Chapter 6
Stochastic Gene Expression and Regulatory Networks

Genetically identical cells exposed to the same environmental conditions can show significant variation in molecular content and marked differences in phenotypic characteristics. This intrinsic variability is linked to the fact that many cellular events at the genetic level involve small numbers of molecules (low copy numbers). We have already encountered intrinsic noise effects within the context of stochastic ion channels (Chap. 3) and biochemical signaling (Chap. 5). Although stochastic gene expression was originally viewed as having detrimental effects on cellular function, with potential implications for disease, it is now seen as being potentially advantageous. For example, intrinsic noise can provide the flexibility needed by cells to adapt to fluctuating environments or respond to sudden stresses and can also support a mechanism by which population heterogeneity is established during cell differentiation and development. Since the demonstration of a functional role for stochastic gene expression in λ-phage [13], there has been an explosion of studies focused on investigating the origins and consequences of noise in gene expression (see the reviews [312, 408, 502, 521, 555, 644]). This typically involves establishing the molecular mechanisms of noise generation at the single gene level and then building on this knowledge to test and predict its effects on larger regulatory networks. *Gene regulation* refers to the cellular processes that control the expression of proteins, dictating under what conditions specific proteins should be produced from their parent DNA. This is particularly crucial for multicellular organisms, where all cells share the same genomic DNA, yet do not all express the same proteins. That is, selective gene expression allows the cells to specialize into different phenotypes (cell differentiation), resulting in the development of different tissues and organs with distinct functional roles.

In this chapter we explore the effects of noise on gene expression and protein synthesis. We begin by reviewing the basic steps in gene expression (Sect. 6.1). We then analyze transcription and translation in some simple unregulated networks and show how translational bursts in the production of protein can occur (Sect. 6.2). Various simple gene regulatory networks are analyzed in Sect. 6.3 using the linear noise (diffusion) approximation (see also Sect. 3.2) and Fourier (spectral) methods. Important examples of nonlinear feedback regulatory networks such as genetic

P.C. Bressloff, *Stochastic Processes in Cell Biology*, Interdisciplinary Applied Mathematics 41, DOI 10.1007/978-3-319-08488-6_6

switches and genetic oscillators are studied in Sect. 6.4, including the *lac* operon and the genetic circuits of the circadian clock. We also discuss some methods for analyzing the effects of noise on biochemical oscillators. The efficacy of gene networks in transmitting information in the presence of molecular noise is investigated in Sect. 6.5, where some basic concepts such as Shannon information and mutual information are introduced. We then look at some models of kinetic proofreading, which is a mechanism for increasing the fidelity of molecular recognition during protein synthesis, for example, and other cellular processes such as T-cell activation in immunology (see Sect. 6.7). Finally, the stochastic simulation algorithm (SSA) introduced by Gillespie to simulate sample trajectories of a gene or biochemical network is described in Sect. 6.8.

6.1 Basics of Gene Expression

In Fig. 6.1a we show the two main stages in the expression of a single gene according to the *central dogma*.

1. *Transcription (DNA → RNA).* The first major stage of gene expression is the synthesis of a *messenger RNA* (mRNA) molecule with a nucleotide sequence complementary to the DNA strand from which it is copied—this serves as the template for protein synthesis. Transcription is mediated by a molecular machine known as RNA polymerase (RNAP). In the case of eukaryotes, transcription takes place in the cell nucleus, whereas subsequent protein synthesis takes place in the cytoplasm, which means that the mRNA has to be exported from the nucleus as an intermediate step.
2. *Translation (RNA → protein).* The second major stage is synthesis of a protein from mRNA. Translation is mediated by a macromolecule known as a *ribosome*, which produces a string of amino acids (polypeptide chains), each specified by a *codon* (represented by three letters) on the mRNA molecule. Since there are four nucleotides (A, U, C, G), there are 64 distinct codons, e.g., AUG and CGG, most of which code for a single amino acid. The process of translation consists of ribosomes moving along the mRNA without backtracking (from one end to the other, technically known as the 5' end to the 3' end) and is conceptually divided into three major stages (as is transcription): initiation, elongation, and termination. Each elongation step invokes translating or "reading" of a codon and the binding of a freely diffusing transfer RNA (tRNA) molecule that carries the specific amino acid corresponding to that codon. Once the chain of amino acids has been generated a number of further processes occur in order to generate a correctly folded protein.

The above simplified picture ignores a major feature of cellular processing, namely, gene regulation. Individual cells frequently have to make "decisions," that is, to express different genes at different spatial locations and times and at different activity levels. One of the most important mechanisms of genetic control is transcriptional regulation, that is, determining whether or not an mRNA molecule

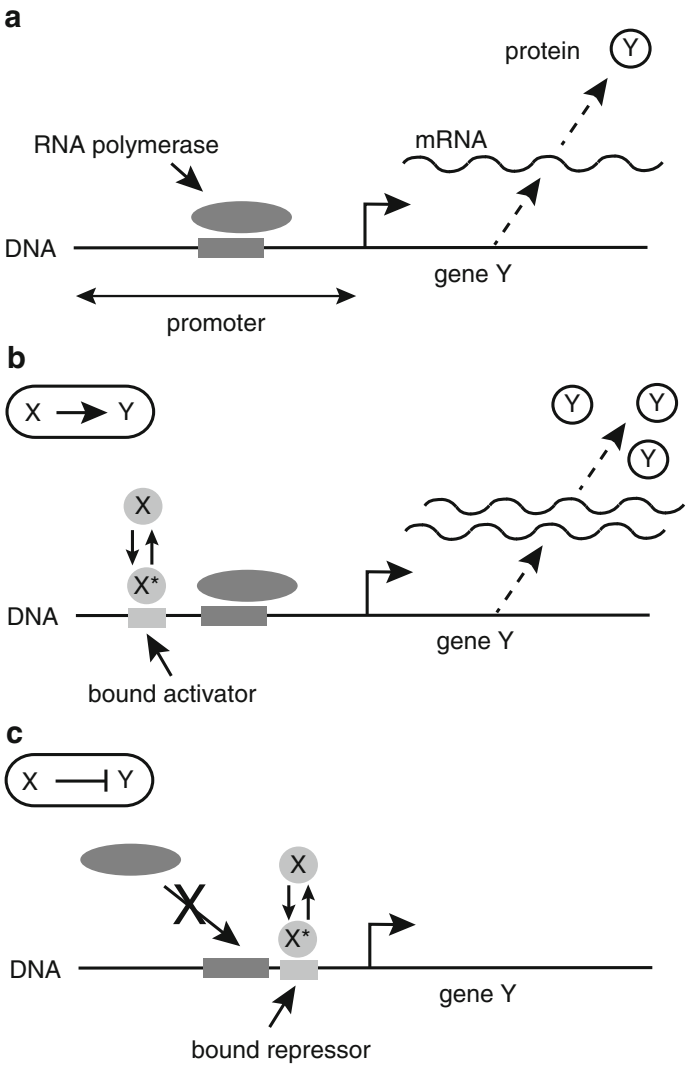

Fig. 6.1: Transcriptional regulation due to the binding of a repressor or activator protein to a promoter region along the DNA. (**a**) Unregulated transcription of a gene Y following binding of RNA polymerase to the promoter region. The resulting mRNA exits the nucleus and is then translated by ribosomes to form protein Y. (**b**) Increased transcription due to the binding of an activator protein *X* to the promoter. An activator typically transitions between inactive and active forms; the active form X^* has a high affinity to the promoter binding site. An external chemical signal can regulate transitions between the active and inactive states. (**c**) Transcription can be stopped by a repressor protein X binding to the promoter and blocking the binding of RNA polymerase

is made. The control of transcription (switching on or off a gene) is mediated by proteins known as transcription factors (see Fig. 6.1b, c). Negative control (or repression) is mediated by repressors that bind to a promoter region along the DNA where RNAP has to bind in order to initiate transcription—it thus inhibits transcription. On the other hand, positive control (activation) is mediated by activators that increase the probability of RNAP binding to the promoter. The presence of

transcription factors means that cellular processes can be controlled by extremely complex gene networks, in which the expression of one gene produces a repressor or activator, which then regulates the expression of the same gene or another gene. This can result in many negative and positive feedback loops, the understanding of which lies at the heart of systems biology [5]. In addition to transcriptional regulation, there are a variety of other mechanisms that can control gene expression including mRNA and protein degradation and translational regulation.

6.1.1 Intrinsic Versus Extrinsic Noise Sources

Following Swain et al. [164], it is useful to distinguish between contributions arising from fluctuations that are inherent to a given system of interest (intrinsic noise) from those arising from external factors (extrinsic noise). In the model of gene expression shown in Fig. 6.1, intrinsic noise is due to fluctuations generated by the binding/unbinding of a repressor or activator and mRNA and protein production and decay—these can be significant due to the small number of molecules involved. Extrinsic noise sources are defined as fluctuations and population variability in the rate constants associated with these events. The classification of a noise source as intrinsic rather than extrinsic is context-dependent, so that intrinsic noise at one level can act as extrinsic noise at another level. Gene-intrinsic noise refers to the variability generated by molecular-level noise in the reaction steps that are intrinsic to the process of gene expression. Network-intrinsic noise is generated by fluctuations and variability in signal transduction and includes gene-intrinsic noise in the expression of regulatory genes. Cell-intrinsic noise arises from gene-intrinsic noise and network-intrinsic noise, as well as fluctuations and variability in cell-specific factors, such as the activity of ribosomes and polymerases, metabolite concentrations, cell size, cell age, and stage of the cell cycle.

An operational definition of gene-intrinsic noise is the difference in the expression of two almost identical genes from identical promoters in single cells averaged over a large cell population. This definition is based on the assumptions that the two genes are affected identically by fluctuations in cell-specific factors and that their expression is perfectly correlated if these fluctuations are the only source of population heterogeneity. The contribution of gene-intrinsic noise can then be investigated experimentally using two-reporter assays (see Sect. 1.2). These assays evaluate, in single cells, the difference in the abundances of two equivalent reporters, such as red and green fluorescent protein, expressed from identical promoters, located at equivalent chromosomal positions. This allows measurements of noise fluctuations generated by the biochemical reaction steps that are intrinsic to the process of gene expression, and how this is affected by mutations or gene deletions. There are, however, some potential limitations. For example, contributions from extrinsic factors, such as imperfect timing in replication and intracellular heterogeneity, might

be measured as gene-intrinsic noise. Moreover, because increased variability in regulatory signals might cause cells to adapt distinct expression states, the measured population-average gene-intrinsic noise and the extrinsic regulatory noise might not always be independent.

6.1.2 Biological Significance of Stochasticity

Stochasticity in gene expression is generally believed to be detrimental to cell function, because fluctuations in protein levels can corrupt the quality of intracellular signals, negatively affecting cellular regulation. One possible benefit of randomness, however, is that it can provide a mechanism for phenotypic and cell-type diversification:

1. Stochasticity in gene expression that generates phenotypic heterogeneity is expected to be particularly beneficial to microbial cells that need to adapt efficiently to sudden changes in environmental conditions. Fluctuations in gene expression provide a mechanism for 'sampling' distinct physiological states and could therefore increase the probability of survival during times of stress, without the need for genetic mutation. A classical example is the infection of *E. coli* by a bacterial virus known as *lambda phage*. Infection is governed by a particular lysis/lysogeny decision circuit, in which only a fraction of infecting phage chooses to lyse (break down) the cell. The remainder become dormant lysogens, in which the bacteriophage nucleic acid is integrated into the host bacterium's genome, awaiting bacterial stress signals to enter the production phase of their life cycle.
2. Switching between phenotypic states with different growth rates might be an important factor in the phenomenon of persistent bacterial infections after treatment with antibiotics. Although most of the population is rapidly killed by the treatment, a small genetically identical subset of dormant 'persistor' cells can survive an extended period of exposure. When the drug treatment is removed, the surviving persistors randomly transition out of the dormant state, causing the infection to reemerge.
3. The primary purpose of the *Saccharomyces cerevisiae* (yeast) galactose-utilization network is to increase the uptake and metabolism of galactose. It involves several positive feedback loops that generate bistability in the network, which endow cells (and their progeny) with long-term epigenetic memory of past galactose-consumption states. It has been suggested that the existence of a negative feedback loop (which appear spurious from a deterministic perspective) reduces this memory by increasing the rate at which cells randomly switch between different phenotypic states that are associated with different expression of the galactose-utilization genes. As a result, the biological function of negative feedback might be to prevent cells from being trapped in suboptimal phenotypic states.

4. Stochasticity may also play a constructive role in development and cellular differentiation in higher organisms. For example, during *Drosophila melanogaster* development, stochastic fluctuations in the turnover of two proteins, Notch and Delta, might underlie the random emergence of neural precursor cells from an initial homogeneous cell population.

6.2 Unregulated Transcription and Translation

The key steps in transcription are binding of RNAP (P) to the relevant promoter region of DNA (D) to form a closed complex (PD_c), the unzipping of the two strands of DNA to form an open complex (PD_o), and finally promoter escape, when RNAP reads one of the exposed strands

$$P + D \underset{k_-}{\overset{k_+}{\rightleftharpoons}} PD_c \xrightarrow{k_{\text{open}}} PD_0 \xrightarrow{k_{\text{escape}}} \text{transcription.}$$

Once the RNAP is reading the strand, the promoter is unoccupied and ready to accept a new polymerase. The binding/unbinding of polymerase is very fast, $k_\pm \gg k_{\text{open}}$ so that the first step happens many times before formation of an open complex. Hence, one can treat the RNAP as in quasi-equilibrium with the promoter characterized by an equilibrium constant $K_P = k_+/k_-$. The rate of transcription will thus be proportional to the fraction of bound RNAP, $k_+/(k_+ + k_-)$. The production of mRNA from a typical gene in *E. coli* occurs at a rate around 10 per minute, while the average lifetime of mRNA due to degradation is around a minute. This implies that on average there are ten mRNA molecules per cell. Generation of the mRNA molecule occurs at a rate of 50 nucleotides per second. Hence, a typical gene of around 1,000 nucleotides will be transcribed in about 20 s. Thus, there are around three RNAP per gene at any one time, suggesting the number fluctuations will be significant.

First, suppose that we ignore any regulation of the promoter as in Fig. 6.1a, and collapse the various stages of transcription into a single step with mRNA production rate k. Letting γ denote the rate of mRNA degradation and $m(t)$ the number of mRNA molecules at time t, we have the reaction

$$m \xrightarrow{k} m+1, \quad m \xrightarrow{\gamma} m-1$$

with corresponding kinetic equation for the concentration $x = m/\Omega$, where Ω is cell volume

$$\frac{dx}{dt} = k - \gamma x.$$

Clearly, given that m is of order 10, the law of mass action breaks down and we have to consider the corresponding birth–death master equation

$$\frac{dp_m(t)}{dt} = -\Omega k p_m(t) + \Omega k p_{m-1}(t) - \gamma m p_m(t) + \gamma(m+1)p_{m+1}(t) \tag{6.2.1}$$

for $m \geq 0$ and $P_{-1} \equiv 0$. This is identical to the autonomous version (3.6.3) of the master equation for a stochastic gating model (Sect. 3.6). We immediately deduce that the resulting probability density is given by the Poisson distribution (3.6.7). Hence, in the limit $t \to \infty$ we obtain a stationary Poisson process with

$$p_m = \mathrm{e}^{-\lambda}\frac{\lambda^m}{m!}, \quad \lambda = \Omega k/\gamma. \tag{6.2.2}$$

It follows that

$$\langle m \rangle = \lambda, \quad \mathrm{var}[m] = \lambda.$$

This is an important result because both the mean and variance in the number of mRNA molecules can be measured experimentally. One commonly used measure of the level of noise in a regulatory network is the so-called Fano factor:

$$\text{Fano factor} = \frac{\langle m^2 \rangle - \langle m \rangle^2}{\langle m \rangle}. \tag{6.2.3}$$

For the unregulated process, the Fano factor is one.

6.2.1 Translational Bursting

In addition to transcription, other steps in the central dogma are also subject to variability including protein translation, which often occurs in bursts [36, 89, 199, 427]. One could add the translation step (mRNA $\to$ protein) to the previous model. However, it is simpler to proceed by exploiting the fact that a single mRNA molecule has a much shorter lifetime than a protein. First, consider a single mRNA molecule with a degradation rate γ, which starts synthesizing a protein at time $t = 0$. Let $p_0(n,t)$ ($p_c(n,t)$) denote the probability that there are n proteins at time t and the mRNA has not (has) decayed. Neglecting protein degradation, we have the master equation

$$\frac{dp_0(n,t)}{dt} = -\gamma p_0(n,t) + r[p_0(n-1,t) - p_0(n,t)] \tag{6.2.4a}$$

$$\frac{dp_c(n,t)}{dt} = \gamma p_0(n,t), \tag{6.2.4b}$$

where r is the rate of protein production and $p_0(-1,t) \equiv 0$. Let

$$P(n) = \lim_{t\to\infty} p_c(n,t).$$

Note that $\lim_{t\to\infty} p_0(n,t) = 0$ due to the decay of mRNA. Integrating Eq. (6.2.4b) with respect to time gives

$$P(n) = \gamma \int_0^\infty p_0(n,t)dt,$$

since $p_c(n,0) = 0$. In order to compute $p_0(n,t)$, integrate Eq. (6.2.4a) with respect to time using $p_0(n,t) = \delta_{n,0}$:

$$-\delta_{n,0} = -P(n) + \frac{r}{\gamma}[P(n-1) - P(n)].$$

Setting $n = 0$ gives

$$P(0) = \frac{\gamma}{r+\gamma}.$$

For $n \geq 1$, we have the recurrence relation

$$P(n) = \frac{r}{r+\gamma}P(n-1) \implies P(n) = \left(\frac{r}{r+\gamma}\right)^n \frac{\gamma}{r+\gamma}.$$

An important quantity is the so-called burst size b, which is the mean number of proteins produced per mRNA. Using generating functions it can be shown that (see Ex. 6.1)

$$b = \frac{r}{\gamma}.$$

The idea of a translational burst refers to the observation that a single mRNA generates a burst of protein production before it decays (see Fig. 6.2a).

Now suppose that there are m mRNA molecules and that translation of each mRNA proceeds independently. The probability of producing N proteins due to bursts from each mRNA molecule can be expressed as a multiple convolution [509]. For example, if $m = 2$, then

$$P_2(N) = \sum_{n=0}^{N} P(n)P(N-n),$$

and

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

Assume that the number of proteins is sufficiently large so that we can approximate the sums by integrals, for example,

$$P_2(N) = \int_0^N P(n)P(N-n)dn.$$

The advantage of the integral formulation is that one can use Laplace transforms and the convolution theorem. Thus, setting

$$\tilde{P}_m(s) = \int_0^\infty P_m(n)\mathrm{e}^{-sn}dn,$$

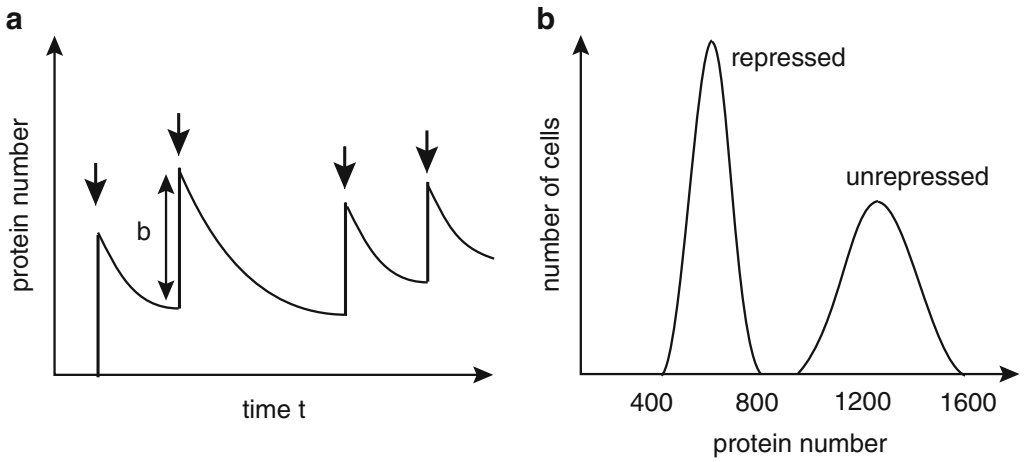

Fig. 6.2: Effects of noise in gene expression. (**a**) Schematic illustration of translational bursting. Each *arrow* represents a burst event where an mRNA transcript releases a burst of proteins of average size b, and proteins decay between bursts. (**b**) Illustration of how negative feedback in an autoregulatory network reduces the mean number of proteins but also reduces the size of fluctuations

we have

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

Calculating $\tilde{P}_m(s)$ and then inverting yields the result (see Ex. 6.1)

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

For $n, b \gg 1$, we can make the approximation

$$\left(\frac{b}{1+b}\right)^n = \mathrm{e}^{-n\ln(1+b^{-1})} \approx \mathrm{e}^{-n/b},$$

which leads to the gamma distribution for n with m fixed:

$$P_m(n) \equiv F(n;m,b^{-1}) = \frac{n^{m-1}\mathrm{e}^{-n/b}}{b^m\Gamma(m)}. \tag{6.2.5}$$

From properties of the gamma distribution, we immediately note that for a given number of mRNA molecules,

$$\langle n\rangle = mb, \quad \mathrm{var}(n) = mb^2.$$

Hence, under the various approximations the Fano factor is of the order of the burst size b. Finally, an estimate for m is $m \approx k/\gamma_0$ where k is the rate of production of mRNAs and γ_0 is the frequency of the cell cycle (assuming that it is higher than the rate of protein degradation).

An alternative approach to analyzing protein bursting is to start from the Chapman–Kolmogorov (CK) equation [199]

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial}{\partial x}[\gamma_0 x p(x)] + k\int_0^x w(x-x')p(x',t)dx', \tag{6.2.6}$$

where $p(x,t)$ is the probability density for x protein molecules (treating x as a continuous variable) at time t, and

$$w(x) = \frac{1}{b}\mathrm{e}^{-x/b} - \delta(x). \tag{6.2.7}$$

The first term on the right-hand side of the CK equation represents protein degradation, where the second term represents the production of proteins from exponentially distributed bursts. The gamma distribution (6.2.5) with $n \to x$ is obtained as the stationary solution of the CK equation, which can be established using Laplace transforms (Ex. 6.2). It is also possible to incorporate autoregulatory feedback into the CK equation by allowing the burst rate to depend on the current level of protein x, which acts as its own transcription factor [199]:

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial}{\partial x}[\gamma_0 x p(x)] + k\int_0^x w(x-x')c(x')p(x',t)dx'. \tag{6.2.8}$$

One possible form of the response function $c(x)$ is a Hill function

$$c(x) = \frac{k^s}{k^s + x^s},$$

with $s > 0$ ($s < 0$) corresponding to negative (positive) feedback. In this case, the stationary density takes the form (Ex. 6.2)

$$p(x) = Ax^{m(1+\varepsilon)-1}\mathrm{e}^{-x/b}[1 + (x/k)^s]^{-m/s}.$$

A more general mathematical analysis of bursting in discrete and continuous models can be found in [407].

6.3 Simple Models of Gene Regulation

One of the simplest gene regulatory networks consists of a gene that can be in one of two states, active or inactive (see Fig. 6.3). In the active state the gene produces protein X at a rate r, which subsequently degrades at a rate γ, whereas no protein is produced in the inactive state. For simplicity, the stages of transcription and translation are lumped together so we do not keep track of the amount of mRNA. Moreover, the transcription factor Y that switches on the gene is independent of protein X, that is, there is no feedback. Since the rate of activation k_+ will be proportional to the concentration c of Y in the nucleus, this simple network can be viewed as an

input/output device that converts the input signal c to an output signal given by the concentration x of protein X. Moreover, if X is a green fluorescent protein, then the output response can be measured. In Sect. 6.5, we will consider how effective such a feedforward networks is in transmitting information in the presence of molecular noise, following the work of Tkacik et al. [631, 632, 634, 665]. Here we will focus on calculating the level of noise.

