

RICE UNIVERSITY

Modeling stochasticity in gene regulation

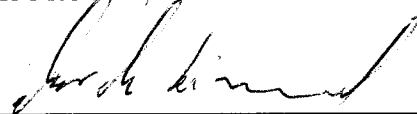
by

Pawel Paszek

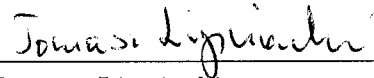
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:



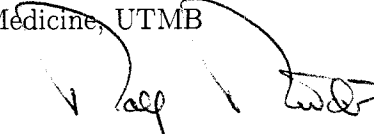
Marek Kimmel, Professor, Chair, Co-advisor
Statistics, RU



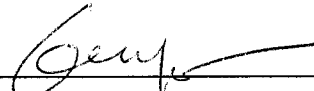
Tomasz Lipniacki, Assistant Professor,
Co-advisor
Fluid Mechanics, IPPT



Allan Brasier, Professor
Medicine, UTMB



Rudolf Riedi, Associate Professor
Statistics, RU



Ka-Yiu San, Professor
Bioengineering and Chemical Engineering, RU

Houston, Texas

May, 2006

UMI Number: 3216759

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI[®]

UMI Microform 3216759

Copyright 2006 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Modeling stochasticity in gene regulation

by

Pawel Paszek

Abstract

Intrinsic stochasticity plays an essential role in gene regulation because of the small number of involved molecules of DNA, mRNA and protein of a given species. To better understand this phenomenon, small gene regulatory systems are mathematically modeled as systems of coupled chemical reactions, but the existing exact description utilizing a Chapman-Kolmogorov equation or simulation algorithms is limited and inefficient. The present work introduces a much more efficient yet accurate modeling approach, which allows analyzing stochasticity in the system in terms of the underlying distribution function.

The novel modeling approach is motivated by the analysis of a single gene regulatory module with three sources of stochasticity: intermittent gene activity, mRNA transcription/decay and protein translation/decay noise. Although the corresponding Chapman-Kolmogorov equation cannot be solved when a large number of molecules are considered, it is used to analytically derive the first two moments of the underlying distribution function. The mRNA and protein variance is found decomposable into additive terms resulting from the respective sources of stochasticity, which allow quantifying their significance in the process.

The variance decomposition is asserted by constructing two approximations that establish

scription and *translation*. First, the genes are transcribed into singlestranded ribonucleic acid (RNA) which is complementary up to some extent to the original DNA strand. Transcription is carried out by large enzymes called RNA polymerases, which repeatedly attach themselves to the transcription initiation site and synthesize RNA while moving along DNA. When RNA polymerase finally reaches the termination site, it releases the transcript and dissociates from DNA. There are three major classes of RNA involved in further protein synthesis: messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). Primary transcript undergoes several modifications to generate mature mRNA molecules. Mature mRNA may carry some functions but in general they are lacking the chemical complexity necessary to be efficient. Instead, in the second step of gene expression, mRNA serves as a template in the protein synthesis. The mRNA transcript is repeatedly translated by ribosomes to form polypeptide chains built from amino acids, while tRNA and rRNA perform structural and catalytic roles in this process. Every mRNA transcript molecule contains at least one coding region that is related to a protein sequence by the genetic code: Each nucleotide triplet of the coding region represents one amino acid ([39], p. 155).

1.3 Gene regulation

The process of gene regulation is, in general, much more involved in eukaryotes than in prokaryotes, however basic regulatory mechanisms include mRNA transcription control, transcript processing and protein translation. The differences arise due to the size and the complexity of the eukaryotic cell. Eukaryotic cells are compartmentalized, which requires that mRNA is synthesized and matures solely in the nucleus and then is exported into the cytoplasm and translated. In prokaryotes, transcription and translation occurs almost simultaneously in the single cellular compartment. Ribosomes attach to bacterial mRNA even

before the transcript has been completed.

Controlling the rate of transcriptional initiation constitutes the major gene regulation mechanism in both prokaryotes and eukaryotes ([57], p. 543). In prokaryotes, genes sharing similar biological functions are grouped together in clusters called operons (e.g. *lac* and *trp* operon) to be simultaneously expressed from the same DNA regulatory sites ([39], [39], p. 339). These sites are referred to as *promoter regions*, *promoter sequences* or *cis-acting elements*, because they promote the recognition of transcription initiation site by the RNA polymerase. Interactions at the promoter site(s) involve binding of regulatory molecules (*trans-acting elements*), which determine whether genes in the cluster are being expressed or not. The most common mode of the transcriptional control in bacteria is negative: A repressor protein prevents a gene from being expressed by targeting another *cis-acting element* called the operator. Due to the repressor bound at the operator site, RNA polymerase cannot initiate transcription. The opposite positive mode of control is implemented when a *transcription factor* (another *trans-acting element*) is required to assist RNA polymerase in transcription initiation. In general, RNA polymerase activity at the specific promoter site can be modified by numerous regulatory proteins called activators or repressors depending whether they favor or permit the polymerase binding.

In eukaryotes, gene regulation is far more complex, however the transcription initiation plays also an essential role ([68], p. 3-7). Eukaryotes employ three types of RNA polymerases, namely polymerase I, II (transcribing majority of the genes) and III. Their binding efficiency and specificity is accomplished through series of mechanisms involving regulatory proteins referred to as *transcription factors* (*trans-acting elements*) ([57], p. 548). Transcription factors recognize regulatory sequences of genes (*cis-acting elements*) by numerous mechanisms including homeodomains, POU domains, bZip domains, bHLH domains, zinc

fingers, TATA-binding proteins, and many other ([68], p. 25-50). There are two types of regulatory elements in eukaryotes: *promoters* and *enhancers*. As in prokaryotes, eukaryotic promoters are located immediately adjacent to the genes they regulate and they require binding of numerous protein factors to initiate transcription. Most of the promoters contain specific binding domains such as TATA, CAAT and GC box, but there is a large variation in their organization and localization among different genes. To assure fast and efficient transcription, the promoter machinery is augmented by series of more remote enhancer sites allowing recognition of specific genes by transcription factors. There is some analogy between enhancer and operator regions in prokaryotes, however the former are much more complex in their structure and function.

