishes. The intersection of nullclines is a fixed point of the phase flow. (A) Nullclines and flow in phase space for the genetic switch with parameter $\alpha = 1$. The intersection of the two nullclines is a stable fixed point, indicating an absence of switch-like behavior. (B) Nullclines and phase flow in the case $\alpha = 3$. In this case, the two nullclines intersect at three positions, two of which are stable fixed points and one of which is unstable, as indicated by the phase flow. The two stable fixed points correspond to the states that the genetic switch can flip between.

In fact, the directions of all of the arrows shown in Figure 19.47 can be figured out by hand by using the nullclines as shown in Figure 19.48. The most important fact about each nullcline is that one of the two components of $(\mathrm{d}u/\mathrm{d}t, \mathrm{d}v/\mathrm{d}t)$ is zero on each of these nullclines by definition. For example, for the blue curve shown in Figure 19.48(A), we see that the $\mathrm{d}v/\mathrm{d}t = 0$, as evidenced by the horizontal arrows. The idea of a figure like this is that we can draw the arrows on the nullclines themselves and then, largely by exploiting the continuity of the vector fields, can figure out what the arrows are doing elsewhere.

The Tricks Behind the Math: Phase Portraits and Vector Fields As we have seen repeatedly in the book, there are many circumstances in which the dynamics of some system of interest involves *coupled* rate equations of the form

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= f(x, y), \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= g(x, y), \end{aligned} \tag{19.139}$$

where, in general, $f(x, y)$ and $g(x, y)$ are nonlinear functions. The idea of the phase portrait is to graphically depict the "flows" implied by the rate equations. In particular, we imagine a velocity vector field $\mathbf{v}(x, y) = (\mathrm{d}x/\mathrm{d}t, \mathrm{d}y/\mathrm{d}t)$, which depicts which way the system will "move" in the next time step. For a given initial condition (x_0, y_0), we can find the subsequent dynamics of the system by following the arrows.

One of the most classic examples of a dynamical system of the form described above is Lotka–Volterra population dynamics, in which a predator and a prey have their populations coupled. If we think of foxes (F) and hares (H), then the dynamics can be written as

$$\frac{dF}{dt} = FH - F, \tag{19.140}$$

$$\frac{dH}{dt} = -FH + H. \tag{19.141}$$

Effectively, what these equations say is that hares make more hares, and that the fox–hare interaction leads to an increase in foxes and a decrease in hares. An example of the nullclines for this system and the corresponding phase portrait are shown in Figures 19.49(A) and (B). In addition, this figure reveals some of the amusing observations on the dynamics of predator–prey systems.

From the standpoint of stability analysis, the most interesting points in a phase portrait are the fixed points. These are the points at which the vector field satisfies the condition $\mathbf{v}(x^*, y^*) = 0$. In other words, if we choose (x^*, y^*) as an initial

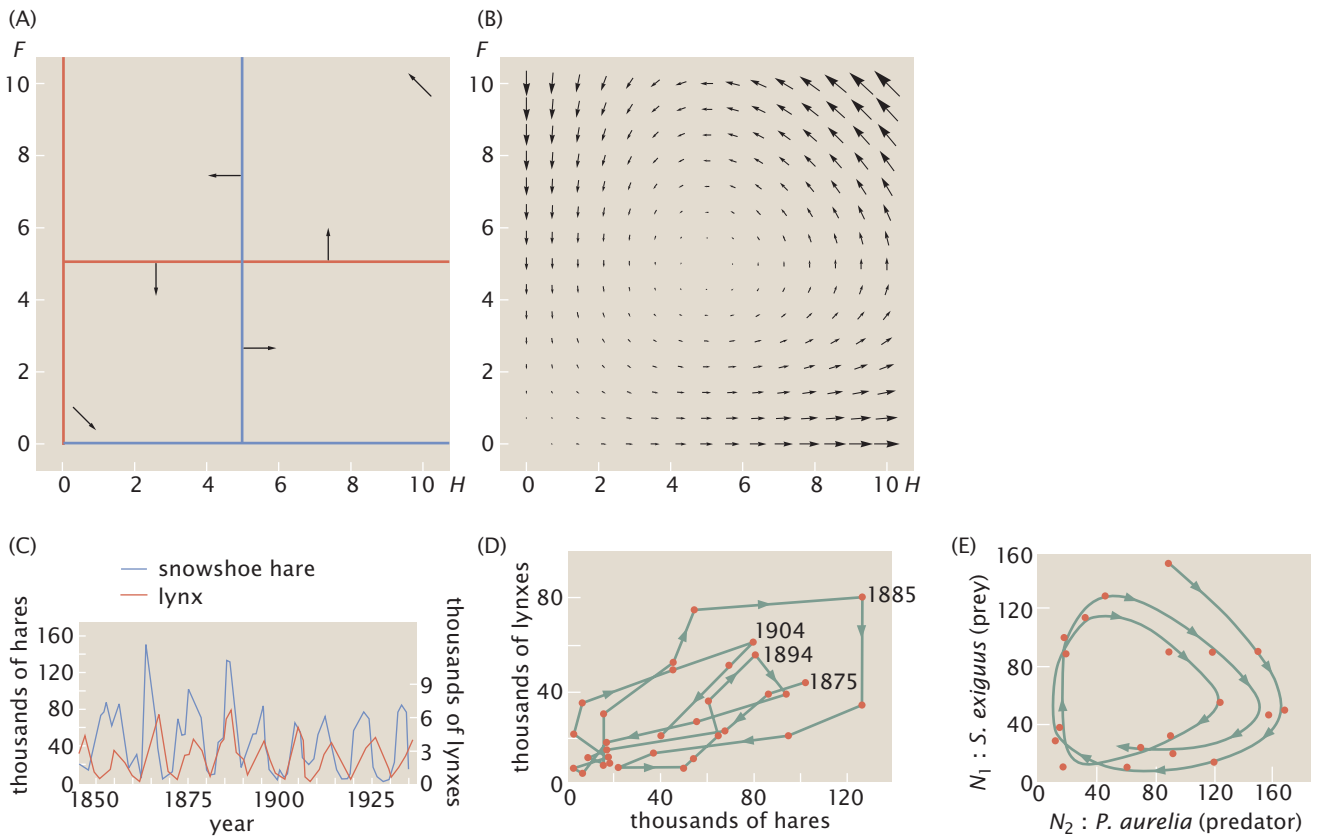


Figure 19.49: Lotka–Volterra model for predator–prey dynamics. (A) Nullclines are used to determine the directions of the flow. (B) Phase portrait for the predator–prey system of differential equations. (C) Population of lynx and hares as a function of time resulting from hunting records. (D) Alternative representation of the lynx and hare population over time, showing the oscillations. (E) Microbial example of population dynamics. (C, D, adapted from T. J. Case, An Illustrated Guide to Theoretical Ecology. Oxford University Press, 2000; E, adapted from G. F. Gause, The Struggle for Existence. Dover Publications, 2003).

condition, the system will stay put. Stability is determined by the directions of the arrows in the neighborhood of the fixed point. If the arrows *all* point back towards that fixed point, the point is said to be a stable fixed point. Otherwise, it is unstable. This type of graphical analysis is a powerful qualitative tool for examining the dynamics of nonlinear coupled equations.

The Math Behind the Models: Linear Stability Analysis for the Genetic Switch One of the key questions we can ask about the solutions found for the switch is the nature of their stability. Of course, one way to characterize the stability is to examine the phase portrait and to look at the directions of all of the little arrows around the various fixed points. We take an alternative approach here in which we search for steady-state solutions of the genetic switch by analyzing the case of α large and α small. First, assume $\alpha \gg 1$. We also assume that the solutions to the steady-state equations, namely,

$$\begin{aligned} u^* &= \frac{\alpha}{1+v^{*n}}, \\ v^* &= \frac{\alpha}{1+u^{*n}}, \end{aligned} \tag{19.142}$$

are such that $u^* \ll 1$. Then $1 + u^{*n} \approx 1$, and the steady-state values for the two concentrations that follow from Equation 19.142, to lowest order in $1/\alpha$, are

$$\begin{aligned} u^* &= \alpha^{1-n}, \\ v^* &= \alpha, \end{aligned} \tag{19.143}$$

consistent with the assumptions we have made. Similarly, by assuming that the solution to Equation 19.142 has the property $v^* \ll 1$, from which $1 + v^{*n} \approx 1$ follows, we find a new solution

$$\begin{aligned} u^* &= \alpha, \\ v^* &= \alpha^{1-n}, \end{aligned} \tag{19.144}$$

for which the roles of u and v are exchanged. Assuming that both u^* and v^* are large leads to $u^* = v^* = \alpha^{1/1+n}$, while the assumption that both are small is inconsistent with Equation 19.142. We conclude that, in addition to the $u^* = v^*$ case, there are two other steady-state protein concentrations. Interestingly, the additional solutions are characterized by very different values for u^* and v^* providing the necessary ingredients for a genetic switch.

Next, we analyze the case $\alpha \ll 1$. Following the same analysis as above, we do not find any additional solutions. Namely, assuming $u^* \ll 1$, we compute from Equation 19.142 $v^* = \alpha$ and $u^* = \alpha$, since now $1 + \alpha^n \approx 1$. The same conclusions are reached assuming $v \ll 1$, while the assumptions $u^* \gg 1$ or $v^* \gg 1$ are not consistent with Equation 19.142. We conclude that there is a critical value of the parameter α, which will be of the order of 1 and dependent on the value of the Hill coefficient n, such that for values of α below the critical value the steady-state solution is unique, while for larger values of α there will be three

steady states. Now, we examine the stability of these solutions, paying particular attention to the case when very different values for u and v are obtained in the steady state.

One of the most important requirements in carrying out an analysis like that given above is to assess the stability of the solutions to a given problem. What this means is that we perturb the system slightly from the steady state (that is, $u = u^* + \delta u$ and $v = v^* + \delta v$) and we ask if the perturbations grow or shrink in time. If the perturbations grow in time, the system is said to be unstable. If the perturbations shrink in time, the system is said to be stable. A favorite example for depicting this idea is to consider a particle on some potential-energy landscape. If the particle is at the bottom of a well (that is, the potential energy is locally of the form $\frac{1}{2}kx^2$), then a small disturbance of the particle from its equilibrium position will result in jiggling around the equilibrium point. Alternatively, if the particle is balanced at the point $x = 0$ on a potential-energy landscape of the form $-\frac{1}{2}kx^2$, then any slight disturbance to the particle will cause it to wander away from the equilibrium. The idea of our stability analysis in this case is the same—we ask whether a slight disturbance away from the steady-state concentration will lead to solutions that grow or decay in time.

To assess the stability of the steady state, we analyze the linear equations for the small deviations $(\delta u, \delta v)$ of the repressor concentrations away from their steady-state values. In particular, in Equation 19.135, we substitute $u = u^* + \delta u(t)$ and $v = v^* + \delta v(t)$ and then exploit the fact that $\delta u(t)$ and $\delta v(t)$ are small and Taylor-expand the nonlinear Hill functions in powers of δu and δv. The result of this analysis is

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\delta u\\ \delta v\end{pmatrix} = \mathbf{A}\begin{pmatrix}\delta u\\ \delta v\end{pmatrix}. \tag{19.145}$$

The matrix $\mathbf{A}$ given by

$$\mathbf{A} = \begin{pmatrix}-1 & f'(v^*)\\ f'(u^*) & -1\end{pmatrix} \tag{19.146}$$

results from linearizing the rate equations, Equation 19.135, around the steady-state solution (u^*, v^*), and

$$f'(x) = -\frac{n\alpha x^{n-1}}{(1+x^n)^2}. \tag{19.147}$$

At this point, the stability of this *linear* set of equations is queried by assuming solutions of the form $\delta u(t) = \delta u_0 \mathrm{e}^{\lambda t}$ and $\delta v(t) = \delta v_0 \mathrm{e}^{\lambda t}$. The essence of the analysis is to examine the sign of the parameter λ. If $\lambda < 0$, the perturbations decay in time, and if $\lambda > 0$, the perturbations grow in time. The behavior of λ is revealed by examining the eigenvalues of the matrix $\mathbf{A}$. The eigenvalues of $\mathbf{A}$ are both real and are given by

$$\lambda_{1,2} = -1 \pm \sqrt{f'(u^*)f'(v^*)}. \tag{19.148}$$

For the steady-state solution to be stable, both λ_1 and λ_2 need to be negative. This will be the case if

$$f'(u^*)f'(v^*) < 1. \tag{19.149}$$

Given this condition for the stability of the solutions, we can now revisit the different solutions found above and explicitly examine their stability. First we consider the single steady state, $u^* = v^* = \alpha$, that we found for $\alpha \ll 1$. In this case, using Equation 19.147, we find $f'(u^*)f'(v^*) = n^2\alpha^{2n} \ll 1$, and the stability condition, Equation 19.149, is satisfied. Next, we consider the three steady-state solutions found for $\alpha \gg 1$. For the solution $u^* = v^* = \alpha^{1/(1+n)}$, we find that $f'(u^*)f'(v^*) = n^2$. Since the Hill coefficient satisfies the condition $n > 1$, we conclude that this solution is unstable. A small perturbation will drive it to one of the other two solutions, which are stable. Namely, for $u^* = \alpha^{1-n}$ and $v^* = \alpha$, we see that $f'(u^*)f'(v^*) = n^2\alpha^{-n(n-1)} \ll 1$, and we conclude that the solution is stable. Since the third solution is obtained by u^* and v^* switching roles, it too will be stable.