6.3.1 Transcriptional Bursting in a Two-State Model

The reaction scheme of the regulatory network shown in Fig. 6.3 is

$$I \underset{k_-}{\overset{k_+}{\rightleftharpoons}} A \xrightarrow{r} \mathrm{p} \xrightarrow{\gamma} \emptyset,$$

where A and I denote the active and inactive states of the gene. We first consider the case in which the number of X proteins is sufficiently large so that we can represent the dynamics in terms of a continuous-valued protein concentration x [318]. The latter evolves according to the (piecewise) deterministic equation

$$\frac{dx}{dt} = rn(t) - \gamma x, \tag{6.3.1}$$

where the discrete random variable $n(t)$ represents the current state of the gene with $n(t) = 1$ (active) or $n(t) = 0$ (inactive). We thus have another example of a stochastic hybrid system. Let $p_j(x,t)$ denote the probability density of the protein concentration for $n(t) = j$, $j = 0, 1$. We then have the differential Chapman–Kolmogorov (CK) equation

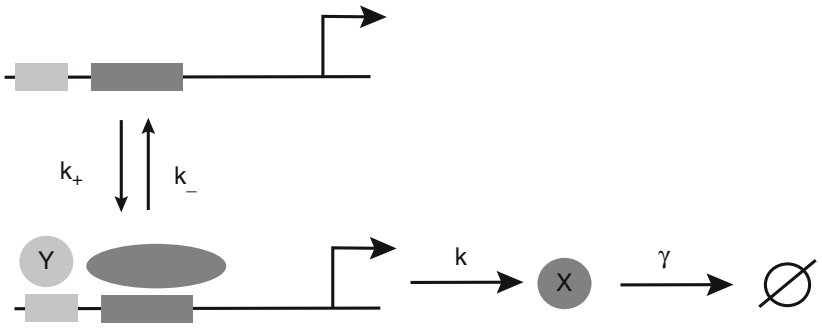

Fig. 6.3: Simple example of a two-state gene regulatory network. The promoter transitions between an active state (bound by a transcription factor protein Y and RNA polymerase) and an inactive state with rates $k_\pm$. The active state produces protein X at a rate r and protein X degrades at a rate γ

$$\frac{\partial p_0}{\partial t} = -\frac{\partial}{\partial x}(-\gamma x p_0(x,t)) + k_- p_1(x,t) - k_+ p_0(x,t) \tag{6.3.2a}$$

$$\frac{\partial p_1}{\partial t} = -\frac{\partial}{\partial x}([r-\gamma x]p_1(x,t)) + k_+ p_0(x,t) - k_- p_1(x,t), \tag{6.3.2b}$$

supplemented by the no-flux boundary conditions $J_s(x) = 0$ at $x = 0, r/\gamma$, where $J_0(x) = -\gamma x p_0(x)$ and $J_1(x) = [r-\gamma x]p_1(x)$. In the limit that the switching between active and inactive states is much faster than the protein dynamics, the probability that the gene is active rapidly converges to the steady state $k_+/(k_+ + k_-)$, and we obtain the deterministic equation

$$\frac{dx}{dt} = r\langle n\rangle - \gamma x = \frac{rk_+}{k_+ + k_-} - \gamma x. \tag{6.3.3}$$

Following [318], we will characterize the long-time behavior of the system in terms of the steady-state solution, which satisfies

$$\frac{d}{dx}(-\gamma x p_0(x)) = k_- p_1(x) - k_+ p_0(x) \tag{6.3.4a}$$

$$\frac{d}{dx}([r-\gamma x]p_1(x)) = k_+ p_0(x) - k_- p_1(x). \tag{6.3.4b}$$

The no-flux boundary conditions imply that $p_0(r/\gamma) = 0$ and $p_1(0) = 0$. First, note that we can take $x \in [0, r/\gamma]$ and impose the normalization condition

$$\int_0^{r/\gamma} [p_0(x) + p_1(x)]dx = 1.$$

Integrating Eq. (6.3.4) with respect to x then leads to the constraints

$$\int_0^{r/\gamma} p_0(x)dx = \frac{k_-}{k_- + k_+}, \quad \int_0^{r/\gamma} p_1(x)dx = \frac{k_+}{k_- + k_+}.$$

Adding Eqs. (6.3.4a) and (6.3.4b) we can solve for $p_0(x)$ in terms of $p_1(x)$ and then generate a closed differential equation for $p_1(x)$. We thus obtain a solution of the form (see Ex. 6.3),

$$p_0(x) = C(\gamma x)^{-1+k_+/\gamma}(r-\gamma x)^{k_-/\gamma}, \quad p_1(x) = C(\gamma x)^{k_+/\gamma}(r-\gamma x)^{-1+k_-/\gamma} \tag{6.3.5}$$

for some constant C. Imposing the normalization conditions then determines C as

$$C = \gamma\left[r^{(k_+ + k_-)/\gamma} B(k_+/\gamma, k_-/\gamma)\right]^{-1},$$

where $B(\alpha,\beta)$ is the beta function:

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt.$$

Finally, setting $r/\gamma = 1$, the total probability density $p(x) = p_0(x) + p_1(x)$ is given by [318]

$$p(x) = \frac{x^{k_+/\gamma - 1}(1-x)^{k_-/\gamma - 1}}{B(k_+/\gamma, k_-/\gamma)}. \tag{6.3.6}$$

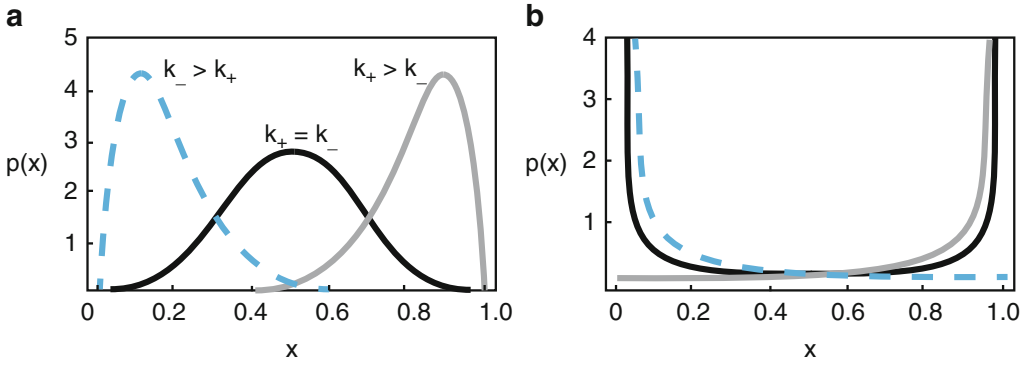

Fig. 6.4: Steady-state protein density $p(x)$ for a simple regulated network in which the promoter transitions between an active and inactive state at rates $k_\pm$. (**a**) Case $k_\pm/k > 1$: there is a graded density that is biased towards $x = 0, 1$ depending on the ratio k_+/k_-. (**b**) Case $k_\pm/k < 1$: there is a binary density that is concentrated around $x = 0, 1$ depending on the ratio k_+/k_-

In Fig. 6.4, we plot $p(x)$, $0 < x < 1$ for various values of $K_\pm = k_\pm/\gamma$. It can be seen that when the rates $k_\pm$ of switching between the active and inactive gene states are faster than the rate of degradation k, then the steady-state density is unimodal (graded), whereas if the rate of degradation is faster, then the density tends to be concentrated around $x = 0$ or $x = 1$, consistent with a binary process. In other words, if switching between promoter states is much slower than other processes, then one can have transcriptional contribution to protein bursting [318]. This scenario tends to occur in eukaryotic gene expression, for which the presence of nucleosomes and the packing of DNA–nucleosome complexes into chromatin generally make promoters inaccessible to the transcriptional machinery. Hence, transitions between open and closed chromatin structures, corresponding to active and repressed promoter states, can be quite slow.

Finally, note that a model identical in form to the above has also been applied to gene expression dynamics in a randomly varying environment [599]. In the latter case, x represents the concentration of mRNA and γ is the rate of degradation. The rate k of mRNA production takes on two values, depending on a binary-valued environmental input $n(t)$, with $k = k_0$ if $n(t) = 0$ and $k = k_1$ if $n(t) = 1$. The environment randomly switches between its two states at the rates $k_\pm$. Equations (6.3.2) thus become

$$\frac{\partial p_0}{\partial t} = -\frac{\partial}{\partial x}([k_0 - \gamma x]p_0(x,t)) + k_- p_1(x,t) - k_+ p_0(x,t) \tag{6.3.7a}$$

$$\frac{\partial p_1}{\partial t} = -\frac{\partial}{\partial x}([k_1 - \gamma x]p_1(x,t)) + k_+ p_0(x,t) - k_- p_1(x,t), \tag{6.3.7b}$$

where $p_j(x,t)$ is the probability density for mRNA concentration x given the environmental input is $n(t) = j, j = 0, 1$. The analysis of the steady-state density proceeds as before and one finds [599]

$$p_0(x) = C(\gamma x - k_0)^{-1+k_+/\gamma}(k_1 - \gamma x)^{k_-/\gamma}, \quad p_1(x) = C(\gamma x - k_0)^{k_+/\gamma}(k_1 - \gamma x)^{-1+k_-/\gamma} \tag{6.3.8}$$

for some constant C. Imposing the normalization conditions, then determines C as

$$C = \gamma\left[(k_1 - k_0)^{(k_+ + k_-)/\gamma} B(k_+/\gamma, k_-/\gamma)\right]^{-1}.$$

It follows from the analog of Fig. 6.4 that if the mRNA degradation rate is faster than the rate of environmental fluctuations, then the steady-state density of mRNA tracks the state of the environment with $p(x)$ localized around $x = 0$ ($x = 1$) when $k_- > k_+$ ($k_+ > k_-$). Stochastic switching has been suggested as a survival strategy used by populations of yeast cells in fluctuating environments [1].

6.3.2 Protein Fluctuations and the Linear Noise Approximation

In the above analysis, we considered the distribution of proteins arising from a single gene, in which the only source of noise came from the random switching of the promoter. We now want to estimate the size of protein fluctuations in a population of $n_{\max}$ genes that takes into account intrinsic noise effects due to a finite number of proteins. Let n_1 denote the number of active genes and n_2 the number of proteins. Setting $x_j = \langle n_j \rangle / \Omega$, where Ω is the system size, the various reactions and the corresponding rate equations based on mass action (valid in the limit $\Omega \to \infty$) are as follows:

1. *Gene activation and inactivation*

$$n_1 \xrightarrow{k_+(n_{\max} - n_1)} n_1 + 1, \quad n_1 \xrightarrow{k_- n_1} n_1 - 1$$

with

$$\frac{dx_1}{dt} = k_+(x_{\max} - x_1) - k_- x_1.$$

2. *Protein production and degradation*

$$n_2 \xrightarrow{r n_1} n_2 + 1, \quad n_2 \xrightarrow{\gamma n_2} n_2 - 1$$

with

$$\frac{dx_2}{dt} = rx_1 - \gamma x_2.$$

In order to take into account the effects of intrinsic noise, it is necessary to turn to the associated master equation. Let $P = P(n_1, n_2, t)$ denote the probability that there are n_1 active genes and n_2 proteins at time t. Then

$$\begin{aligned}\frac{dP}{dt} &= k_+(n_{\max} - n_1 + 1)P(n_1 - 1, n_2, t) + k_-(n_1 + 1)P(n_1 + 1, n_2, t) \\ &\quad + rn_1 P(n_1, n_2 - 1, t) + \gamma(n_2 + 1)P(n_1, n_2 + 1, t) \\ &\quad - [k_+(n_{\max} - n_1) + k_- n_1 + rn_1 + \gamma n_2] P(n_1, n_2, t). \end{aligned} \tag{6.3.9}$$

Since the transition rates are linear in n_1 and n_2, one could determine the means and variances by taking moments. However, this method is not applicable to master equations with nonlinear transition rates. Therefore, we will follow the approximation method introduced in Sect. 3.2, whereby the master equation is reduced to a Fokker–Planck (FP) equation by carrying out a system-size expansion. The resulting FP equation can then be linearized about a stable fixed point of the deterministic rate equations, resulting in a multivariate OU process that can be used to calculate means and variances [162, 164, 625]. One of the useful features of the *linear-noise approximation* is that it can be applied systematically, once the mass-action kinetic equations are expressed in the general form (6.3.17) as described in Box 6A. For the given regulatory network, there are two chemical species ($N = 2$) and four single-step reactions ($R = 4$). For $a = 1, 2$ (gene activation and inactivation), we have $S_{i,1} = \delta_{i,1}, S_{i,2} = -\delta_{i,1}, f_1(\mathbf{x}) = k_+(x_{\max} - x_1)$, and $f_2(\mathbf{x}) = k_- x_1$. Expressing the master equation as (6.3.18) and carrying out a diffusion approximation then leads to the FP equation (6.3.20) with drift terms

$$V_1(\mathbf{x}) = k_+(x_{\max} - x_1) - k_- x_1, \quad V_2(\mathbf{x}) = rx_1 - \gamma x_2$$

and a diagonal diffusion matrix $\mathbf{D}$ with nonzero components

$$D_{11} = k_+(x_{\max} - x_1) + k_- x_1, \quad D_{22} = rx_1 + \gamma x_2.$$

In the deterministic limit, we recover the kinetic equations expressed as

$$\frac{dx_i}{dt} = V_i(\mathbf{x}).$$

It immediately follows that there is a unique fixed point given by

$$x_1^* = \frac{k_+}{k_+ + k_-} x_{\max}, \quad x_2^* = \frac{r}{\gamma} x_1^*.$$

Linearizing the corresponding Langevin equation about this fixed point by setting $X_i(t) = x_i^* + \Omega^{-1/2} Y_i(t)$ then yields the OU process (6.3.25) for Y_i, which takes the explicit form

$$dY_1 = -(k_+ + k_-)Y_1 dt + dW_1, \quad dY_2 = [rY_1 - \gamma Y_2]dt + dW_2, \tag{6.3.10}$$

with $W_1(t)$ and $W_2(t)$ independent Wiener processes satisfying

$$\langle dW_1(t) dW_1(t') \rangle = [k_+(x_{\max} - x_1^*) + k_- x_1^*]\delta(t-t')dtdt' = 2k_- x_1^* \delta(t-t')dtdt',$$
$$\langle dW_2(t) dW_2(t') \rangle = [r x_1^* + \gamma x_2^*]\delta(t-t')dtdt' = 2r x_1^* \delta(t-t')dtdt'.$$

Introducing the stationary covariance matrix

$$\Sigma_{ij} = \langle [Y_i(t) - \langle Y_i(t) \rangle][Y_j(t) - \langle Y_j(t) \rangle] \rangle$$

one sees that $Y_i(t)$ is a Gaussian process with zero mean and covariances determined from the matrix equation

$$\mathbf{A}\Sigma + \Sigma \mathbf{A}^T = -\mathbf{D}, \tag{6.3.11}$$

with

$$\mathbf{A} = \begin{pmatrix} -(k_+ + k_-) & 0 \\ r & -\gamma \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} 2k_- x_1^* & 0 \\ 0 & 2r x_1^* \end{pmatrix}.$$

Finally, solving the matrix equation (6.3.11) for the covariance gives the Fano factors (see Ex. 6.4):

$$\frac{\mathrm{var}[n_1]}{\langle n_1 \rangle} = \frac{k_-}{k_+ + k_-} = 1 - \langle n_1 \rangle / n_{\max}, \tag{6.3.12a}$$

$$\frac{\mathrm{var}[n_2]}{\langle n_2 \rangle} = 1 + \langle n_2 \rangle \frac{\gamma}{k_+ + k_- + \gamma} \frac{\mathrm{var}[n_1]}{\langle n_1 \rangle^2}. \tag{6.3.12b}$$

Note that $\langle n_j \rangle = \Omega x_j^*$ and $\mathrm{var}[n_j] = \Omega \Sigma_{jj}$. We immediately see that the presence of a transcription factor increases the Fano factor of the protein above one.

An alternative approach to analyzing the multivariate OU process derived from the linear noise approximation is to Fourier transform the corresponding multivariate Langevin equation (6.3.10) and calculate the spectrum of the protein concentration [328, 624]. First, since we are ultimately interested in protein number fluctuations, we rescale the Langevin equation by setting $\Delta n_j = \sqrt{\Omega} Y_j = n_j - \langle n_j \rangle$ and use the white noise formulation (see Sect. 2.2.5):

$$\frac{d\Delta n_1}{dt} = -(k_+ + k_-)\Delta n_1 + \eta_1, \quad \frac{d\Delta n_2}{dt} = r\Delta n_1 - \gamma \Delta n_2 + \eta_2, \tag{6.3.13}$$

with $\eta_1(t)$ and $\eta_2(t)$ independent Gaussian white noise processes satisfying

$$\langle \eta_1(t)\eta_1(t')\rangle = [k_+(n_{\max} - \langle n_1\rangle) + k_-\langle n_1\rangle]\delta(t-t') = 2k_-\langle n_1\rangle\delta(t-t'),$$
$$\langle \eta_2(t)\eta_2(t')\rangle = [r\langle n_1\rangle + \gamma\langle n_2\rangle]\delta(t-t') = 2r\langle n_1\rangle\delta(t-t').$$

Fourier transforming the linear equations (6.3.13) with

$$\Delta n_j(t) = \int_{-\infty}^{\infty} \widetilde{\Delta n}_j(\omega) \mathrm{e}^{-i\omega t}\frac{d\omega}{2\pi}$$

yields

$$-i\omega\widetilde{\Delta n}_1 = -(k_+ + k_-)\widetilde{\Delta n}_1 + \widetilde{\eta_1}, \quad -i\omega\widetilde{\Delta n}_2 = r\widetilde{\Delta n}_1 - \gamma\widetilde{\Delta n}_2 + \tilde{\eta}_2. \tag{6.3.14}$$

It follows that

$$\begin{aligned}\widetilde{\Delta n}_2 &= \frac{r\widetilde{\Delta n}_1}{\gamma - i\omega} + \frac{\tilde{\eta}_2}{\gamma - i\omega} \\ &= \frac{r\tilde{\eta}_1}{(k_+ + k_- - i\omega)(\gamma - i\omega)} + \frac{\tilde{\eta}_2}{\gamma - i\omega}.\end{aligned}$$

From the spectral analysis of Sect. 2.2.5, we have

$$\langle \tilde{\eta}_1(\omega)\tilde{\eta}_1(\omega')\rangle = 2k_-\langle n_1\rangle \cdot 2\pi\delta(\omega+\omega'), \quad \langle \tilde{\eta}_2(\omega)\tilde{\eta}_2(\omega')\rangle = 2r\langle n_1\rangle \cdot 2\pi\delta(\omega+\omega').$$

Hence, the spectrum of the protein fluctuations, defined by $\langle \widetilde{\Delta n}_2(\omega)\widetilde{\Delta n}_2(\omega')\rangle = S_2(\omega)\delta(\omega+\omega')$, is

$$S_2(\omega) = \frac{r^2(2k_-\langle n_1\rangle)}{(\omega^2 + (k_+ + k_-)^2)(\omega^2+\gamma^2)} + \frac{2r\langle n_1\rangle}{\omega^2+\gamma^2}. \tag{6.3.15}$$

It follows that

$$\begin{aligned}\mathrm{var}[n_2] = \langle(\Delta n_2)^2\rangle &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\langle \widetilde{\Delta n}_2(\omega)\widetilde{\Delta n}_2(\omega')\rangle \mathrm{e}^{-i\omega t}\mathrm{e}^{-i\omega' t}\frac{d\omega}{2\pi}\frac{d\omega'}{2\pi} \\ &= \int_{-\infty}^{\infty} S_2(\omega)\frac{d\omega}{2\pi}.\end{aligned}$$

The integral can be evaluated using partial fractions and the identity

$$\int_{-\infty}^{\infty}\frac{d\omega}{\omega^2+a^2} = \frac{\pi}{a},$$

which gives

$$\begin{aligned}\text{var}[n_2] &= \frac{r\langle n_1\rangle}{\gamma}+\frac{r^2k_-\langle n_1\rangle}{(k_++k_-)^2-\gamma^2}\left(\frac{1}{\gamma}-\frac{1}{k_++k_-}\right)\\ &= \frac{r\langle n_1\rangle}{\gamma}+\frac{r^2\langle n_1\rangle/\gamma}{k_++k_-+\gamma}\frac{k_-}{(k_++k_-)}\\ &= \langle n_2\rangle+(r/\gamma)^2\frac{\gamma}{k_++k_-+\gamma}\langle n_1\rangle(1-\langle n_1\rangle/n_{\max}).\end{aligned} \tag{6.3.16}$$

This agrees with Eq. (6.3.12b). Finally, note that there are two contributions to the size of protein fluctuations. First, there is the output noise $\langle n_2\rangle$ arising from the production of a finite number of proteins in which the variance equals the mean, reflecting a pure Poisson process. The second contribution arises from the random switching of the promoter and is proportional to the binomial variance $p_1(1-p_1)$ where $p_1=\langle n_1\rangle/n_{\max}$ is the mean fraction of active genes. For further applications of frequency domain analysis to feedforward gene networks see Exs. 6.5 and 6.6.

Box 6A. Linear noise approximation.

Suppose that the mass-action kinetics of a general biochemical or gene network is written in the form

$$\frac{dx_i}{dt}=\sum_{a=1}^{R}S_{ia}f_a(\mathbf{x}),\quad i=1,\ldots,N \tag{6.3.17}$$

where a labels a single-step reaction and $\mathbf{S}$ is the so-called $N\times R$ stochiometric matrix for N molecular species and R reactions. Thus S_{ia} specifies the change in the number of molecules of species i in a given reaction a. The functions f_a are known transition intensities or *propensities*. Given this notation, the corresponding master equation is

$$\frac{dP(\mathbf{n},t)}{dt}=\Omega\sum_{a=1}^{R}\left(\prod_{i=1}^{N}\mathbb{E}^{-S_{ia}}-1\right)f_a(\mathbf{n}/\Omega)P(\mathbf{n},t), \tag{6.3.18}$$

where Ω represents the system size. Typically, Ω is the volume of the well-mixed compartment where reactions occur or the total number of molecules in cases where there is number conservation. Here $\mathbb{E}^{-S_{ia}}$ is a step or ladder operator such that for any function $g(\mathbf{n})$,

$$\mathbb{E}^{-S_{ia}}g(n_1,\ldots,n_i,\ldots,n_N)=g(n_1,\ldots,n_i-S_{ia},\ldots,n_N). \tag{6.3.19}$$

A diffusion approximation of the master equation can now be obtained along identical lines to Sect. 3.2 (see also [162]). That is, set $f_a(\mathbf{n}/\Omega)P(\mathbf{n},t) \to f_a(\mathbf{x})\mathbf{p}(\mathbf{x},t)$ and use the fact that

$$\prod_{i=1}^{N}\mathbb{E}^{-S_{ia}}h(\mathbf{x}) = h(\mathbf{x}-\mathbf{S}_a/\Omega)$$
$$= h(\mathbf{x}) - \Omega^{-1}\sum_{i=1}^{N}S_{ia}\frac{\partial h}{\partial x_i} + \frac{1}{\Omega^2}\sum_{i,j=1}^{N}S_{ia}S_{ja}\frac{\partial^2 h(\mathbf{x})}{\partial x_i \partial x_j} + O(\Omega^{-3}).$$

Carrying out a Taylor expansion of the master equation to second order thus yields the multivariate FP equation

$$\frac{\partial p}{\partial t} = -\sum_{i=1}^{N}\frac{\partial V_i(\mathbf{x})p(\mathbf{x},t)}{\partial x_i} + \frac{1}{2\Omega}\sum_{i,j=1}^{N}\frac{\partial^2 D_{ij}(\mathbf{x})p(\mathbf{x},t)}{\partial x_i \partial x_j}, \tag{6.3.20}$$

where

$$V_i(\mathbf{x}) = \sum_{a=1}^{R}S_{ia}f_a(\mathbf{x}), \quad D_{ij}(\mathbf{x}) = \sum_{a=1}^{R}S_{ia}S_{ja}f_a(\mathbf{x}). \tag{6.3.21}$$

The FP equation (6.3.20) corresponds to the multivariate Langevin equation

$$dX_i = V_i(\mathbf{X})dt + \frac{1}{\sqrt{\Omega}}\sum_{a=1}^{R}B_{ia}(\mathbf{X})dW_a(t), \tag{6.3.22}$$

where $W_a(t)$ are independent Wiener processes and $\mathbf{D} = \mathbf{B}\mathbf{B}^T$, that is,

$$B_{ia} = S_{ia}\sqrt{f_a(\mathbf{x})}. \tag{6.3.23}$$

Now suppose that the deterministic system, written as

$$\frac{dx_i}{dt} = V_i(\mathbf{x}),$$

has a unique stable fixed point $\mathbf{x}^*$ for which $V_i(\mathbf{x}^*) = 0$ and introduce the Jacobian matrix $\mathbf{A}$ with

$$A_{ij} = \left.\frac{\partial V_i}{\partial x_j}\right|_{\mathbf{x}=\mathbf{x}^*}. \tag{6.3.24}$$

The Langevin equation suggests that, after a transient phase, the stochastic dynamics is characterized by Gaussian fluctuations about the fixed point. Substituting $X_i(t) = x_i^* + Y_i(t)/\sqrt{\Omega}$ into the Langevin equation (6.3.20) and keeping only lowest order terms in $\Omega^{-1/2}$ yields the Ornstein–Uhlenbeck (OU) process

$$dY_i = \sum_{j=1}^{N} A_{ij} Y_j dt + \sum_{a=1}^{R} B_{ia}(\mathbf{x}^*) dW_a(t). \tag{6.3.25}$$

Introducing the stationary covariance matrix

$$\Sigma_{ij} = \langle [Y_i(t) - \langle Y_i(t) \rangle][Y_j(t) - \langle Y_j(t) \rangle] \rangle$$

it immediately follows from the analysis of the multivariate OU process (see Ex. 2.7), that

$$\mathbf{A}\Sigma + \Sigma \mathbf{A}^T = -\mathbf{B}\mathbf{B}^T. \tag{6.3.26}$$

6.3.3 Autoregulatory Network

So far we have considered a simple feedforward regulatory network. However, much of the complexity in gene networks arises from feedback, in which proteins influence their own synthesis directly or indirectly by acting as transcription factors within a regulatory network. A common example is autoregulation, in which a gene is directly regulated by its own gene product [625] (see Fig. 6.5a). A simple kinetic model of negative autoregulatory feedback is

$$\frac{dx_1}{dt} = -\gamma x_1 + F(x_2), \quad \frac{dx_2}{dt} = r x_1 - \gamma_p x_2, \tag{6.3.27}$$

where $x_1(t)$ and $x_2(t)$ denote the concentrations (or number) of mRNA and protein molecules at time t. The parameters γ, γ_p represent the degradation rates, r represents the translation rate of proteins, and $F(y)$ represents the nonlinear feedback effect of the protein on the transcription of mRNA. A typical choice for F in the case of a repressor is the Hill function

$$F(y) = \frac{k}{1 + (y/K)^n}. \tag{6.3.28}$$

We will assume that the network acts in a regime where the Hill function is approximately linear with $F(y) = k_0 - ky$. The analysis of intrinsic noise proceeds along similar lines to regulated gene transcription.

