To explain the origin of binary gene expression, we have to consider the fact that the intracellular rules for chemistry are quite different from the descriptions of ordinary test tube chemistry. In a cell there exists an enormous number of different reactants but sometimes only very few molecules of a kind [9, 10]. The RREs will give an incomplete description of such biochemical systems as the number of molecules may not change at a continuous rate. For a biochemical event, say, a reaction to occur, the biomolecules have to collide with each other. There is an uncertainty in the time of collision due to the probabilistic nature of intracellular diffusion [11]. The uncertainty becomes significant if the number of biomolecules participating in the reaction kinetics is small. The reaction time is a random variable varying from cell to cell and resulting in differing molecular numbers in different cells. In the case of gene expression, the process involves several biochemical reactions (e.g., the binding of RNAP at the promoter site and regulatory molecules at the appropriate regulatory sequences, ribosome binding at the translation initiation site, the degradation of the mRNA and protein etc.). The randomness in the occurrence of the biochemical events gives rise to fluctuations (commonly referred as noise) around the mean mRNA and protein levels. Thus, in a population of cells, the protein levels in individual cells at a specific instant of time are not identical but spread around a mean value. In the case of a bistable system, a transition from the low to the high expression state, say, is brought about once the fluctuations or noise associated with the low expression level cross the threshold set by the protein concentration in the unstable steady state (which falls on the boundary between the basins of attraction of
2.2 Exploring Stochastic Gene Expression

2.2.1 Factors Affecting Noise

As explained in the last section, the process of gene expression is fundamentally stochastic, with randomness inherent in the biomolecular events resulting in cell-to-cell variations in the mRNA and protein levels. Fluctuations in each component potentially affect the performance of the entire system. Let us discuss briefly the mechanisms underlying the generation of fluctuations in the expression levels of single genes. In the minimal model of gene expression, a gene can be in two possible states: inactive (G) and active (G*). Random transitions occur between the states G and G* according to the first-order kinetics shown in Figure 2.2. In the active state G*, transcription is initiated followed by translation and protein synthesis. In the inactive state G, a leaky production of mRNA occurs at a low basal rate. In certain circumstances, the amount of noise is small so that a deterministic description of the time evolution is adequate. We discuss two cases in which noise is negligible.

Large System Size

The system size should be large, i.e., large numbers of expressed mRNA/protein molecules and large cell volumes are required, with the concentrations being finite. Noise tends to increase when the size of the system is decreased. With larger numbers of the mRNA and
protein molecules, the fluctuations around the expression levels have lower amplitudes (Figure 2.3(a)). This observation has the following explanation: consider a protein that can move freely between the nucleus and the cytoplasm. At equilibrium, the concentrations of the protein in the nucleus and the cytoplasm are more or less equal. However, because the volume of the nucleus is less than that of the cytoplasm, when a protein molecule leaves the nucleus and enters the cytoplasm, the nuclear protein concentration is far more affected than the concentration of the protein in the cytoplasm. Suppose 20 molecules are present in the nucleus and 2,000 in the cytoplasm. When 1 protein molecule crosses the nuclear membrane to enter the cytoplasm, it causes a 5% change in the nuclear protein concentration, but only a 0.05% change in the cytoplasmic protein concentration. This difference arises from the difference in volumes and different number of molecules existing in the two compartments and is thus, referred to as the finite-number effect. The usual measures of noise include the coefficient of variation (CV) \( \eta \) and the Fano Factor (FF). The CV is defined to be the ratio of the standard deviation \( \sigma \) and the mean \( N \) whereas the FF = \( \frac{\sigma^2}{N} \). The value of \( \eta \) typically scales as \( 1/\sqrt{N} \) so that noise is large when the number of molecules is small.