The analysis above leads to the phase portrait shown in Figure 19.47 in terms of the parameter α. For α less than some critical value (which is of the order of 1), the rate equations at long times lead to a unique steady state in which the concentrations of the two repressor proteins are equal. On the other hand, for α larger than the critical value, at long times the system will settle into one of two stable states, with the concentration of one repressor dominating over the other. Which of the two steady states is reached depends on the initial conditions. In this regime the rate equations, Equation 19.134, describe a genetic switch.

19.3.6 Genetic Networks That Oscillate

In addition to switches, another dynamical element that is ubiquitous in cell dynamics is an oscillator where one or more chemical species in the cell vary in time in a periodic fashion. There are numerous ways that an oscillator can be built up from a collection of interacting genes and proteins, and here we examine a very simple example of a relaxational oscillator that makes use of two transcription factors, namely, a repressor and an activator. The repressor binds as a dimer and represses the production of the activator, while the activator increases its own production and that of the repressor, also binding as a dimer to the promoter DNA. The states and weights corresponding to this architecture are shown in Figure 19.50.

To write the chemical rate equation for the repressor and activator proteins, we assume that they are produced at a constant rate that depends on the particular state of the promoter, and that the probability of finding the promoter in one of its possible states is given by equilibrium considerations discussed earlier in this chapter. Using the states-and-weights diagrams in Figure 19.50 to compute the equilibrium probabilities for the different promoter states, we obtain the following rate equations for the evolution of the concentration of activator and repressor:

$$\frac{dc_A}{dt} = -\gamma_A c_A + r_{0A}\frac{1}{1 + (c_A/K_d)^2 + (c_R/K_d)^2} + r_A\frac{(c_A/K_d)^2}{1 + (c_A/K_d)^2 + (c_R/K_d)^2}, \tag{19.150}$$

$$\frac{dc_R}{dt} = -\gamma_R c_R + r_{0R}\frac{1}{1 + (c_A/K_d)^2} + r_R\frac{(c_A/K_d)^2}{1 + (c_A/K_d)^2}, \tag{19.151}$$

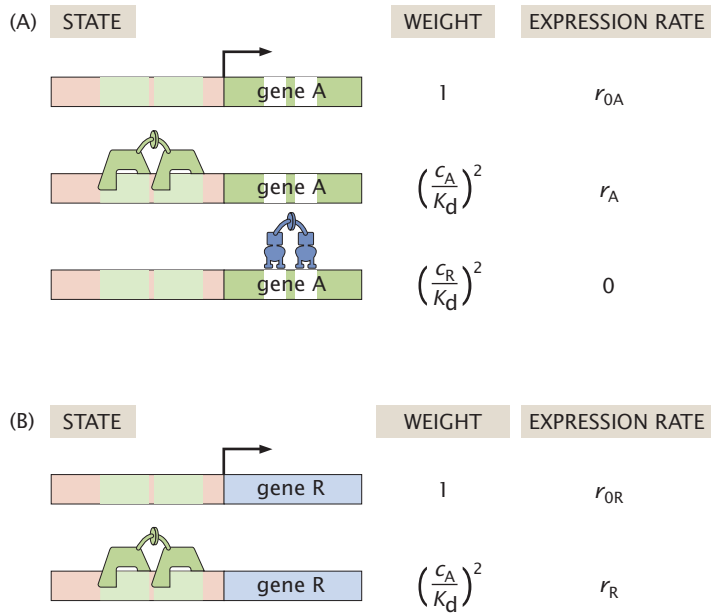


Figure 19.50: States and weights for the promoters that control the expression of activator and repressor proteins. (A) Regulation of the activator gene. The gene product has positive feedback and activates its own expression. (B) Regulation of the repressor gene. The presence of activator stimulates production of repressor.

where γ_R and γ_A are the degradation rates of the receptor and activator proteins, r_{0R} and r_{0A} are their basal rates of production, while r_R and r_A are the rates of production of these two protein species in the presence of activator bound to the promoter DNA. To simplify the analysis of the rate equations, we make the assumption that the same equilibrium dissociation constant K_d describes activator and repressor binding to promoter DNA.

As in the case of the genetic switch, we begin the analysis of the rate equations by writing them in dimensionless form. To that end, we use $1/\gamma_R$ as the unit of time and K_d as a unit of concentration. The rate equations for the dimensionless activator and receptor concentration are then given by

$$\begin{aligned} \frac{d\tilde{c}_A}{dt} &= -\tilde{\gamma}_A \tilde{c}_A + \frac{\tilde{r}_{0A} + \tilde{r}_A \tilde{c}_A^2}{1 + \tilde{c}_A^2 + \tilde{c}_R^2}, \\ \frac{d\tilde{c}_R}{dt} &= -\tilde{c}_R + \frac{\tilde{r}_{0R} + \tilde{r}_R \tilde{c}_A^2}{1 + \tilde{c}_A^2}. \end{aligned} \tag{19.152}$$

Oscillations can arise when there is a separation of time scales between the activator and repressor dynamics. To gain intuition about this, we plot the nullclines for the activator and repressor shown in Figure 19.51(A). The nullclines are the steady-state values of repressor and activator for fixed amount of activator and repressor, respectively. They are obtained by setting the time derivatives of the activator and repressor concentration in Equation 19.152 to zero and solving for the corresponding concentration of activator and repressor. This yields

$$\tilde{c}_R = \sqrt{-1 - \tilde{c}_A^2 + \frac{\tilde{r}_{0A} + \tilde{r}_A \tilde{c}_A^2}{\tilde{\gamma}_A \tilde{c}_A}} \tag{19.153}$$

and

$$\tilde{c}_R = \frac{\tilde{r}_{0R} + \tilde{r}_R \tilde{c}_A^2}{1 + \tilde{c}_A^2} \tag{19.154}$$

for the two nullclines.

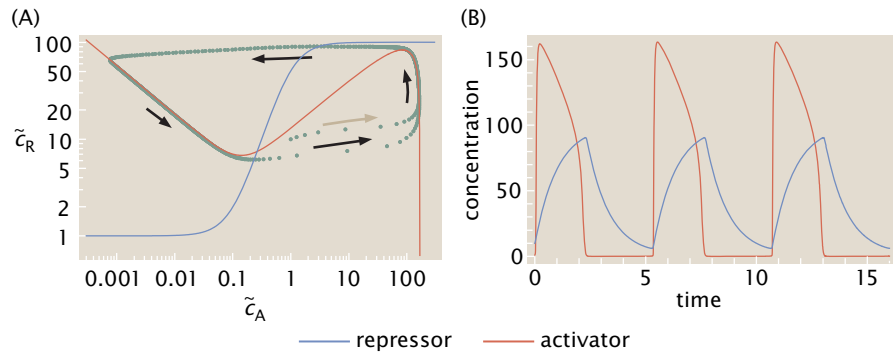


Figure 19.51: Dynamics of a genetic oscillator. (A) Nullclines for the two-component genetic oscillator for parameter values $\tilde{r}_{0R} = 1$, $\tilde{r}_R = 100$, $\tilde{r}_{0A} = 100$, $\tilde{r}_A = 5000$, and $\tilde{\gamma}_A = 30$. The light arrow indicates the initial transient and the dark arrows illustrate the limit cycle. (B) Solutions to the rate equations with initial conditions $\tilde{c}_A = 1$ and $\tilde{c}_R = 10$ using the same parameters as in (A).

If the repressor dynamics is much slower than the activator dynamics, then the activator dynamics will quickly reach its steady-state value for a given repressor concentration. In other words, at any instant in time, the amount of activator can be read off from its nullcline given the current concentration of repressor. Keeping this in mind, we can follow the progression of the dynamical system by starting initially with a small amount of repressor and activator, as shown in Figure 19.51(A). The activator concentration quickly reaches its steady state, which for a small amount of repressor is a large concentration of activator. Note that the green points in Figure 19.51(A) represent positions in phase space at equal time intervals, so a dense interval of points indicates slow phase flow, while a sparse one corresponds to fast flow. Once a large concentration of activator is obtained, this leads to a slow increase in the repressor concentration and the phase trajectory follows the right portion of the activator nullcline, as shown in Figure 19.51(A). When the repressor concentration rises above a critical value for which the steady-state activator concentration is small, the activator concentration quickly drops to this very small value, as indicated by the switch of the trajectory from the right to the left side of the activator nullcline. In response to this sudden drop in activator concentration, repressor concentration drops as well, but slowly, and the trajectory tracks the left side of the activator nullcline. Eventually, the repressor concentration drops below a critical value and the activator concentration jumps to a large, steady-state value (corresponding to the fast switch from the left to the right part of the activator nullcline) and the cycle repeats. Precisely this kind of progression, which is generally characteristic of relaxation oscillators, is shown in Figure 19.51(B) where we plot the concentrations of both activator and repressor over a few cycles.

19.4 Cellular Fast Response: Signaling

Gene regulatory networks are clearly of central importance to the functioning of organisms of all types. Of course, there are many aspects of biology where the dynamics of regulation is critical that do not involve gene transcription as an ultimate outcome. This is particularly obvious for biological behaviors that simply occur too quickly for transcription of new genes to have any useful impact. Rather, these *signaling* networks involve batteries of proteins and their partner ligands connected together such that their interactions affect the activity of some enzyme. For example, a membrane-spanning receptor might bind a ligand in the extracellular space. As a result of this binding

event, there will be a concomitant structural change on the intracellular domain of this same protein, activating a protein kinase enzyme activity, which results in the phosphorylation of some other protein, rendering it active. The goal of the remainder of this chapter is to examine some examples of this kind of signaling and to construct simple models of their behavior.

19.4.1 Bacterial Chemotaxis

One fascinating and fairly well-understood example of signal transduction that we have mentioned briefly is the case of bacterial chemotaxis. Bacteria import small nutrients such as sugars and amino acids to use as building blocks, as we calculated in Chapter 3. A bacterial cell must take up a huge number (in excess of 10^9) of glucose molecules to go through a cycle of cell division. Obviously, this can be done more rapidly in areas of higher ambient glucose concentration. It therefore behooves the bacterium to actively seek out regions of its watery environment that contain the highest accessible concentration of glucose. An elegant and extraordinarily efficient system has evolved for this purpose. Several highlights of how we came to understand the workings of this system are sketched in the Experiments Behind the Facts below.

As we mentioned in Section 4.4.4 (p. 159), the motor used for swimming by the class of bacteria including *E. coli* and *Salmonella* is a rotary propellor that spins a long flagellum (each bacterial cell has several flagella that all work in synchrony). The only known control point the bacterium has for the rotor is to alter its direction of spin to be either clockwise or counterclockwise. Counterclockwise rotation of the flagella drives the bacterium forward in a nearly straight "run," while clockwise rotation causes the flagellar bundle to become disorganized and the bacterium "tumbles," randomly changing its direction. The chemotactic signal transduction machinery regulates this directional switching. If desirable nutrients are present at high concentrations, the bacterium tends to keep moving in a straight line, tumbling less frequently. If nutrient concentrations are low, the bacterium tends to tumble more frequently. *E. coli* is able to use the patterns of directional switching generated by this signal transduction network to swim up gradients of desirable nutrients. Some of the key elements of how this important and fascinating network works were indicated schematically in Figure 4.16 (p. 160).

How can a binary switch be used to detect the direction of a gradient? We can imagine at least two possibilities. First, the bacteria might be able to compare the signal coming from receptors located at the opposite poles of the cell, and switch in such a way as to swim toward the end with the higher signal, that is, sensing the gradient in space. Alternatively, the bacteria might be able to compare the signal being received at a given moment in time with the strength of the signal it received in the recent past, that is, sensing the gradient in time. As we will discuss below, the bacteria appear to use the time-based mechanism. The reader will have a chance to explore and compare these two possible schemes in the problems at the end of the chapter.

The cellular decision-making that attends chemotaxis is mediated by a signal transduction network that has been extremely well characterized. Our comments will center on the particular features of the *E. coli* chemotaxis network, which is an example of the two-component signaling systems introduced in Section 7.2.3 (p. 292). The key elements

in this system are (i) membrane-spanning receptors that interact with the molecules in the environment (sugars, amino acids, etc.); (ii) CheW and CheA, proteins that bind to the intracellular domain of the receptor and change their activity depending on whether or not the receptor has a ligand bound (CheA is a protein kinase that can catalyze the attachment of phosphate groups to other target proteins, and CheW modulates CheA activity); (iii) a messenger molecule known as CheY that, when phosphorylated by CheA, can interact with the flagellar rotary motor to induce it to switch to clockwise (tumbling) rotation; (iv) CheZ, a phosphatase that can remove the phosphate from CheY; and (v) a pair of enzymes known as CheR and CheB that can respectively methylate and demethylate the receptors themselves, effectively tuning their affinity for their binding partners.