Let $P = P(m,n,t)$ denote the probability that there are m mRNA and n proteins at time t. Then

$$\begin{aligned} \frac{dP}{dt} &= \Omega k_0 P(m-1,n,t) + [kn + \gamma(m+1)]P(m+1,n,t) \\ &\quad + rmP(m,n-1,t) + \gamma_p(n+1)P(m,n+1,t) \\ &\quad - [\Omega k_0 + (kn + \gamma m) + rm + \gamma_p n] P(m,n,t). \end{aligned} \tag{6.3.29}$$

In order to carry out a linear noise approximation, we first rewrite the kinetic equations in the general form (6.3.17) with two chemical species ($N=2$) and four single-step reactions ($R=4$). For $a=1,2$ (mRNA production and degradation/repression), we have $S_{i,1}=\delta_{i,1}, S_{i,2}=-\delta_{i,1}, f_1(\mathbf{x})=k_0$, and $f_2(\mathbf{x})=kx_2+\gamma x_1$. Expressing the master equation as (6.3.18) and carrying out a diffusion approximation then leads to the FP equation (6.3.20) with drift terms

$$V_1(\mathbf{x})=k_0-kx_2-\gamma x_1, \quad V_2(\mathbf{x})=rx_1-\gamma_p x_2$$

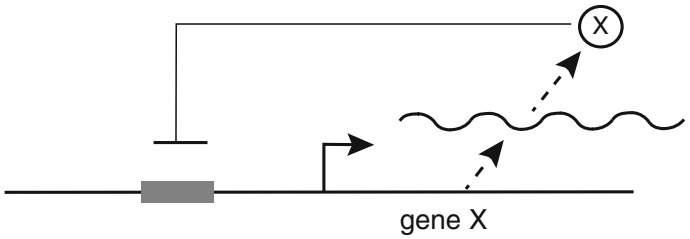

Fig. 6.5: Negative autoregulatory network. A gene X is repressed by its own protein product

and a diagonal diffusion matrix $\mathbf{D}$ with nonzero components

$$D_{11}=k_0+kx_2+\gamma x_1, \quad D_{22}=rx_1+\gamma_p x_2.$$

There is a unique fixed point of the deterministic dynamics (in the linear regime)

$$x_1^*=\frac{k_0\gamma_p}{\gamma\gamma_p+kr}, \quad x_2^*=\frac{r}{\gamma_p}x_1^*.$$

Linearizing the corresponding Langevin equation about this fixed point by setting $X_i(t)=x_i^*+\Omega^{-1/2}Y_i(t)$ then yields the OU process (6.3.25) for Y_i. Introducing the stationary covariance matrix

$$\Sigma_{ij}=\langle[Y_i(t)-\langle Y_i(t)\rangle][Y_j(t)-\langle Y_j(t)\rangle]\rangle$$

one sees that $Y_i(t)$ is a Gaussian process with zero mean and covariances determined from the matrix equation

$$\mathbf{A}\Sigma+\Sigma\mathbf{A}^T=-\mathbf{D} \tag{6.3.30}$$

with

$$\mathbf{A}=\begin{pmatrix}-\gamma & -k\\ r & -\gamma_p\end{pmatrix}, \quad \mathbf{D}=\begin{pmatrix}kx_2^*+\gamma x_1^* & 0\\ 0 & rx_1^*+\gamma_p x_2^*\end{pmatrix}.$$

Solving the matrix equation (6.3.30) yields (see Ex. 6.7)

$$\Sigma_{12} = \Sigma_{21} = \frac{\eta}{1+\eta}\left(1 - \frac{\phi}{1+b\phi}\right)x_2^*, \quad \Sigma_{22} = x_2^* + \frac{r}{\gamma_p}\Sigma_{12},$$

where

$$b = \frac{r}{\gamma}, \quad \eta = \frac{\gamma_p}{\gamma}, \quad \phi = \frac{k}{\gamma_p}.$$

Here b is the burst size, η is the ratio of degradation rates, and ϕ describes the strength of the negative feedback. It follows that the Fano factor for proteins is

$$\frac{\text{var}[n]}{\langle n\rangle} = 1 + \frac{b}{1+\eta}\left(1 - \frac{\phi}{1+b\phi}\right). \tag{6.3.31}$$

The above analysis establishes the negative feedback can reduce fluctuations in protein number (see Fig. 6.2b). That is, in the absence of feedback ($\phi = 0$), the Fano factor is $1 + b/(1+\eta)$, which is clearly larger than the case $\phi > 0$. Also note that when $\eta \ll 1$ and $b \gg 1$, we recover the result obtained from the protein translation model of Sect. 6.2.

6.4 Genetic Switches and Oscillators

Once feedback and nonlinearities are included in gene networks, a rich repertoire of dynamics can occur. Here we briefly consider two important classes of dynamical gene networks, namely, switches and oscillators.

6.4.1 Mutual Repressor Model of a Genetic Switch

Considerable insight into genetic switches has been obtained by constructing a synthetic version of a switch in *E. coli*, in which the gene product of the switch is a fluorescent reporter protein [206]. This allows the flipping of the switch to be observed by measuring the fluorescent level of the cells. The underlying gene circuit is based on a mutual repressor model (see Fig. 6.6). It consists of two repressor proteins whose transcription is mutually regulated. That is, the protein product of one gene binds to the promoter of the other gene and represses its output. For simplicity, the explicit dynamics of transcription and translation are ignored so that we only model the mutual effects of the proteins on protein production. Denoting the concentrations of the proteins by $x(t), y(t)$, the resulting kinetic equations are

$$\frac{dx}{dt} = -\gamma x + \frac{r}{1+Ky^n}, \quad \frac{dy}{dt} = -\gamma y + \frac{r}{1+Kx^n}. \tag{6.4.1}$$

Here γ is the rate of protein degradation, r is the rate of protein production in the absence of repression, and K is a binding constant for the repressors. As in the model of autoregulation, negative feedback is modeled in terms of a Hill function with Hill coefficient n. It is convenient to rewrite the equations in nondimensional form by measuring x and y in units of $K^{-1/n}$ and time in units of γ^{-1}:

$$\frac{du}{dt} = -u + \frac{\alpha}{1+v^n}, \quad \frac{dv}{dt} = -v + \frac{\alpha}{1+u^n}, \tag{6.4.2}$$

with $\alpha = rK^{1/n}/\gamma$. Analysis of the fixed point solutions of this pair of equations establishes that the mutual repressor model acts as a bistable switch. For simplicity, consider the case $n = 2$ (protein dimerization). The fixed point equation for u is

$$u = \alpha \left[1 + \left(\frac{\alpha}{1+u^2} \right)^2 \right]^{-1},$$

which can be rearranged to yield a product of two polynomials:

$$(u^2 - \alpha u + 1)(u^3 + u - \alpha) = 0.$$

The cubic is a monotonically increasing function of u and thus has a single root given implicitly by

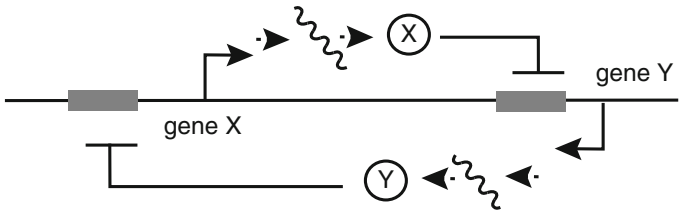

Fig. 6.6: Mutual repressor model of a genetic switch. A gene X expresses a protein X that represses the transcription of gene Y and the protein Y represses the transcription of gene X

$$u = \frac{\alpha}{1+u^2} = v.$$

This solution is guaranteed by the exchange symmetry of the underlying equations. The roots of the quadratic are given by

$$u = U_{\pm} \equiv \frac{1}{2} \left[\alpha \pm \sqrt{\alpha^2 - 4} \right],$$

with $v = U_{\mp}$. It immediately follows that there is a single fixed point when $\alpha < 2$ and three fixed points when $\alpha > 2$. Moreover, linear stability analysis establishes that the symmetric solution is stable when $\alpha < 2$ and undergoes a pitchfork bifurcation at the critical value $\alpha_c = 2$ where it becomes unstable and a pair of stable fixed points emerge.

Given that the deterministic system is bistable, one can now investigate the effects of intrinsic noise by constructing a master equation along the lines of Sect. 6.3. We will construct the master equation for a slightly simplified mutual repressor model consisting of a single promoter site; if a dimer of one protein is bound to the site then this represses the expression of the other [328, 470]. Thus the promoter can be in three states O_j, $j = 0,1,2$: no dimer is bound to the promoter (O_0); a dimer of protein X is bound to the promoter (O_1); a dimer of protein Y is bound to the promoter (O_2). Suppose that the number of proteins X and Y are n and m, respectively. The state transition diagram for the three promoter states is then

$$O_1 \underset{n(n-1)\kappa}{\overset{\beta\kappa}{\rightleftharpoons}} O_0 \underset{\beta\kappa}{\overset{m(m-1)\kappa}{\rightleftharpoons}} O_2,$$

where κ is a rate and β is a nondimensional dissociation constant. Protein X (Y) is produced at a rate α when the promoter is in the states $O_{0,1}$ ($O_{0,2}$), and both proteins are degraded at a rate γ in all three states. Let $p_j(n,m,t)$, $j = 0,1,2$, be the probability that there are n (m) proteins X (Y) and the promoter is in state j at time t. The master equation for $\mathbf{p} = (p_0, p_1, p_2)^T$ is given by

$$\frac{d}{dt}p_j(n,m,t) = \sum_{j=0,1,2}\sum_{n',m'}\left[\delta_{n,n'}\delta_{m,m'}A_{jk} + \delta_{j,k}W^j_{nm,n'm'}\right]p_k(n',m',t), \tag{6.4.3}$$

where

$$\mathbf{A} = \kappa\begin{pmatrix} -n(n-1)-m(m-1) & \beta & \beta \\ n(n-1) & -\beta & 0 \\ m(m-1) & 0 & -\beta \end{pmatrix}, \tag{6.4.4}$$

and

$$\begin{aligned}&\sum_{n',m'} W^0_{nm,n'm'}p_0(n',m',t)\\ &= \gamma[(n+1)p_0(n+1,m,t) + (m+1)p_0(n,m+1,t) - (n+m)p_0(n,m,t)]\\ &\quad + \alpha(p_0(n-1,m,t) + p_0(n,m-1,t) - 2p_0(n,m,t))\end{aligned} \tag{6.4.5a}$$

$$\begin{aligned}&\sum_{n',m'} W^1_{nm,n'm'}p_1(n',m',t)\\ &= \gamma[(n+1)p_1(n+1,m,t) + (m+1)p_1(n,m+1,t) - (n+m)p_1(n,m,t)]\\ &\quad + \alpha(p_1(n-1,m,t) - p_1(n,m,t))\end{aligned} \tag{6.4.5b}$$

$$\begin{aligned}&\sum_{n',m'} W^2_{nm,n'm'}p_1(n',m',t)\\ &= \gamma[(n+1)p_2(n+1,m,t) + (m+1)p_2(n,m+1,t) - (n+m)p_2(n,m,t)]\\ &\quad + \alpha(p_2(n,m-1,t) - p_2(n,m,t)).\end{aligned} \tag{6.4.5c}$$

Kepler and Elston [328] consider two approximations of the master equation, one based on a system-size expansion of the W^j terms with respect to the mean number $N = \alpha/\gamma$ of proteins when the promoter is in state O_0 and the other based on a QSS approximation. The latter assumes that the rates of protein production and degradation are much slower than the rates of switching between promoter states. First, introduce the rescaling $t \to t\gamma$ and set $x = n/N$, $y = m/N$. The master equation for the resulting probability densities $p_j(x,y,t)$ takes the form

$$\frac{\partial}{\partial t} p_j(x,y,t) = \sum_{j=0,1,2} \left[\frac{1}{\varepsilon} A_{jk} + N\delta_{j,k} \mathbb{W}^j \right] p_k(x,y,t), \tag{6.4.6}$$

where $\varepsilon = \gamma^3/\kappa\alpha^2$ and $b = \beta\gamma^2/\alpha^2$ are dimensionless parameters,

$$\mathbf{A} = \begin{pmatrix} -x(x-1/N) - y(y-1/N) & b & b \\ x(x-1/N) & -b & 0 \\ y(y-1/N) & 0 & -b \end{pmatrix}, \tag{6.4.7}$$

and $\mathbb{W}^j$ are differential shift operators

$$\mathbb{W}^0 = \left(e^{\partial_x/N} - 1\right) x + \left(e^{\partial_y/N} - 1\right) y + \left(e^{-\partial_x/N} + e^{-\partial_y/N} - 2\right) \tag{6.4.8a}$$

$$\mathbb{W}^1 = \left(e^{\partial_x/N} - 1\right) x + \left(e^{\partial_y/N} - 1\right) y + \left(e^{-\partial_x/N} - 1\right) \tag{6.4.8b}$$

$$\mathbb{W}^2 = \left(e^{\partial_x/N} - 1\right) x + \left(e^{\partial_y/N} - 1\right) y + \left(e^{-\partial_y/N} - 1\right). \tag{6.4.8c}$$

The latter are a way of representing a Taylor expansion. That is, for any smooth function $f(x)$,

$$\begin{aligned} f(x \pm \Delta x) &= f(x) \pm f'(x)\Delta x + f''(x)\Delta x^2/2! \pm \dots \\ &= \left(1 \pm \Delta x\, \partial_x + \frac{\Delta x^2}{2!}\partial_x^2 \pm \dots\right) f(x) = e^{\pm \Delta x \partial_x} f(x). \end{aligned}$$

If the promoter transitions are fast and the expected number of protein molecules is large, then there are two small parameters in the model, ε and $1/N$. Taking the limits $\varepsilon \to 0$ and $N \to \infty$ in either order, one obtains the kinetic equations (see also Ex. 6.8)

$$\frac{dx}{dt} = f(x,y), \quad \frac{dy}{dt} = f(y,x), \quad \text{with } f(x,y) = \frac{1}{1 + \frac{y^2}{b+x^2}} - x. \tag{6.4.9}$$

One finds that the deterministic system is bistable for $0 < b < b_c = 4/9$ (see Fig. 6.7). At the critical point $b = b_c$ there is a saddle-node bifurcation in which a stable/unstable pair annihilate so that for $b > b_c$ there is a single stable fixed point. There are then two approximations of the full master equation that can be used to explore the effects of noise-induced transitions in the bistable regime, depending on

whether one considers the system-size expansion in $1/N$ for fixed ε or the QSS expansion in ε for fixed N. For the sake of illustration, we focus on the former. Taylor expanding the differential operators $\mathbb{W}^j$ and keeping only the leading order terms yields the multivariate differential Chapman–Kolmogorov (CK) equation [328, 470]

$$\frac{\partial p_j}{\partial t} = -\frac{\partial F_j(x)p_j}{\partial x} - \frac{\partial G_j(y)p_j}{\partial y} + \frac{1}{\varepsilon}\sum_{k=0,1,2} A_{jk}(x,y)p_k \tag{6.4.10}$$

with

$$\begin{aligned} F_0(x) &= 1-x, \quad F_1(x) = 1-x, \quad F_2(x) = -x \\ G_0(y) &= 1-y, \quad G_1(y) = -y, \quad G_2(y) = 1-y, \end{aligned} \tag{6.4.11}$$

and

$$\mathbf{A} = \begin{pmatrix} -x^2-y^2 & b & b \\ x^2 & -b & 0 \\ y^2 & 0 & -b \end{pmatrix}. \tag{6.4.12}$$

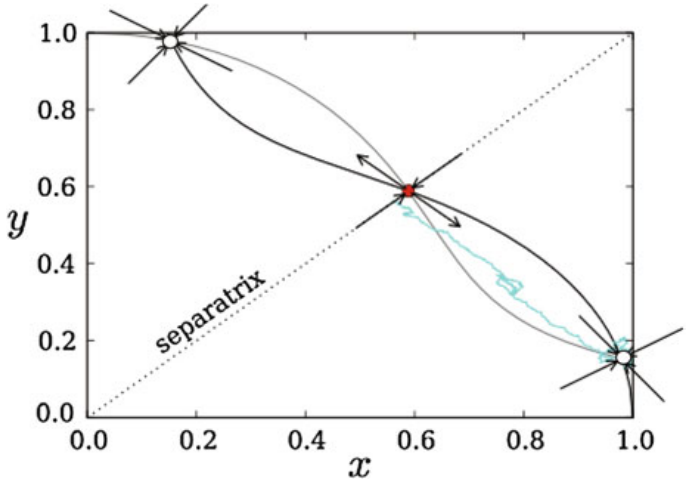

Fig. 6.7: Phase-plane dynamics of mutual repressor model analyzed by Kepler and Elston [328] and Newby [470] with $b = 0.15$. The *black curve* shows the y-nullcline and the *gray curve* shows the x-nullcline. The *open circles* show the stable fixed points; the *filled circle* shows the unstable saddle. The *irregular curve* shows a stochastic trajectory leaving the lower basin of attraction to reach the separatrix

The CK equation (6.4.10) describes an effective stochastic hybrid system in which the concentration of proteins X and Y play the role of the piecewise deterministic continuous variables, and the state of the promoter is the discrete variable that evolves according to a continuous-time Markov process. We have previously encountered stochastic hybrid systems in our analysis of voltage-gated ion channels (Sect. 3.5). One could now use a QSS approximation to obtain a Fokker–Planck

(FP) equation for the total probability density $p(x,y,t) = \sum_{j=0,1,2} p_j(x,y,t)$ along the lines outlined in Sect. 7.4 (see also Kepler and Elston [328]). However, a diffusion approximation of the full master equation based on an FP representation can generate exponential errors in the mean time of noise-induced escape from the basin of attraction of one of the metastable fixed points (see also Sects. 3.4 and 3.5). A more accurate estimate can be obtained using large deviation theory and the WKB methods outlined in Chap. 10, as has been shown for the mutual repressor model by Newby [470].

6.4.2 *The* lac *Operon*

The idea of a genetic switch was first proposed over 40 years ago by Jacob and Monod [296], in their study of the *lac* operon. When there is an abundance of glucose, *E. coli* uses glucose exclusively as a food source irrespective of whether or not other sugars are present. However, when glucose is unavailable, *E. coli* can feed on other sugars such as lactose, and this occurs via the *lac* operon switch that induces the expression of various genes. A variety of mathematical models of the *lac* operon

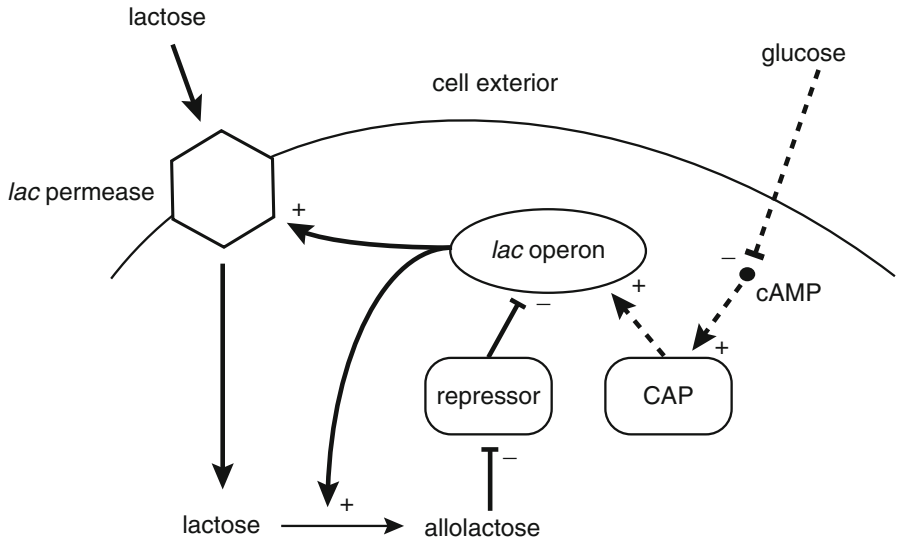

Fig. 6.8: Feedback control circuit of the *lac* operon. See text for details

have been developed over the years [239, 240, 557, 688, 693, 694]. Here we briefly describe a simplified model presented in Chap. 10 of Keener and Sneyd [322]. The basic feedback control mechanism is illustrated in Fig. 6.8. There are two control sites on the *lac* operon: (see Fig. 6.9), a repressor site that blocks RNAP from binding to the promoter site and a preceding control site to which a dimeric catabolic activator protein (CAP) molecule can bind provided it forms a complex with cyclic AMP (cAMP). Bound CAP promotes the binding of RNAP to the promoter region.

When there is sufficient glucose in the cell exterior, the action of cAMP is inhibited so that CAP cannot bind and the *lac* operon is repressed. On the other hand, when glucose is removed, the CAP–cAMP complex can bind to the activator site and activate the *lac* operon. The latter consists of several genes that code for the proteins responsible for lactose metabolism. One of these proteins is *lac* permease, which allows the entry of lactose into the cell that is enhanced by a positive feedback loop. The feedback mechanism involves another protein, β-galactosidase, which converts lactose into allolactose. Allolactose can bind to the repressor protein and prevent its binding to the repressor binding site. This further activates the *lac* operon, resulting in the further production of allolactose and increased entry of lactose via the *lac* permease.