In addition to transcription initiation, gene regulation can be possibly controlled at other different stages, which include primary RNA processing, mRNA stability, protein translation initiation and termination, as well as protein posttranslational modifications. In addition, mRNA export from the nucleus can be actively regulated in eukaryotes.

Complex organization of the eukaryotic cell requires another level of control of gene regulation. Eukaryotic DNA is condensed to literally fit inside the nucleus (stretched human DNA is about 2 meters long, while nucleus has 6-8 μm in diameter, [68], p. 14): About 120-160 bp long pieces of duplex DNA are wrapped around histones and other regulatory proteins and tightly packed into chromatin structure. Such structure is not accessible to neither RNA polymerase nor transcription factors. Therefore expressing any gene requires remodeling and unfolding of the corresponding DNA region, which is accomplished by histone acetylation [69], [21] and [11].

1.4 Sources of stochasticity

Intrinsic stochasticity in gene regulation is well recognized because of a small numbers of involved molecules of DNA, mRNA and protein of a given species.

In prokaryotes, stochasticity is attributable primarily to mRNA polymerase binding, followed by mRNA transcription and protein translation [1], [41], [3], [29]. Prokaryotic cells are relatively small and haploid. Bacterial mRNA is typically very unstable (half-life time at the order of 1 min), therefore at a given time, there usually exist only several copies of corresponding mRNA transcript (sometimes as few as one or two) and tens of copies of protein. This implies that the production or degradation of a single mRNA or protein molecule has a significant effect on the cell's behavior [13].

On the other hand, eukaryotes are diploid: Each gene has two homologous copies, which can be independently activated and repressed. In some cases one of these copies may become transcriptionally inactive. Moreover, some cells may have gene or chromosomal duplications which lead to a larger number of homologous gene copies. Corresponding mRNA and protein levels are much higher than in prokaryotes, with up to hundreds of mRNA molecules and hundreds of thousands of protein molecules of a given species (e.g. molecules involved in NF- κ B regulatory pathway [36], [37]). This implies that stochastic effects due to intermittent gene activity (gene activation and repression) followed by pulses of mRNA production are much stronger than the stochastic effects caused by production or degradation of single mRNA or protein molecule. A gene may be activated even by a single *trans*-activating regulatory protein, which binds to the promoter region and allows very fast and efficient mRNA transcription and production of a burst of protein molecules. As long as the transcription factor is not bound to the promoter region, the resulting transcription initiation frequency remains low since it is triggered only by erratic RNA polymerase binding. Recently, a grow-

ing number of experiments on eukaryotic cells supports this hypothesis: Fluorescence in situ hybridization analysis of β -actin transcription sites reveals cyclical mRNA transcription from a single gene [15]. Dual-reporter measurements of intrinsic noise in gene expression in budding yeast *Saccharomyces cerevisiae* disclose effects due to the intermittent gene activity [49]. Oscillations in NF- κ B-dependent gene products measured at the single cell level in HeLa (human ovarian carcinoma) and SK-N-AS (human S-type neuroblastoma) cells are demonstrated to arise through the action of a transcription factor [45], [4], [46] and [37]. Additional stochasticity in eukaryotic gene regulation corresponds to the DNA-chromatin interactions. Normally condensed and tightly packed DNA must be remodeled to become accessible to either RNA polymerase or regulatory proteins, which is accomplished by histone acetylation [49].

1.5 Overview of the thesis

The present chapter (Chapter 1) briefly presents a biological background of the problem stressing the basis of the gene regulation and the significance of various stochastic effects in the process.

Chapter 2 introduces basic notions of Markov processes, which are extensively applied to mathematically describe gene regulatory systems. In particular it provides a derivation of the differential Chapman-Kolmogorov equation, which captures the time evolution of the underlying distribution function. In addition, it provides a literature overview summarizing recent attempts in modeling stochasticity in gene regulation and their biological implications.

Chapter 3 presents analysis of the single gene regulatory module with three major sources of stochasticity, namely: intermittent gene activity, mRNA transcription/decay and protein translation/decay noise. Based on the analysis of the corresponding Chapman-Kolmogorov

equation two approximations to the exact stochastic description are introduced: First, the continuous model, which considers only the stochasticity due to the intermittent gene activity. Second, the mixed model, which considers stochasticity due to the intermittent gene activity as well as the mRNA transcription/decay noise.

Chapter 4 includes the comparison between the marginal protein distribution resulting from the continuous approximation and the corresponding distribution given by the Kepler-Elston model [30]. In the case when the Kepler-Elston approximation is satisfactory, it is used to analyze two-gene systems.

Chapter 5 includes analysis of the regulatory system in the case when the process of gene activity is governed by the collective actions of multiple regulatory factors. Based on the gene expression data, developed models are applied to hypothesize the existence of a sequential activation mechanism of NF- κ B dependent genes important in cell survival and inflammation.

Finally, the thesis are concluded in Chapter 6 with discussion that stresses mathematical and biological implications of the presented work.

Chapter 2

Previous modeling of stochastic effects in gene regulation

2.1 Gene regulation as a Markov process

Gene regulatory systems describing the interactions between the DNA, mRNA and protein molecules at the single cell level may be considered as systems of coupled chemical reactions. Under assumption of spatial homogeneity the stochastic process governing chemical reactions is a Markov process.