Experiments Behind the Facts: Measuring the Process of Chemotaxis Quantitative measurement of the behavior of bacteria engaged in chemotaxis has been performed in many elegant ways. Here we highlight a few of the key experiments that form the backdrop for our discussion. Video tracking microscopy was introduced to make it possible to perform single-cell analyses of bacteria engaged in their chemotactic response as shown in Figures 19.52(A) and (B). The idea of such experiments is easily stated, but the easy words mask what was an experimental *tour de force* when first introduced. Stated simply, the microscope stage is shifted constantly so that the cell of interest always stays in focus in the center of the field of view. A more recent version of the same experiment elects to hold the bacterium in an optical trap, with the runs and tumbles characterized by changes in the way the trapped bacterium jiggles about as shown in Figures 19.52(D) and (E).

Using tracking microscopy, it was possible to ask precise questions such as how fast are cells moving, how often do they tumble and what is their angular reorientation after a tumbling event? Figure 19.52 shows the outcome of such experiments. The advent of fluorescent proteins made it possible to observe cells engaged in these kinds of behaviors while simultaneously measuring the quantities (and even dynamics) of the molecules such as CheY-P that mediate the behavior. For example, in the experiment shown in Figure 19.53, the amount of CheY-P was monitored using fluorescence correlation spectroscopy (FCS) as introduced in Section 13.1.2 (p. 511). At the same time, as shown in the figure, by monitoring a fluorescent bead attached to one of the flagella on the immobilized bacterium, the direction of rotation could be observed, resulting in the ability to characterize the fraction of time (the so-called motor bias) the motor spends rotating in the opposite direction.

Recent FRET measurements have provided the kind of systematic, quantitative dissection of the chemotactic response that can really drive theoretical understanding forward. In the experiments shown in Figure 19.54(A), CheY-P and CheZ were each labeled with fluorescent molecules that serve as a FRET pair. As a result, the level of FRET serves as a direct readout of the amount of CheY-P as a function of the chemoattractant concentration because when there is lots of phosphorylated CheY, the interaction between CheY and CheZ is increased. As shown in Figure 19.54(B), the time history after stimulation

EXPERIMENTS

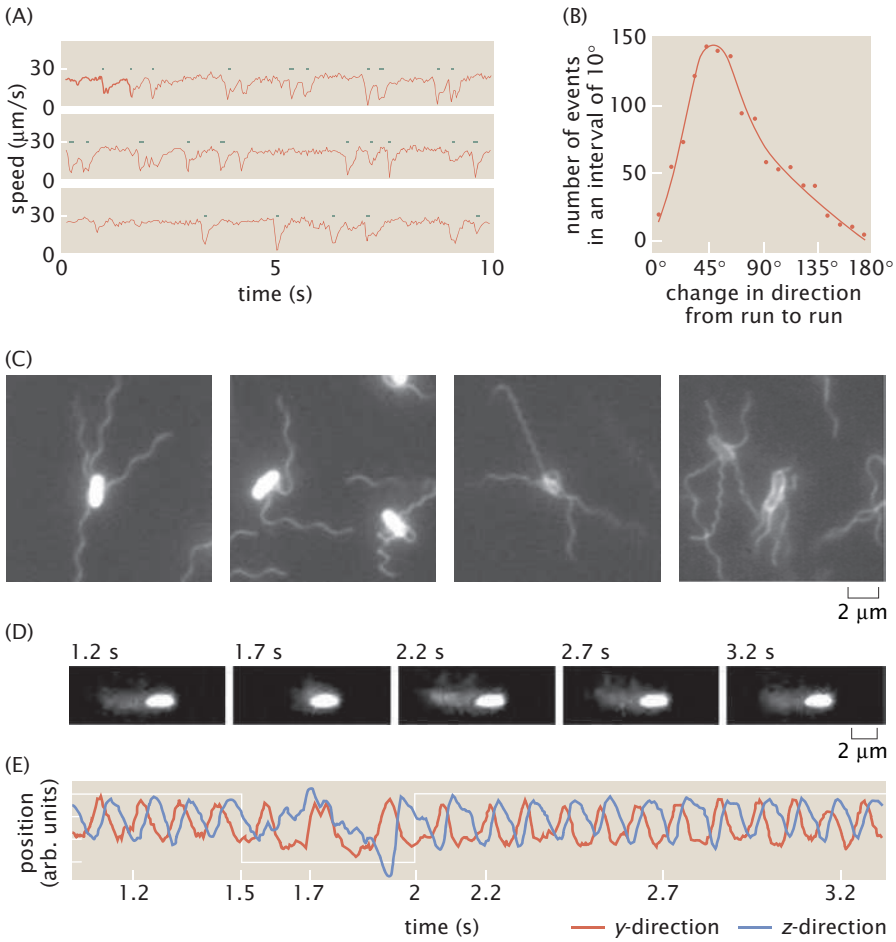


Figure 19.52: Chemotactic dynamics as observed using tracking microscopy and optical trapping. (A) Measurement of the speed of a bacterium as a function of time. The individual tumble events are shown by horizontal bars and are reflected by a marked reduction in the speed for a short interval. (B) Angular distribution of tumbles. (C) Images of tumbling bacteria illustrating the spreading apart of the flagella during the tumbling process. (D) Images of a bacterium held in an optical trap at various observation times. The fluorescently labeled flagella look different during the run and tumble events. (E) *x*- and *y*-positions of the bacterium as observed in the optical trap as a function of time. (A, B, adapted from H. C. Berg and D. A. Brown, *Nature* 239:500, 1972; C, adapted from L. Turner, W. S. Ryu, and H. C. Berg, *J. Bacteriol.* 182:2793, 2000; D, E, adapted from T. L. Min et al., *Nat. Methods*, 6:831, 2009.)

with a pulse of chemoattractant can be monitored directly with these experiments. Figure 19.54(C) shows how the activity of the chemoreceptor depends upon the concentration of chemoattractant for a number of different mutants that have their ability to adapt altered.

Examination of the tumbling frequency after exposure to a shift in concentration makes it possible to explore the question of adaptation. In particular, the time scale and precision of adaptation can be measured by watching cells after such a concentration jump and keeping track of their tumbling frequency. The results of such experiments are shown in Figure 19.55, where it is seen that the idea of "precise adaptation" is not a misnomer. It is also interesting to see how the adaptation time depends upon chemical details such as the concentration of CheR, while the precision itself does not.

Even for the relatively simple network that governs bacterial chemotaxis, it is hard to avoid getting lost in the alphabet soup of names, so we try to examine how the network works conceptually without focusing on the names of the molecules. In addition, we will take a hierarchical view, first explaining the overall functioning of the network and then taking up the fancy bells and whistles that make it work over such a wide range of concentrations, in the phenomenon

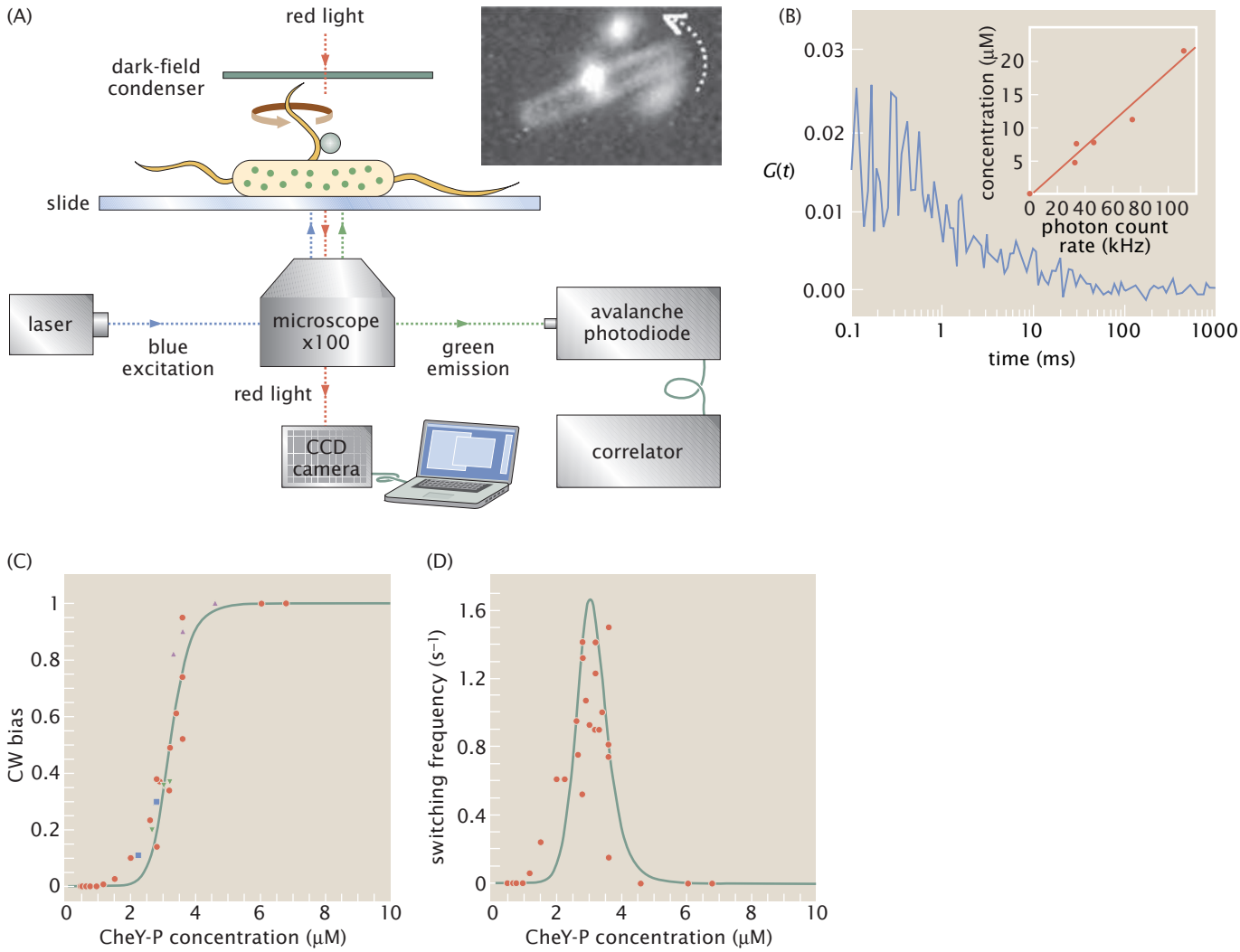


Figure 19.53: Tumbling frequency and CheY-P concentration. (A) Schematic of the experimental setup used to simultaneously quantify the amount of protein and the flagellar dynamics. (B) Measured correlation function as a function of time. This is related in turn to the concentration of protein (CheY-P-GFP). (C) Motor bias as a function of concentration of CheY-P. (D) Switching frequency and concentration of CheY-P. (Adapted from P. Cluzel, M. Surette, and S. Leibler, *Science*, 287:1652, 2000.)

known as adaptation. In simplest terms, the question of whether or not the cell will tumble (and hence change direction) comes down to the state of phosphorylation of the messenger molecule CheY. In order to be responsive to changes in the environment, the phosphorylation of CheY must be sensitive to whether or not there is a ligand bound to the receptor. In the presence of desirable attractant molecules, such as glucose or aspartate, the cell should repress tumbling, so we expect that the ligand-bound receptor will tend to be in the "off" form, where CheY is not phosphorylated, and the unbound receptor will tend to be in the "on" form, where CheY is phosphorylated. (Although *E. coli* is actually able to use the same chemotactic network to swim away from noxious chemicals, here we will only consider the happier problem of swimming toward delicious ones.) An idealization of these elements is shown in Figure 19.56(A), where we have combined the transmembrane receptor, CheW, and CheA into a single unit, and for the moment are ignoring the other components of the pathway.