**Fast Promoter Kinetics**

Noise can also be negligible when the transitions between the inactive and active states of the promoter occur at fast rates. The system should have a kinetics in which \( k_{on} \) and \( k_{off} \) become large, with their ratio remaining constant. Considering a simple model of gene expression in which the stochasticity is associated with only the random transition between the two gene states, \( G \) and \( G^* \), Karmakar and Bose [16] have shown that the probability density function \( \rho(x) \) describing the distribution of protein levels \( x \) (normalised by the maximum protein concentration \( x_{max} \)) in the steady state is given by the beta distribution

\[
\rho(x) = N x^{r_1-1} (1-x)^{r_2-1}
\]

where \( r_1 = \frac{k_{on}}{\gamma_p}, \ r_2 = \frac{k_{off}}{\gamma_p} \) with \( \gamma_p \) being the protein degradation rate constant. The physical interpretation of the result is as follows. If the gene is always in the inactive state, the mean protein level in the steady state is given by \( x = 0 \). If the gene is always in the active state, the steady state protein level is \( x = 1 \) which corresponds to maximal protein synthesis. When the random activation and deactivation processes are taken into account, two major possibilities arise. If the activation and deactivation rates are faster than the protein decay rate \( (r_1 > 1, r_2 > 1) \), an average intermediate protein level is obtained due to the accumulation of proteins over the random transitions between the inactive and active states of the gene. In the opposite case, i.e., when the activation and deactivation rates are slower than the protein decay rate \( (r_1 < 1, r_2 < 1) \), the mean protein level is either \( x = 0 \) or \( x = 1 \) depending on whether the gene is in the
2.2. Exploring Stochastic Gene Expression

Figure 2.3: Time series of protein concentrations observed without and with stochasticity. The blue line corresponds to the deterministic case. (a) Low-amplitude fluctuations are associated with high numbers of expressed mRNA/protein molecules and a large cell volume. Large fluctuations are observed as the cell volume is decreased. (b) Binary gene expression due to slow promoter transition.

inactive or the active state. This is because a sufficient time is available in the active state for the protein to reach the steady state level \( x = 1 \). Similarly, the residence time of the gene in the inactive state is sufficiently long so that the accumulated proteins decay fully during the period of inactivity resulting in the attainment of the steady state level \( x = 0 \). The time series and the associated histogram in the case of slow promoter transitions are shown in Figure 2.3(b). It clearly shows that the agreement between deterministic and stochastic behaviors does not hold good for slow promoter transitions. It should be noted that in this case the protein distribution has a binary nature and the origin of this binary gene expression is purely stochastic in nature. In the deterministic case, the dynamics result in monostability rather than bistability. There are now several examples of stochastic binary gene expression in eukaryotes brought about by slow promoter transitions [6, 17]. Simple models of stochastic gene expression, incorporating the essential aspects have been developed, the predictions of which are in very good agreement with experimental observations [18, 19]. The deterministic RREs provide a valid approximation of the stochastic description when the two conditions of large system size and fast promoter transitions are satisfied. If these are not fulfilled, the effects of fluctuations cannot be neglected and should be taken into account.
2.2. Exploring Stochastic Gene Expression

Recent Developments

Several experiments combined with theoretical studies provide important new insight on the stochastic aspects of gene expression \([1-8, 16, 19-23]\). The stochastic effects lead to fluctuating protein levels in individual cells and cell to cell variations in an identical population of cells kept in the same environment. The cell has evolved mechanisms to suppress/control noise conferring robustness on important cellular functions. Recent studies, however, suggest that noise is beneficial in certain cellular functions. Some organisms exploit the inherent stochasticity in the system to introduce diversity into their population. Bacteria and other microorganisms are subjected to a number of stresses during their lifetime. These single-cell creatures have to cope with stresses of various kinds, which could be environmental (like, alterations in temperature, salinity, osmolarity, pH and the availability of nutrients) or due to adverse agents (being challenged by antibiotics, bacteriophage, mutagens, toxins, radiation etc.). Microorganisms take recourse to a number of strategies for survival under stress and adapting to changed circumstances \([12, 15, 24, 25]\). A prominent feature of such strategies is the generation of phenotypic heterogeneity in an isogenic microbial population \([26-29]\). The heterogeneity is advantageous as it gives rise to variant subpopulations which are better suited to persist under stress. Some prominent examples include: lysis/lysogeny in bacteriophage, lactose utilization in \textit{E. coli} \([30]\), galactose utilization in \textit{S. cerevisiae} \([31]\) and competence development in \textit{B. subtilis} \([14, 32, 33]\).