2.1.1 Basic introduction to Markov processes

In general, a stochastic process describes dynamics of certain time-dependent random variable $\mathbf{X}(t)$. The process, which governs the evolution of $\mathbf{X}(t)$, can be completely described by a set of joint probability densities

$$p(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \mathbf{x}_3, t_3, \dots)$$

given the realization of $\mathbf{X}(t)$, i.e., $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots$ at times t_1, t_2, t_3, \dots [17]. One can also define conditional probability densities given the joint probability density function:

$$\begin{aligned} & p(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \dots | \mathbf{y}_1, \tau_1; \mathbf{y}_2, \tau_2; \dots) \\ = & p(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \dots; \mathbf{y}_1, \tau_1; \mathbf{y}_2, \tau_2; \dots) / p(\mathbf{y}_1, \tau_1; \mathbf{y}_2, \tau_2; \dots), \end{aligned}$$

given that $p(\mathbf{y}_1, \tau_1; \mathbf{y}_2, \tau_2; \dots) > 0$ and assuming the time ordering $t_1 \geq t_2 \geq t_3 \geq \dots \geq \tau_1 \geq \tau_2 \dots$. This interprets conditional probabilities as predictions of $\mathbf{X}(t)$, i.e., $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots$ at times t_1, t_2, t_3, \dots , given the past $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \dots$ at times $\tau_1, \tau_2, \tau_3, \dots$.

One of the most important examples of stochastic processes are processes satisfying the *Markov property* and known as *Markov processes*. The *Markov property* is defined based on the conditional probability function and assumes that the future values of $\mathbf{X}(t)$ depend entirely on the present and are not affected by the past, i.e.,

$$p(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \dots | \mathbf{y}_1, \tau_1; \mathbf{y}_2, \tau_2; \mathbf{y}_3, \tau_3 \dots) = p(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \dots | \mathbf{y}_1, \tau_1).$$

Using the Markov property, an arbitrary joint probability density function can be expressed in the terms of the conditional probability densities, i.e.,

$$p(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2; \dots; \mathbf{x}_n, t_n) = p(\mathbf{x}_1, t_1 | \mathbf{x}_2, t_2) p(\mathbf{x}_2, t_2 | \mathbf{x}_3, t_3) \dots p(\mathbf{x}_{n-1}, t_{n-1} | \mathbf{x}_n, t_n).$$

Therefore, the process governing $\mathbf{X}(t)$ can be uniquely defined by all conditional probabilities $p(x_i, t_i | x_j, t_j)$, where $i, j = 1, 2, 3, \dots$ and $t_1 \geq t_2 \geq t_3 \geq \dots$. Relationship between such conditional probabilities is given by the following equation and is valid for all stochastic processes:

$$\begin{aligned} p(\mathbf{x}_1, t_1 | \mathbf{x}_3, t_3) &= \int p(\mathbf{x}_1, t_1; \mathbf{x}_2, t_2 | \mathbf{x}_3, t_3) d\mathbf{x}_2 \\ &= \int p(\mathbf{x}_1, t_1 | \mathbf{x}_2, t_2; \mathbf{x}_3, t_3) p(\mathbf{x}_2, t_2 | \mathbf{x}_3, t_3) d\mathbf{x}_2, \end{aligned}$$

but introducing the Markov property yields the fundamental equation, which is referred

to as the *Chapman-Kolmogorov equation*:

$$p(\mathbf{x}_1, t_1 | \mathbf{x}_3, t_3) = \int p(\mathbf{x}_1, t_1 | \mathbf{x}_2, t_2) p(\mathbf{x}_2, t_2 | \mathbf{x}_3, t_3) d\mathbf{x}_2. \quad (2.1)$$

The Chapman-Kolmogorov equation is an identity, satisfied by the conditional probability densities of any Markov process. The time ordering is essential, i.e., $t_1 \geq t_2 \geq t_3$ and thus Eq. (2.1) follows by integrating out all possible paths leading from the state \mathbf{x}_3 to state \mathbf{x}_1 .

In the case of the discrete state space, when the underlying random variable takes on integer values and is denoted with $\mathbf{N}(t)$, the Chapman-Kolmogorov equation yields:

$$P(\mathbf{n}_1, t_1 | \mathbf{n}_3, t_3) = \sum_{\mathbf{n}_2} P(\mathbf{n}_1, t_1 | \mathbf{n}_2, t_2) P(\mathbf{n}_2, t_2 | \mathbf{n}_3, t_3), \quad (2.2)$$

and corresponds to matrix multiplication, with possibly infinite matrices. $P(\mathbf{n}_1, t_1 | \mathbf{n}_2, t_2)$ is the matrix of all possible transition probabilities from the state \mathbf{n}_2 to state \mathbf{n}_1 .

In the case of the systems of reacting molecules, such as gene regulatory networks, the discrete state space is a natural domain for the process. The state variable $\mathbf{N}(t)$ corresponds to the vector of amounts of considered molecular species at a given instant of time, which may only assume positive integer values.

2.1.2 Derivations of the differential Chapman-Kolmogorov equation

By imposing certain restrictions on the conditional probability function connected with the intuitive notion of continuous movement of the system, the Chapman-Kolmogorov equation (2.1) can be reduced to a differential equation. This idea was introduced by Kolmogorov (1931) [35], however the following derivations are based on Gardiner (2003) [17].

First, note that with probability one, sample paths of a Markov process are continuous

functions of time t if for any $\varepsilon > 0$

$$\lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{|\mathbf{x}-\mathbf{z}|>\varepsilon} p(\mathbf{x}, t + \Delta t | \mathbf{z}, t) d\mathbf{x} = 0 \quad (2.3)$$

uniformly in \mathbf{x} , \mathbf{z} and t . Eq. (2.3) is known as the *Lindenberg condition* and means that the probability that the final state \mathbf{x} is different from the initial state \mathbf{z} tends to zero faster than Δt , as $\Delta t \rightarrow 0$.