Suppose that the CAP dynamics is ignored, so that we can focus on the positive feedback loop indicated in Fig. 6.8 by solid arrows. Let A denote the concentration of allolactose and similarly for lactose (L), the permease (P), the protein product β-galactosidase (B), mRNA (M), and the repressor (R). Let p_{on} and p_{off} denote the probabilities that the operon is on and off, respectively, with $p_{\text{on}} + p_{\text{off}} = 1$. Ignoring the effect of the CAP site, we have the simple kinetic scheme

$$\frac{dp_{\text{on}}}{dt} = k_{-r}(1 - p_{\text{on}}) - k_r R^* p_{\text{on}},$$

where R^* is the concentration of repressor in the activated state. Each activated repressor protein interacts with two molecules of allolactose to become inactivated, so from mass-action kinetics,

$$\frac{dR^*}{dt} = k_{-a}R - k_a A^2 R^*,$$

where the binding/unbinding of a single repressor molecule to the operon has a negligible effect on the total concentration $R_T = R + R^*$. The next simplification is to take these reactions to be much faster than those associated with gene expression so that p_{on} and R^* take the steady-state values

$$R^* = \frac{R_T}{1 + K_a A^2}, \quad p_{\text{on}} = \frac{1}{1 + K_r R^*},$$

with $K_a = k_a/k_{-a}$ and $K_r = k_r/k_{-r}$. Combining these two results gives the steady-state probability

$$p_{\text{on}} = \frac{1 + K_a A^2}{1 + K_r R_T + K_a A^2} \equiv \Gamma(A).$$

It follows that the concentration of mRNA is determined by the equation

$$\frac{dM}{dt} = \alpha_M \Gamma(A) - \gamma_M M, \tag{6.4.13a}$$

where α_M and γ_M are the rates of mRNA production and degradation. This is the first of the model equations. The next two equations represent the dynamics of the enzymes directly produced by the on-state of the operon, namely, permease and β-galactosidase:

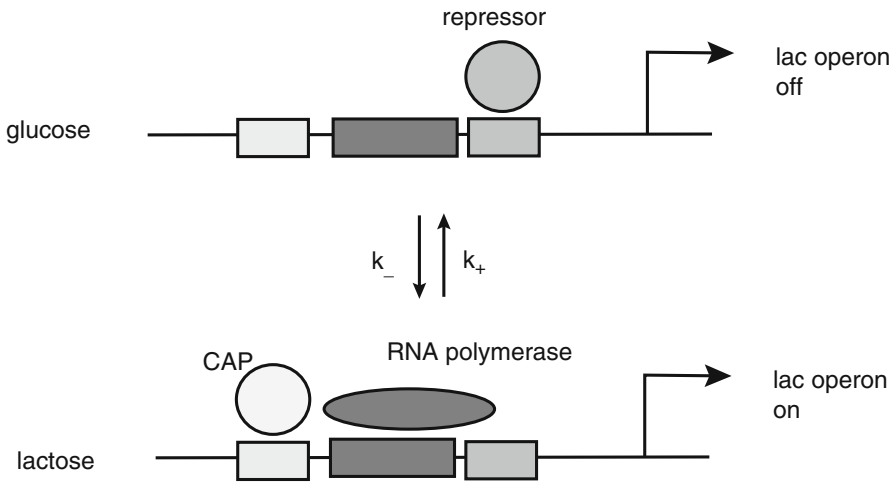

Fig. 6.9: Repressor and CAP sites for the *lac* operon

$$\frac{dP}{dt} = \alpha_P M - \gamma_P P, \tag{6.4.13b}$$

$$\frac{dB}{dt} = \alpha_B M - \gamma_B B. \tag{6.4.13c}$$

Note that although both enzymes are produced by different parts of the same mRNA, the effective production rates differ due to different times of production (permease is produced after β-galactosidase) and the time delay associated with permease migrating to the cell membrane. The final two equations specify the dynamics of lactose and allolactose based on Michaelis–Menten kinetics (see Box 6B). Let L_e be a fixed concentration of lactose exterior to the cell. Lactose enters the cell at a Michaelis–Menten rate proportional to the permease concentration P, where it is converted to allolactose via the enzymatic action of β-galactosidase; the latter also breaks down allolactose into glucose and galactose. Thus

$$\frac{dL}{dt} = \alpha_L P \frac{L_e}{K_{Le} + L_e} - \alpha_A B \frac{L}{K_L + L} - \gamma_L L \tag{6.4.13d}$$

$$\frac{dA}{dt} = \alpha_A B \frac{L}{K_L + L} - \beta_A B \frac{A}{K_A + A} - \gamma_A A. \tag{6.4.13e}$$

Keener and Sneyd [322] show that for physiologically based parameter values, the system of Eq. (6.4.13) exhibits bistability in the interior lactose concentration as a function of the exterior lactose concentration L_e. Note that the stochastic analysis

outlined in Sect. 6.4.1 for the mutual repressor model could be extended to the more complicated model of the *lac* operon in order to investigate the effects of intrinsic noise on the bistable switch.

6.4.3 Genetic Oscillator Network

There are numerous examples of gene circuits that support oscillations. Here we consider a relaxation oscillator consisting of an activator that increases its own production and that of a repressor, which in turn represses the production of the activator (see Fig. 6.10). Let x denote the concentration of the activator and y denote the concentration of the repressor. The resulting kinetic equations take the form

$$\frac{dx}{dt} = -\gamma_x x + r_{0x}\frac{1}{1+(x/K_d)^2+(y/K_d)^2} + r_x\frac{(x/K_d)^2}{1+(x/K_d)^2+(y/K_d)^2} \tag{6.4.14a}$$

$$\frac{dy}{dt} = -\gamma_y y + r_{0y}\frac{1}{1+(x/K_d)^2} + r_y\frac{(x/K_d)^2}{1+(x/K_d)^2}, \tag{6.4.14b}$$

where γ_x, γ_y are the degradation rates of the two proteins, r_{0x}, r_{0y} are protein production rates when respective promoters are not bound by transcription factors, and r_x, r_y are the enhanced production rates when the promoter sites are activated. (It is assumed that when the promoter of gene X is repressed, production of protein X is blocked.) The production terms are based on the equilibrium binding probabilities of the X and Y promoter domains. The Hill coefficient $n = 2$ arises because the transcription factors bind as dimers. In the case of the unbound promoter A of activator gene X, the binding reactions are

$$2X + A \rightleftharpoons A^+, \quad 2Y + A \rightleftharpoons A^-,$$

where $A^\pm$ denote the activated and repressed promoter states. In terms of the equilibrium law of mass action (see Sects. 1.4 and 4.1), the concentrations of the various reactants and products satisfy

$$\frac{[A^+]}{[X^2][A]} = \frac{1}{K_d}, \quad \frac{[A^-]}{[Y^2][A]} = \frac{1}{K_d},$$

with the dissociation constant K_d taken to be the same for both binding reactions. Denoting the total concentration of promoter domains of gene X by $T_A = [A] + [A^+] + [A^-]$, we have

$$\frac{[A^+]}{T_A - [A^+] - [A^-]} = \frac{[X^2]}{K_d}, \quad \frac{[A^-]}{T_A - [A^+] - [A^-]} = \frac{[Y^2]}{K_d}.$$

These can be solved to give the equilibrium probabilities that the X promoter domain is activated or repressed:

$$\frac{[A^+]}{T_A} = \frac{([X]/K_d)^2}{1+([X]/K_d)^2+([Y]/K_d)^2}, \quad \frac{[A^-]}{T_A} = \frac{([Y]/K_d)^2}{1+([X]/K_d)^2+([Y]/K_d)^2}.$$

A similar analysis of the single binding reaction

$$2X + R \rightleftharpoons R^+,$$

where R, R^+ are the unbound and activated states of the Y gene promoter, yields

$$\frac{[R^+]}{T_R} = \frac{([X]/K_d)^2}{1+([X]/K_d)^2},$$

where T_R is the total concentration of the Y gene promoter.

As in the case of the genetic switch, it is useful to nondimensionalize the equations by taking time to be in units of γ_y^{-1} and concentrations in units of K_d:

$$\frac{dx}{dt} = -\gamma x + \frac{R_{0x} + R_x x^2}{1+x^2+y^2} \tag{6.4.15a}$$

$$\frac{dy}{dt} = -y + \frac{R_{0y} + R_y x^2}{1+x^2}, \tag{6.4.15b}$$

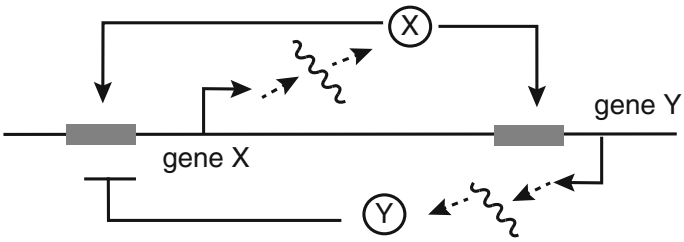

Fig. 6.10: Activator–repressor model of a genetic oscillator. A gene X expresses a protein X that activates the transcription of genes X and Y and protein Y represses the transcription of gene X

where $\gamma = \gamma_x/\gamma_y$ and $R_{0x} = r_{0x}/\gamma$, etc. If $\gamma \ll 1$, then we have a slow–fast system with the repressor acting as the slow variable. The existence of a relaxation oscillator can then be established using phase-plane analysis.

6.4.4 The Circadian Clock and Molecular Noise

The circadian rhythm plays a key physiological role in the adaptation of living organisms to the alternation of night and day [214, 492]. Experimental studies of a wide range of plants and animals has established that in almost all cases, autoregulatory feedback on gene expression plays a central role in the molecular mechanisms

underlying circadian rhythms [335, 501]. Based on experimental data, a variety of models of increasing complexity have been developed, which show how regulatory feedback loops in circadian gene networks generate sustained oscillations under conditions of continuous darkness [188, 222, 381, 382, 602, 648]. The resulting circadian oscillator has a natural period of approximately 24 h, which can be entrained to the external light–dark cycle. Given that the circadian rhythm is controlled by gene networks, this immediately raises the issue regarding the extent to which such oscillations are robust to intrinsic noise arising from small numbers of molecules [21, 187, 189, 227]. Here we review the analysis of Gonze et al. [228], who considered the effects of molecular noise on a minimal model of the circadian clock in the fungus *Neurospora* [382]. A schematic diagram of the basic model is shown in Fig. 6.11. A clock gene X (*frq* in Neurospora, *per* in Drosophila) is transcribed to form mRNA (M), which exits the nucleus and is subsequently translated into cytoplasmic clock protein (X_C). The resulting protein either degrades or enters the nucleus (X_N) where it inhibits its own gene expression.

The governing equations for the concentrations m, x_C, x_N of mRNA, cytosolic protein, and nuclear protein, respectively, are

$$\frac{dm}{dt} = k\frac{K_m^n}{K_m^n + x_N^n} - \gamma\frac{m}{K_m' + m} \tag{6.4.16a}$$

$$\frac{dx_C}{dt} = rm - \gamma_P\frac{x_C}{K_p + x_C} - k_1 x_C + k_2 x_N \tag{6.4.16b}$$

$$\frac{dx_N}{dt} = k_1 x_C - k_2 x_N. \tag{6.4.16c}$$

Here k is the unregulated rate of transcription, r is the rate of translation, and γ, γ_p are the rates of mRNA and protein degradation; degradation is assumed to obey Michaelis–Menten kinetics. The negative regulation of transcription is taken to be cooperative with a Hill coefficient of n. Finally, the rate constants k_1, k_2 characterize the transport of protein into and out of the nucleus. It can be shown that the above model exhibits limit cycle oscillations in physiologically reasonable parameter regimes and thus provides a molecular basis for the sustained oscillations of the circadian clock under constant darkness [382]. In order to explore the robustness of such oscillations to molecular noise, it is necessary to turn to a master equation formulation of the gene network. One can then approximate the master equation by an FP equation as outlined in Box 6A, but now one has to linearize the FP equation about a limit cycle rather than a fixed point.

As in previous examples of gene regulation, it is convenient to rewrite this system of equations in the form of Eq. (6.3.17), which involves a sum over $R = 6$ single-step reactions labeled by a, whose transition rates $f_a(\mathbf{x})$ and stochiometric coefficients S_{ia} are listed in the table below with $\mathbf{x} = (m, x_C, x_N)^T$ [228]. Given this decomposition, one can now write down the FP equation obtained under the diffusion approximation [see also Eq. (6.3.20)]:

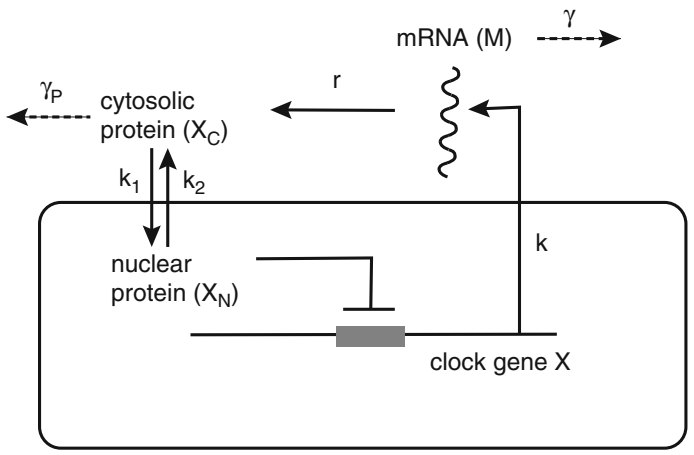

Fig. 6.11: Minimal model for a negative autoregulation network underlying circadian rhythms. Transcription of a clock gene (X) produces mRNA (M), which is transported outside the nucleus and then translated into cytosolic clock protein (X_C). The protein is either degraded or transported into the nucleus (X_N) where it exerts negative feedback on the gene expression

$$\frac{\partial p}{\partial t} = -\sum_{i=1}^{3} \frac{\partial V_i(\mathbf{x}) p(\mathbf{x},t)}{\partial x_i} + \frac{1}{2\Omega} \sum_{i,j=1}^{3} \frac{\partial^2 D_{ij}(\mathbf{x}) p(\mathbf{x},t)}{\partial x_i \partial x_j}, \tag{6.4.17}$$

where Ω is the total number of molecules that can be present in the system, say,

$$V_i(\mathbf{x}) = \sum_{a=1}^{R} S_{ia} f_a(\mathbf{x}), \quad D_{ij}(\mathbf{x}) = \sum_{a=1}^{R} S_{ia} S_{ja} f_a(\mathbf{x}). \tag{6.4.18}$$

From Table 6.1, we deduce that

$$V_1(\mathbf{x}) = f_1(\mathbf{x}) - f_2(\mathbf{x}), V_2(\mathbf{x}) = f_3(\mathbf{x}) - f_4(\mathbf{x}) - f_5(\mathbf{x}) + f_6(\mathbf{x}), V_3(\mathbf{x}) = f_5(\mathbf{x}) - f_6(\mathbf{x}),$$

and

$$\begin{aligned}
D_{11}(\mathbf{x}) &= \frac{1}{2}(f_1(\mathbf{x}) + f_2(\mathbf{x}), \quad D_{12} = D_{21} = D_{13} = D_{31} = 0, \\
D_{22}(\mathbf{x}) &= \frac{1}{2}(f_3(\mathbf{x}) + f_4(\mathbf{x}) + f_5(\mathbf{x}) + f_6(\mathbf{x})) \\
D_{23}(\mathbf{x}) &= D_{32}(\mathbf{x}) = -\frac{1}{2}(f_5(\mathbf{x}) + f_6(\mathbf{x})), \quad D_{33}(\mathbf{x}) = \frac{1}{2}(f_5(\mathbf{x}) + f_6(\mathbf{x})).
\end{aligned}$$

In the deterministic limit $\Omega \to \infty$, we recover the deterministic model (6.4.16), which can be rewritten in the more compact form

$$\frac{dx_i}{dt} = V_i(\mathbf{x}), \quad i = 1,2,3. \tag{6.4.19}$$

Reaction	Transition rate	Transition
$X \to X + M$	$f_1 = k\frac{K_m^n}{K_m^n + x_N^n}$	$M \to M + 1$
$M \to \emptyset$	$f_2 = \gamma\frac{m}{K_m' + m}$	$M \to M - 1$
$M \to X_C + M$	$f_3 = rm$	$X_C \to X_C + 1$
$X_C \to \emptyset$	$f_4 = \gamma_P\frac{x_C}{K_p + x_C}$	$X_C \to X_C - 1$
$X_C \to X_N$	$f_5 = k_1 x_C$	$X_C \to X_C - 1,\ X_N \to X_N + 1$
$X_N \to X_C$	$f_6 = k_2 x_N$	$X_C \to X_C + 1,\ X_N \to X_N - 1$

Table 6.1: Single-step reactions of the minimal circadian clock gene network

One way to investigate the effects of molecular noise on the circadian clock is to linearize the FP equation about the limit cycle solution, analogous to the linear noise approximation for Gaussian-like fluctuations about fixed points (see Box 6A). However, the linear noise approximation requires that perturbations remain small for all times, which is not the case for limit cycles, since they are marginally stable with respect to phase shifts around the limit cycle. Therefore, one needs to separate out the effects of longitudinal and transverse fluctuations of the limit cycle [56, 578]. The basic intuition is that Gaussian-like transverse fluctuations are distributed in a tube of radius $1/\sqrt{\Omega}$, whereas the phase around the limit cycle undergoes Brownian diffusion. Thus, consider the Langevin equation corresponding to the FP equation (6.4.17):

$$dX_i(t) = V_i(\mathbf{X}(t))dt + \frac{1}{\sqrt{\Omega}}\sum_{j=1}^{n} D_{ij}(\mathbf{X}(t))dW_j(t), \tag{6.4.20}$$

where n is the number of chemical species ($n = 3$ in the case of the circadian clock) and $W_j(t)$ are independent Wiener processes. We can then decompose the stochastic vector $\mathbf{X}(t)$ according to

$$\mathbf{X}(t) = \mathbf{x}^*(t + S(t)) + \mathbf{T}(t), \tag{6.4.21}$$

where the scalar random variable $S(t)$ represents the undamped random phase shift along the limit cycle and $\mathbf{T}(t)$ is a transversal perturbation (see Fig. 6.12). Since there is no damping of fluctuations along the limit cycle, the random phase $S(t)$ is taken to undergo Brownian motion. The associated phase diffusion coefficient D_θ is an effective time constant that characterizes the robustness of the oscillator to intrinsic noise. However, it is important to note that the decomposition (6.4.21) is not unique, so that the precise definition of the phase depends on the particular method of analysis.

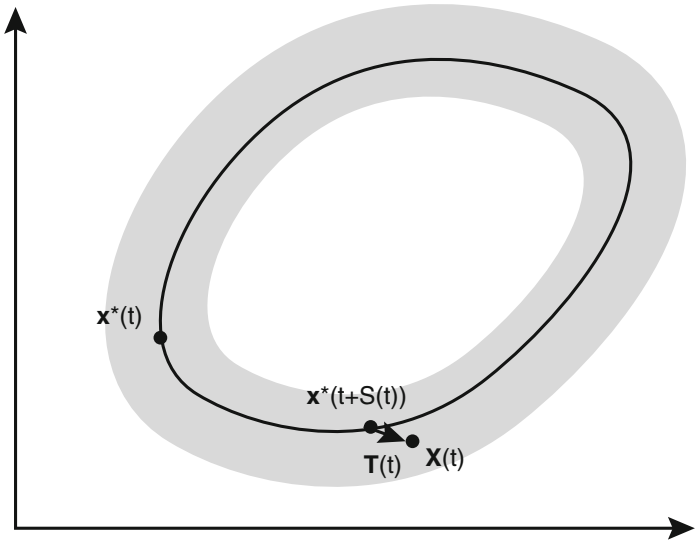

Fig. 6.12: Decomposition of a stochastic limit cycle $\mathbf{X}(t)$ into a random phase shift $S(t)$ along the deterministic limit cycle $\mathbf{x}^*(t)$ and a random transversal component $\mathbf{T}(t)$

For example, one recent study defines the phase in order to ensure that the mean size of transversal fluctuations remains small [343]. On the other hand, Gonze et al. [228] estimate D_θ for their circadian clock model using an alternative approach based on a WKB approximation of solutions to the FP equation (see also [210, 650]). In particular, they find $D_\theta \sim 1/\Omega$ so that larger systems are more robust to noise as one would expect. Gonze et al. also determine the rate of decay of correlations. Let $\mathbf{x}_0$ be a point on the deterministic limit cycle and suppose $\mathbf{X}(0) = \mathbf{x}_0$. Define the rth return time τ_r of a trajectory to be when an arbitrarily chosen component $X_j(t)$ returns to $x_{j,0}$ for the rth time, $X_j(\tau_r) = x_{j,0}$. In the deterministic case, $\tau_r = rT$ where T is the minimal period of the limit cycle oscillation. For sufficiently large Ω (small noise), we expect the distribution of first return times $\tau_1 = \tau$ to be approximately given by the Gaussian

$$P(\tau) \sim \frac{1}{\sqrt{2\pi D_\theta T}} \exp\left[-\frac{(\tau - T)^2}{2D_\theta T}\right].$$

It can also be shown that the autocorrelation function for each concentration takes the form of damped oscillations [210, 228],

$$\begin{aligned} C_j(t) &= \langle X_j(t) X_j(0) \rangle \\ &\approx C_{j,0} + C_{j,1} \mathrm{e}^{-t/\gamma} \cos(\omega t + \alpha_j), \quad t \to \infty, \end{aligned} \tag{6.4.22}$$

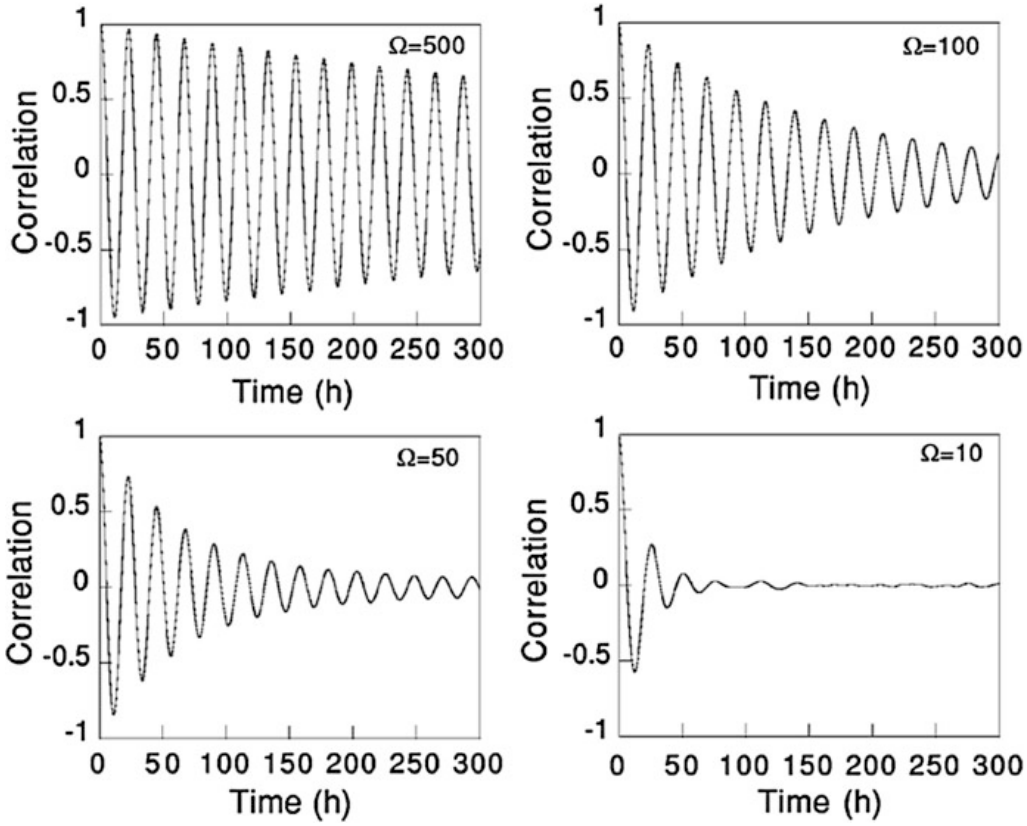

Fig. 6.13: Illustration of the time evolution of one of the autocorrelation functions of the stochastic circadian clock model considered by Gonze et al. [228]. As the system size Ω is decreased, the rate of decay of correlations becomes more rapid. Parameter values can be found in [228]

for some coefficients $C_{j,0}, C_{j,1}$ and phases α_j, with

$$\omega \approx \frac{2\pi}{T}, \quad \gamma \approx \frac{T^2}{2D_\theta \pi^2}. \tag{6.4.23}$$

It follows that the rate of decay γ^{-1} of correlations is inversely proportional to the system size (see Fig. 6.13).