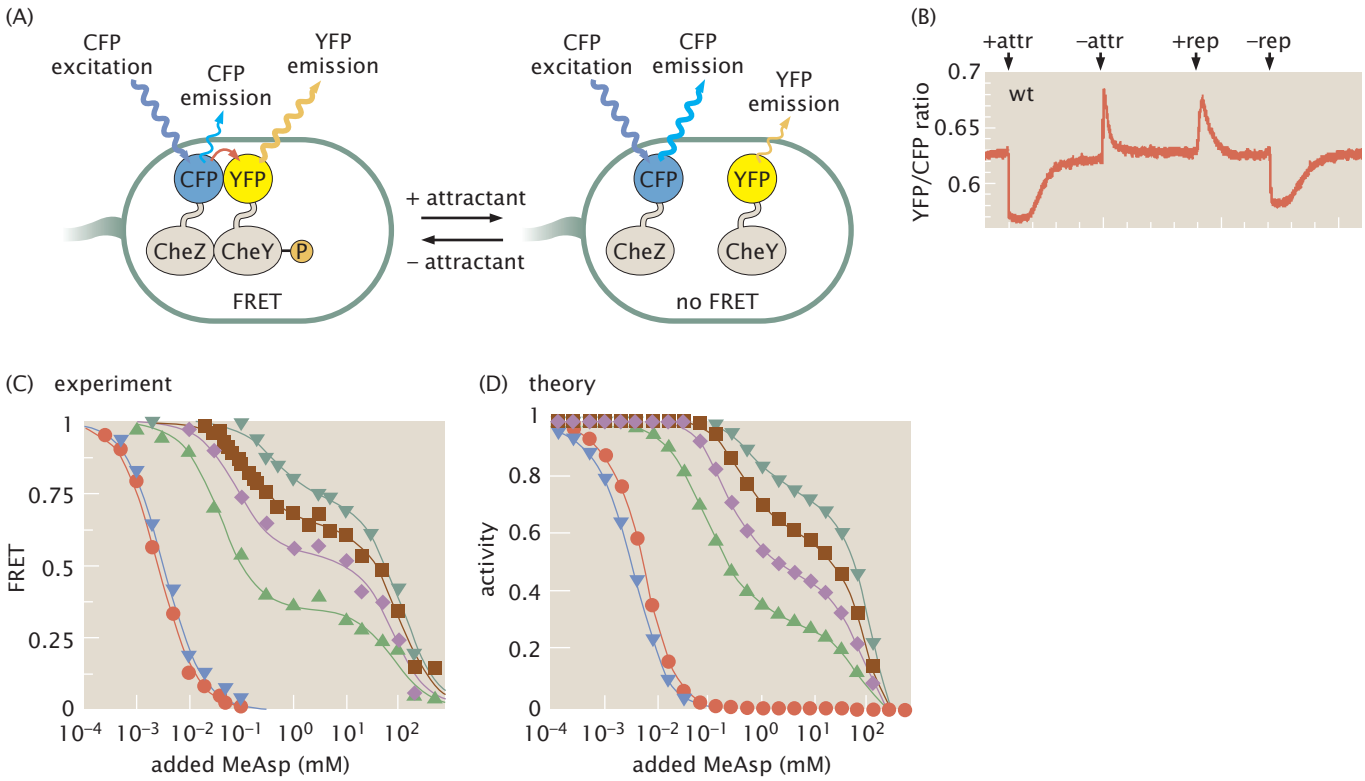


Figure 19.54: FRET measurements of chemotactic response. (A) Schematic of how the FRET measurements report on the chemotactic response of the system in the presence of chemoattractant. (B) FRET signal as a function of time. Addition of chemoattractant results in a decrease in the FRET signal corresponding to a reduction in the frequency with which the direction of rotation of the motor changes. After a certain adaptation time, the FRET signal (and the motor bias) go back to their unperturbed value. (C) Graph of concentration dependence of the "on" probability based on *in vivo* fluorescence resonance energy transfer (FRET) measurements. The different curves correspond to different bacterial strains. The wild-type response is shown as orange circles. The other symbols are for mutants that correspond to different states of receptor methylation, increasing from left to right. (D) The results of a calculation of the probability of the receptor being active as a function of the concentration of chemoattractant. The model reproduces many aspects of the living cell responses, including the complex behaviors of the methylation mutants. (A, B, Adapted from V. Sourjik and H. C. Berg, *Proc. Natl Acad. Sci. USA* 99:123, 2002; C, D Adapted from J. E. Keymer et al., *Proc. Natl Acad. Sci. USA* 103:1786, 2006.)

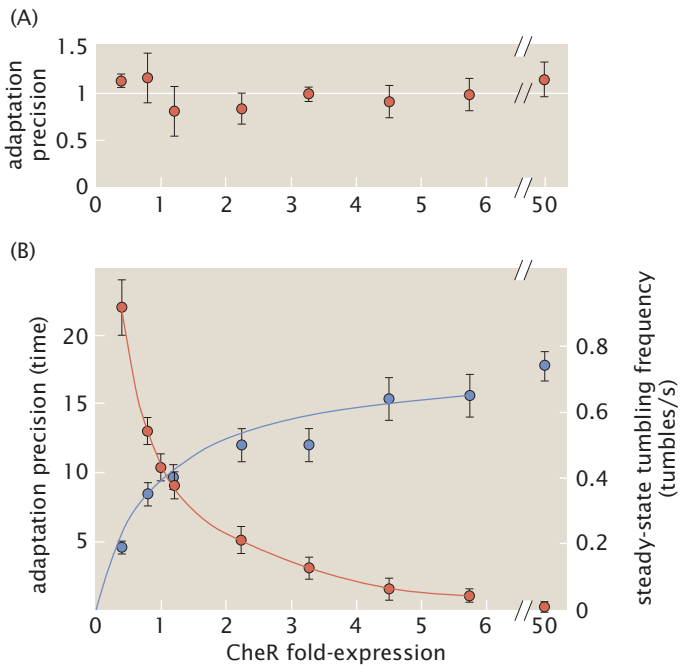


Figure 19.55: Tumbling frequency and adaptation. (A) Precision of the adaptation as measured by how precisely the rotational frequency returns to its original value. Rather than tuning the chemoattractant concentration in these experiments, the level of expression of CheR is controlled. (B) Tumbling frequency of the cells before stimulation is shown with blue data points and refers to the vertical axis on the right. Average adaptation time is shown by the red circles with reference to the vertical scale on the left. Both quantities are plotted as a function of CheR fold-expression. (Adapted from U. Alon, M. G. Surette, N. Barkai, and S. Leibler, *Nature*, 397:168, 1999.)

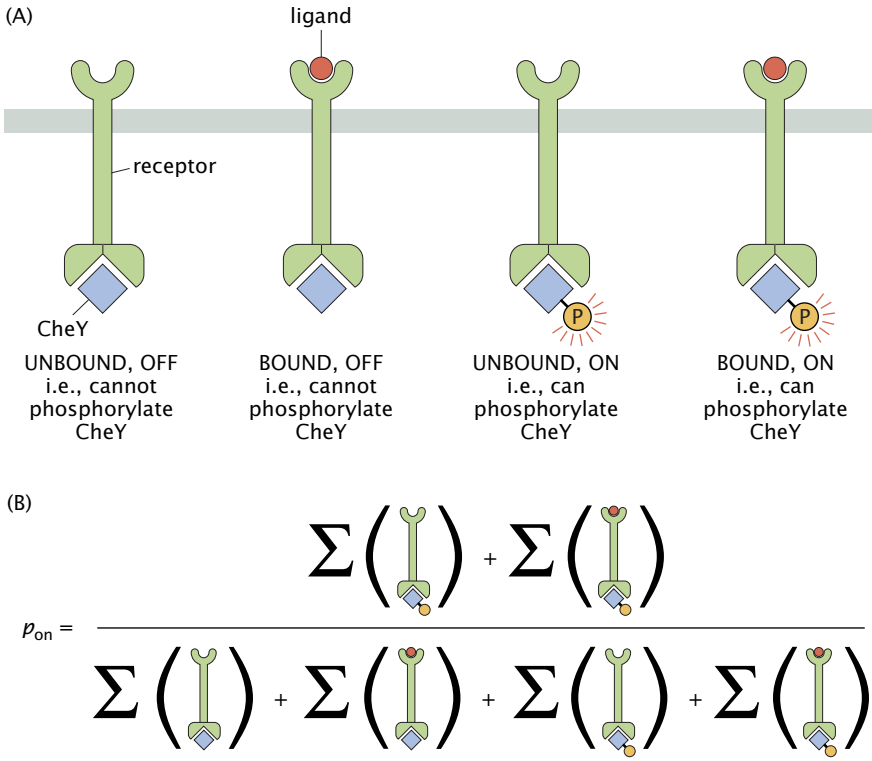


Figure 19.56: Probability that a receptor will be "on." (A) The receptor and its states of occupancy and activity. The receptor can either have a bound ligand or not. Similarly, the receptor can either be "on" or "off," where this state of activity determines whether or not it is able to phosphorylate the messenger CheY. (B) The probability that the receptor will be "on" is constructed as a ratio of the "on" states, appropriately weighted by their Boltzmann factors to the sum over the statistical weights of all states.

The MWC Model Can Be Used to Describe Bacterial Chemotaxis

We can treat this complex process approximately by appealing to our usual statistical mechanical formulation in which we imagine a rapid preequilibrium of the state of activity of the receptor. In particular, the quantity p_{on} measures the ability of the receptor to produce phosphorylated CheY, resulting in a change in the motor's direction of rotation. As we have done throughout the book, the statistical mechanics of this system can be examined by appealing to a states-and-weights diagram like that shown in Figure 19.57. The probability that the receptor will be active is obtained by constructing the ratio

$$p_{on} = \left[\frac{\Omega^L}{L!} e^{-\beta L \varepsilon_{sol}} e^{-\beta \varepsilon_{on}} + \frac{\Omega^{L-1}}{(L-1)!} e^{-\beta (L-1)\varepsilon_{sol}} e^{-\beta \varepsilon_{on}} e^{-\beta \varepsilon_b^{on}} \right] \Bigg/ \left[\frac{\Omega^L}{L!} e^{-\beta L \varepsilon_{sol}} \left(e^{-\beta \varepsilon_{off}} + e^{-\beta \varepsilon_{on}} \right) + \frac{\Omega^{L-1}}{(L-1)!} e^{-\beta (L-1)\varepsilon_{sol}} \left(e^{-\beta \varepsilon_{off}} e^{-\beta \varepsilon_b^{off}} + e^{-\beta \varepsilon_{on}} e^{-\beta \varepsilon_b^{on}} \right) \right]. \quad (19.155)$$

This result can be simplified by multiplying through the top and bottom of the equation by $L!/\Omega^L$, resulting in

$$p_{on} = \frac{e^{-\beta \varepsilon_{on}} [1 + (L/\Omega) e^{-\beta \Delta \varepsilon_{on}}]}{e^{-\beta \varepsilon_{on}} [1 + (L/\Omega) e^{-\beta \Delta \varepsilon_{on}}] + e^{-\beta \varepsilon_{off}} [1 + (L/\Omega) e^{-\beta \Delta \varepsilon_{off}}]}. \quad (19.156)$$

Here we have defined $\Delta \varepsilon_{on}$ as the difference in energy between a single ligand bound to the "on" state of the receptor and the same ligand in solution, and $\Delta \varepsilon_{off}$ equivalently for ligand binding to the receptor in the "off" state. Throughout the book, we have repeatedly translated back and forth between the statistical mechanical language

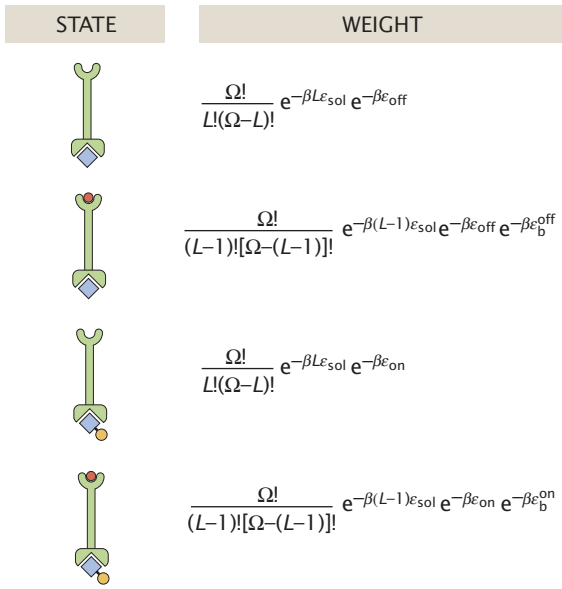


Figure 19.57: States and weights for a simple model of bacterial chemotaxis. The lower two states correspond to the case when the receptor is "on." The prefactors in front of the exponential terms correspond to the number of ways of rearranging the ligands in the lattice model of the solution.

used above and the thermodynamical language using equilibrium constants. By exploiting the relationship between energy differences and biochemical dissociation constants derived in Section 6.4.1 (p. 270), our expression for the probability that the receptor will be "on" can be rewritten using the dissociation constants as

$$p_{on} = \frac{1}{1 + e^{-\beta(\varepsilon_{off}-\varepsilon_{on})}\dfrac{1+[L]/K_d^{off}}{1+[L]/K_d^{on}}}. \tag{19.157}$$

This formula suggests that the probability of the "on" state depends on a few biologically important variables: the energy difference between the "on" and "off" states of the receptor in the absence of ligand, the affinities of the ligand for the "on" state and the "off" state of the receptor, and the amount of ligand itself. For attractive substances, binding of the ligand will tend to favor the "off" state (where CheY is not phosphorylated), that is, $K_d^{off} < K_d^{on}$.

Let us consider the implications of this result. In the absence of ligand (if $[L] = 0$), the equation simplifies to the familiar result for a two-state system such as an ion channel with the active and inactive states controlled by the relative values of ε_{off} and ε_{on}. Since in the absence of ligand the receptor is active for phosphorylation, we know that ε_{off} is larger than ε_{on}, thus favoring the "on" state. On the other hand, we expect that with increasing ligand concentration, the inactive state will predominate. This means within this model that $K_d^{off} < K_d^{on}$.

In order to modulate its response over a wide range of ligand concentrations and conditions, *E. coli* is actually able to move around in the parameter space of ε_{on} and ε_{off} by performing regulated covalent modifications of the receptor protein itself. This is the job of the methylase CheR and the demethylase CheB, which add and remove methyl groups on a series of glutamate residues present in the intracellular domain of the membrane-spanning receptor protein. The more highly methylated the receptor protein, the more likely it is to be in the "on" state. These modifications permit two impressive consequences. First, as mentioned above, *E. coli* can detect gradients of chemoattractants by comparing the strength of the signal it currently

senses with the strength of the signal it detected in the recent past. Second, the bacterium is able to detect gradients in concentration over many orders of magnitude of absolute concentrations, a phenomenon known as adaptation. This corresponds to our own ability to whisper to someone else even in a crowded and noisy room, or our ability to see our surroundings either inside a darkened room or after stepping out into the bright sunshine. For the bacteria, both adaptation and time-sensing depend on the fact that the demethylase, CheB, is itself regulated by phosphorylation by CheA, and therefore depends on ligand binding to the receptor. If CheB is phosphorylated (that is, if the receptor is "on"), CheB will be more active as a demethylase, and will tend to convert the receptor into an "off" state, damping the response. Conversely, if CheB is dephosphorylated (that is, the receptor is "off"), more methyl groups will accumulate, tending to switch the receptor "on." This sequence of events takes some time, a few seconds. At the same time, ligand binding influences the activity state of the receptor. Therefore, receptor occupancy by ligand reflects current conditions, and the methylation state of the receptor reflects the conditions of a few seconds ago. The cell is able to swim up concentration gradients essentially by comparing these two signals.