Based on the Lindenberg condition, derivations of the differential Chapman-Kolmogorov equation rely on the method of dividing the differentiability conditions into parts: One connected with the continuous motion of a representative particle and the other with the discontinuous motion. Therefore, assume that for all $\varepsilon > 0$:

$$\lim_{\Delta t \rightarrow 0} \frac{p(\mathbf{x}, t + \Delta t | \mathbf{z}, t)}{\Delta t} = W(\mathbf{z} | \mathbf{x}, t) \text{ uniformly in } \mathbf{x}, \mathbf{z} \text{ and } t \text{ for all } |\mathbf{x} - \mathbf{z}| \geq \varepsilon, \quad (2.4)$$

$$\lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{|\mathbf{x}-\mathbf{z}|<\varepsilon} (x_i - z_i) p(\mathbf{x}, t + \Delta t | \mathbf{z}, t) d\mathbf{x} = A_i(\mathbf{z}, t) + O(\varepsilon), \quad (2.5)$$

$$\lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{|\mathbf{x}-\mathbf{z}|<\varepsilon} (x_i - z_i)(x_j - z_j) p(\mathbf{x}, t + \Delta t | \mathbf{z}, t) d\mathbf{x} = B_{ij}(\mathbf{z}, t) + O(\varepsilon), \quad (2.6)$$

where the last two being uniform in \mathbf{z} , ε and t . It can be shown that the higher-order coefficients of the form (2.5) and (2.6) must vanish ([17], p. 48). According to the Lindenberg condition the process has continuous paths only when $W(\mathbf{z} | \mathbf{x}, t)$ vanishes for all $\mathbf{z} \neq \mathbf{x}$, therefore the function $W(\mathbf{z} | \mathbf{x}, t)$ describes the discontinuities in the path of the process, while

A_i and B_{ij} are connected with the continuous part of the path.

Consider time evolution of the expectation of a twice continuously differentiable function $f(\mathbf{z})$. Then,

$$\partial_t \int f(\mathbf{x})p(\mathbf{x}, t|\mathbf{y}, t')d\mathbf{x} = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int f(\mathbf{x})[p(\mathbf{x}, t + \Delta t|\mathbf{y}, t') - p(\mathbf{x}, t|\mathbf{y}, t')]d\mathbf{x},$$

but from the Chapman-Kolmogorov equation, Eq. (2.1):

$$p(\mathbf{x}, t + \Delta t|\mathbf{y}, t') = \int p(\mathbf{x}, t + \Delta t|\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t')d\mathbf{z},$$

therefore, following a change in the integration order,

$$\begin{aligned} & \partial_t \int f(\mathbf{x})p(\mathbf{x}, t|\mathbf{y}, t')d\mathbf{x} \\ &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \left\{ \int \int f(\mathbf{x})p(\mathbf{x}, t + \Delta t|\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t')d\mathbf{x}d\mathbf{z} - \int f(\mathbf{z})p(\mathbf{z}, t|\mathbf{y}, t')d\mathbf{z} \right\}. \end{aligned} \quad (2.7)$$

The integral with respect to \mathbf{x} can be divided into two regions $|\mathbf{x} - \mathbf{z}| \geq \varepsilon$ and $|\mathbf{x} - \mathbf{z}| < \varepsilon$. For $|\mathbf{x} - \mathbf{z}| < \varepsilon$, since $f(\mathbf{z})$ is assumed twice continuously differentiable, $f(\mathbf{x})$ is given by:

$$f(\mathbf{x}) = f(\mathbf{z}) + \sum_i \frac{\partial f(\mathbf{z})}{\partial z_i}(x_i - z_i) + \sum_{i,j} \frac{1}{2} \frac{\partial^2 f(\mathbf{z})}{\partial z_i \partial z_j}(x_i - z_i)(x_j - z_j) + |\mathbf{x} - \mathbf{z}|^2 R(\mathbf{x}, \mathbf{z}), \quad (2.8)$$

where $|R(\mathbf{x}, \mathbf{z})| \rightarrow 0$ as $|\mathbf{x} - \mathbf{z}| \rightarrow 0$. Now, Eq. (2.7) reads,

$$\begin{aligned}
\partial_t \int f(\mathbf{x})p(\mathbf{x}, t|\mathbf{y}, t')d\mathbf{x} &= \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \left\{ \iint_{|\mathbf{x}-\mathbf{z}|<\varepsilon} \left[\sum_i \frac{\partial f(\mathbf{z})}{\partial z_i} (x_i - z_i) + \sum_{i,j} \frac{1}{2} \frac{\partial^2 f(\mathbf{z})}{\partial z_i \partial z_j} (x_i - z_i)(x_j - z_j) \right] \right. \\
&\quad \times p(\mathbf{x}, t + \Delta t|\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t')d\mathbf{x}d\mathbf{z} \\
&\quad + \iint_{|\mathbf{x}-\mathbf{z}|<\varepsilon} |\mathbf{x} - \mathbf{z}|^2 R(\mathbf{x}, \mathbf{z})p(\mathbf{x}, t + \Delta t|\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t')d\mathbf{x}d\mathbf{z} \\
&\quad + \iint_{|\mathbf{x}-\mathbf{z}|<\varepsilon} f(\mathbf{z})p(\mathbf{x}, t + \Delta t|\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t')d\mathbf{x}d\mathbf{z} \\
&\quad + \iint_{|\mathbf{x}-\mathbf{z}|\geq\varepsilon} f(\mathbf{x})p(\mathbf{x}, t + \Delta t|\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t')d\mathbf{x}d\mathbf{z} \\
&\quad \left. - \iint_{|\mathbf{x}-\mathbf{z}|\geq\varepsilon} f(\mathbf{z})p(\mathbf{x}, t + \Delta t|\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t')d\mathbf{x}d\mathbf{z} \right\}, \tag{2.9}
\end{aligned}$$

where the last line gives the negative term in Eq. (2.7) since the integral of $p(\mathbf{x}, t + \Delta t|\mathbf{z}, t)$ with respect to \mathbf{x} is equal to 1. Eq. (2.9) can be inspected line by line:

Lines 1 and 2: By assumed uniform convergence, interchange of the limit and integration with assumption (2.5) and (2.6) gives

$$\int \left[\sum_i A_i(\mathbf{z}, t) \frac{\partial f(\mathbf{z})}{\partial z_i} + \sum_{i,j} \frac{1}{2} B_{ij}(\mathbf{z}, t) \frac{\partial^2 f(\mathbf{z})}{\partial z_i \partial z_j} \right] p(\mathbf{z}, t|\mathbf{y}, t')d\mathbf{z} + O(\varepsilon). \tag{2.10}$$

Line 3: We have,

$$\begin{aligned}
& \left| \frac{1}{\Delta t} \int_{|\mathbf{x}-\mathbf{z}|<\varepsilon} |\mathbf{x}-\mathbf{z}|^2 R(\mathbf{x}, \mathbf{z}) p(\mathbf{x}, t + \Delta t | \mathbf{z}, t) d\mathbf{x} \right| \\
& \leq \left[\frac{1}{\Delta t} \int_{|\mathbf{x}-\mathbf{z}|<\varepsilon} |\mathbf{x}-\mathbf{z}|^2 p(\mathbf{x}, t + \Delta t | \mathbf{z}, t) d\mathbf{x} \right] \underset{|\mathbf{x}-\mathbf{z}|<\varepsilon}{Max} |R(\mathbf{x}, \mathbf{z})| \\
& \rightarrow \left[\sum_i B_{ii}(\mathbf{z}, t) + O(\varepsilon) \right] \left\{ \underset{|\mathbf{x}-\mathbf{z}|<\varepsilon}{Max} |R(\mathbf{x}, \mathbf{z})| \right\}, \tag{2.11}
\end{aligned}$$

but as $\varepsilon \rightarrow 0$ $\underset{|\mathbf{x}-\mathbf{z}|<\varepsilon}{Max} |R(\mathbf{x}, \mathbf{z})|$ vanishes.

Lines 4, 5 and 6: Assumption (2.4) yields

$$\iint_{|\mathbf{x}-\mathbf{z}|\geq\varepsilon} f(\mathbf{z}) \left[W(\mathbf{z}|\mathbf{x}, t) p(\mathbf{x}, t | \mathbf{y}, t') - W(\mathbf{x}|\mathbf{z}, t) p(\mathbf{z}, t | \mathbf{y}, t') \right] d\mathbf{x} d\mathbf{z}. \tag{2.12}$$

Incorporating terms (2.10-2.12) into Eq. (2.9) and taking the limit $\varepsilon \rightarrow 0$ gives

$$\begin{aligned}
\partial_t \int f(\mathbf{z}) p(\mathbf{z}, t | \mathbf{y}, t') d\mathbf{z} &= \int \left[\sum_i A_i(\mathbf{z}, t) \frac{\partial f(\mathbf{z})}{\partial z_i} + \sum_{i,j} \frac{1}{2} B_{ij}(\mathbf{z}, t) \frac{\partial^2 f(\mathbf{z})}{\partial z_i \partial z_j} \right] p(\mathbf{z}, t | \mathbf{y}, t') d\mathbf{z} \tag{2.13} \\
&+ \int f(\mathbf{z}) \left\{ \int \left[W(\mathbf{z}|\mathbf{x}, t) p(\mathbf{x}, t | \mathbf{y}, t') - W(\mathbf{x}|\mathbf{z}, t) p(\mathbf{z}, t | \mathbf{y}, t') \right] d\mathbf{x} \right\} d\mathbf{z}.
\end{aligned}$$

Integration by parts of Eq. (2.13) yields the following:

$$\begin{aligned}
\int f(\mathbf{z})\partial_t p(\mathbf{z}, t|\mathbf{y}, t')d\mathbf{z} &= \int f(\mathbf{z})\left\{-\sum_i \frac{\partial}{\partial z_i} A_i(\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t') + \sum_{i,j} \frac{1}{2} \frac{\partial^2}{\partial z_i \partial z_j} B_{ij}(\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t')\right. \\
&\quad \left. + \int \left[W(\mathbf{z}|\mathbf{x}, t)p(\mathbf{x}, t|\mathbf{y}, t') - W(\mathbf{x}|\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t') \right] d\mathbf{x}\right\}d\mathbf{z}. \\
&\quad + \text{surface terms} \tag{2.14}
\end{aligned}$$

Now, suppose that the process is restricted to a region R with surface S . In addition, one can choose $f(\mathbf{z})$ to be arbitrary but nonvanishing only in a region R' entirely contained in R . Therefore based on Eq. (2.14) one can deduce that for all \mathbf{z} in the interior of R , the following equation holds

$$\begin{aligned}
\partial_t p(\mathbf{z}, t|\mathbf{y}, t') &= -\sum_i \frac{\partial}{\partial z_i} \left[A_i(\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t') \right] + \sum_{i,j} \frac{1}{2} \frac{\partial^2}{\partial z_i \partial z_j} \left[B_{ij}(\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t') \right] \\
&\quad + \int \left[W(\mathbf{z}|\mathbf{x}, t)p(\mathbf{x}, t|\mathbf{y}, t') - W(\mathbf{x}|\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t') \right] d\mathbf{x}. \tag{2.15}
\end{aligned}$$

with the initial conditions $p(\mathbf{z}, t|\mathbf{y}, t) = \delta(\mathbf{y} - \mathbf{z})$. Note that the surface terms necessary vanish.