Experiments monitoring the dynamics of gene expression in single cells provide evidence of phenotypic heterogeneity in cell populations. Several experiments combined
2.2. Exploring Stochastic Gene Expression

Figure 2.5: (a) Two reporter proteins (yellow and cyan) expressed from the same promoter; time series when (b) only intrinsic noise is present, (c) only extrinsic noise is present.

with theoretical studies are providing important new insight on the stochastic aspects of gene expression. A wide variety of tools are now available to monitor gene expression at the single cell level. The techniques commonly used include fluorescence microscopy, flow cytometry, time-lapse fluorescence microscopy etc. These methods rely on the use of reporter genes synthesizing fluorescent proteins as tags for quantitative measurements of gene expression levels in single cells. The reporter gene is usually fused at the end of the coding sequence of the gene whose expression level is to be measured, the two genes thus share the same promoter (Figure 2.4). The fluorescence intensity is a measure of the amount of proteins synthesized. The reporter gene is often placed under the direct control of a promoter instead of being fused with the gene of interest. In fluorescence microscopy, cells containing the reporter gene construct are fixed on a glass slide and visualized with the help of a phase-contrast microscope. Images of the cells are obtained using digital cameras. Flow cytometry can analyze thousands of cells in a short period of time. One thus obtains a snapshot of gene expression patterns in individual cells. Let us note that flow cytometry measures the protein levels at a single time point in each cell. Time lapse fluorescence microscopy, on the other hand, is used to study dynamic gene expression, i.e., the temporal evolution of the protein levels in individual cells.

2.2.3 Classification of Noise

Let us consider a specific gene of interest in the stochastic scenario. The amount of proteins it produces varies from cell to cell in a population. The fluctuations in the
protein levels originate in two ways: first, even if all the cells are precisely in the same state, the reaction events leading to transcription and translation of the gene would still occur at different times in different cells. Such stochastic effects are set **locally** by the gene sequence and the properties of the protein it encodes, and is referred to as **intrinsic noise**. On the other hand, there are other molecular species in the cell, e.g., RNA polymerase (RNAP), which are themselves gene products, and therefore will vary over time and from cell to cell. This variation causes additional fluctuations in the expression of the gene of interest and is termed the **extrinsic noise**. Thus, extrinsic sources of noise arise independently of the gene but influence its expression. In general, the total variation in the gene expression level has both intrinsic and extrinsic components. Elowitz et al. [8] pioneered the dual-reporter method to quantify noise in gene expression. Two fluorescent proteins (say, cyan and yellow) are expressed from identical promoters on the same chromosome (Figure 2.5(a)). Intrinsic noise gives rise to differences in the two reporter protein amounts in a single cell (Figure 2.5(b)). Extrinsic noise has the same effect on the amount of the two reporter proteins at any time instant in a single cell but causes cell-to-cell differences or changes in a single cell over time (Figure 2.5(c)). Other than this classification, noise can also be distinguished by its dependence on the state variable. If the noise does not itself depend on the state variable $x$ then it is referred to as **additive noise**; stochastic variations in the different parameters, say, the rate constants constitute **multiplicative noise**. This discrimination becomes useful in some mathematical treatments which we will discuss in detail in the coming section.