6.4.5 *Quasi-Cycles in a Biochemical Oscillator*

In Sect. 6.4.4, we discussed the effects of intrinsic noise on a biochemical limit cycle oscillator that exists in the absence of noise. One also finds that stochastic biochemical and gene networks can exhibit noise-induced oscillations (quasi-cycles) in parameter regimes for which the underlying deterministic kinetic equations have only fixed point solutions. These quasi-cycles are characterized by a peak in the

power spectrum obtained using a linear noise approximation of the chemical master equation. Following Boland et al. [55, 56], we will illustrate this using a stochastic version of the Brusselator. (A spatially extended version of the model will be considered in Sect. 9.3.) The Brusselator is an idealized model of an autocatalytic reaction, in which at least one of the reactants is also a product of the reaction [220]. The model consists of two chemical species X and Y interacting through the following reaction scheme:

$$\begin{aligned} \emptyset &\xrightarrow{a} X \\ X &\xrightarrow{b} Y \\ 2X + Y &\xrightarrow{c} 3X \\ X &\xrightarrow{d} \emptyset \end{aligned}$$

These reactions describe the production and degradation of an X molecule, an X molecule spontaneously transforming into a Y molecule, and two molecules of X reacting with a single molecule of Y to produce three molecules of X. The corresponding mass-action kinetic equations for $u = [X], v = [Y]$ are (after rescaling so that $c = d = 1$)

$$\frac{du}{dt} = a - (b+1)u + u^2 v, \tag{6.4.24a}$$

$$\frac{dv}{dt} = bu - u^2 v. \tag{6.4.24b}$$

The system has a fixed point at $u^* = a, v^* = b/a$, which is stable when $b < a^2 + 1$ and unstable when $b > a^2 + 1$ (see below). Moreover, the fixed point undergoes a Hopf bifurcation at the critical value $b = a^2 + 1$ for fixed a, leading to the formation of a stable limit cycle (see Box 4B).

Following the examples of Sect. 6.3, it is straightforward to write down a stochastic version of the Brusselator model. Let $n_1(t)$ and $n_2(t)$ denote the number of X and Y molecules at time t, respectively, and take Ω to be cell volume. The various state transitions are

$$(n_1, n_2) \xrightarrow{T_1} (n_1 + 1, n_2), \tag{6.4.25a}$$

$$(n_1, n_2) \xrightarrow{T_2} (n_1 - 1, n_2 + 1) \tag{6.4.25b}$$

$$(n_1, n_2) \xrightarrow{T_3} (n_1 + 1, n_2 - 1) \tag{6.4.25c}$$

$$(n_1, n_2) \xrightarrow{T_4} (n_1 - 1, n_2), \tag{6.4.25d}$$

with $\mathbf{n} = (n_1, n_2)$ and

$$T_1 = \Omega a, \quad T_2 = b n_1, \quad T_3 = n_1^2 n_2 / \Omega^2, \quad T_4 = n_1. \tag{6.4.26}$$

It is convenient to rewrite the kinetic equations (6.4.24) in the generic form (6.3.17):

$$\frac{du_i}{dt} = \sum_{r=1}^{p} S_{ir} f_r(u_1, u_2), \quad i = 1, 2 \tag{6.4.27}$$

where $u_j = n_j/\Omega$, r labels the single-step reaction, p is the number of single-step reactions, and $\mathbf{S}$ is the stochiometric matrix. In the case of the Brusselator model, Eq. (6.4.24) shows that $p = 4$,

$$S_{11} = 1, S_{21} = 0, \quad S_{12} = -1, S_{22} = 1, \quad S_{13} = 1, S_{23} = -1, \quad S_{14} = -1, S_{24} = 0,$$

and $\Omega f_r(n_1/\Omega, n_2/\Omega) = T_r(n_1, n_2)$. The corresponding master equation is then given by [see Eq. (6.3.18)],

$$\frac{dP(n_1, n_2, t)}{dt} = \Omega \sum_{r=1}^{p} \left(\prod_{i=1,2} \mathbb{E}^{-S_{ir}} - 1 \right) f_r(n_1/\Omega, n_2/\Omega) P(n_1, n_2, t). \tag{6.4.28}$$

Now suppose that Ω is sufficiently large so that we can carry out a linear noise approximation and obtain a Langevin equation for a multivariate OU process (see Box 6A). That is, we approximate the master equation (6.4.28) by an FP equation and then linearize about the fixed point (u^*, v^*) by setting

$$\frac{n_j}{\Omega} = u_j = u_j^* + \frac{1}{\sqrt{\Omega}} v_j.$$

This yields the Langevin equation

$$\frac{dv_j(t)}{dt} = \sum_{j'} A_{jj'} v_{j'}(t) + \eta_j(t), \tag{6.4.29}$$

with white noise terms satisfying

$$\langle \eta_j(t) \rangle = 0, \quad \langle \eta_j(t) \eta_{j'}(t') \rangle = D_{jj'} \delta(t - t').$$

Here

$$A_{jj'} = \sum_{r=1}^{p} S_{jr} \frac{\partial f_r(u_1^*, u_2^*)}{\partial u_{j'}^*} \tag{6.4.30}$$

and

$$D_{jj'} = \sum_{r=1}^{p} S_{jr} S_{j'r} f_r(u_1^*, u_2^*). \tag{6.4.31}$$

Fourier transforming the Langevin equation with respect to time using

$$V_i(\omega) = \int_{-\infty}^{\infty} \mathrm{e}^{i\omega t} v_i(t) dt$$

etc. gives

$$\sum_l \Phi_{jl}(\omega)V_l(\omega) = \eta_j(\omega)$$

with

$$\Phi_{jl}(\omega) = -i\omega\delta_{j,l} - A_{jl}$$

and

$$\langle \eta_j(\omega)\rangle = 0, \quad \langle \eta_j(\omega)\eta_{j'}(\omega')\rangle = D_{jj'}\delta(\omega+\omega').$$

Hence,

$$\begin{aligned}\langle V_i(\omega)V_i(\omega')\rangle &= \left\langle \left[\sum_l \Phi_{il}^{-1}(\omega)\eta_l(\omega)\right]\left[\sum_j \Phi_{ij}^{-1}(\omega')\eta_j(\omega')\right]\right\rangle \\ &= \delta(\omega+\omega')\sum_{l,j}\Phi_{il}^{-1}(\omega)D_{lj}\Phi_{ij}^{-1}(-\omega').\end{aligned}$$

Defining the power spectrum of the ith chemical species by (see Sects. 2.2.4 and 6.3)

$$\langle V_i(\omega)V_i(\omega')\rangle = S_i(\omega)\delta(\omega+\omega'),$$

we deduce that

$$S_i(\omega) = \sum_{l,j}\Phi_{il}^{-1}(\omega)D_{lj}(\Phi^\dagger)_{ji}^{-1}(\omega). \tag{6.4.32}$$

Note that the above analysis applies to any two-species master equation of the form (6.4.28) and can be extended to multiple species. In the case of the Brusselator model system, whose deterministic kinetic equations are given by Eq. (6.4.24), we have

$$\sum_r S_{1r}f_r(u_1,u_2) = a-(b+1)u_1+u_1^2u_2, \tag{6.4.33}$$

$$\sum_r S_{2r}f_r(u_1,u_2) = bu_1-u_1^2u_2. \tag{6.4.34}$$

It follows that

$$\mathbf{A} = \begin{pmatrix} b-1 & a^2 \\ -b & -a^2 \end{pmatrix}, \tag{6.4.35}$$

and

$$\mathbf{D} = \begin{pmatrix} 2(b+1)a & -2ba \\ -2ba & 2ba \end{pmatrix} \tag{6.4.36}$$

on setting $u_1^* = a, u_2^* = b/a$. The corresponding power spectra are [55]

$$S_1(\omega) = 2a((1+b)\omega^2+a^4)\Gamma(\omega)^{-1}, \quad S_2(\omega) = 2ab(\omega^2+1+b)\Gamma(\omega)^{-1}, \tag{6.4.37}$$

where

$$\Gamma(\omega) = (a^2-\omega^2)^2+(1+a^2-b)^2\omega^2.$$

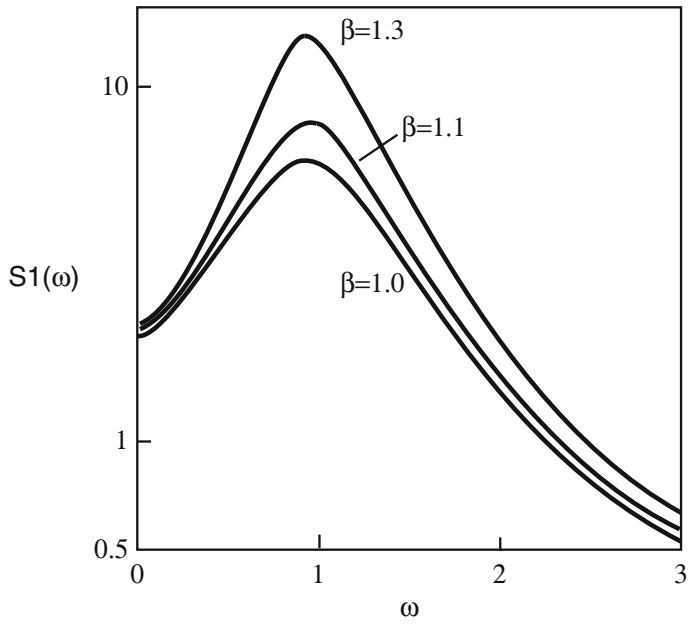

Fig. 6.14: Power spectrum $S_1(\omega)$ of fluctuations in the concentration of X molecules in Brusselator system for parameter values in the fixed point regime of the kinetic equations (6.4.24): $a = 1$ and $b = 1.0, 1.1, 1.3$

In Fig. 6.14 we plot the power spectrum $S_1(\omega)$ in a parameter regime where the deterministic kinetic equation (6.4.24) support a stable fixed point. It can be seen that there is a peak in the power spectrum at $\omega = \omega_c \neq 0$, indicating the presence of stochastic oscillations (quasi-cycles) even though the deterministic system operates below the Hopf bifurcation point. Moreover the frequency ω_c is approximately equal to the Hopf frequency of limit cycle oscillations beyond the bifurcation point. Thus, intrinsic noise can extend the parameter regime over which a biochemical system can exhibit oscillatory behavior. An analogous result holds for Turing pattern formation, as discussed in Sect. 9.3.

6.5 Information Transmission in Regulatory Networks

So far in this chapter we have focused on methods for calculating the level of molecular noise in gene networks. However, as we indicated in the introduction, one important consequence of noise is that it can limit the ability of a network to transmit information. In this section we show how to mathematically quantify this idea, following closely the recent review by Tkacik and Walczak [631]. In order to motivate the analysis, let us return to the simple feedforward regulatory network considered in Sect. 6.3.2 and Fig. 6.3. We now make explicit the fact that the rate at which the gene switches to its active state will depend on the background concentration $c(t)$ of the transcription factor Y. Therefore, we set $k_+ \to k_+ c$ and treat the network

as an input/output device $c(t) \to x(t)$, where $x(t)$ is the protein concentration (see Fig. 6.15). In the case of promoter switching that is faster than the protein kinetics and time variation of c, the probability p_1 that the gene is active reaches the QSS $p_1(c) = k_+c/(k_+c+k_-)$. Therefore, ignoring intrinsic fluctuations, the protein concentration evolves according to the kinetic equation

$$\frac{dx}{dt} = rk_+c(t)/(k_+c(t)+k_-) - \gamma x(t). \tag{6.5.1}$$

In the case of a constant input, the steady-state solution is

$$\bar{x}(c) \equiv \frac{r}{\gamma}\frac{k_+c}{k_+c+k_-}, \tag{6.5.2}$$

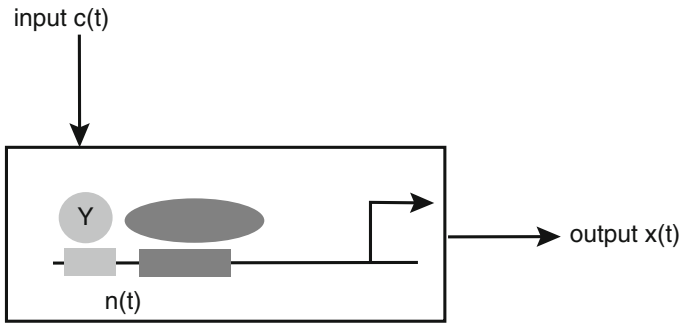

Fig. 6.15: Simple regulatory network represented as a noisy input/output channel. The input signal is the concentration $c(t)$ of transcription factor and the output signal is the concentration $x(t)$ of expressed protein. The internal state $n(t)$ of the channel specifies whether the gene is in the active ($n = 1$) or inactive ($n = 0$) state

which is an invertible function. Hence, we can faithfully reconstruct the input from the output—we have lossless channel. However, this no longer holds when the effects of intrinsic noise are included. Given a fixed input and weak noise, the output response can be characterized in terms of the stationary Gaussian distribution

$$P(x|c) = \frac{1}{\sqrt{2\pi\sigma_X^2(c)}} \mathrm{e}^{-(x-\bar{x}(c))^2/2\sigma_X^2(c)}, \tag{6.5.3}$$

where $\sigma_X^2(c)$ is the variance of the protein concentration. In Sect. 6.3.2 we calculated the variance in the case of a large but finite population of genes using a linear noise approximation. We found that there were two contributions to the variance—input noise arising from the stochastic binding and unbinding of transcription factors and output Poisson noise due to the finite number of proteins [see Eq. (6.3.16)]. Both terms depend on c under the substitution $k_+ \to k_+c$. It turns out that analogous terms arise in the case of a single gene. However, the input noise tends to be smaller

than diffusion noise, which arises from the passive transport of transcription factors within the nucleus [632]. Therefore, applying the Berg–Purcell theory of diffusion-limited reactions to estimate the size of fluctuations in the input concentration c (see Sect. 5.1), we have $\sigma_c^2 = c/(Da\tau)$ where D is the diffusivity, a is the size of the promoter binding site, and τ is the detection time (taken to be the lifetime of a protein). The total output variance is then

$$\sigma_X^2(c) = \bar{x}(c) + \left(\frac{d\bar{x}(c)}{dc}\right)^2 \sigma_c^2, \tag{6.5.4}$$

where the factor $(d\bar{x}(c)/dc)^2$ converts input fluctuations to output fluctuations in the small noise limit.

Suppose that the inputs are drawn from some stationary distribution $P(c)$. A fundamental issue is finding a way to quantify how much information one can extract, in principle, about the values of the input c based on measurements of the output x, given the conditional probability $P(x|c)$. Since the joint probability distribution is $P(x,c) = P(x|c)P(c)$, it follows that if x and c were statistically independent (c-independent $\bar{x}$ and σ_X^2), then $P(c,x) = P(c)P(x)$ and $P(x|c) = P(x)$. In this case the channel cannot transmit any information. This suggests that quantifying the amount of information transmitted involves some measure of the statistical interdependence of the inputs and outputs. A first guess might be the covariance

$$\mathrm{Cov}(c,x) = \int\int (c - \langle c\rangle)(x - \langle x\rangle)dx\,dc.$$

However, this only captures linear correlations, whereas many gene networks are nonlinear input/output devices. A much more general measure of statistical interdependence, which is based on a minimal number of assumptions about the underlying stochastic process, was introduced by Shannon [582] and is known as the *mutual information* between c and x.

6.5.1 Entropy and Mutual Information

In Sect. 1.4 we introduced the notion of entropy in terms of the number of different configurations M that a macromolecule with given energy E can realize, that is, $S = k_B \ln M$. (In statistical mechanics this is referred to as the entropy of a microcanonical ensemble.) An implicit assumption is that each of these microstates i is equally likely so the probability distribution over the set of microstates for fixed E is uniform, $p_i = 1/M$. Hence, we could rewrite the entropy as

$$S = -k_B \sum_i p_i \ln p_i. \tag{6.5.5}$$

It turns out that such a formula also applies to systems with different constraints than fixed energy, where the distribution p_i is not uniform. One example is the Boltzmann–Gibbs distribution itself, where one allows the energy to fluctuate but the mean energy is fixed by the temperature of the surrounding environment. The basic idea is that the entropy still counts the number of accessible states but weights them according to the probability of observing a given state. Shannon subsequently introduced an information theoretic notion of entropy as a measure of the uncertainty of a random variable; the larger the entropy, the greater the amount of information that is generated by observing the state of the system. The convention in information theory is to set $k_B = 1$ and to use base 2 logarithms. Thus, the *Shannon entropy* is defined according to

$$S = -\sum_i p_i \log_2 p_i \tag{6.5.6}$$

so that S is measured in bits. One bit is the entropy of a binary variable that has two equally accessible states. In the case of M possible states, the entropy takes values in the range $0 \leq S \leq \log_2 M$. If $S = 0$, then there is no uncertainty and making a measurement yields no new information. On the other hand, the entropy is maximal when $p_i = 1/M$. It is also possible to define Shannon entropy for a continuous random variable such as the concentration c of a protein:

$$S = -\int p(c) \log_2 p(c) dc. \tag{6.5.7}$$

(Note that certain care has to be taken, since the entropy depends on the units chosen for the continuous random variable. However, we will be interested in changes in entropy where this is no longer an issue.)

Recall that we want to find some measure of the statistical interdependence of an input c and an output x. From the perspective of information theory, one can quantify this in terms of how much one's uncertainty in x is reduced by knowing c. Prior to knowing c, the entropy is $S[P_X] = -\int P(x) \log_2 P(x) dx$, whereas after c is specified the entropy becomes $S[P_{X|C}] = -\int P(x|c) \log_2 P(x|c) dx$. Thus, a measure of the reduction in uncertainty is the entropy difference

$$\Delta S = S[P_X] - S[P_{X|C}].$$

Now imagine measuring the entropy difference over an ensemble of different input concentration regimes distributed according to $P(c)$. The resulting quantity is called the *mutual information*

$$I(c;x) = \int P(c)(S[P_X] - S[P_{X|C}]) dc. \tag{6.5.8}$$

One important property of the mutual information is that it is symmetric with respect to exchanging input and output. In order to see this, we use the definition of Shannon entropy, and the relation between joint and conditional probabilities:

$$
\begin{aligned}
I(c;x) &= -\int\int P(c)\,[P(x)\log_2 P(x) - P(x|c)\log_2 P(x|c)]\,dx\,dc \\
&= \int\int \left[P(c,x)\log_2 \frac{P(c,x)}{P(c)} - P(c,x)\log_2 P(x)\right] dx\,dc \\
&= \int\int \left[P(c,x)\log_2 \frac{P(c,x)}{P(c)P(x)}\right] dx\,dc = \int\int \left[P(c,x)\log_2 \frac{P(c|x)}{P(c)}\right] dx\,dc \\
&= \int\int [P(c|x)P(x)\log_2 P(c|x) - P(c,x)\log_2 P(c)]\,dx\,dc \\
&= -\int\int P(x)\,[P(c)\log_2 P(c) - P(c|x)\log_2 P(c|x)]\,dc\,dx.
\end{aligned}
$$

Hence, the mutual information measures how much, on average, our uncertainty in one variable is reduced by knowing the value of the complementary variable.

Example 6.1 (Gaussian noise). As an illustration of the above ideas, suppose that an input signal is corrupted by additive Gaussian noise

$$x = \gamma c + \xi.$$

This means that the conditional probability distribution is

$$P(x|c) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\gamma c)^2}{2\sigma^2}\right).$$

Also suppose that the input c is also drawn from a Gaussian distribution,

$$P(c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{c^2}{2\sigma_c^2}\right).$$

It follows that x is also a Gaussian with

$$P(x) = \int P(x|c)P(c)dc = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{x^2}{2\sigma_x^2}\right)$$

with

$$\sigma_x^2 = \gamma^2\sigma_c^2 + \sigma^2.$$

The mutual information is

$$
\begin{aligned}
I(c;x) &= \frac{1}{\ln 2}\int\int \left[P(c,x)\ln\frac{P(x|c)}{P(x)}\right] dx\,dc \\
&= \frac{1}{\ln 2}\int\int P(c,x)\left[\ln\left(\sqrt{\frac{2\pi\sigma_x^2}{2\pi\sigma^2}}\right) - \frac{(x-\gamma c)^2}{2\sigma^2} + \frac{x^2}{2\sigma_x^2}\right] dx\,dc,
\end{aligned}
$$

on using the identity $\log_2 x = \ln x / \ln 2$. Using the results $\int\int P(c,x)x^2 dx dc = \langle x^2 \rangle$, $\int\int P(x,c)dx dc = 1$, and

$$\int\int P(c,x)(x-\gamma c)^2 dx dc = \int\int P(x|c)P(c)(x-\gamma c)^2 dx dc = \sigma^2,$$

we deduce that

$$I(c;x) = \frac{1}{\ln 2} \ln \left(\sqrt{\frac{\sigma_x^2}{\sigma^2}} \right) = \frac{1}{2} \log_2 \left(1 + \frac{\gamma^2 \sigma_c^2}{\sigma^2} \right). \tag{6.5.9}$$

The mutual information depends on the so-called SNR σ_c^2/σ^2. It can also be shown that in the case of Gaussian additive noise, information transmission or mutual information is maximized for a given input variance when the input is drawn from a Gaussian distribution [631].

6.5.2 *Optimizing Mutual Information in a Simple Regulatory Network*

A basic goal of information theory within the context of gene regulatory networks is to determine the distribution of inputs for which a network with a given form of intrinsic noise maximizes its information transmission as measured by mutual information. This program has been developed for a range of networks with ever increasing complexity, including multiple gene products that may interact, and self-regulatory feedback [632–635, 665]. For related studies on information transmission, see [105, 383, 638, 705]. In order to make analytical progress, the system is usually assumed to operate in the small noise regime so that the various conditional probability distributions can be approximated by Gaussians. Here we will illustrate the basic ideas by focusing on the simple regulatory network of Fig. 6.15, following the analysis of Tkacik et al. [634]. Since we are optimizing with respect to $P(c)$, we use the following version of the mutual information:

$$I(x;c) = -\int P(c)\log_2 P(c) dc + \int\int P(c|x)\log_2 P(c|x) dc\, dx. \tag{6.5.10}$$

However, to use this formula, it is necessary to determine $P(c|x)$ given $P(x,c)$. Exploiting the small noise approximation, we model $P(c|x)$ as the Gaussian

$$P(c|x) = \frac{1}{\sqrt{2\pi\sigma_C^2(x)}} e^{-(c-\bar{c}(x))^2/2\sigma_C^2(x)}, \tag{6.5.11}$$

where $\bar{c}(x)$ is the most likely value of c given the output x and $\sigma_C^2(x)$ is the corresponding variance about the expected value. Substituting into the expression for the mutual information yields

$$I(x;c) = -\int P(c)\log_2 P(c)dc - \frac{1}{2}\int P(x)\log_2[2\pi e\sigma_C^2(x)]dx.$$

It remains to determine $\bar{c}(x)$ and $\sigma_C^2(x)$. Using Bayes' theorem (Sect. 1.3), we have

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} = \frac{1}{Z(x)}\mathrm{e}^{-F(c,x)},$$

where all c-independent terms have been lumped together in the normalization factor Z and, from Eq. (6.5.3),

$$F(c,x) = -\ln P(c) + \frac{1}{2}\ln\sigma_X^2(c) + \frac{1}{2}\frac{(x-\bar{x}(c))^2}{\sigma_X^2(c)}.$$

Comparison with Eq. (6.5.11) shows that $\bar{c}(x)$ and $\sigma_C^2(x)$ are defined by

$$\left.\frac{\partial F(c,x)}{\partial c}\right|_{c=\bar{c}(x)=0} = 0, \quad \left.\frac{\partial^2 F(c,x)}{\partial c^2}\right|_{c=\bar{c}(x)=0} = \frac{1}{\sigma_C^2(x)}.$$

Expanding the solution for $1/\sigma_C^2(x)$ to leading order in $1/\sigma_X^2(c)$ (small noise approximation) then gives [634]

$$\frac{1}{\sigma_C^2(x)} = \frac{1}{\Sigma^2(c)} \equiv \frac{1}{\sigma_X^2(c)}\left(\frac{d\bar{x}(c)}{dc}\right)^2 \tag{6.5.12}$$

and hence

$$I(x;c) = -\int P(c)\log_2 P(c)dc + \frac{1}{2}\int P(c)\log_2\left[\frac{1}{2\pi e}\frac{1}{\Sigma^2(c)}\right]dc. \tag{6.5.13}$$

We have used the fact that, to leading order, $P(x)dx = P(c)dc$ in the small noise limit.