Our calculations so far illustrate the key ideas, but they will not suffice to capture the full complexity of chemotactic behavior as revealed in Figure 19.54(C). In addition to the precise adaptation already discussed, the system exhibits a high degree of cooperativity. To account for cooperativity, our previous results can be amended to the form

$$p_{\text{on}} = \frac{1}{1 + \mathrm{e}^{-n\beta(\varepsilon_{\text{off}} - \varepsilon_{\text{on}})} \dfrac{(1 + [L]/K_{\text{d}}^{\text{off}})^n}{(1 + [L]/K_{\text{d}}^{\text{on}})^n}}. \tag{19.158}$$

To see how this result emerges, Figure 19.58 resorts to our usual states-and-weights procedure in which we imagine a cluster of N receptors. The fate of the one is the fate of the many. Either all receptors are inactive or all are active. As in the usual MWC mentality, the relative energies of the inactive and active states are different and the K_{d} for the binding of ligands depends upon which of the two states the receptors are in. To see how the statistical weight of the active state arises, note that the number of *bound* ligands can be anything between 0 and N. The generic weight for the active state when it has n ligands bound is of the form

$$w_n = \mathrm{e}^{-\beta\varepsilon_{\text{on}}} \frac{N!}{(N-n)!n!} \mathrm{e}^{-n\beta(\varepsilon_{\text{b}}^{\text{on}} - \varepsilon_{\text{sol}})}. \tag{19.159}$$

However, we note that this is just the nth term in a binomial of order N (except for the prefactor $\mathrm{e}^{-\beta\varepsilon_{\text{on}}}$), and hence, when we sum together all such terms, we find the overall statistical weight for the active state shown in the figure.

The inclusion of cooperativity sharpens the response of the system. Previously, we have considered cases of cooperativity such as oxygen binding to hemoglobin (Section 7.2.4, p. 298), where a single protein has multiple ligand-binding sites. In chemotaxis, the *E. coli* cell clusters essentially all its membrane-spanning receptors together in a single patch at one pole in a tight cluster as shown in Figure 13.23 (p. 537), such that binding of one ligand to one receptor can influence the conformational state of many other receptors, including distinct receptors that are able to detect different substances. A fully

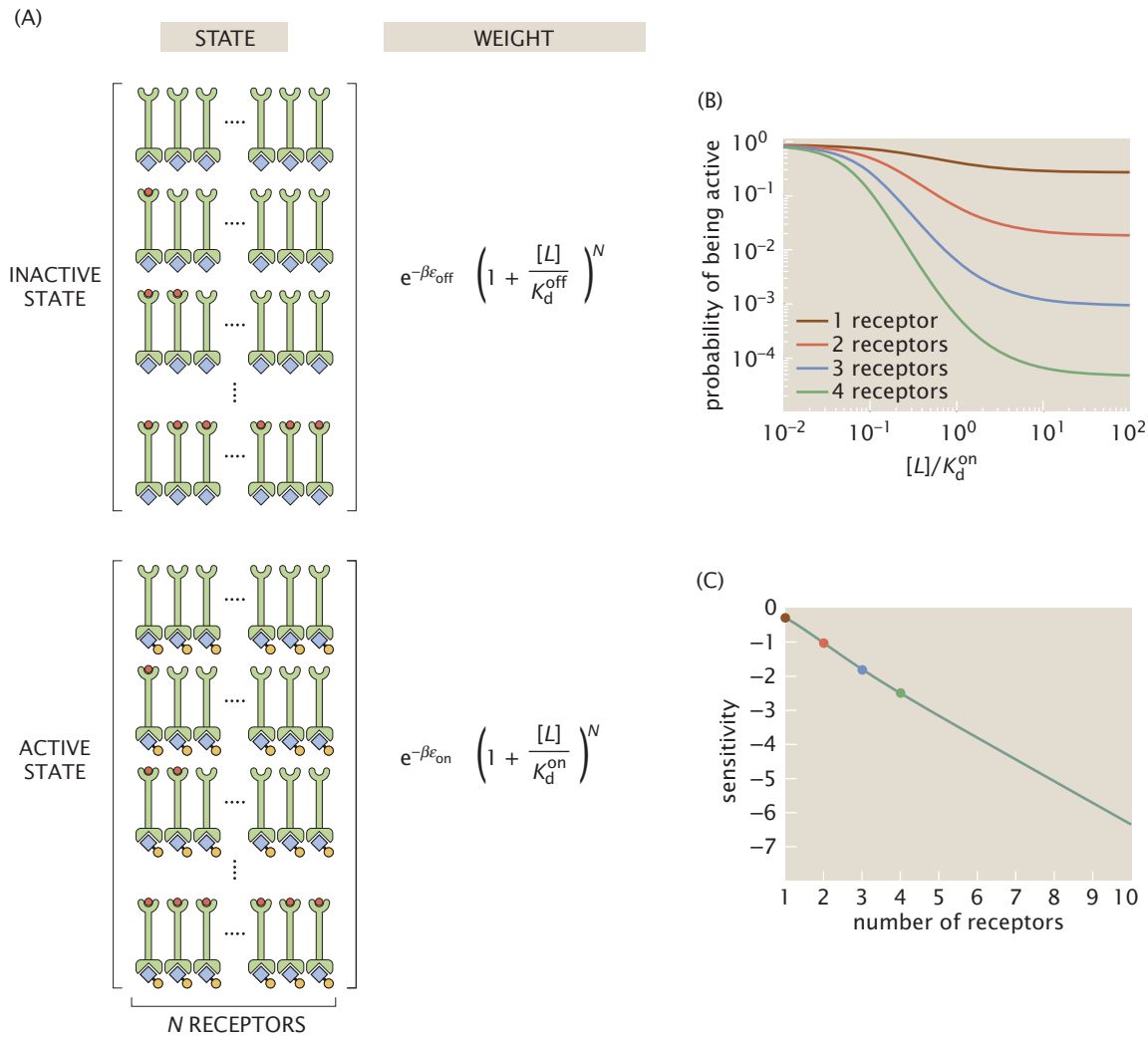


Figure 19.58: MWC model of bacterial chemotaxis. (A) States and weights for the MWC model in which there are N receptors in a cluster. (B) Probability that the receptors will be on as a function of the concentration of chemoattractant for different choices of the number of receptors in a cluster. (C) Sensitivity of the chemotactic response. For (B) and (C) a value of $K_d^{off}/K_d^{on} = 1/20$ was used.

detailed mathematical model that incorporates adaptation and cooperativity in mixed receptor clusters along with the basic two-state model derived above is able to reproduce many of the complex features of chemotactic receptor response, as illustrated in Figure 19.54(D).

Precise Adaptation Can Be Described by a Simple Balance Between Methylation and Demethylation

As already illustrated in Figure 19.55, bacteria that are exposed to a uniform change in concentration will temporarily respond as though they have been subjected to a gradient of chemoattractant. This response is characterized by a change in tumbling frequency. However, as seen in the figure, after some transient response time, they will faithfully return to their original tumbling frequency. An idea for how this takes place is illustrated schematically in Figure 19.59, where it is seen that the state of methylation of the receptor (here we examine a minimal model with only one site of methylation) is

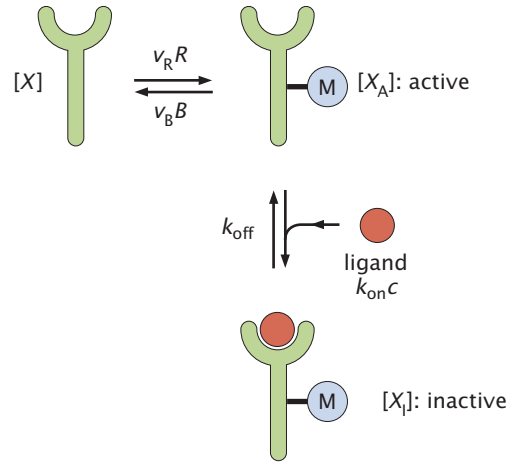


Figure 19.59: Kinetic scheme for a toy model of precise adaptation. The concentration of unmethylated receptors is given by X, the concentration of active methylated receptors by X_A and the concentration of inactive methylated receptors is given by X_I. R refers to the concentration of CheR and B to the concentration of CheB.

constantly tuned by the presence of CheR (methylation) and CheB (demethylation).

One of the key qualitative features of the model is captured by the topology of the various reactions and specifically by the fact that the demethylation can only take place from the active state. Hence, if the concentration of chemoattractant changes, it will temporarily change the number of active chemoreceptors, and this means that the balance between CheR and CheB will be perturbed, resulting in a net change in the number of methylated receptors depending upon whether the chemoattractant was decreased or increased. For example, if the amount of chemoattractant goes up, this will increase the number of inactive receptors. However, this will result in a concomitant increase in active receptors over time since CheR will win out over CheB in the coupled reactions they mediate.

The dynamics of this system of reactions is captured by three coupled dynamical equations. By inspection of Figure 19.59 we can read off these equations as follows. First, for the X concentration, we have

$$\frac{dX}{dt} = \nu_B B \frac{X_A}{K_A + X_A} - \nu_R R, \tag{19.160}$$

where we assume that CheR is working at saturation (that is, there is an excess of X such that all of the CheR molecules are engaged in methlyation) and we have adopted the Michaelis–Menten form (see Section 15.2.7, p. 596) for the reaction of CheB. For the active state, we have

$$\frac{dX_A}{dt} = \nu_R R - \nu_B B \frac{X_A}{K_A + X_A} - k_{on} c X_A + k_{off} X_I, \tag{19.161}$$

where we have included ligand binding and unbinding. Finally, for the inactive state, in this model, the only way to enter or exit the state is through ligand binding and unbinding, and this is captured through the dynamical equation

$$\frac{dX_I}{dt} = k_{on} c X_A - k_{off} X_I. \tag{19.162}$$

As we have already seen, one of the most useful ways to examine the dynamics of low-dimensional dynamical systems like this is by appealing to the phase portrait. In this case, note that the equations for X_A and X_I make no reference to the concentration X. As a result, we can consider the dynamics of X_A and X_I independently of the dynamics

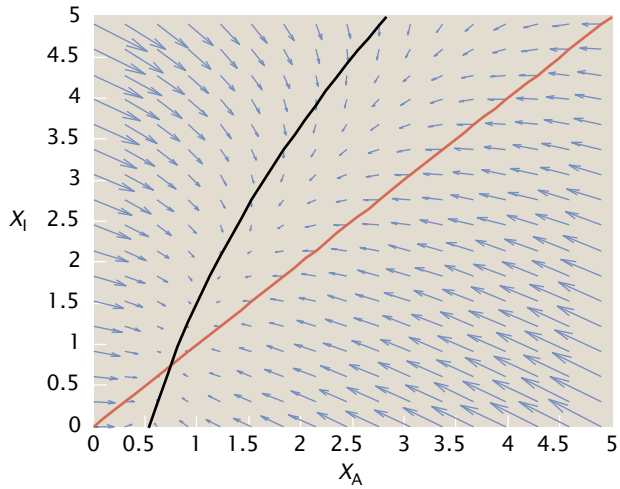


Figure 19.60: Phase portrait for simple model of precise adaptation. The straight line is the nullcline on which $dX_I/dt = 0$ and the other curve is the nullcline that shows the locus of points in the $X_A - X_I$ plane where $dX_A/dt = 0$.

of X, which means we can resort to a two-dimensional phase portrait of the vector field $(dX_A/dt, dX_I/dt)$ as shown in Figure 19.60. The two nullclines, corresponding to $dX_I/dt = 0$ and $dX_A/dt = 0$, are shown on the phase portrait and their point of intersection is the fixed point.

The position of this fixed point can be solved for explicitly, resulting in

$$X_A^* = \frac{K_A \nu_R R}{\nu_B B - \nu_R R} \tag{19.163}$$

and

$$X_I^* = \frac{k_{on} c}{k_{off}} X_A^* = \frac{k_{on} c}{k_{off}} \frac{K_A \nu_R R}{\nu_B B - \nu_R R}. \tag{19.164}$$

In particular, by inspecting the functional form of X_A^*, we see that the concentration of the active form of the receptor does not depend upon the overall concentration c. As a result, when the concentration suffers an overall shift, the fixed point will shift up and down, but not to the left or right, illustrating that the fixed-point concentration of X_A^* is invariant, corresponding to the precise adaptation seen experimentally.