### 2.3 The Stochastic Framework

Stochasticity has significant consequences for cellular function and it is therefore important to quantify it. The formal relationships between all the different stochastic models referred to in the systems biology literature (e.g., Chapman-Kolmogorov equation, differential Chapman-Kolmogorov equation, Liouville equation, Master equation etc.) are beautifully reviewed in Ref. [34]. The stochastic time evolution of the system can be studied using the Master equation (ME) approach [35, 36]. The ME is a differential equation describing the temporal rate of change of the probability that the system is in a specific state at time $t$. The state at time $t$ is described in terms of the number of biomolecules (mRNAs, proteins etc.) present in the system at $t$. The solution of the ME gives a knowledge of the probability distribution the lower two moments of which yield the mean and the variance. One is often interested in the steady state solution of the ME, i.e., when the temporal rate of change of the probability is zero. The ME has exact, analytical solutions only in the cases of simple biochemical kinetics. A rigorous simulation technique based on the Gillespie algorithm [37, 38] provides a numerical solution of the ME. The computational cost in terms of time and computer memory can, however,
become prohibitive as the complexity of the system increases. Two approximate methods for the study of stochastic processes are based on the Langevin and Fokker-Planck (FP) equations [35, 36]. These equations are strictly valid in the case of a large numbers of molecules so that a continuous approximation is justified and the system state is defined in terms of the concentrations of molecules rather than numbers. In fact, both the equations are obtained from the ME in the large molecular number limit. In the Langevin equation (LE), additive and multiplicative stochastic terms are added to the rate equation governing the deterministic dynamics. The corresponding FP equation is a rate equation for the probability distribution [35, 36].

2.3.1 Langevin & Fokker-Planck Equations

The general stochastic formalism based on the Langevin and FP equations for the steady state analysis of a bistable system is described in Refs. [35, 36, 39-42]. We consider an one-variable LE containing a multiplicative and an additive noise term:

\[ \frac{dx}{dt} = f(x) + g_1(x)\epsilon(t) + \Gamma(t) \]  

(2.2)

where \( \epsilon(t) \) and \( \Gamma(t) \) represent Gaussian white noises with mean zero and correlations given by

\[ \langle \epsilon(t)\epsilon(t') \rangle = 2D_1\delta(t-t') \]
\[ \langle \Gamma(t)\Gamma(t') \rangle = 2D_2\delta(t-t') \]
\[ \langle \epsilon(t)\Gamma(t') \rangle = < \Gamma(t)\epsilon(t') > = 2\lambda\sqrt{D_1D_2}\delta(t-t') \]  

(2.3)

\( D_1 \) and \( D_2 \) are the strengths of the two types of noise \( \epsilon(t) \) and \( \Gamma(t) \) respectively and \( \lambda \) is the degree of correlation between them. The first term, \( f(x) \), in Eq. (2.2) represents the deterministic dynamics, i.e., it is the function in the r.h.s. of the deterministic RRE, \( \frac{dx}{dt} = f(x) \).

The additive noise \( \Gamma(t) \) represents noise arising from an external perturbative influence or originating from some missing information embodied in the rate equation approximation [13]. On the other hand, in many of the models of gene expression, some of the elementary processes are lumped together and an effective rate constant associated with the combined process. It is, however, expected that the rate constants fluctuate in time due to a variety of stochastic influences like fluctuations in the number of regulatory molecules and RNA polymerases. In the LE, the fluctuations in the rate constants are taken into account through the inclusion of multiplicative noise terms like \( g_1(x)\epsilon(t) \) in Eq. (2.2). To understand the structure of \( g_1(x) \), let us consider an example where, \( f(x) = a + \frac{bx}{1+e^x} \), where \( a \) and \( b \) are certain parameters/rate constants. If we consider the parameter \( b \) to vary stochastically, i.e., \( b \rightarrow b + \epsilon(t) \), we have
2.3. The Stochastic Framework

\begin{equation}
g_1(x) = \frac{x}{1 + x}
\end{equation}

As mentioned before, \( e(t) \) represents random fluctuations of the Gaussian white noise-type. There are alternative versions of the LE in which the noise terms added to the deterministic part have an explicit structure in terms of gene expression parameters \[38, 43, 44\]. Gillespie [38] has shown how to derive the Chemical LE starting with the Master equation describing the stochastic time evolution of a set of elementary reactions with the noise terms depending on the number of molecules as well as the gene expression parameters. Similarly, for an effective kinetic equation of the form

\begin{equation}
\frac{dn}{dt} = f(n) - g(n)
\end{equation}

where \( n \) is the number of molecules and \( f(n), g(n) \) are the synthesis and decay rates respectively, one can write the LE as \[43, 44\]

\begin{equation}
\frac{dn}{dt} = f(n) - g(n) + \sqrt{f(n) + g(n)} \Gamma(t)
\end{equation}

While these alternative approaches have their own merit, the majority of stochastic gene expression studies, based on the Langevin formalism, start with equations of the type shown in (2.2). The choice is dictated by the simplicity of the calculational scheme with specific focus on the separate effects of additive and multiplicative noise (fluctuating rate constants) on the gene expression dynamics.