We now have the variational problem of finding the input distribution $P^*(c)$ that maximizes the mutual information. However, it is first necessary to specify certain constraints regarding the optimization procedure. First, there is a maximum number of proteins involved, which can be imposed by restricting the allowed range of the input concentration c and the corresponding expected number of proteins $\bar{x}(c)$; the latter is typically implemented by normalizing $\bar{g}(c)$ to lie in the interval $[0,1]$. The constraint that $\int_0^{c_{\max}} P(c)dc = 1$ can be incorporated into the variational problem using a Lagrange multipler λ, so that the optimization problem takes the form

$$\frac{\delta}{\delta P(c)}\left[I(x;c) - \lambda\int P(c)dc\right] = 0. \tag{6.5.14}$$

Using the expression for $I(x;c)$ and properties of functional derivatives[1], one finds that the optimal input distribution $P^*(c)$ satisfies

$$0 = \frac{1}{2}\ln\left[\frac{1}{2\pi e}\frac{1}{\Sigma^2(c)}\right] - \ln P^*(c) - 1 - \lambda \ln 2.$$

Rearranging and exponentiating gives the optimal distribution

$$P^*(c) = \frac{1}{Z}\frac{1}{\sqrt{2\pi e}}\frac{1}{\Sigma(c)}, \tag{6.5.15}$$

where $Z = \mathrm{e}^{1+\lambda \ln 2}$ is a normalization factor with

$$Z = \int_0^{c_{\max}} \frac{1}{\sqrt{2\pi e}}\frac{dc}{\Sigma(c)}. \tag{6.5.16}$$

The corresponding optimal mutual information is simply

$$I^* = \log_2 Z. \tag{6.5.17}$$

Note that the resulting expression for I^* will still depend on the various parameters of the underlying regulatory network. These include the parameters associated with the kinetics of binding and unbinding of transcription factors as in Eq. (6.5.4) and, in more complex networks, interactions between multiple gene products. Thus, there is an additional optimization step in which one maximizes the mutual information I^* with respect to these network parameters. Some of the predictions regarding the structure of optimal networks can be found elsewhere [105, 634, 635, 665, 705]. Here we illustrate the theory with the simple regulatory network given by Fig. 6.15. From Eqs. (6.5.4) and (6.5.16),

$$Z = \int_0^{c_{\max}} \frac{1}{\sqrt{2\pi e}}\frac{(d\bar{x}(c)/dc)^2}{\bar{x}(c) + c\,(d\bar{x}(c)/dc)^2}. \tag{6.5.18}$$

---

[1] The mutual information is expressed as a functional of $P(c)$, that is, $I = I[P]$. By analogy with the least-action principle of classical mechanics (see Chap. 10), we can take the functional derivative according to

$$\delta I = \lim_{\varepsilon \to 0}\frac{1}{\varepsilon}(I[P + \varepsilon\delta P] - I[P]),$$

where $\delta P(c)$ is an arbitrary smooth function. Evaluating the first term in I, after converting to natural logarithms, we have

$$\int [P(c) + \varepsilon\delta P(c)]\ln(P(c) + \varepsilon\delta P(c))dc - \int P(c)\ln P(c)dc = \varepsilon\int \delta P(c)[\ln P(c) + 1]dc + O(\varepsilon^2).$$

Combining this with the other terms and using the fact that $\delta P(c)$ is arbitrary, we can set the total factor multiplying $\delta P(c)$ to be zero, which yields Eq. (6.5.15).

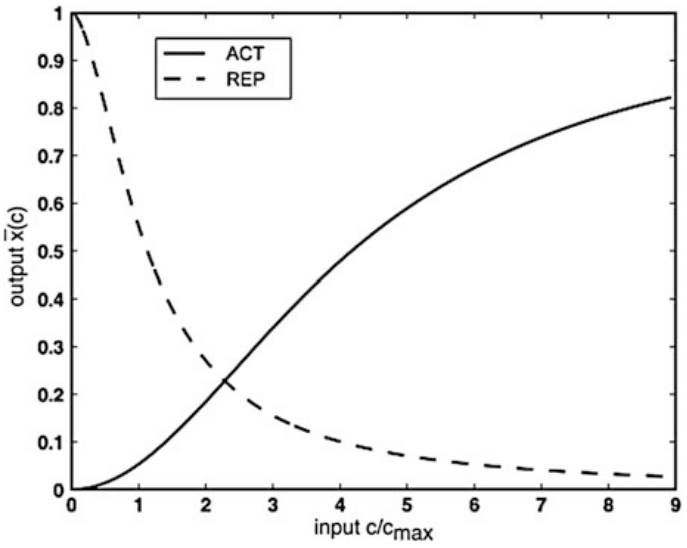

Fig. 6.16: The optimal input/output relations for repressors (*dashed line*) and activators (*solid line*) for the simple regulatory network shown in Fig. 6.15 (Adapted from Tkacik and Walczak [631])

Here c has been nondimensionalized by fixing $Da\tau$ and normalizing $\bar{x}$. Suppose that $\bar{x}(c)$ is given by the Hill function

$$\bar{x}(c) = \frac{c^n}{K^n + c^n}.$$

For fixed $c_{\max}$ the only free parameters are K and n. Optimizing the mutual information with respect to these parameters leads to the results shown in Fig. 6.16.

We end this brief detour into the world of information theory by briefly noting some potential limitations of the above approach, as highlighted in the review by Tkacik and Walczak [631]. First, there is no a priori reason why information should be identified as an appropriate measure of biological function. Second, even if it is an appropriate measure, it is possible that gene networks and biochemical signaling pathways have not yet become optimized for biological function through evolution. That is, the networks observed today simply reflect their particular evolutionary history rather than some history-independent optimization scheme. On the other hand, recent experiments concerning specific gene networks active during early development suggest that at least these networks operate close to the limits imposed by intrinsic noise [237]. Finally, in order to strengthen the links between theory and experiment, it will be necessary to confront the following theoretical challenges: (i) dealing with information transmission in time-dependent nonlinear networks; (ii) understanding information transmission in spatially inhomogeneous systems; (iii) extending analytical methods beyond the small noise limit; (iv) linking information transmission to other important measures of network function such as metabolic cost.

6.6 Fluctuations in DNA Transcription

One of the simplifications in the models of gene expression discussed so far is that the multistage processes underlying transcription and translation have been reduced to single-step processes with exponential waiting times (Poisson approximation). However, transcription (and also translation) can be broken up into at least three stages called initiation, elongation, and termination [128, 238]. During the initiation stage, RNAP binds to a promoter site on the DNA and unzips the double helix so that the strand of DNA to be transcribed is made accessible. Following the transcription of the first few nucleotides, the so-called transcription elongation complex (TEC) is formed, which consists of the RNAP, the DNA, and the emerging mRNA. This signals the beginning of the elongation phase where the TEC slides along the DNA, extending the transcript one nucleotide at a time. The process is terminated when a specific site is reached, for example, and the nascent mRNA is released. An implicit assumption of the single-step Poisson approximation of gene transcription is that the rate-limiting step is initiation. However, there is growing evidence from single-molecule experiments that initiation can be much faster than elongation [18]. Moreover, in vitro studies of *E. coli* RNAP have established that processive mRNA synthesis is often disrupted by transcriptional pauses that can last anything from 1s to more than 20s [258, 469]. In some cases, the pauses are linked with the reverse translocation of the RNAP along the DNA, a process known as *backtracking* [482]. These observations suggest that the distribution of transcription times might be non-exponential with heavy tails arising from the long transcriptional pauses.

Recently, there have been a number of stochastic models of the elongation stage and backtracking [413, 535, 552, 661, 690]. For the sake of illustration, we will focus on the model of Voliotis et al. [661], which is based on a master equation description of the dual processes of the TEC translocating along the DNA and the extension of the nascent mRNA via polymerization. A schematic diagram of the basic kinetics is shown in Fig. 6.17 in the simpler case that backtracking cannot occur. Suppose that n nucleotides have been transcribed and the active site of RNAP is at the end of the precursor mRNA chain. Denote this so-called pretranslocated state of the active site by $m = 0$. The active site can then shift one step beyond the precursor mRNA to form a posttranslocated state denoted by $m = 1$. It is now in a position to add the next nucleotide to the precursor mRNA by polymerization so that $n \to n+1$ and m resets to 0. The rates of polymerization and depolymerization are given by k_f and k_b, while the rates of forward and backward translocation are given by a and b. Let $P_{n,m}(t)$ be the probability of finding the TEC in state (n,m) at time t. The corresponding master equation is given by [661]

$$\frac{dP_{n,0}}{dt} = k_f P_{n-1,1} + bP_{n,1} - (k_b + a)P_{n,0} \tag{6.6.1a}$$

$$\frac{dP_{n,1}}{dt} = k_b P_{n+1,0} + aP_{n,0} - (k_f + b)P_{n,1}, \tag{6.6.1b}$$

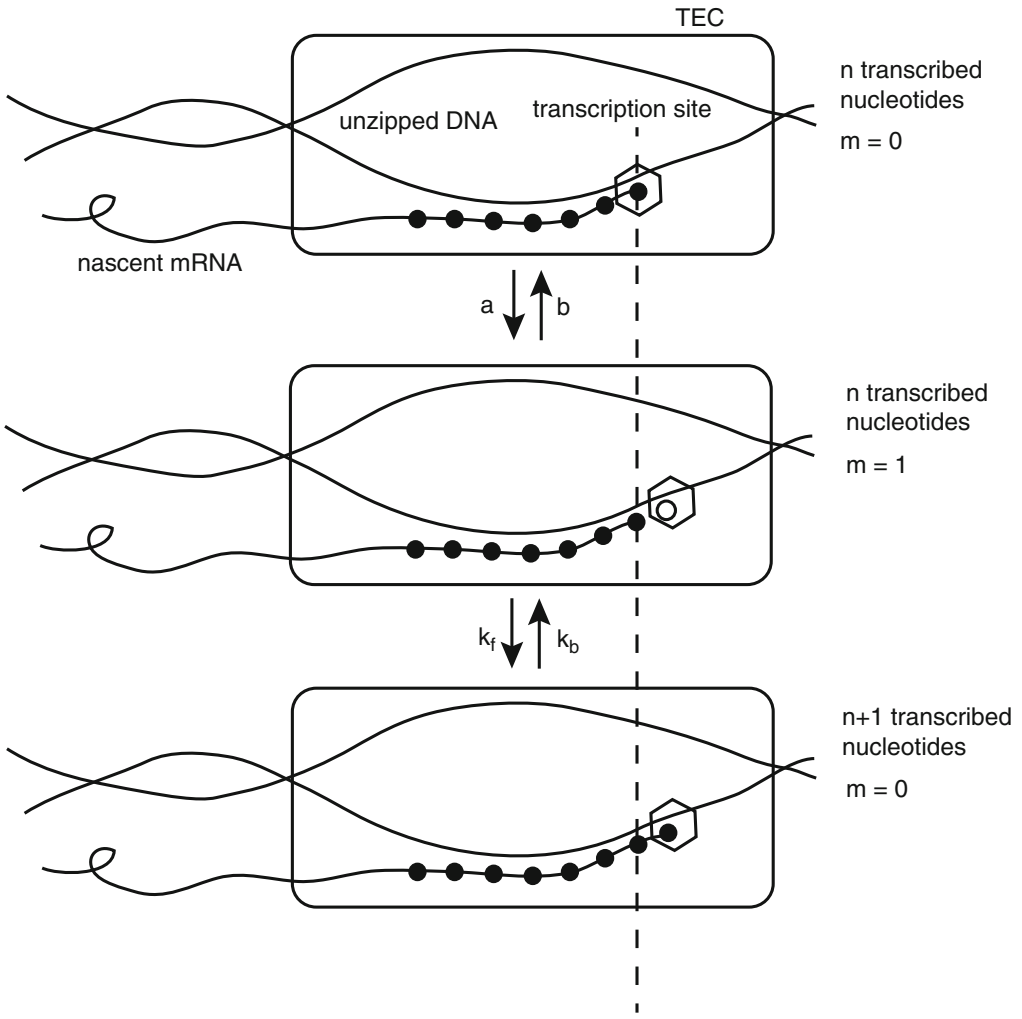

Fig. 6.17: Schematic illustration of the transcription elongation complex (*TEC*). In the absence of fluctuations, the active transcription site in the pretranslocated state ($m = 0$) takes one step beyond the nascent mRNA to enter the posttranslocated state ($m = 1$). Translocation is a reversible reaction with transition rates a and b. The site can then add one nucleotide to the mRNA via polymerization so that $n \to n+1$ and $m = 1 \to m = 0$. This step is also reversible with forward and backward transition rates k_f, k_b, respectively (Redrawn from [661])

with $n = 0, 1, \ldots N-1$. There is a reflecting boundary condition at $n = 0$, which can be implemented by introducing a fictitious state $n = -1$ and setting $k_b P_{0,0} = k_f P_{-1,1}$. Similarly, there is an absorbing boundary condition $P_{N,0} = 0$, since the process terminates when $n = N$ is reached.

The first step in the analysis is to introduce the mean occupancy for each translocation step ($m = 0, 1$) by summing over all nucleitide positions $n = 0, \ldots, N-1$. Setting $\Pi_m(t) = \sum_{n=0}^{N-1} P_{n,m}(t)$ and using the boundary conditions, we have

$$\frac{d\Pi_0}{dt} = (k_f + b)\Pi_1 - (k_b + a)\Pi_0, \quad \Pi_1 = 1 - \Pi_0$$

with the initial condition $\Pi_0(0) = 1$. There is convergence to the steady-state solution

$$\Pi_0^* = (k_f + b)\tau, \quad \Pi_1^* = (k_b + a)\tau, \quad \tau = \frac{1}{k_f + k_b + a + b},$$

with τ as the relaxation time. Assuming polymerization/depoylmerization is much slower than translocation ($k_f, k_b \ll a, b$), we can make the QSS approximation $P_{m,n}(t) = \Pi_m^* P_n(t)$ (see also Sect. 7.4), with $P_n(t)$ satisfying the birth–death master equation

$$\frac{dP_n}{dt} = \omega_- P_{n+1} + \omega_+ P_{n-1} - (\omega_+ + \omega_-)P_n, \tag{6.6.2}$$

and the effective polymerization/depolymerization rates are

$$\omega_+ = k_f(k_b + a)\tau \approx \frac{k_f a}{a+b}, \quad \omega_- = k_b(k_f + b)\tau \approx \frac{k_b b}{a+b}.$$

The boundary conditions become $\omega_- P_0 = \omega_+ P_{-1}$ (reflecting) and $P_N = 0$ (absorbing). The elongation time is defined as the time for the TEC to reach position $n = N$ starting from $n = 0$. In terms of the mean-field model given by the birth–death process, calculating the mean and variance of the elongation time requires solving a FPT for the discrete Markov process. This can be achieved by following analogous steps to the analysis of continuous process in Sect. 2.3.

Suppose that the TEC starts at position $n(0) = n_0$. Define the survival probability that the TEC has not yet reached the absorbing boundary at $n = N$ by

$$S(n_0, t) = \sum_{n=0}^{N-1} P(n, t | n_0, 0), \tag{6.6.3}$$

where we have made the initial condition explicit by setting $P_n(t) \to P(n, t|n_0, 0)$. If T is the (stochastic) elongation time, then $S(n_0, t)$ is the probability that $T \geq t$. This implies that the cumulative distribution function of the elongation time is $1 - S(n_0, 0)$. Hence the first and second moments of the elongation time are

$$T(n_0) = \langle T \rangle = \int_0^\infty t \frac{\partial S(n_0, t)}{\partial t} dt = \int_0^\infty S(n_0, t) dt \tag{6.6.4}$$

and

$$T_2(n_0) = \langle T^2 \rangle = \int_0^\infty t^2 \frac{\partial S(n_0, t)}{\partial t} dt = 2 \int_0^\infty t S(n_0, t) dt. \tag{6.6.5}$$

Equations for S and the moments of T can be obtained by considering the backward master equation

$$\begin{aligned} \frac{dP(n, t|n_0, 0)}{dt} &= \omega_+ \left[P(n, t|n_0 + 1, 0) - P(n, t|n_0, 0) \right] \\ &\quad + \omega_- \left[P(n, t|n_0 - 1, 0) - P(n, t|n_0, 0) \right]. \end{aligned} \tag{6.6.6}$$

The backward equation follows from differentiating with respect to t both sides of the Chapman–Kolmogorov equation

$$P(n_1,s|n_0,0)=\sum_n P(n_1,s|n,t)P(n,t|n_0,0).$$

Summing Eq. (6.6.6) from $n=0$ to $n=N-1$ shows that

$$\frac{dS(n_0,t)}{dt}=\omega_+\left[S(n_0+1,t)-S(n_0,t)\right]+\omega_-\left[S(n_0-1,t)-S(n_0,t)\right], \tag{6.6.7}$$

supplemented by the boundary conditions $S(N,t)=0$ and $S(0,t)=S(-1,t)$ and the initial condition $S(n_0,0)=1$.

Let us now calculate the mean elongation time. Integrating Eq. (6.6.7) with respect to t gives

$$-1=\omega_+[T(n_0+1)-T(n_0)]+\omega_-[T(n_0-1)-T(n_0)] \tag{6.6.8}$$

with $T(N)=0$ and $T(0)=T(-1)$. Setting $U(n_0)=T(n_0)-T(n_0-1)$ we obtain the first-order difference equation

$$\omega_+U(n_0+1)-\omega_-U(n_0)=-1,$$

which has the solution $U(0)=0$, $U(1)=-1/\omega_+$, $U(2)=-1/\omega_+-\omega_-/\omega_+^2$, etc. that is

$$U(n)=-\frac{1}{\omega_+}\left[1+\frac{\omega_-}{\omega^+}+\left(\frac{\omega_-}{\omega^+}\right)^2+\ldots+\left(\frac{\omega_-}{\omega^+}\right)^{n-1}\right].$$

Since $-T(n)=U(N)+U(N-1)+\ldots+U(n+1)$, we deduce that

$$T(n_0)=\sum_{n_0+1}^{N}\frac{1}{\omega_+}\sum_{m=0}^{n-1}\left(\frac{\omega_-}{\omega^+}\right)^m. \tag{6.6.9}$$

Introducing $K=\omega_-/\omega_+$ and noting that $0\leq K<1$, we can sum the geometric series to give [204, 661]

$$T(n_0)=\frac{1}{\omega_+}\sum_{n=n_0+1}^{N}\frac{1-K^n}{1-K}=\frac{1}{\omega_+(1-K)}\left[N-n_0-\frac{K^{n_0+1}-K^{N+1}}{1-K}\right].$$

Finally, setting $n_0=0$ we obtain the mean elongation time $\mu=T(0)$ with

$$\mu=\frac{1}{\omega_+(1-K)}\left[N-\frac{K(1-K^N)}{1-K}\right]. \tag{6.6.10}$$

The variance can be calculated in a similar fashion (see Ex. 6.9). Here we simply note that when chain lengthening is dominant, $K\ll 1$, both the mean and variance are linear functions of the chain length N:

$$\mu=\frac{N}{\omega_+}+K\frac{N-1}{\omega_+}+O(K^2), \tag{6.6.11}$$

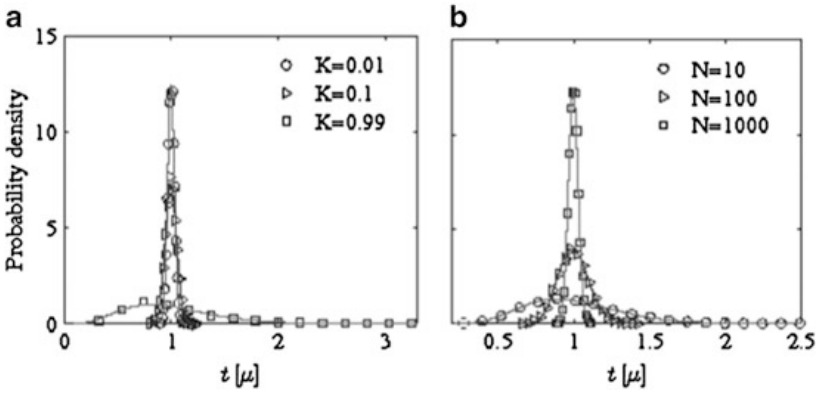

Fig. 6.18: Distribution of elongation times (in units of the mean elongation time) for the TEC model without backtracking. Results from mean-field theory are given by *solid curves* and superimposed with stochastic simulations results. (**a**) Results for $N = 1,000\,\text{bp}$, $\omega_+ = 20\,\text{s}^{-1}$, and various values of $K = \omega_-/\omega_+$. (**b**) Results for $K = 0.01$, $\omega_+ = 20\,\text{s}^{-1}$, and different template lengths N (Adapted from Voliotis et al. [661])

and

$$\sigma^2 = \frac{N}{\omega_+^2} + 4K\frac{N-1}{\omega_+^2} + O(K^2). \tag{6.6.12}$$

For sufficiently long sequences $N \gg 1$, one finds that the distribution of elongation times is given by a narrow Gaussian with fluctuations scaling as $1/\sqrt{N}$. This adds a characteristic delay to the Poisson-like distribution of initiation times (see Fig. 6.18).

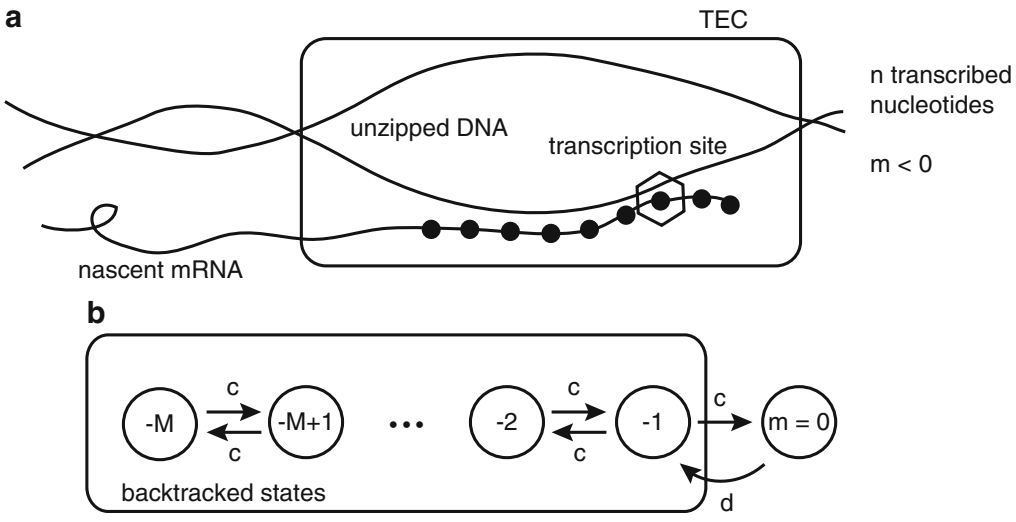

Fig. 6.19: Schematic illustration of TEC backtracking. (**a**) Example of a backtracked state of the TEC with $m = -2$. (**b**) Unbiased random walk model of backtracking. Transitions between backtracked states occur at a rate c. The TEC enters the backtracking regime from the state $m = 0$ at a rate d and exits the backtracking regime at the rate c. (Redrawn from [661].)