19.4.2 Biochemistry on a Leash

One of the most fundamental features of living organisms is movement. As noted in our discussion of chemotaxis, cells make "decisions" about where to go and these decisions in eukaryotes are implemented in the form of polymerization of actin filaments. Examples of actin polymerization organized in both space and time were shown in Figures 15.2 (p. 576) and 15.3 (p. 577). What chains of events link the detection of some external cue and the formation of new actin filaments in a motile cell? The advent of video microscopy in conjunction with a host of different classes of fluorescent markers has made the study of cell motility one of the most exciting areas of current research. As a particular case study that will allow us to flex several sets of muscles that we have developed throughout the book, we consider molecules that have the interesting feature that they include a tethered ligand and receptor pair that compete with free ligands. These tethering motifs are a common feature of signaling molecules.

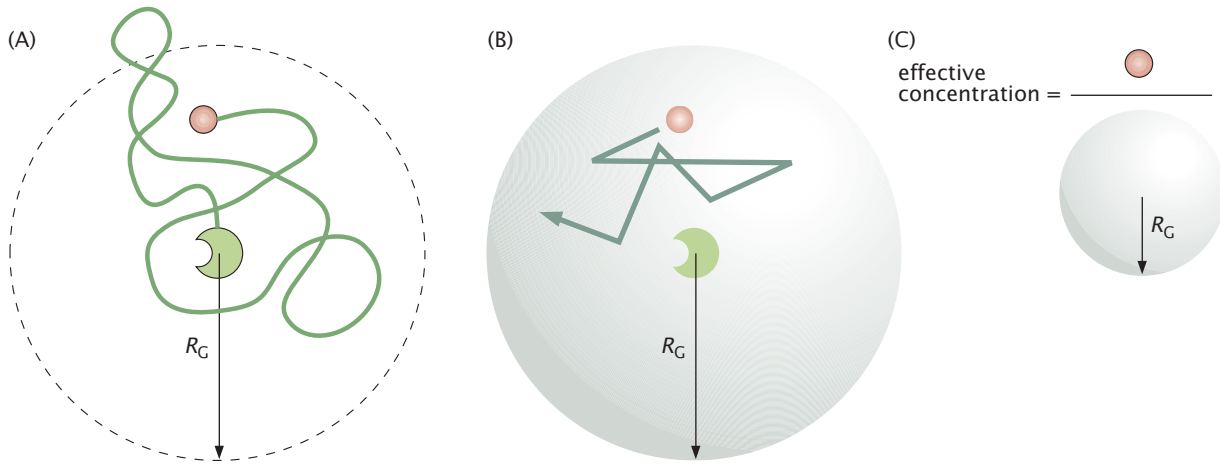


Figure 19.61: Tethering and effective concentration. (A) As a result of tethering, the ligand can only explore a limited region of space. (B) The concentration of the tethered ligand can be estimated by considering a sphere with a radius given by the radius of gyration of the tether. (C) To compute the effective concentration due to tethering, consider one ligand per volume given by a sphere with a radius equal to that of the radius of gyration.

Tethering Increases the Local Concentration of a Ligand

One simple way to see the significance of tethering is illustrated in Figure 19.61. The idea is that the tethered ligand is confined to a volume dictated by the length of the tether. In particular, if the tether has a length L resulting in a radius of gyration R_G, then the effective concentration of the tethered ligand can be estimated as

$$\text{effective concentration} = \frac{1}{\frac{4}{3}\pi R_G^3}. \tag{19.165}$$

To develop an intuitive sense of the significance of this tethering, this estimate can be used to roughly determine the concentration at which the free ligands compete with the tethered ligand. In particular, for the case in which a tethered ligand competes with free ligands for the attention of a tethered receptor, clearly at high enough concentrations, the free ligands will dominate the binding.

Signaling Networks Help Cells Decide When and Where to Grow Their Actin Filaments for Motility

The case of bacterial chemotaxis described above is but one of many examples where the motility of cells is dictated by the presence of environmental cues. In many cases, these environmental cues have the effect of inducing actin polymerization, which leads to changes in cell shape that are then coupled to motility. From the standpoint of cell signaling, a small signaling molecule can relay information to N-WASP, a protein that can interface with a complex of proteins called the Arp2/3 complex to create new actin filaments. The way in which this works is shown in Figure 19.62(A). In particular, the presence of two ligands, Cdc42 and PIP_2, activates N-WASP by binding to this protein in a way that then permits it to activate Arp2/3. The presence of Cdc42 and PIP_2 leads to the unbinding of GDB and B domains from the C domain and Arp2/3, and N-WASP begins to stimulate actin polymerization by recruiting (and perhaps appropriately orienting) actin monomers to the proximity of the Arp2/3. With the help of activated N-WASP, Arp2/3 promotes actin polymerization by

providing heterogeneous nucleation sites. Here, our aim is to study this process quantitatively.

Synthetic Signaling Networks Permit a Dissection of Signaling Pathways

As with the analysis of genetic networks, one exciting way in which signaling pathways have been dissected is by rewiring such pathways to form various synthetic signaling networks. Figure 19.62(B) shows a synthetic activator of Arp2/3 in which a domain known as a PDZ domain is attached to the output domain that activates Arp2/3. On the other end of the construct is a peptide sequence that binds to PDZ. This synthetic protein mimics N-WASP and can be activated by soluble ligands that bind to the PDZ domain.

To analyze the function of this signaling process, we invoke statistical mechanics in the same spirit as we have earlier for considering gene regulation. The goal of our statistical mechanical model of the synthetic switch is to work out the probability that the molecule is in the active state. In particular, the active state corresponds to the case in which the tethered receptor is not bound to the tethered ligand. That is, the tethered ligand and receptor are separately flopping around freely. As usual, we resort to a states-and-weights diagram to work out the probability of the active state. As shown in Figure 19.63, there are three classes of states, each with their own corresponding statistical weights: (i) the switch is in the autoinhibitory state and the tethered ligand and receptor are bound to each other; (ii) the tethered ligand and receptor are both flopping around freely and the receptor has no bound free ligands; (iii) the tethered ligand and receptor are both flopping around freely, and the receptor has bound one of the free ligands. Our aim is to make falsifiable predictions for the signal dependence on, for example, the linker length and ligand concentration.

To develop an intuitive sense of how this situation plays out, the probability of finding the switch in the active state is represented

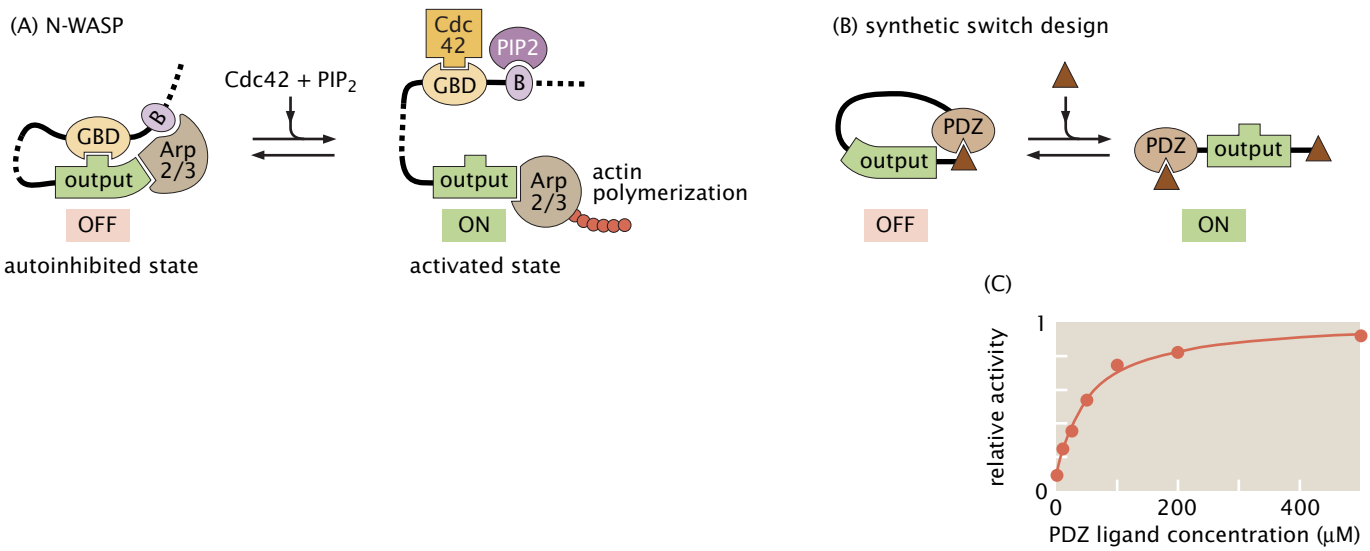


Figure 19.62: Schematic of the signaling process leading to actin polymerization. (A) Activation of Arp2/3 by ligands Cdc42 and PIP_2. (B) Synthetic switch constructed to activate Arp2/3 as a result of the presence of an alternative ligand. (C) Activity of the synthetic switch as a function of the signaling ligand. (Adapted from J. E. Dueber et al., *Science* 301:1904, 2003.)

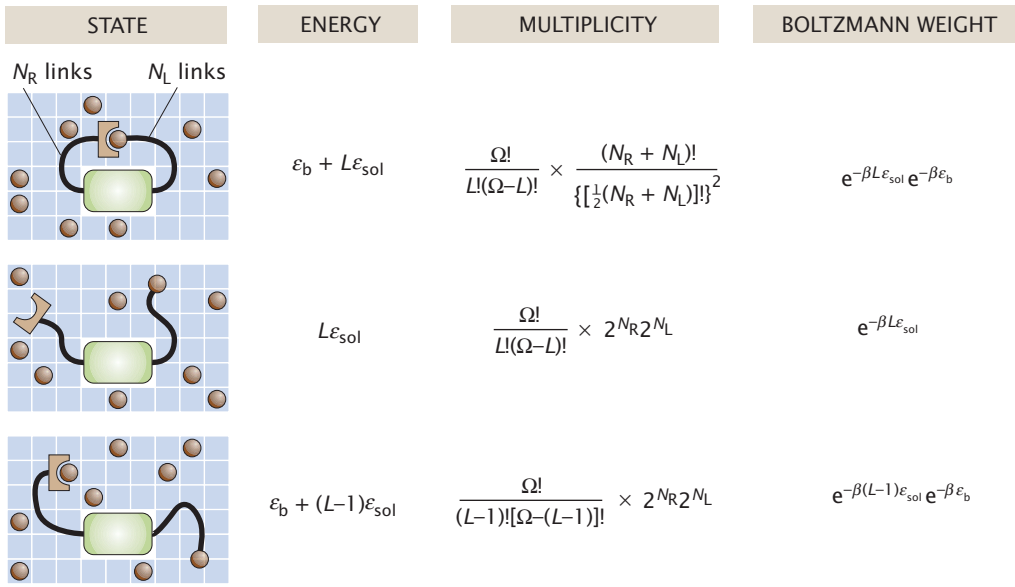


Figure 19.63: States and weights for the synthetic signaling problem.

schematically in Figure 19.64. The essence of the situation is that as the concentration of free ligand is increased, the probability that the receptor will be bound by one of the free ligands will increase until this outcome dominates the probability. From the standpoint of testing our understanding of such systems, one of the other design parameters that can be varied is the length of the flexible tethers. As will be shown explicitly when we demonstrate the contributions of the autoinhibitory state to the overall partition function, the length of the tether is a significant part of the overall free energy budget.

To make this calculation concrete, we resort here to simple one-dimensional ideas on the random walk introduced in Chapter 8 and show how the calculation generalizes to three dimensions, but leave the details for the reader as a problem at the end of the chapter. Our strategy will be to break the *total* partition function for this system down into three parts as reflected in Figure 19.63, where the sum can be written as

$$Z_{tot}(L, N_R, N_L) = \underbrace{Z_1(L, N_R, N_L)}_{\text{autoinhibitory state}} + \underbrace{Z_2(L, N_R, N_L)}_{\text{free tethers}} + \underbrace{Z_3(L, N_R, N_L)}_{\text{tether with ligand}}. \tag{19.166}$$

The parameter L is the number of ligands in the system, N_R is the number of Kuhn segments in the polymer tether that has the tethered receptor, and N_L is the number of Kuhn segments in the polymer tether that has the tethered ligand. Given these decompositions, we can then write the probability that the switch will be in the active state as

$$p_{active} = \frac{Z_2 + Z_3}{Z_1 + Z_2 + Z_3}. \tag{19.167}$$

The separate contributions to the total partition function can be worked out in much the way we have done in similar problems

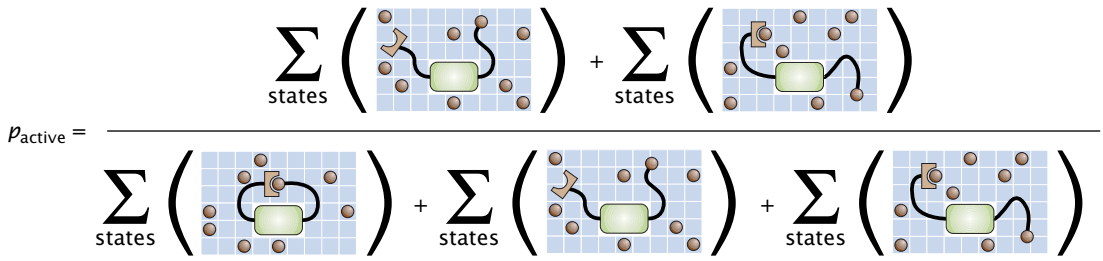


Figure 19.64: Probability of activation of Arp2/3. The numerator is the sum of the statistical weights of the active states.