The FP equation can be developed from the LE following usual procedure [13, 36, 42]. The FP equation corresponding to Eq. (2.2) is [13, 36, 42]

\begin{equation}
\frac{\partial P(x,t)}{\partial t} = - \frac{\partial}{\partial x}[A(x)P(x,t)] + \frac{\partial^2}{\partial x^2}[B(x)P(x,t)]
\end{equation}

where

\begin{equation}
A(x) = f(x) + D_1 g_1(x) + \lambda \sqrt{D_1 D_2} g_1(x)
\end{equation}

and

\begin{equation}
B(x) = D_1 [g_1(x)]^2 + 2\lambda \sqrt{D_1 D_2} g_1(x) + D_2
\end{equation}

The steady state probability distribution (SSPD), from Eq. (2.7), is given by \[36, 39, 40\]
where \( N \) is the normalization constant. Eq. (2.10) can be recast in the form

\[
P_{st}(x) = Ne^{-\phi_F(x)}
\]  

(2.11)

with

\[
\phi_F(x) = \frac{1}{2} \ln[D_1|g_1(x)|^2 + 2\lambda \sqrt{D_1 D_2} g_1(x) + D_2] - \int_x^\infty \frac{f(y)dy}{D_1[g_1(y)]^2 + 2\lambda \sqrt{D_1 D_2} g_1(y) + D_2}
\]  

(2.12)

where \( \phi_F(x) \) defines the stochastic potential corresponding to the FP equation.

Let us now consider the case when only the additive noise term is present in Eq. (2.2), i.e., the second term on the r.h.s. is missing. From Eqs. (2.10) and (2.12), the steady state probability distributions, \( P_{st}(x) \), and the associated expression potentials, \( \phi_F(x) \), can be computed. The stochastic potential \( \phi_F(x) \) has the form

\[
\phi_F(x) = \frac{1}{2} \ln D_2 - \frac{1}{D_2} \int_x^\infty f(y)dy
\]  

(2.13)

\[
= \frac{1}{2} \ln D_2 + \frac{1}{D_2} \phi_D(x)
\]  

(2.14)

where \( \phi_D(x) \) is the deterministic potential, i.e., \( f(x) = -\frac{\partial \phi_D(x)}{\partial x} \), as described in Section 1.2.4. It implies that in the presence of only additive noise, the functional forms of the stochastic and deterministic potentials are similar except some additive/multiplicative constants.

2.3.2 Gillespie Simulation

The stochastic time evolution of a dynamical system can be studied through computer simulation based on the Gillespie algorithm (GA) [37]. We now discuss briefly the salient features of the GA, more detailed discussions can be obtained from Refs. [37, 45, 46].
Algorithm

Let $M$ be the number of reactions controlling the time evolution of molecular numbers in a dynamical system. Each reaction is characterized by a reaction propensity $a_\mu (\mu = 1, 2, \ldots, M)$ with $a_\mu (t) dt$ defining the probability that the reaction $\mu$ occurs in volume $V$ in the time interval $(t, t+dt)$ given the state of the system at time $t$. The propensity $a_\mu (t)$ is a product of two parts, the reaction rate $c_\mu$ for reaction $\mu$ and the number of possible reactions $\mu$ in volume $V$. One also defines a reaction stoichiometry matrix $S$ which is an $N \times M$ matrix where $N$ is the number of reactants (different species of molecules) in the dynamical system. The element $S_{ij}$ ($i = 1, \ldots, N, \ j = 1, \ldots, M$) denotes the change in the number of reactant $i$ molecules, from $X_i$ to $X_i + S_{ij}$, when the $j$th reaction takes place. The deterministic rate equation can be written in a compact form

$$\frac{dX}{dt} = S\nu$$

(2.15)