Moreover, if initiation is much faster than elongation, then the transcription time is much more regular than if initiation dominates.

Voliotis et al. [661] show that the above picture persists when backtracking pauses are included, provided that they are sufficiently rare. However, the distribution of elongation times is drastically altered when backtracking becomes significant. The simplest way to incorporate backtracking into the model is to treat it as a separate process. That is, one can introduce additional translocation states of the TEC given by $m = -1, \ldots, -M$, which represent backtracked states shifted by $|m|$ steps from state $m = 0$. The duration of backtracking pauses can also be analyzed in terms of a FPT problem—in this case, a random walk on a finite lattice with a reflecting boundary at $m = -M$ and an absorbing boundary at $m = 0$ (see Fig. 6.19). One finds that there is a broad distribution of pause durations that exhibits power law behavior at intermediate duration times. Consequently, the distribution of elongation times is significantly altered. Numerical simulations of the full model also suggest that the distribution of elongation times with long pauses naturally exhibits switching between high and low mRNA product rats, resulting in transcriptional bursting.

6.7 Kinetic Proofreading

A major requirement for proper cell function is that the genetic code is "read" with few mistakes during protein synthesis or DNA replication. For example, both transcription and translation involve the incorporation of specific molecular substrates at particular times, namely, a specific mRNA nucleotide during the production of mRNA or a specific amino acid during production of a protein. The incorporation of each substrate involves some recognition site within an RNAP or a ribosome, respectively, that is more energetically disposed to bind the correct substrate C, say, rather than an incorrect substrate D. In a simple reaction scheme, the frequency of errors is of the order $e^{-\Delta G_{CD}/k_B T}$, where ΔG_{CD} is the smallest difference in binding energies between the correct substrate and an incorrect substrate. The basic problem is that typical values of ΔG_{CD} cannot account for the small error rates observed in protein synthesis. For example, the maximum frequency at which a wrong but similar amino acid is inserted during protein translation is 10^{-4}, which means that even smaller error rates must occur in each recognition step. The error rates are smaller still in the case of DNA transcription, taking values around 10^{-9}. *Kinetic proofreading* is a mechanism for error correction in biochemical processes, which was first introduced by Hopfield [273] and independently by Ninio [479]. The proofreading mechanism increases specificity of biochemical interactions by including a number of intermediate steps that can undo errors at the cost of increased reaction time and free energy expenditure.

Box 6B. Enzyme kinetics.

Enzymes are generally protein catalysts that help convert other molecules called substrates into products, without themselves being changed by the reaction. In contrast to single-step reactions, the rate of reaction does not increase linearly with the concentration of substrate, since it saturates at high concentrations. A simple model to explain this behavior was first proposed by Michaelis and Menten. The basic reaction scheme involves an enzyme E converting a substrate S to a complex C, which then breaks down to form the product P together with the original enzyme. This can be represented by the following two-step process:

$$S + E \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C \xrightarrow{k_2} P + E. \tag{6.7.1}$$

Although all the reactions are reversible, reaction rates are typically measured under conditions in which the product P is continually removed from the system, which effectively prevents the final reverse reaction from occurring. Setting $s = [S], c = [C], e = [E]$ and $p = [P]$, we have the system of kinetic equations

$$\frac{ds}{dt} = k_{-1}c - k_1 se, \tag{6.7.2a}$$

$$\frac{de}{dt} = (k_{-1} + k_2)c - k_1 se, \tag{6.7.2b}$$

$$\frac{dc}{dt} = -(k_{-1} + k_2)c + k_1 se, \tag{6.7.2c}$$

$$\frac{dp}{dt} = k_2 c. \tag{6.7.2d}$$

Note that the total concentration of enzyme is conserved, $e + c = e_0$ for some constant e_0. Hence, we can eliminate e such that

$$\frac{dc}{dt} = k_1 e_0 s - (k_{-1} + k_2 + k_1 s)c.$$

If the total concentration of enzyme is small, then s changes relatively slowly, which suggests that c reaches steady state before s changes significantly. Thus,

$$c = e_0 \frac{k_1 s}{k_{-1} + k_2 + k_1 s}.$$

Under this so-called equilibrium approximation the overall rate of generating product (or depleting substrate) is

$$\frac{dp}{dt} = k_2 e_0 \frac{k_1 s}{k_{-1} + k_2 + k_1 s} = k_2 e_0 \frac{s}{s + K_M}, \tag{6.7.3}$$

where K_M is the Michaelis constant

$$K_M = \frac{k_{-1} + k_2}{k_1}.$$

Hence, the rate of production is linear s when the substrate concentration is low but saturates when s is sufficiently large. The resulting behavior is referred to as Michaelis–Menten kinetics. For a more general discussion of various kinetic schemes including Michaelis–Menten, see the books by Siegel and Edelstein-Keshet [579] and Keeener and Sneyd [322].

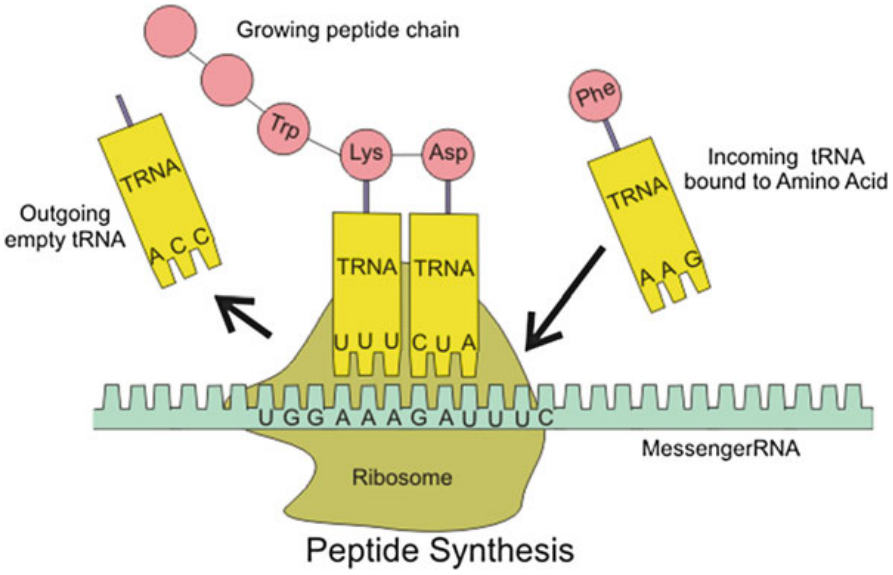

Fig. 6.20: Ribosomes can bind to an mRNA chain and use it as a template for determining the correct sequence of amino acids in a particular protein. Amino acids are selected, collected, and carried to the ribosome by transfer RNA (*tRNA*) molecules, which enter one part of the ribosome and bind to the messenger RNA chain. The attached amino acids are then linked together by another part of the ribosome. Once the protein is produced, it can then "fold" to produce a specific functional three-dimensional structure. Specificity is achieved through the interaction between the codon (triplet of nucleotides) in mRNA and the anti-codon in the tRNA. (Public domain figure from Wikipedia)

6.7.1 Kinetic Proofreading in Protein Synthesis

Consider the binding interaction between a codon E of mRNA and the anti-codon of a tRNA during protein synthesis (see Fig. 6.20). Let C denote the correct tRNA and D an incorrect tRNA. Here E may be viewed as an enzyme acting on a substrate C or D according to a classical *Michaelis–Menten* scheme (see Box 6B). That is,

$$E \underset{k_{\text{off}}}{\overset{k_{\text{on}}[C]}{\rightleftharpoons}} EC \xrightarrow{W} E + \text{correct amino acid incorporated} \tag{6.7.4a}$$

$$E + D \underset{k'_{\text{off}}}{\overset{k'_{\text{on}}[D]}{\rightleftharpoons}} ED \xrightarrow{W} E + \text{incorrect amino acid incorporated.} \tag{6.7.4b}$$

(It is assumed that the catalytic step has no selectivity, that is, the rate of catalysis W is the same for both substrates.) The corresponding kinetic equations are

$$\frac{d[EC]}{dt} = k_{\text{on}}[E][C] - (k_{\text{off}} + W)[EC] \tag{6.7.5a}$$

$$\frac{d[ED]}{dt} = k'_{\text{on}}[E][D] - (k'_{\text{off}} + W)[ED], \tag{6.7.5b}$$

$$[E]_{\text{Total}} = [EC] + [ED] + [E]. \tag{6.7.5c}$$

The last equation ensures that the total concentration of ribosomes or enzymes is fixed. At steady state, we have

$$[EC] = [E]\frac{k_{\text{on}}[C]}{k_{\text{off}} + W}, \quad [ED] = [E]\frac{k'_{\text{on}}[D]}{k'_{\text{off}} + W}.$$

It follows that the rates of correct and incorrect translation are

$$R_{\text{correct}} = W[EC], \quad R_{\text{incorrect}} = W[ED],$$

and the error rate is

$$F_0 = \frac{R_{\text{incorrect}}}{R_{\text{correct}}} = \left[\frac{k'_{\text{on}}[D]}{k'_{\text{off}} + W}\right]\left[\frac{k_{\text{on}}[C]}{k_{\text{off}} + W}\right]^{-1}. \tag{6.7.6}$$

Typically, one finds that the on rates are approximately the same for all tRNAs (being diffusion limited) and the tRNAs have similar concentrations, that is, $k_{\text{on}} \approx k'_{\text{on}}$ and $[D] \approx [C]$. Hence, the error rate F_0 is minimized by taking the catalytic rate W to be much smaller than the off rates. Introducing the dissociation constants $K_C = k_{\text{off}}/k_{\text{on}}, K_D = k'_{\text{off}}/k'_{\text{on}}$, we have

$$F_0 \approx \frac{K_C}{K_D} = e^{-\Delta G_{CD}/k_B T}. \tag{6.7.7}$$

The above simple binding model neglects the fact that when tRNA binds to a codon, it is chemically altered via the hydrolysis of GTP; an analogous process occurs during the polymerization of microtubules (see Sect. 4.1). The transition to the new state is irreversible and in this state the tRNA can also dissociate from the mRNA. This leads to the new reaction scheme

$$E \underset{k_{\text{off}}}{\overset{k_{\text{on}}[C]}{\rightleftharpoons}} EC \xrightarrow{r} \underset{\downarrow E}{EC^*} \xrightarrow{W} E + \text{correct amino acid incorporated} \tag{6.7.8a}$$

$$E \underset{k'_{\text{off}}}{\overset{k'_{\text{on}}[D]}{\rightleftharpoons}} ED \xrightarrow{r} \underset{\downarrow E}{ED^*} \xrightarrow{W} E + \text{incorrect amino acid incorporated.} \tag{6.7.8b}$$

Let the on and off rates of the modified substrates C^* and D^* be $q_{\text{off}}, q_{\text{on}}[C^*] \approx 0$ and $q'_{\text{off}}, q'_{\text{on}}[D^*]$, respectively, with $[C^*], [D^*] \approx 0$. The steady-state concentrations of the modified substrate then satisfy (see Ex. 6.10),

$$[EC^*] = \frac{1}{q_{\text{off}} + W} \frac{rk_{\text{on}}}{k_{\text{off}} + r} [E][C].$$

Again assuming that the rates of catalysis W, r are much smaller than the on and off rates, and taking the concentrations of all tRNAs to be the same, we obtain the new error rate

$$F \approx \frac{q_{\text{off}}}{q'_{\text{off}}} \frac{k'_{\text{on}}}{k'_{\text{off}}} \frac{k_{\text{off}}}{k_{\text{on}}}.$$

Finally, taking the on rates to be tRNA nonspecific,

$$F = \frac{Q_C}{Q_D} \frac{K_C}{K_D} = e^{-\Delta G_{CD}/k_B T} e^{-\Delta G_{C^*D^*}/k_B T} < F_0. \tag{6.7.9}$$

In particular, if the difference in biding energies of the two substrates is the same for the native and modified states, then

$$F = \left[e^{-\Delta G_{CD}/k_B T} \right]^2. \tag{6.7.10}$$

In summary, the inclusion of an irreversible step into the kinetic scheme, $EA \to EA^*$, which necessitates the expenditure of energy, provides an additional opportunity for the incorrect substrate to dissociate and leads to a reduction in the error rate. An even higher level of accuracy can be achieved by having a sequence of n irreversible proofreading stages:

$$E \underset{k_{\text{off}}}{\overset{k_{\text{on}}[C]}{\rightleftharpoons}} EC \xrightarrow{r_1} \underset{\downarrow E}{EC^*} \xrightarrow{r_2} \underset{\downarrow E}{EC^{**}} \dots \xrightarrow{r_m} \underset{\downarrow E}{EC^{**\dots *}} \xrightarrow{W} E + \text{product.}$$

6.7.2 Kinetic Proofreading in T-Cell Activation

T cells, which mature in the thymus, are one of two key cell types of the adaptive immune system, whose basic function is the detection and destruction of intracellular pathogens such as certain bacteria and all viruses (see the review by Coombs and Goldstein [125]). (The other cell type consists of B cells, which mature in the bone marrow and are mainly concerned with the detection and destruction of extracellular pathogens.) In order to execute their function, T cells scan the surfaces of cells for molecular markers of infection. Detection of the appropriate marker activates the T cell which then responds to the pathogen, either by killing the infected cell (effector T cells) or by signaling other parts of the immune system such as B cells (helper T cells). Since T cells only scan the surface of other cells, it is necessary that some cells are able to present information regarding their internal contents to the surface. This is achieved by cutting intracellular proteins into peptide fragments and transporting these fragments to the surface for surveillance by T cells. If a pathogen is present within the cell, then signature peptide groups known as antigens will be made accessible. A major challenge for the pathogen recognition machinery is that the vast majority of peptides on a given antigen-presenting cell do not signify the presence of a pathogen. Thus, a T cell has to recognize an antigen against a noisy background of these so-called self-peptides, just as a ribosome has to recognize the correct tRNA during each stage of protein synthesis. It is not too surprising, therefore, that a kinetic proofreading model has been developed for T-cell activation by McKeithan [430].

The model of McKeithan considers the interaction of a T-cell receptor (TCR) with a ligand consisting of a peptide fragment that is bound to a specialized molecule in the surface of an antigen-presenting cell, known as a major histocompatibility complex (MHC) molecule (see Fig. 6.21a). The peptide–MHC complex that forms the ligand is denoted by pMHC. There are two basic assumptions of the model: (i) In order to respond to an antigen, a TCR in an inactive state T has to undergo a sequence of N modifications to form the activated state B_N. (ii) Dissociation of pMHC from the TCR can occur at any stage, after which the receptor quickly returns to its inactive state (see Fig. 6.21b). Suppose that the off rate back to the inactive state T is the same for all intermediate states. We then have the following hierarchy of kinetic equations for the concentrations $[T], [B_j]$, $j = 0, \ldots, N$:

$$\frac{d[T]}{dt} = -k_{\text{on}}[T][P] + k_{\text{off}} \sum_{i=0}^{N} [B_i], \tag{6.7.11a}$$

$$\frac{d[B_0]}{dt} = k_{\text{on}}[T][P] - k_{\text{off}}[B_0] - k_p[B_0], \tag{6.7.11b}$$

$$\frac{d[B_i]}{dt} = k_p([B_{i-1}] - [B_i]) - k_{\text{off}}[B_i], \tag{6.7.11c}$$

$$\frac{d[B_N]}{dt} = k_p[B_{N-1}] - k_{\text{off}}[B_N], \tag{6.7.11d}$$

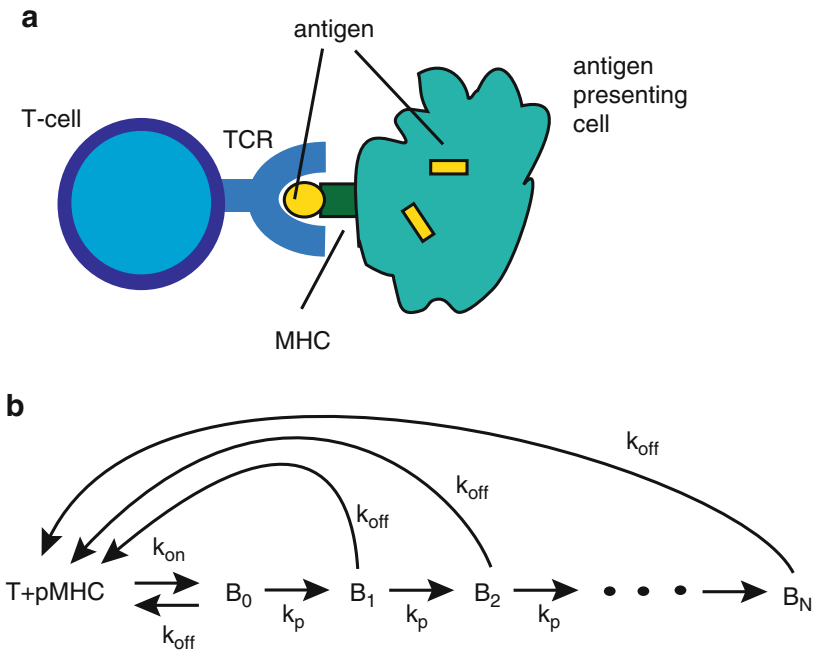

Fig. 6.21: Kinetic proofreading model of T-cell activation. (**a**) Schematic diagram of a T-cell receptor (*TCR*) binding to an antigen that is attached to a major histocompatibility complex molecule (*MHC*) in the surface of an antigen-presenting cell. (**b**) Reaction diagram, see text for details

where $[P]$ is the concentration of a specific pMHC complex. Solving these equations in steady state shows that the fraction of activated complexes is (see Ex. 6.10)

$$\frac{[B_N]}{\sum_{i=0}^{N}[B_i]} = \left(\frac{k_p}{k_p + k_{\text{off}}}\right)^N . \tag{6.7.12}$$

Note that $k_p/(k_p + k_{\text{off}})$ is the probability that in any intermediate step i, the T cell is modified before dissociation of the pMHC. Assuming that k_p is independent of the particular substrate, it follows that the off rate k_{off} is the only parameter whose variation can distinguish between peptides. Even for small values of N, the fraction of activated T cells is sensitive to small changes in k_{off}, reflecting the objective of the kinetic proofreading mechanism. However, it comes at a cost, namely that the actual value of the activity level in response to the correct antigen reduces as N increases, so that an increase in selectivity coincides in a decrease in sensitivity.

6.8 Stochastic Algorithms for Chemical Kinetics

6.8.1 The Stochastic Simulation Algorithm

The SSA, which was originally developed by Gillespie [217–219], is an efficient numerical scheme for generating exact sample paths of a continuous-time Markov process whose probability distribution evolves according to a chemical master equation. Following Sect. 6.2, suppose that the mass-action kinetics of a general biochemical network is written in the form

$$\frac{dx_i}{dt} = \sum_{a=1}^{R} S_{ia} f_a(\mathbf{x}), \quad i = 1,\ldots,N, \tag{6.8.1}$$

where a labels a single-step reaction, f_a are the transition intensities or propensities, and $\mathbf{S}$ is the $N \times R$ stochiometric matrix. Given this notation, the corresponding master equation is

$$\frac{dP(\mathbf{n},t)}{dt} = \Omega \sum_{a=1}^{R} \left(\prod_{i=1}^{N} \mathbb{E}^{-S_{ia}} - 1 \right) f_a(\mathbf{n}/\Omega) P(\mathbf{n},t), \tag{6.8.2}$$

where Ω represents the system size. Typically, Ω is the volume of the well-mixed compartment where reactions occur or the total number of molecules in cases where there is number conservation. Here $\mathbb{E}^{-S_{ia}}$ is a step or ladder operator such that for and function $g(\mathbf{n})$,

$$\mathbb{E}^{-S_{ia}} g(n_1,\ldots,n_i,\ldots,n_N) = g(n_1,\ldots,n_i - S_{ia},\ldots,n_N). \tag{6.8.3}$$

In the following we eliminate the global factor of Ω by rescaling time $t \to \Omega t$.

The starting point for constructing the SSA is to define a new probability function $p(\tau,a|\mathbf{x},t)$, which is the probability, given $\mathbf{X}(t) = \mathbf{x}$, that the next reaction in the system will occur in the time interval $[t+\tau, t+\tau+\Delta\tau)$ and will be the reaction a. From this perspective, both τ and a are random variables conditioned on $\mathbf{X}(t) = \mathbf{x}$. An analytical expression for $p(\tau,a|\mathbf{x},t)$ can be obtained by introducing another probability function $P_0(\tau|\mathbf{x},t)$, which is the probability, given $\mathbf{X}(t) = \mathbf{x}$, that no reaction of any kind occurs in the time interval $[t,t+\tau)$. It follows from the definitions of P_0 and the propensities f_a that P_0 satisfy the equation

$$P_0(\tau + d\tau|\mathbf{x},t) = P_0(\tau|\mathbf{x},t) \left[1 - \sum_{a=1}^{R} f_a(\mathbf{x}) d\tau \right],$$

which is the product of the probability that no reaction occurs in $[t,\tau)$ and the probability that there are no transitions in the infinitesimal interval $[t+\tau, t+\tau+d\tau)$. Rearranging and taking the limit $d\tau \to 0$ yields

$$\frac{dP_0(\tau|\mathbf{x},t)}{d\tau} = -F(\mathbf{x}) P_0(\tau|\mathbf{x},t), \quad F(\mathbf{x}) = \sum_{a=1}^{R} f_a(\mathbf{x}).$$

Under the initial condition $P(0|x,t) = 1$, we have the solution

$$P_0(\tau|\mathbf{x},t) = \exp(-F(\mathbf{x})\tau).$$

We now note

$$p(\tau,a|\mathbf{x},t)d\tau = P_0(\tau|\mathbf{x},t)f_a(\mathbf{x})d\tau,$$

which implies that p can be written in the form

$$p(\tau,a|\mathbf{x},t) = F(\mathbf{x})\exp(-F(\mathbf{x})\tau)\frac{f_a(\mathbf{x})}{F(\mathbf{x})}. \tag{6.8.4}$$

Hence, τ is an exponential random variable with mean and standard deviation $1/F(\mathbf{x})$, while a is a statistically independent integer random variable with $\mathbf{x}$-dependent probability $f_a(\mathbf{x})/F(\mathbf{x})$.

One exact Monte Carlo method for generating samples of the random variables τ, a is to draw two random numbers r_1, r_2 from the uniform distribution on $[0,1]$ and take

$$\tau = -\frac{1}{F(\mathbf{x})}\ln r_1 \tag{6.8.5a}$$

$$a = \text{ the smallest integer for which } \sum_{s=1}^{a} f_a(\mathbf{x}) > r_2 F(\mathbf{x}). \tag{6.8.5b}$$

The direct method of implementing the SSA is as follows:

1. Initialize the time $t = t_0$ and the chemical state $\mathbf{x} = \mathbf{x}_0$.
2. Given the state $\mathbf{x}$ at time t, determine the $f_a(\mathbf{x})$ for $a = 1,\ldots,R$ and their sums $F(\mathbf{x})$.
3. Generate values for τ and a using Eq. (6.8.5).
4. Implement the next reaction by setting $t \to t' = t + \tau$ and $x_j \to x_j' = x_j + S_{ja}/\Omega$.
5. Return to step 2 with $(\mathbf{x},t)$ replaced by $(\mathbf{x}',t')$, or else stop.

There have been a variety of subsequent algorithms that differ in the implementation of step 2, including the next reaction method [215] and the modified next reaction method [8]. The latter is based on the random time-change representation of Kurtz, which will be considered in Chap. 11 after developing the theory of martingales.