Eq. (2.15) is called the *differential Chapman-Kolmogorov equation*. It can be shown that under certain conditions a non-negative solution to the differential Chapman-Kolmogorov equation exists and satisfies a Chapman-Kolmogorov equation (2.1).

Each of the assumptions (2.4), (2.5), (2.6) leads to a distinctive part of the equation, which corresponds to one of the three processes: jumps, drift and diffusion.

Jump processes. When $A_i(\mathbf{z}, t) = B_{ij}(\mathbf{z}, t) = 0$, Eq. (2.15) simplifies to the following form:

$$\partial_t p(\mathbf{z}, t | \mathbf{y}, t') = \int [W(\mathbf{z} | \mathbf{x}, t) p(\mathbf{x}, t | \mathbf{y}, t') - W(\mathbf{x} | \mathbf{z}, t) p(\mathbf{z}, t | \mathbf{y}, t')] d\mathbf{x}, \quad (2.16)$$

and is often referred to as the *master equation*.

To show that the process described by Eq. (2.16) is a jump process, i.e., its paths are discontinuous at discrete points, one can solve master equation (2.16) to the first order in Δt . Since

$$\partial_t p(\mathbf{z}, t + \Delta t | \mathbf{y}, t) \simeq \frac{p(\mathbf{z}, t + \Delta t | \mathbf{y}, t) - p(\mathbf{z}, t | \mathbf{y}, t)}{\Delta t}, \quad (2.17)$$

one gets that:

$$\begin{aligned} p(\mathbf{z}, t + \Delta t | \mathbf{y}, t) &= p(\mathbf{z}, t | \mathbf{y}, t) + \partial_t p(\mathbf{z}, t + \Delta t | \mathbf{y}, t) \Delta t = p(\mathbf{z}, t | \mathbf{y}, t) \\ &+ \Delta t \int [W(\mathbf{z} | \mathbf{x}, t + \Delta t) p(\mathbf{x}, t + \Delta t | \mathbf{y}, t) - W(\mathbf{x} | \mathbf{z}, t + \Delta t) p(\mathbf{z}, t + \Delta t | \mathbf{y}, t)] d\mathbf{x}. \end{aligned} \quad (2.18)$$

With the initial conditions $p(\mathbf{z}, t | \mathbf{y}, t) = \delta(\mathbf{y} - \mathbf{z})$, the former yields

$$p(\mathbf{z}, t + \Delta t | \mathbf{y}, t) = \delta(\mathbf{y} - \mathbf{z}) [1 - \Delta t \int d\mathbf{x} \mathbf{W}(\mathbf{x} | \mathbf{y}, t)] + \Delta t \mathbf{W}(\mathbf{z} | \mathbf{y}, t). \quad (2.19)$$

Therefore for any Δt there is a finite probability for the particle to stay at the position \mathbf{y} , given by the coefficient of the $\delta(\mathbf{y} - \mathbf{z})$. The distribution of those particles, which do not remain at \mathbf{y} is given by $\mathbf{W}(\mathbf{z} | \mathbf{y}, t)$ after normalization. The typical path of $X(t)$ will consist of sections of straight lines with discontinuous jumps, where the distribution of jumps is

given by $W(\mathbf{z}|\mathbf{y}, t)$.

In the case of discrete state space, the differential Chapman-Kolmogorov equation for the jump process is given by:

$$\partial_t P(\mathbf{n}, t|\mathbf{n}', t') = \sum_{\mathbf{m}} [W(\mathbf{n}|\mathbf{m}, t)P(\mathbf{m}, t|\mathbf{n}', t') - W(\mathbf{m}|\mathbf{n}, t)P(\mathbf{n}, t|\mathbf{n}', t')]. \quad (2.20)$$

In this case, $W(\mathbf{n}|\mathbf{m}, t)$ and $W(\mathbf{m}|\mathbf{n}, t)$ are matrices, which include all possible transition into the accessible states.

Diffusion processes. When $W(\mathbf{z}|\mathbf{x}, t) = 0$, Eq. (2.15) reads:

$$\partial_t p(\mathbf{z}, t|\mathbf{y}, t') = - \sum_i \frac{\partial}{\partial z_i} [A_i(\mathbf{z}, t)p(\mathbf{z}, t|\mathbf{y}, t')] + \sum_{i,j} \frac{1}{2} \frac{\partial^2}{\partial z_i \partial z_j} [B_{ij}(\mathbf{z})p(\mathbf{z}, t|\mathbf{y}, t')]. \quad (2.21)$$

Eq. (2.21) is often referred to as the *Fokker-Planck equation* and describes a process known as a diffusion, characterized by a continuous sample paths (with $W(\mathbf{z}|\mathbf{x}, t) = 0$ the Lindenberg Condition, Eq. (2.3) is satisfied). The vector $A_i(\mathbf{z}, t)$ defines a drift and the matrix $B_{ij}(\mathbf{z})$ a diffusion component.

The distribution function captured by the Chapman-Kolmogorov equation (note that the adjective "differential" is going to be omitted from now on) describes exactly the underlying stochastic process. Unfortunately, this description is very limited and inefficient. The solution to the Chapman-Kolmogorov equation (even numerical) can be obtained only in some very simple cases. Usually, the Chapman-Kolmogorov equation describing systems of chemically reacting molecules such as gene expression systems cannot be solved due to non-

linearities, large numbers of considered molecular species and possibly infinite state space. Fortunately, it has many advantages: In the case of linear equations it allows one to analytically derive the moments of the underlying distribution function using generating function techniques [63], [61], [62]. In addition, it can serve as a departure point for various approximations which greatly simplify the analysis [67], [30], [66].