throughout the book. The key point is that each class of state has a number of microscopically equivalent configurations and to find their contribution to the overall partition function, we need to multiply the Boltzmann weight for each class of state by its corresponding microscopic degeneracy (obtained by adding up all of the different ways of arranging the system). For example, the contribution from the states in which the tethers are flopping around freely and there is no free ligand bound is given by

$$Z_2 = \underbrace{\frac{N!}{L!(N-L)!}}_{\text{solution ligands}} \times \underbrace{2^{N_R} 2^{N_L}}_{\text{tether configs.}} \times \underbrace{e^{-\beta L \varepsilon_{sol}} e^{-\beta \varepsilon_{sol}}}_{\text{Boltzmann weight}}. \tag{19.168}$$

The treatment of the tether degrees of freedom is based on the simplest one-dimensional random walk in which we imagine that every segment in the tether can point either to the left or right and we do not worry about self-avoidance. It is straightforward to use a more robust model of the tethers, but we use this one for simplicity. What this means precisely is that each tether can be in one of 2^N different configurations, where N is the number of Kuhn segments in the tether of interest. We have also introduced the energy ε_{sol} for the energy of the ligands when they are free in solution and the parameter ε_{sol}^{lig} for the energy of the tethered ligand when it is in solution. The most interesting class of states are those that are associated with the autoinhibition of the switch and that involve the tethering ligand and receptor being linked. In this case, the contribution to the partition function is

$$Z_1 = \underbrace{\frac{N!}{L!(N-L)!}}_{\text{solution ligands}} \times \underbrace{\frac{(N_R + N_L)!}{\left\{\left[\frac{1}{2}(N_R + N_L)\right]!\right\}^2}}_{\text{tether closure}} \times \underbrace{e^{-\beta L \varepsilon_{sol}} e^{-\beta \varepsilon_b}}_{\text{Boltzmann weight}}, \tag{19.169}$$

where we have used the result from Section 8.2.4 (p. 333). The contribution from tether closure is the number of ways of making a closed loop out of a polymer of length $N_R + N_L$ Kuhn segments. The last contribution to the total partition function arises from those microstates in which one of the free ligands attaches to the tethered receptor. This means that the solution contribution to the partition function will only

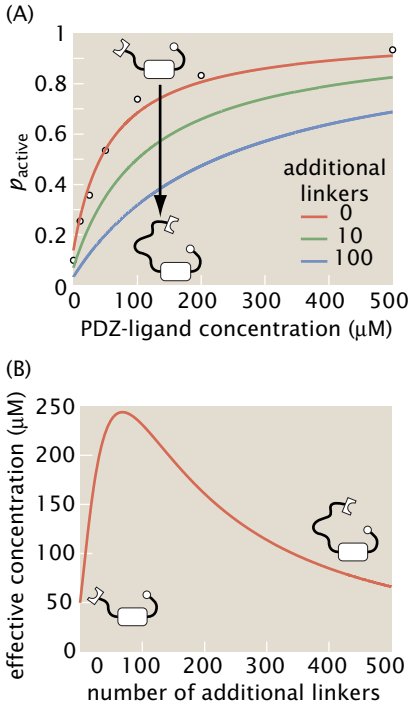


Figure 19.65: Prediction of dependence of activation on effective tail length. (A) p_{active} as a function of ligand concentration for different tether lengths. Experimental data are shown as small circles. (B) The effective concentration of tethered ligand as seen by the tethered PDZ domain as a function of tether length. (Data from J. E. Dueber et al., *Science* 301:1904, 2003.)

involve $L-1$ ligands. This term can be written as

$$Z_3 = \underbrace{\frac{N!}{(L-1)![N-(L-1)]!}}_{\text{solution ligands}} \times \underbrace{2^{N_R} 2^{N_L}}_{\text{tether configs.}} \times \underbrace{e^{-\beta(L-1)\varepsilon_{sol}} e^{-\beta\varepsilon_b}}_{\text{Boltzmann weight}}. \quad (19.170)$$

The actual formula for p_{active} can now be obtained by substituting the values for Z_1, Z_2, and Z_3 obtained above into Equation 19.167. The resulting expression is considerably simpler if we use an alternative form of this equation, namely,

$$p_{active} = \frac{1 + (Z_3/Z_2)}{1 + (Z_1/Z_2) + (Z_3/Z_2)}. \quad (19.171)$$

This leads to an expression for p_{active} of the form

$$p_{active} = \frac{1 + (c/c_0)e^{-\beta\Delta\varepsilon_1}}{1 + p_{loop}e^{-\beta\Delta\varepsilon_2} + (c/c_0)e^{-\beta\Delta\varepsilon_1}}, \quad (19.172)$$

where we have introduced $c = L/Nv$, $c_0 = 1/v$, and p_{loop}, which is the probability of forming a loop, and where $\Delta\varepsilon_1$ is the binding energy for a free ligand and $\Delta\varepsilon_2$ is the binding energy for the tethered ligand–receptor pair. For the one-dimensional model considered above, we have

$$p_{loop} = \frac{(N_R + N_L)!/\left\{\left[\frac{1}{2}(N_R + N_L)\right]!\right\}^2}{2^{N_R+N_L}}, \quad (19.173)$$

which amounts to the ratio of the number of closed configurations for the polymer of length $N_R + N_L$ to the *total* number of configurations. However, the one-dimensional model has outlived its usefulness and we can just as well use the result of a full three-dimensional analysis of p_{loop} using the Gaussian model of a polymer, for example. This calculation is left as an exercise for the reader.

The outcome of this kind of analysis is shown in Figure 19.65. There are several subtleties that were not accounted for in the calculation described above. First, as shown in the figure, the tethers do not emanate from the same point. This results in a fundamental difference in the behavior of p_{loop} as a function of tether length as shown in Figure 19.65(B). Second, in the figure, we used a three-dimensional Gaussian model for the tethers rather than the one-dimensional example worked out above.

19.5 Summary and Conclusions

Regulation and signaling are two of the most important ways in which cells orchestrate their behavior in space and time. The goal of this chapter has been to take stock of some of the key architectures of regulatory and signaling networks and to show how simple models using statistical mechanics and rate equations can be put forth to develop intuition and to make predictions about how these networks work. The so-called "thermodynamic models" of gene expression are predicated on the idea of using equilibrium statistical mechanics to examine the probability of promoter occupancy. A dynamical interpretation of these same questions uses rate equations to compute the concentration of both mRNA and their associated proteins.

19.6 Problems

Key to the problem categories: • Model refinements and derivations, • Estimates, • Data interpretation, • Computational simulations, • Model construction

• 19.1 Strong and weak promoters

In the chapter, we introduced repression as a quantitative measure of the reduction in the level of gene expression due to the action of a repressor molecule. For the simple model of repression introduced on p. 814, make a plot comparing repression in the case of a weak and a strong promoter. Show that, unlike the weak promoter case, in the case of the strong promoter, the repression depends upon the number of polymerase molecules in the cell.

• 19.2 Lac Repressor and the *lac* operon

A beautiful set of quantitative experiments on the *lac* operon were done by the Müller-Hill group in the 1990s, where repression of expression of the *lacZ* gene was measured in a population of different mutant *E. coli* cells. The mutant cells differed in the number, sequence, and position of the operator sites that bound the Lac repressor. In this problem, we explore how, using thermodynamic models of gene expression, these data can be used to obtain a number of quantities characterizing the Lac repressor–DNA interaction as well as DNA looping.

(a) Using the data from Oehler et al. (1994) shown in Figure 19.22 determine the *in vivo* binding energy of Lac repressor to each one of its operators and reproduce Figure 19.23.

(b) Use your results from (a), and the repression measured by Oehler et al. (1994) in cells with two operators present, which leads to DNA looping, in order to determine the looping energy and to reproduce Figure 19.27.

(c) As mentioned many times throughout the book, Müller et al. (1996) performed an experiment where the repression level was measured as a function of the distance between operators. The experiment and its results are shown in Figure 1.11 (p. 19). Based on their repression data and the thermodynamic models from the chapter, make a plot of the looping energy as a function of the interoperator distance. Show analytically that a maximum in repression corresponds to a minimum in looping energy. At what interoperator distance is the inferred looping free energy at a minimum? Is this consistent with the measured persistence length of DNA *in vitro*, which is 50 nm?

(d) Fit the looping energy obtained in (c) to the functional form $\Delta F_{\text{loop}} = a/N_{\text{bp}} + b \ln N_{\text{bp}} + cN_{\text{bp}} + e$. Use this looping energy to make predictions about the outcome of a hypothetical experiment similar to the one performed by Müller et al. (1996), but now using cells bearing 10, 200, and 900 Lac repressor molecules per cell.

Relevant data for this problem are provided on the book's website.

• 19.3 Sensitivity of the regulation factor

An important concept in gene regulation is the sensitivity, that is, how steep is the change in gene expression (for example, the steepness of the transition from the "off" to the "on" state in activation) in response to a change in the number of transcription factors. It can be quantified by obtaining the slope on a log–log plot of the level of gene expression versus the number of transcription factors at this transition. Using thermodynamic models of gene regulation, determine how the sensitivity depends on the relevant parameters for the following regulatory motifs in the case of a weak promoter:

(a) Simple activation.

(b) Simple repression.

(c) Two binding sites where the same species of repressor can bind. They can recruit each other and repress RNA polymerase independently. What happens when the interaction is turned off? For simplicity, assume that both binding sites have the same binding energy.

(d) Repression in the presence of DNA looping.

• 19.4 Plasmid copy number and gene expression

In this problem, we work out an expression for the repression for the case in which there are N plasmids, each harboring the same promoter subjected to repression by the simple repression motif.

(a) Write a partition function for P RNA polymerase molecules that can bind to the plasmids, resulting in expression of our gene of interest. Take into account the cell's nonspecific reservoir and assume that $P \gg N$. Calculate the mean number of plasmids occupied by RNA polymerase, $\langle N \rangle$. Could you just have predicted this result based on what you know about the $N = 1$ case?

(b) Work out an expression for the repression defined as

$$\text{repression} = \frac{\langle N \rangle (R = 0)}{\langle N \rangle (R \neq 0)}. \tag{19.174}$$

Make sure to take into account the distinct cases where $N < R$ and $N > R$, where R is the number of repressors, and assume that you are dealing with a weak promoter, namely $(P/N_{\text{NS}})e^{-\beta \Delta \varepsilon_{\text{pd}}} \ll 1$.

(c) Show that your result yields the same expression for simple repression in the case where $N = 1$ that we found in the chapter.

(d) Consider the case where there are two plasmids (that is, $N = 2$) and work out the repression as a function of the number of repressors and make a corresponding plot.

• 19.5 The transcriptional machinery in eukaryotes

In the thermodynamic models of gene regulation discussed in the chapter, the RNA polymerase is treated as a single molecular species. While this might be a reasonable assumption for transcription in prokaryotes, in eukaryotes tens of different molecules need to come together in order to form the transcriptional machinery. The objective of this problem is to develop intuition about the requirements for our simple model to apply in such a complex case by assuming that the transcriptional machinery is made out of

two different subunits, X and Y, that come together at the promoter.

(a) Calculate the probability of finding the complex X + Y bound to the promoter in the case where unit X binds to DNA and unit Y binds to X. Can you reduce this to an effective one-molecule problem such as in the bacterial case?

(b) Calculate the fold-change in gene expression for simple repression using transcriptional machinery such as that proposed in (a). Explore the weak promoter assumption in order to reduce the expression to that corresponding to the bacterial case. Repeat this for the case where an activator can contact Y.

(c) Repeat (a) and (b) for a case where Y binds to a site on the DNA that is near the X-binding site, and there is an interaction energy between X and Y.

• 19.6 Induction of transcription factors

Even though experiments where the concentration of a transcription factor is varied are easier to interpret in terms of models, like those described in this chapter, the experiments that are the easiest to perform are those where the affinity of the transcription factor to its specific binding sites on the DNA is regulated by an inducer molecule. In the case of Lac repressor, allolactose or any of its analogs (IPTG, for example) can be used to reduce its specific binding energy to values similar to its nonspecific binding to DNA.

Assume a simple model of induction where one inducer molecule binds to the repressor, which then loses its ability to bind specifically to its operator site. Calculate repression in this case and plot it as a function of the number of inducer molecules in the cell.

• 19.7 Solving the unregulated promoter master equation

Solve the master equation for the unregulated promoter shown in Equation 19.39 in steady state by proposing a solution in terms of a generating function given by $f(s) = \sum_{m=0}^{+\infty} p(m) s^m$. In order to do this, you will have to multiply both sides of the equation by s^m and sum over all values of m in order to obtain a differential equation for $f(s)$.

• 19.8 Cell-to-cell variability as a function of fold-change

In the chapter, we derived the Fano factor for a promoter architecture regulated by a repressor that binds to a single site overlapping the promoter. In this case, the Fano factor depends on the mean absolute number of mRNA molecules per cell. An alternative way of looking at the Fano factor is as a function of the fold-change in gene expression, which, under the weak promoter approximation, is just the regulation factor. Reproduce the plot shown in Figure 19.37(A) by calculating the Fano factor as a function of the corresponding fold-change in the mean level of gene expression.