6.8.2 Tau-Leaping

In many applications the mean time between reactions, $1/F(\mathbf{x})$, is very small so that simulating every reaction becomes computationally infeasible, irrespective of the version of the SSA chosen. Gillespie [218] introduced *tau-leaping* in order to

address this problem by sacrificing some degree of exactness of the SSA in return for a gain in computational efficiency. The basic idea is to "leap" the system forward by a pre-selected time τ (distinct from the τ of the SSA), which may include several reaction events. Given $\mathbf{X}(t) = \mathbf{x}$, τ is chosen to be large enough for efficient computation but small enough so that

$$f_a(\mathbf{x}) \approx \text{ constant in } [t, t+\tau) \text{ for all } a.$$

Let $\mathscr{N}(\lambda)$ denote a Poison counting process with mean λ. During the interval $[t, t+\tau)$ there will be approximately $\mathscr{N}(\lambda_a)$ reactions of type a with $\lambda_a = f_a(\mathbf{x})\tau$. Since each of these reactions increases x_j by S_{ja}/Ω, the state at time $t+\tau$ will be

$$X_j(t+\tau) = \mathbf{x} + \sum_{a=1}^{R} \mathscr{N}_a(f_a(\mathbf{x})\tau) S_{ja}, \tag{6.8.6}$$

where the $\mathscr{N}_a$ are independent Poisson processes. This equation is known as the *tau-leaping formula*. However, there are two fundamental problems with the original formulation of tau-leaping. First, it is difficult to choose the appropriate value of τ at each iteration of the algorithm—occasionally large changes in propensities occur that cause one or more components x_j to become negative. Second, although tau-leaping becomes exact in the limit $\tau \to 0$, the inefficiency becomes prohibitive since the R generated Poisson random numbers will be zero most of the time resulting in no change of state. These two issues have been addressed in various modifications in the tau-leaping procedure (see for example Cao et al. [93]).

6.9 Exercises

Problem 6.1 (Bursting in protein translation I).

(a) Consider a single mRNA molecule which produces n proteins with probability

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

Use a generating function to show that the burst size $b \equiv \langle n \rangle = r/\gamma$.

(b) Calculate the Laplace transform

$$\tilde{P}(s) = \int_0^\infty P(n) \mathrm{e}^{-ns} dn$$

with n treated as a continuous variable (for large protein number).

(c) Evaluate the inverse Laplace transform of $\tilde{P}_m(s) = [\tilde{P}(s)]^m$ to obtain the result

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

Problem 6.2 (Bursting in protein translation II). Consider the Chapman–Kolmogorov equation (6.2.8) for protein bursting:

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial}{\partial x}[\gamma_0 x p(x)] + k\int_0^x w(x-x')c(x')p(x',t)dx',$$

with

$$w(x) = \frac{1}{b}\mathrm{e}^{-x/b} - \delta(x).$$

(a) Suppose that $c(x) = 1$ (no autoregulatory feedback). Laplace transforming the steady-state equation with respect to the protein number x, show that the stationary distribution is given by the gamma distribution

$$p(x) = \frac{1}{b^m \Gamma(m)} x^{m-1}\mathrm{e}^{-x/b}, \quad m = \frac{k}{\gamma_0}.$$

(For a more challenging problem, Laplace transform the full time-dependent equation, solve the resulting quasilinear PDE in Laplace space using the method of characteristics, and show that the system converges to the gamma distribution in the limit $t \to \infty$.)

(b) Suppose that $c(x)$ is given by the Hill function

$$c(x) = \frac{k^s}{k^s + x^s} + \varepsilon.$$

Using Laplace transforms along similar lines to part (a), show that the stationary probability density is

$$p(x) = Ax^{m(1+\varepsilon)-1}\mathrm{e}^{-x/b}[1 + (x/k)^s]^{-m/s},$$

where A is a normalization factor.

(c) Plot the stationary density of part (b) for the parameter values $m = 10, b = 20,$ and $k = 70nM$ and the following four cases: (i) $c \equiv 1$ (no feedback); (ii) $s = +1$, $\varepsilon = 0.05$; (iii) $s = -1, \varepsilon = 0.2$; (iv) $s = -4, \varepsilon = 0.2$. Hence show that negative feedback reduces noise, whereas positive feedback enhances noise and can lead to bistability.

Problem 6.3 (Binary response in stochastic gene expression). Consider the stochastic model of a gene expression in which the gene randomly switches between an active and inactive state. The steady-state probability densities $p_{0,1}(x)$ for protein concentration x when the gene is in an active ($j = 1$) or inactive ($j = 0$) state satisfy the pair of equations

$$\frac{d}{dx}(-\gamma x p_0(x)) = k_- p_1(x) - k_+ p_0(x)$$
$$\frac{d}{dx}([r - \gamma x]p_1(x)) = k_+ p_0(x) - k_- p_1(x)$$

with boundary conditions $p_0(r/\gamma) = 0$ and $p_1(0) = 0$.

(a) Derive the normalization conditions

$$\int_0^{r/\gamma} p_0(x)dx = \frac{k_-}{k_- + k_+}, \quad \int_0^{r/\gamma} p_1(x)dx = \frac{k_+}{k_- + k_+}.$$

(b) By adding the pair of steady-state equations show that one solution is

$$p_0(x) = \frac{r - \gamma x}{\gamma x} p_1(x).$$

(c) Substituting for $p_0(x)$, solve the resulting differential equation for $P(x) = (r - \gamma x)p_1(x)$, and thus obtain the solution

$$p_0(x) = C(\gamma x)^{-1+k_+/\gamma}(r - \gamma x)^{k_-/\gamma}, \quad p_1(x) = C(\gamma x)^{k_+/\gamma}(r - \gamma x)^{-1+k_-/\gamma}.$$

(d) Using part (c), show that

$$\int_0^{r/\gamma} p_0(x)dx = \frac{C}{\gamma} r^{(k_+ + k_-)/\gamma} B(k_+/\gamma, 1 + k_-/\gamma),$$
$$\int_0^{r/\gamma} p_1(x)dx = \frac{C}{\gamma} r^{(k_+ + k_-)/\gamma} B(1 + k_+/\gamma, k_-/\gamma),$$

where $B(\alpha, \beta)$ is the beta function:

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt.$$

(e) Using the standard property

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)},$$

show that the solution in part (c) satisfies the normalization conditions provided that

$$C = \gamma \left[r^{(k_+ + k_-)/\gamma} B(k_+/\gamma, k_-/\gamma) \right]^{-1}.$$

Problem 6.4 (Linear noise approximation of a two-state gene regulatory network). Consider the simple kinetic model of gene expression given by equations

$$\frac{dx_1}{dt} = k_+(1 - x_1) - k_- x_1, \quad \frac{dx_2}{dt} = r x_1 - \gamma x_2.$$

Here x_1 is the density of active genes and x_2 is the density of protein. There is a unique fixed point

$$x_1^* = \frac{k_+}{k_+ + k_-} x_{\max}, \quad x_2^* = \frac{r}{\gamma} x_1^*.$$

Applying the linear noise approximation to the corresponding master equation for finite copy numbers yields an OU process whose stationary covariance matrix Σ satisfies the matrix equation

$$\mathbf{A}\Sigma + \Sigma\mathbf{A}^T = -\mathbf{B}\mathbf{B}^T \equiv -\mathbf{D},$$

with

$$\mathbf{A} = \begin{pmatrix} -(k_+ + k_-) & 0 \\ r & -\gamma \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} 2k_- x_1^* & 0 \\ 0 & 2rx_1^* \end{pmatrix}.$$

By solving the matrix equation in component form, determine the variances σ_1^2 and σ_2^2 for $Y_i = (X_i - x_i^*)/\sqrt{\Omega}$, where Ω is the system size.

Problem 6.5 (Frequency domain analysis of a simple gene network). Consider a simple model of protein translation given by the stochastic kinetic equations

$$\frac{dx}{dt} = k - \gamma x + \eta(t), \quad \frac{dy}{dt} = rx - \gamma_p y + \eta_p(t),$$

where x and y are concentrations of mRNA and protein, γ, γ_p are degradation rates, k is the rate of mRNA production, and r is the rate of protein production. Moreover $\eta(t)$ and $\eta_p(t)$ are independent white noise terms with $\langle \eta \rangle = \langle \eta_p \rangle = 0$, and

$$\langle \eta(t)\eta(t') \rangle = q\delta(t-t'), \quad \langle \eta_p(t)\eta_p(t') \rangle = q_p\delta(t-t'), \quad \langle \eta(t)\eta_p(t') \rangle = 0.$$

(a) Introducing the Fourier transforms

$$\tilde{\eta}(\omega) = \int_{-\infty}^{\infty} \mathrm{e}^{i\omega t}\eta(t)dt, \qquad \eta(t) = \int_{-\infty}^{\infty} \mathrm{e}^{-i\omega t}\tilde{\eta}(\omega)\frac{d\omega}{2\pi},$$

show that

$$\langle \eta(\omega)\eta(\omega') \rangle = 2q\pi\delta(\omega + \omega').$$

(b) By linearizing about the steady state $x^* = k/\gamma$, $y^* = rk/(\gamma\gamma_p)$ and using Fourier transforms show that the power spectra of the fluctuations $X(t) = x(t) - x^*$ and $Y(t) = y(t) - y^*$ are given by

$$S_{XX}(\omega) = \frac{q}{\omega^2 + \gamma^2}, \quad S_{YY}(\omega) = \frac{q_p}{\omega^2 + \gamma_p^2} + \frac{r^2 q}{(\omega^2 + \gamma^2)(\omega^2 + \gamma_p^2)}.$$

(c) Using the definition of the power spectrum, written in the form

$$\langle X(t)^2 \rangle = \int_{-\infty}^{\infty} S_{XX}(\omega)\frac{d\omega}{2\pi},$$

show that

$$\langle X(t)^2 \rangle = \frac{q}{2\gamma}.$$

Similarly, show that

$$\langle Y(t)^2\rangle = \frac{q_p}{2\gamma_p} + \frac{r^2 q}{2\gamma_p \gamma^2} + \mathcal{O}(\gamma^{-3}).$$

Hint: you should assume that $\gamma \gg \gamma_p$ and use the result

$$\int_{-\infty}^{\infty} \frac{1}{\omega^2 + a^2} \frac{d\omega}{2\pi} = \frac{1}{2a}.$$

(d) From the linear noise approximation, one obtains the following Fano factors for mRNA (m) and proteins (n):

$$\frac{\text{var}[m]}{\langle m\rangle} = 1, \quad \frac{\text{var}[n]}{\langle n\rangle} = 1 + b,$$

where $b = r/\gamma$. Use this to determine q and q_p.

Problem 6.6 (Attenuation of noise in signaling cascades). Consider a generic model of a stochastic signaling cascade consisting of molecular species labeled $i = 0, \ldots, n$ with corresponding concentrations y_i [624] (see Fig. 6.22). Suppose that the rate of production of species i only depends on the concentration y_{i-1} of the species at the previous level of the cascade and that it degrades at a fixed rate γ_i. In the deterministic limit, we have a system of first-order kinetic equations:

$$\dot{y}_i + \gamma_i y_i = f_{i-1}(y_{i-1}),$$

where f_{i-1} is the corresponding production rate function and $f_{-1} = 0$. Suppose that there exists a unique stable steady state. Linearizing about the steady state and adding white noise terms to take into account intrinsic fluctuations, we have the system of linear equations

$$\delta\dot{y}_i + \gamma_i \delta y_i = c_{i-1}\delta y_{i-1} + \eta_i,$$

where

$$\langle \eta_i\rangle = 0, \quad \langle \eta_i(t)\eta_j(t')\rangle = q_i \delta_{i,j}\delta(t - t'),$$

and c_i is the derivative of f_i at the steady state (with $c_{-1} = 0$). That is, c_i can be interpreted as a differential gain or amplification factor. Both γ_i and c_i have units of t^{-1}. For convenience, set $\gamma_i = 1$ for all i.

(a) Using Fourier transforms show that

$$\langle \delta y_n^2(\omega)\rangle = \alpha_n + \beta_{n-1}\alpha_{n-1} + \beta_{n-1}\beta_{n-2}\alpha_{n-2} + \ldots + \beta_{n-1}\ldots\beta_0\langle \delta y_0^2(\omega)\rangle,$$

where

$$\alpha_j = \frac{q_j}{1+\omega^2}, \quad \beta_j = \frac{c_j^2}{1+\omega^2}.$$

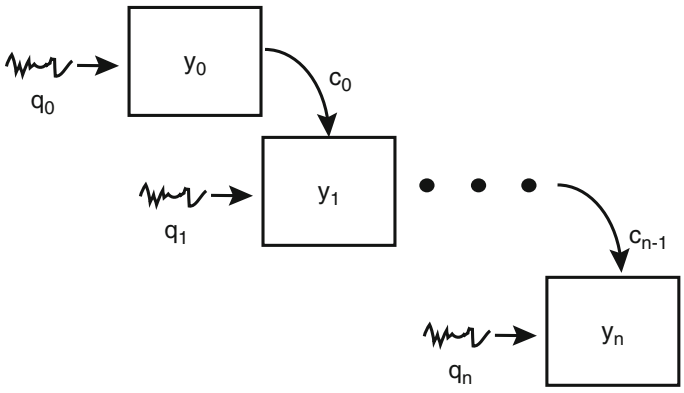

Fig. 6.22: Schematic diagram of a linearized stochastic cascade

(b) Let $q = \max_i\{q_i\}$, $c = \max_i\{c_i\}$ and define

$$\alpha = \frac{q}{1+\omega^2}, \quad \beta = \frac{c^2}{1+\omega^2}.$$

Show that

$$\lim_{n\to\infty}\langle \delta y_n^2(\omega)\rangle \le \frac{\alpha}{1-\beta}.$$

Taking the inverse Fourier transform, obtain the result

$$\lim_{n\to\infty}\langle \delta y_n^2(t)\rangle \le q\int_{-\infty}^{\infty}\frac{1}{1+\omega^2-c^2}\frac{d\omega}{2\pi} = \frac{q}{2\sqrt{1-c^2}}.$$

This establishes that fluctuations in the output of the signaling cascaded will be bounded provided that $|c_i| \le |c| < 1$.

(c) Now consider a finite cascade of length n with $c_i = c < 1$ for all $i = 0,\ldots,n$, $q_i = q$ for all $i = 1,\ldots,n$, and $q_0 > q$. Thus the noise at the input level is higher than in successive levels of the cascade. Using part (a), show that

$$\langle \delta y_n^2(\omega)\rangle = q\sum_{j=0}^{n-1}\frac{c^{2j}}{(1+\omega^2)^{j+1}} + q_0\frac{c^{2n}}{(1+\omega^2)^{n+1}}.$$

Taking inverse Fourier transforms and using contour integration, establish that

$$\langle \delta y_n^2(t)\rangle \le q\sum_{j=0}^{n-1}\frac{(2j)!}{j!j!}\frac{c^{2j}}{2^{2j+1}} + q_0\frac{(2n)!}{n!n!}\frac{c^{2n}}{2^{2n+1}}.$$

(d) Applying Stirling's approximation to part (c),

$$j! \approx \left(\frac{j}{e}\right)^j\sqrt{2\pi j},$$

obtain the inequality

$$\langle \delta y_n^2(t) \rangle \leq \frac{q}{2}\left(1+\sum_{j=1}^{n-1}\frac{c^{2j}}{\sqrt{\pi j}}\right)+\frac{q_0}{2}\frac{c^{2n}}{\sqrt{\pi n}}.$$

The first term represents an increase in intrinsic fluctuations with cascade length n, whereas the second represents a faster than exponential decrease in the input noise with cascade length. Show that the optimal cascade length for minimizing the total noise is

$$n_{\rm opt} = \left[\frac{1}{1-(q_0-q)^2/(q_0^2c^4)}\right]_-,$$

with $[x]_-$ denoting the greatest integer less than x.

Problem 6.7 (Linear noise approximation of autoregulation). Consider a simple kinetic model of gene autoregulation given by

$$\frac{dx_1}{dt} = -\gamma x_1 + F(x_2), \quad \frac{dx_2}{dt} = r x_1 - \gamma_p x_2,$$

with $F(x) = k_0 - kx$. Here x_1 is the concentration of mRNA and x_2 is the concentration of protein. There is a unique fixed point

$$x_1^* = \frac{k_0\gamma_p}{\gamma\gamma_p + kr}, \quad x_2^* = \frac{r}{\gamma_p}x_1^*.$$

Applying the linear noise approximation to the corresponding master equation for finite copy numbers yields an OU process whose stationary covariance matrix Σ satisfies the matrix equation

$$\mathbf{A}\Sigma + \Sigma\mathbf{A}^T = -\mathbf{B}\mathbf{B}^T \equiv -\mathbf{D},$$

with

$$\mathbf{A} = \begin{pmatrix} -\gamma & -k \\ r & -\gamma_p \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} kx_2^* + \gamma x_1^* & 0 \\ 0 & rx_1^* + \gamma_p x_2^* \end{pmatrix}.$$

Solving the matrix equation show that the Fano factor for proteins is

$$\frac{\mathrm{var}[n]}{\langle n \rangle} = 1 + \frac{b}{1+\eta}\left(1 - \frac{\phi}{1+b\phi}\right),$$

where $b = r/\gamma, \eta = \gamma_p/\gamma, \phi = k/\gamma_p$.

Problem 6.8 (Mutual repressor model). Consider the mutual repressor model with a single promoter site, whose stochastic version is described by the master equation (6.4.6).

(a) Suppose that the kinetics of promoter transitions are much faster than the production and degradation of proteins. Write down the continuous-time Markov equations describing the evolution of the probabilities p_j, assuming the concentrations x, y of proteins X, Y are fixed. (Recall that the proteins bind to the promoter site as dimers.) Solve for the steady-state probabilities p_j^* in terms of x and y.

(b) Under the adiabatic approximation $p_j = p_j^*(x,y)$, with $p_0^* + p_1^*$ ($p_0^* + p_2^*$) interpreted as the rate of production of protein x (y), write down the kinetic equations for x, y and use the solutions of part (a) to derive the deterministic equations

$$\frac{dx}{dt} = f(x,y), \quad \frac{dy}{dt} = f(y,x), \quad \text{with } f(x,y) = \frac{1}{1+\frac{y^2}{b+x^2}} - x.$$

Use phase-plane analysis to construct a bifurcation diagram for this planar system with b treated as a bifurcation parameter.

(c) Derive the Chapman–Kolmogorov equation (6.4.10) by carrying out a system-size expansion of the master equation expressed in the form (6.4.6).

Problem 6.9 (Model of transcriptional elongation). Consider the birth–death master equation for the elongation phase of transcription in the absence of backtracking (Sect. 6.6).

(a) Starting from the backward master equation (6.6.7) derive a difference equation for the second moment $T_2(n_0)$ of the elongation time, where n_0 is the starting position along the chain, analogous to the difference equation (6.6.8) for the first moment.

(b) Solve the difference equation in part (a) recursively by introducing the variable $U_2(n_0) = T_2(n_0) - T_2(n_0 - 1)$.

(c) Using the result from part (b) and the formula (6.6.10) for the mean elongation time, determine the variance σ^2 of the elongation time in terms of $K = \omega_-/\omega_+$ and ω_+, where $\omega_\pm$ are the effective polymerization/depolymerization rates, and show that when $K \ll 1$,

$$\sigma^2 = \frac{N}{\omega_+^2} + 4K\frac{N-1}{\omega_+^2} + O(K^2).$$

Problem 6.10 (Kinetic proofreading). Consider the kinetic proofreading model given by the modified Michaelis–Menten reaction kinetics of Eq. (6.7.8).

(a) Write down the kinetic equations for the evolution of the concentrations $[EC]$ and $[EC^*]$.

(b) Show that the steady-state concentration of the modified enzyme–substrate complex $[EC^*]$ is

$$[EC^*] = \frac{1}{q_{\text{off}} + W}\frac{rk_{\text{on}}}{k_{\text{off}} + r}[E][C].$$

(c) Repeating the analysis for the incorrect substrate D, derive the following approximation for the error rate:

$$F = \frac{Q_C}{Q_D}\frac{K_C}{K_D} = e^{-\Delta G_{CD}/k_BT} e^{-\Delta G_{C^*D^*}/k_BT}.$$

(d) Now consider the kinetic proofreading model of T-cell activation given by Eq. (6.7.11). Show that the steady-state fraction of active TCRs is

$$\frac{[B_N]}{\sum_{i=0}^{N}[B_i]} = \left(\frac{k_p}{k_p + k_{\rm off}}\right)^N.$$

Problem 6.11 (Computer simulations: gene networks). Write MatLab programs based on the Gillespie algorithm (see Sect. 6.8) that generate trajectories for each of the following gene networks.

(a) The model of regulated transcription whose master equation is given by Eq. (6.3.9). There are two discrete variables (number of active genes n_1 and number of mRNA molecules n_2) and four reactions (gene activation and deactivation, mRNA production and degradation). Take the parameter values $k_+ = 0.03\,\text{min}^{-1}, k_- = 0.2\,\text{min}^{-1}, k = 10\,\text{min}^{-1}$, and $\gamma = 0.2\,\text{min}^{-1}$ and consider the two cases $n_{\max} = 10$, and $n_{\max} = 100$. Run the simulations for sufficient time to reach steady state. Plot a histogram of $n_1(T)$ and $n_2(T)$ based on 100 simulations, say, where T is the final time. Determine the mean and variance, and compare the numerical Fano factor with the theoretical expressions based on the diffusion approximation.

(b) The mutual repressor model whose master equation is given by Eq. (6.4.3). You will have to determine the stochiometric matrix and the propensities. There are three discrete variables (number of X proteins n, number of Y proteins m, and state of the promoter) and four reactions involving changes in the promoter state, and each promoter state involves degradation and production reactions. Take the parameter values $\alpha = 1{,}000\,\text{s}^{-1}$, $\beta = 5 \times 10^5$, and $\kappa = 5 \times 10^{-5}\,\text{s}^{-1}$ and consider the two cases (i) $\gamma = 1\,\text{s}^{-1}$ (monostable) and (ii) $\gamma = 0.75\,\text{s}^{-1}$ (bistable). Plot sample trajectories over a time interval of length $T = 10\,\text{min}$ and histogram $m(T)$.

(c) The circadian clock model with stochiometry and propensities listed in Sect. 5.1. Use the following parameter values taken from [228]: $k = 0.5\,\text{nM}\,\text{h}^{-1}, \gamma = 0.3\,\text{nM}\,\text{h}^{-1}, K_m = 2.0\,\text{nM}, K'_m = 0.2\,\text{nM}$, $r = 2.0\,\text{h}^{-1}, \gamma_p = 1.5\,\text{nM}\,\text{h}^{-1}, K_p = 0.1\,\text{nM}$, and $k_1 = k_2 = 0.2\,\text{h}^{-1}$. Plot a sample trajectory of the number of mRNA M and the number of cytosolic clock proteins X_C as a function of time, and check that the oscillation period is around 22 h. Compare with solutions of the deterministic kinetic rate equations. Also plot several sample trajectories in the (M, X_C) phase plane superimposed on the deterministic limit cycle.

Problem 6.12 (Keizer's paradox.). Consider the following autocatalytic reaction scheme:

$$X \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} 2X, \quad X \xrightarrow{k_2} \emptyset.$$

(a) Write down the deterministic kinetic equation for the concentration x of the chemical species X. Show that it has an unstable fixed point at $x_1 = 0$ and a stable fixed point at $x_2 = (k_1 - k_2)/k_{-1}$.

(b) Construct the master equation for the probability $P(n,t)$ that there are n molecules of X at time t. By writing out the explicit equations for $dP(0,t)/dt$, $dP(1,t)/dt$ etc., use induction to show that the unique steady-state solution $P^*(n)$ is

$$P^*(0) = 1, \quad P^*(n) = 0, \quad n > 0.$$

Hence the stochastic model shows that there will be no X left in the system – $X = 0$ is an absorbing state. This appears to contradict the deterministic limit, which is known as Keizer's paradox.

(c) Use the Gillespie algorithm to explore the evolution of the probability distribution $P(m,t)$ as a function of time. In particular, demonstrate that at intermediate times the distribution localizes around the deterministic steady-state x_2 before eventually forming a peak around zero. Hence, provide an explanation of Keizer's paradox in terms of the non-commutativity of the operations $\lim t \to \infty$ and $\lim \Omega \to \infty$ where Ω is the system size.