2.1.3 Stochastic simulations

Instead of directly solving the Chapman-Kolmogorov equation one may numerically simulate the Markov process it describes by use of the Gillespie algorithm (otherwise known as a stochastic simulation algorithm) [18], [19]. The Gillespie algorithm is a computer-oriented Monte-Carlo simulation method, which assigns probability to every single possible reaction and by means of a Markovian random walk in the space of molecular species generates trajectories of the underlying stochastic process. The approach enables to follow the transient behavior of reacting species but also, but taking appropriate samples, to estimate the distribution of the considered molecular populations at each instant of time. Unfortunately, stochastic simulations become computationally inefficient when a number of reacting molecules is large.

During the last few years, several approximation methods were proposed to accelerate the Gillespie algorithm. Among other methods, τ -leap procedure introduced a division of the time domain into τ long intervals. The length of the interval τ was chosen to obtain approximately constant propensity functions for the considered chemical reactions. With this requirement satisfied, the number of reactions fired in each channel during interval τ becomes a Poisson random variable [20], [8]. Such treatment is less computationally intense since it requires counting only the reactions fired within each of the contiguous time intervals τ

instead of determining the precise firing times of each single reaction. Another approximation can be applied if it is possible to separate chemical reactions into "fast" and "slow" [23], [55], [9]. The "fast" reactions can be approximated by the deterministic rate equations or the stochastic Langevin equations, while the "slow" reactions are treated stochastically according to the Gillespie scheme. Unfortunately, even the accelerated stochastic simulations become computationally inefficient, especially when estimating multivariate distributions.

2.2 Review of the prior modeling and its implications

At the beginning, the research was focused on gene regulation in prokaryotes, because the development of model biological systems such as *lacZ* operon and bacteriophage λ allowed extensive experimental investigations. Early attempts to explain the regulation in bacteria (phage λ repressor) relied on the deterministic approach assuming a rapid equilibrium between regulatory proteins and gene promoters [28], [1], [56]. More recently, Monte Carlo techniques [18], [19] were applied to explore effects of small number of mRNA and protein molecules in bacteria, sometimes in quite complicated scenarios. For example, Arkin et al. (1998) [3] considered a detailed stochastic model of the phage λ lysis-lysogeny decision circuit in *Escherichia coli*. In this system, intrinsic stochasticity plays an essential role in directing the cells into two different phenotypes as they follow different paths.

Analysis of model bacterial systems resulted in more theoretical investigations of noise propagation in prokaryotic regulatory networks. McAdams and Shapiro (1995) [41] proposed a hybrid modeling approach that integrated conventional biochemical kinetics with a framework of circuit simulations to model in vivo behavior of phage λ . Using Monte Carlo simulation, McAdams and Arkin (1997) [42] analyzed chemical reactions controlling transcript initiation and translation termination in a small prokaryotic genetic networks.

They incorporated a competition between mRNA translation and degradation, which created switching mechanisms that selected between alternative regulatory pathways. McAdams and Arkin (1999) [43] in their following publication hypothesized that cells use redundancy and extensive feedback mechanisms to achieve regulatory reliability to compensate for the transcriptional noise. Thattai and Oudenaarden (2001) [63] used a more rigorous approach to investigate stochasticity in small (one- and two-) gene networks. In addition to Monte Carlo simulations, they derived expressions for expected values and variances of mRNA and protein number based on the analysis of the corresponding Chapman-Kolmogorov equation. They demonstrated that a negative feedback mechanism efficiently decreases system noise.

Neither of the above investigations incorporated the intermittent gene activity as a potential source of stochasticity in gene regulation, although this idea was introduced a decade before. To explain heterogeneous levels of individual gene expression in steroid-inducible mouse mammary tumor virus system, Ko (1991, 1992) [32], [33] postulated that the stochasticity in eukaryotic gene regulation is driven by interactions between transcription factors and DNA (gene promoters). More precisely, at a given instant of time a gene copy is thought to be either "*switched on*" by having transcription complex bound to its promoter, or "*switched off*" by having transcription complex not bound. In the proposed mathematical model, which was simulated numerically, Ko assumed that transcription and translation proceeds deterministically when a gene is turned on. Monte Carlo simulations of this model were applied by Cook et al. (1998) [10] who showed that haploinsufficiency diseases may arise from the stochasticity caused by intermittent gene activity. Recently, a growing number of experiments on eukaryotic cells including *Saccharomyces cerevisiae* [49], [5], rat NRK [15] as well as human HeLa and SK-N-AS [45], [37] complemented with numerical simulations supports Ko's hypothesis contributing stochasticity in eukaryotic regulation to the intermittent gene

activity. Lately, stochasticity due to the intermittent gene activity was incorporated into bacterial studies. Kierzek et al. (2001) [29] simulated random fluctuations in the number of protein molecules in a very detailed model of transcription initiation in *LacZ* gene in *E. coli*. Similarly, using Monte Carlo techniques, Bundschuh et al. (2003) [7] analyzed effects of protein dimerization on the noise reduction in the control circuit for the λ repressor protein cI of phage λ in *E. coli*.

The past few years of research resulted in more theoretical and rigorous approaches departing from the analysis of the Chapman-Kolmogorov equation. Tao (2004) [61] investigated effects of the negative and positive feedbacks on the intrinsic and external noise in a single-gene regulatory networks. He employed a Chapman-Kolmogorov equation for the underlying distribution function and calculated the first two moments of the gene product marginal distributions, i.e., expected value and variance of the number of mRNA and protein molecules. Tao assumed that the transcription rate depends on the amount of the protein and he neglected stochasticity due to the switching of the gene status. While considering two-gene networks Tao (2004a) [62] disregarded mRNA transcript as an intermediate gene product. To simplify the corresponding Chapman-Kolmogorov equation he introduced a Fokker-Planck approximation and analyzed statistics of the protein production/decay noise. The noise in one- and two-gene regulatory networks was also analyzed by Tomioka et al. (2004) [66]. The authors analyzed a protein fluctuations using a linear noise approximation to the corresponding Chapman-Kolmogorov equation. Tomioka et al. (2004) assumed that a network is close to the deterministic stable equilibrium and disregarded stochastic effects due to the intermittent gene activity.