The GA monitors time evolution by obtaining information firstly on the time of occurrence of the next reaction given the state of the system in terms of molecular numbers at time $t$ and secondly the reaction type. Let us define a quantity $a_0$ as, $a_0 = \sum_{\mu=1}^{M} a_\mu$. With knowledge of the state of the system at time $t$, the probability that the next reaction occurs in the time interval $t + \tau$ and $t + \tau + d\tau$ and is of type $\mu$ is $P_\mu(\tau)d\tau$ where

$$P_\mu(\tau) = a_\mu \exp(-a_0 \tau)$$

$$= \frac{a_\mu}{a_0} a_0 \exp(-a_0 \tau)$$

$$= P_2(\mu) P_1(\tau)$$

(2.16)

with $P_1(\tau) = a_0 \exp(-a_0 \tau)$ and $P_2(\mu) = \frac{a_\mu}{a_0}$. The GA generates two random numbers $r_1$ and $r_2$ using a standard uniform random number generator. The time $\tau$ is then given by [37]

$$\tau = \frac{1}{a_0} \ln \left( \frac{1}{r_1} \right)$$

(2.17)

The reaction type $\mu$ is taken to be the integer for which the condition

$$\sum_{\nu=1}^{\mu-1} a_\nu < r_2 a_0 \leq \sum_{\nu=1}^{\mu} a_\nu$$

(2.18)

is satisfied. The random number $\tau$ obtained from Eq. (2.17) is generated according to the probability distribution $P_1(\tau)$ whereas Eq. (2.18) generates the random integer $\mu$ according to the probability distribution $P_2(\mu)$. Once the pair $(\tau, \mu)$ is determined, the
Figure 2.6: Flowchart of Stochastic Simulation Algorithms.

time \( t \) is advanced by \( \tau \), i.e., \( t \rightarrow t + \tau \) and the molecular numbers, \( X_i \)'s, \( (i = 1, \cdots, N) \) are adjusted according to the reaction \( \mu \). The SSPD \( P(X_i) \) of a molecular type can be computed by combining the data over a sufficiently large interval of time after the steady state conditions are achieved.

**Disadvantages & Alternatives**

Although the Gillespie algorithm solves the Master Equation describing the reaction networks in cells exactly, a disadvantage is that many trajectories are needed for an accurate estimation of the time-dependent solution of the Master equation and only long time-integrations yield accurate results for its steady state solution. It requires substantial amounts of computational effort to simulate a complex system. Three situations cause an increase in the computational complexity of the Gillespie algorithm:

- Increase in the number of possible reactions,
- Increase in the number of molecules of the species,
- Faster reaction rate.

These conditions decrease the step time of each iteration thus forcing the algorithm to run for a larger number of iterations to simulate a given experiment. These factors also cause great disparities in the timescales of reaction channels in the system. In fact, in order
to maintain the exactness of the simulation, the algorithm requires shorter time steps to capture the fast dynamics of the system. Thus, the difference in the time-scales between different reaction channels may sometimes cause computational difficulties. Since the development of the direct method in [37], many variants of the stochastic simulation algorithm have been proposed in order to improve the performance of the algorithm. For example, while the direct method is an exact procedure, the Tau-Leaping Method, developed by Gillespie [47, 48] is an approximation based on taking larger time leaps in the time evolution of the system. More than one reaction are executed in the time interval $\tau$ and the order in which they are executed is not important. Most of these variants are compiled in Figure 2.6 in the form of a flowchart.

2.4 Summary

Cellular biology is characterized by random dynamics and chance processes. Many vital molecules and pieces of the cellular machinery diffuse within cells, moving along random trajectories as they collide with the other biomolecular inhabitants of the cell. Cellular components may block each other's progress, be produced or degraded at random times, and become unevenly separated as cells grow and divide. Cellular behaviour is profoundly influenced by the noise associated with gene expression levels. Theoretical models of stochastic processes and the explosion of results on stochastic gene expression in single cell/molecule experiments provide new insights on the origins, control and consequences of noise. We have discussed in this chapter some of the basic concepts and techniques for probing the random nature of cellular processes which are integral to the studies carried out in the thesis.
Bibliography