• 19.9 Separation of time scales and transcriptional regulation

For transcription to start, the RNA polymerase bound to the promoter needs to undergo a conformational change to the so-called open complex. The rate of open complex formation is often much smaller than the rates for the polymerase binding and falling off the promoter. Here, we investigate within a simple model how this state of affairs might justify the equilibrium assumption underlying thermodynamic models of gene regulation, namely that the equilibrium probability that the promoter is occupied by the RNA polymerase determines the level of gene expression.

(a) Write down the chemical kinetics equation for this situation. Consider three states: RNA polymerase bound nonspecifically on the DNA (N); RNA polymerase bound to the promoter in the closed complex (C); and RNA polymerase bound to the promoter in the open complex (O). To simplify matters, take both the rate for N $\rightarrow$ C and the rate for C $\rightarrow$ N to be k. Assume that the transition C $\rightarrow$ O is irreversible, with rate Γ.

(b) For $\Gamma = 0$, show that in the steady state there are equal numbers of RNA polymerases in the N and C states. What is the steady state in the case $\Gamma \neq 0$?

(c) For the case $\Gamma \neq 0$, show that for times $1/k \ll t \ll 1/\Gamma$, the numbers of RNA polymerases in the N and C states are equal, as would be expected in equilibrium.

• 19.10 Copy number and the Poisson promoter

The model of the Poisson promoter considered in the chapter assumed that the number of copies of the gene of interest was fixed at one. However, as a result of the replication of the chromosomal DNA, during some part of the cell cycle there will be two (or even more for rapidly dividing cells) copies of the gene of interest. In this problem, we imagine that during a fraction f of the cell cycle, there is one copy of our gene of interest and during the rest of the cell cycle there are two such copies.

(a) Write down the appropriate distribution $p(m)$ for m mRNA molecules as a function of the parameter f.

(b) Find $\langle m \rangle$.

(c) Find $\langle m^2 \rangle$ and use it to find the Fano factor.

(d) Plot the Fano factor as a function of f for different choices of the mean mRNA copy number for a single promoter. How "Poissonian" do you expect an unregulated promoter to be?

(Problem courtesy of Rob Brewster and Daniel Jones.)

• 19.11 Gillespie algorithm revisited

In the computational exploration, we showed how the mRNA evolves as a function of time.

(a) Plot the bias of the reaction-choice coin flip (that is, production or decay) as a function of time. Explain intuitively what is happening.

(b) Plot the time step as a function of time.

• 19.12 Mean protein burst size for a single mRNA

Using the probability distribution for a protein burst of size n from Equation 19.113 and the definition of the mean burst size as

$$\langle n \rangle = \sum_{n=0}^{\infty} n P(n), \tag{19.175}$$

demonstrate that the mean burst size is given by the ratio of the protein translation rate to the rate of mRNA decay as

in Equation 19.114:

$$\langle n \rangle = \frac{r_p}{\gamma} = b. \tag{19.176}$$

• 19.13 A minimal genetic switch

In this problem, we consider a simpler switch than that considered in the chapter. For this switch, we consider an activator that activates its own production.

(a) Make a figure of the states, weights, and rates for an activator that activates itself by binding as a dimer and such that its binding to the DNA is characterized by a Hill function with Hill coefficient 2. Your states and weights should be analogous to those shown in Figure 19.45. Given these states and weights, write a rate equation for the time evolution of the activator. Include a term for a basal rate of production even in the absence of activator.

(b) Make a one-dimensional phase portrait by performing a graphical analysis of the differential equation based on the plot of $\mathrm{d}A/\mathrm{d}t$ versus A. Use this phase portrait to characterize the existence of fixed points and their stability. Is it appropriate to refer to this as a switch?

• 19.14 Chemotaxis of *E. coli*

In chemotaxis experiments, a source of nutrient molecules can be introduced into the medium containing bacteria via a micropipette. The outward diffusion of the nutrient molecules creates a position-dependent concentration gradient, and the chemotactic response of the bacteria can be observed under a microscope.

(a) Estimate the nutrient gradient in steady state as a function of the distance from the micropipette r by assuming that it keeps the concentration fixed at c_0 for distances $r < r_0$. Make a plot of the concentration gradient as a function of r for typical values $c_0 = 1\,\mathrm{mM}$ and $r_0 = 1\,\mu\mathrm{m}$.

(b) Assuming that the bacterium makes two measurements of the concentration using one array of receptor proteins at one of its ends and another array at the other, estimate the maximum distance from the nutrient source for which the bacterium is still able to detect a gradient. Assume that the receptor array counts the number of molecules present in a cubic volume with side $a = 100$ nm. To solve this problem, you should recall that the counting error for N molecules is roughly $\sqrt{N}$, and in order to detect the difference in concentration between the two ends of the bacterium, the measurement error should be less than the difference itself.

(c) Now assume a different strategy, where one receptor is employed but the bacterium compares the concentration at two different positions along a run, separated by a distance of $10\,\mu\mathrm{m}$. Compute the maximum distance from the nutrient source at which the bacterium will be able to detect the gradient in this case.

• 19.15 MWC model for heterogeneous receptor clusters

Develop an MWC model for the response of chemotactic receptor clusters where there are M molecules of one type of receptor and N molecules of the other type in a given cluster. The entire cluster is either active or inactive and the two different receptors are characterized by different affinities for the chemoattractant of interest. Specifically, derive an equation that is analogous to Equation 19.158 for the probability that the receptor cluster will be in the on state. To do so, construct a states-and-weights diagram like that shown in Figure 19.58.

• 19.16 N-WASP and biochemistry on a leash

In the last section of the chapter, we considered the action of N-WASP using a simple one-dimensional random walk model to treat the statistical mechanics of looping. Redo that analysis by using the Gaussian model of a polymer chain. First, assume that the loop has to close on itself and then account for the finite size of the protein domain. Compare your results with those obtained in the chapter.

19.7 Further Reading

Alon, U (2007) An Introduction to Systems Biology: Design Principles of Biological Circuits, Chapman & Hall/CRC. Alon's book gives a comprehensive and thoughtful discussion of regulation.

Bintu, L, Buchler, NE, Garcia, HG, et al. (2005) Transcriptional regulation by the numbers: applications, *Curr. Opin. Genet. Dev.* **15**, 125. Application of thermodynamic models to several different regulatory architectures.

Buchler, NE, Gerland, U, & Hwa, T (2003) On schemes of combinatorial transcription logic, *Proc. Natl Acad. Sci. USA* **100**, 5136. Excellent general discussion of thermodynamic models of gene regulation.

Walsh, CT (2006) Posttranslational Modification of Proteins: Expanding Nature's Inventory, Roberts and Company Publishers. This book is full of interesting insights into the phenomenology of post-translational modification. This is a reminder that there is more to regulation than transcriptional control.

Cherry, JL, & Adler, FR (2000) How to make a biological switch, *J. Theor. Biol.* **203**, 117. This article presents an interesting discussion of the issues that arise in designing biological switches.

Ptashne, M (2004) A Genetic Switch, 3rd ed., Cold Spring Harbor Laboratory Press. A beautiful book that focuses on ideas as opposed to facts and paints a picture of how gene regulation works.

Ptashne, M, & Gann, A (2002) Genes and Signals, Cold Spring Harbor Laboratory Press. This book provides an excellent overview of transcriptional regulation.

Michel, D (2010) How transcription factors can adjust the gene expression floodgates, *Prog. Biophys. Mol. Biol.* **102**, 16. This article gives a comprehensive discussion of the physics of gene expression.

Müller-Hill, B (1996) The Lac Operon: A Short History of a Genetic Paradigm, Walter de Gruyter. Müller-Hill's book is a fascinating and idiosyncratic account of the development of thinking on gene regulation in general and the *lac* operon in particular. The book is full of interesting touches such as Figure 3, which illustrates the ways in which synthetic analogs of lactose have played a role in the development of molecular biology.

Davidson, EH (2001) Genomic Regulatory Systems, Academic Press. Davidson's book is full of both interesting facts and provocative ideas. We particularly recommend it as a way to

explore the complexity associated with eukaryotic gene regulation.

Ellner, SP, & Guckenheimer, J (2006) Dynamic Models in Biology, Princeton University Press. This book examines dynamical models and their relevance to biology and has a treatment of both the genetic switch and the repressilator.

Berg, HC, & Brown, DA (1972) Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking, *Nature* **239**, 500. This paper uses a three-dimensional tracking technique to follow individual bacterial cells during chemotaxis and demonstrates how bacteria find their way by altering the timing of runs and tumbles.

Keymer, JE, Endres, RG, Skoge, M, Meir, Y, & Wingreen, NS (2006) Chemosensing in *Escherichia coli*: two regimes of two-state receptors, *Proc. Natl Acad. Sci. USA* **103**, 1786. This interesting paper is the basis of the model presented in Section 19.4.

19.8 References

Alon, U, Surette, MG, Barkai, N, & Leibler, S (1999) Robustness in bacterial chemotaxis, *Nature* **397**, 168.

Ben-Tabou de-Leon, S, & Davidson, EH (2007) Gene regulation: gene control network in development, *Annu. Rev. Biophys. Biomol. Struct.* **36**, 191.

Cai, L, Friedman, N, & Xie, XS (2006) Stochastic protein expression in individual cells at the single molecule level, *Nature* **440**, 358.

Case, TJ (2000) An Illustrated Guide to Theoretical Ecology, Oxford University Press.

Cluzel, P, Surette, M, & Leibler, S (2000) An ultrasensitive bacterial motor revealed by monitoring signalling proteins in single cells, *Science* **287**, 1652.

Driever, W, Thoma, G, & Nüsslein-Volhard, C (1989) Determination of spatial domains of zygotic gene expression in the *Drosophila* embryo by the affinity of binding sites for the bicoid morphogen, *Nature* **340**, 363.

Dueber, JE, Yeh, BJ, Chak, K, & Lim, WA (2003) Reprogramming control of an allosteric signaling switch through modular recombination, *Science* **301**, 1904.

Elowitz, MB, & Leibler, S (2000) A synthetic oscillatory network of transcriptional regulators, *Nature* **403**, 335.

Gardner, TS, Cantor, CR, & Collins, JJ (2000) Construction of a genetic toggle switch in *Escherichia coli*, *Nature* **403**, 339.

Gause, GF (1934) The Struggle for Existence, Williams & Wilkins (reprinted by Dover Publications, 2003).

Gillespie, DT (1977) Exact stochastic simulation of coupled chemical reactions, *J. Phys. Chem.* **81**, 2340.

Golding, I, Paulsson, J, Zawilski, SM, & Cox, EC (2005) Real-time kinetics of gene activity in individual bacteria, *Cell* **123**, 1025.

Kim, HD, & O'Shea, EK (2008) A quantitative model of transcription factor-activated gene expression, *Nat. Struc. Mol. Biol.* **15**, 1192.

Kuhlman, T, Zhang, Z, Saier, MHS, & Hwa, T (2007) Combinatorial transcriptional control of the lactose operon of *Escherichia coli*, *Proc. Natl Acad. Sci. USA* **104**, 6043.

Min, TL, Mears, PJ, Chubiz, LM, et al. (2009) High-resolution, long-term characterization of bacterial motility using optical tweezers, *Nat. Methods* **6**, 831.

Müller, J, Oehler, S, & Müller-Hill, B (1996) Repression of *lac* promoter as a function of distance, phase and quality of an auxiliary *lac* operator, *J. Mol. Biol.* **257**, 21.

Myasnikova, E, Samsonova, A, Kozlov, K, Samsonova, M, & Reinitz, J (2001) Registration of the expression patterns of *Drosophila* segmentation genes by two independent methods, *Bioinformatics* **17**, 3.

Oehler, S, Alberti, S, & Müller-Hill, B (2006) Induction of the *lac* promoter in the absence of DNA loops and the stoichiometry of induction, *Nuc. Acids Res.* **34**, 606.

Oehler, S, Amouyal, M, Kolkhof, P, von Wilcken-Bergmann, B, & Müller-Hill, B (1994) Quality and position of the three *lac* operators of *E. coli* define efficiency of repression, *EMBO J.* **13**, 3348.

Rosenfeld N, Young JW, Alon U, et al. (2005) Gene regulation at the single-cell level. *Science* **307**, 1962.

Small, S, Blair, A, & Levine, M (1992) Regulation of *even-skipped* stripe 2 in the *Drosophila* embryo, *EMBO J.* **11**, 4047.

Small, S, Blair, A, & Levine, M (1996) Regulation of two pair-rule stripes by a single enhancer in the *Drosophila* embryo, *Developmental Biology* **175**, 314.

Taniguchi, Y, Choi, PJ, Li, G-W, et al. (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells, *Science* **329**, 533.

Turner, L, Ryu, WS, & Berg, HC (2000) Real-time imaging of fluorescent flagellar filaments, *J. Bacteriol.* **182**, 2793.

Yu, J, Xiao, J, Ren, X et al. (2006) Probing gene expression in live cells, one protein molecule at a time, *Science* **311**, 1600.

Zenklusen, D, Larson, DR, & Singer, RH (2008) Single-RNA counting reveals alternative modes of gene expression in yeast, *Nat. Struc. Mol. Bio.* **15**, 1263.