The stochasticity caused by a switching of a gene status recognized by Ko (1991, 1992) [32], [33], was first rigorously analyzed by Kepler and Elston (2001) [30]. In their influential

paper, Kepler and Elston (2001) considered synthesis of protein oligomers in the process, however assumed a direct protein translation from DNA. The approach involved a Chapman-Kolmogorov equation for the underlying probability distribution function approximated by a Fokker-Planck equation. In the case of a single gene without feedback regulation Kepler and Elston (2001) derived a steady state expected value and variance of the protein number in the system. In the case of a single self-activating gene, they further simplified the Fokker-Planck equation by neglecting the diffusion term, which lead to the first order system of PDEs. While analyzing a system of two mutual repressors, Kepler and Elston (2001) used Monte Carlo simulations to obtain a marginal protein distributions.

Stochasticity due to the intermittent gene activity was also analyzed by Raser and O'Shea (2004) [49] in more general model incorporating in addition the mRNA/protein production/decay noise. The authors analyzed the corresponding Chapman-Kolmogorov equation with moment generating functions and derived a normalized steady state protein variance. Raser and O'Shea (2001) proposed that the balance between gene promoter activation and transcription influences the variability in the mRNA level and confirmed this hypothesis by matching a Monte Carlo simulations of the model with the measurements of the intrinsic noise in the genetically engineered cells of *E.coli* and budding yeast.

For the sake of simplicity it is usually assumed that the transcriptional gene activity is due to the actions of a single *trans*-acting regulatory molecule (transcription factor) and a single *cis*-acting regulatory element, i.e., operator in bacteria or promoter in eukaryotes [32], [30], [29], [63], [7], [61], [66]. In fact, the specific patterns of gene expression are determined by combinatorial interactions of series of transcription factors that may bind to various regulatory sites within gene promoters and enhancers ([68], p.72). In this context, Louis et al. (2003) [40] developed a theoretical model of regulation of *Sex-lethal* gene that controls

sex determination and dosage compensation in *Drosophila melanogaster*. These authors considered stochastic effects caused by interactions between multiple regulatory molecules and a series of regulatory sites. Pirone and Elston (2004) [53] considered three binding sites that controlled transcription of a *lacZ* gene. They used the Fokker-Planck equations to calculate the first two moments of the underlying distribution function. In this work they focused on oligomerization reactions leading to the formation of protein dimers and tetramers, while disregarding mRNA in the model.

The aforementioned investigations relied on the exact stochastic description in the terms of the Chapman-Kolmogorov equation or Monte Carlo simulations. The alternative approach utilizes a Langevin equation, which is a stochastic differential equation driven by white noise [67]. In this approach, the system is described by deterministic rate equation augmented with an additive of multiplicative stochastic terms, which mimic the external noise. The approach leads to the Fokker-Planck equations for the underlying distribution function. Using this approach Hasty et al. (2000) [24] demonstrated that small deviations in the transcription rate can lead to large fluctuations in the protein number. Similarly, Ozbudak et al. (2002) [50] concluded that the level of phenotypic variation in an isogenic *Bacillus subtilis* population can be regulated by genetic parameters. In addition, Simpson et al. (2004) [58] considered transcriptional regulation involving switching between discrete high and low transcriptional rates. The approach included a frequency analysis of a spectral density and provided a distribution of noise associated with mRNA production/decay noise and fluctuations at the operator state.

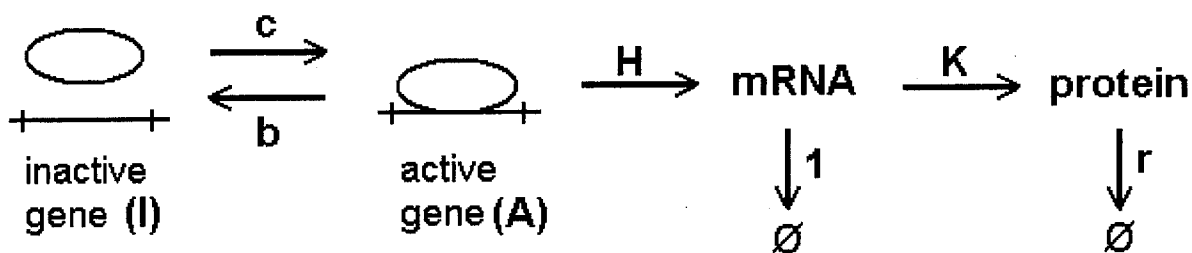
Chapter 3

Analysis of a single gene regulatory module

3.1 Exact stochastic description

Consider a system of a single haploid gene without feedback regulation into its own transcriptional activity with three sources of stochasticity: intermittent gene activity, mRNA transcription/decay, and protein translation/decay noise (Fig. 3.1). Since the gene activation and inactivation is due to the binding and dissociation of regulatory proteins it is natural to assume that activation and inactivation rates (intensities) depend on the amounts (concentrations) of regulatory factors. In the case of regulation without the feedback, it is assumed that a gene is activated at a constant rate c and deactivated at a constant rate b , which follows from the constant amount of regulatory protein. While the gene promoter is bound by the regulatory protein, mRNA transcript molecules are produced at a constant

Figure 3.1 : Schematic representation of a single gene regulatory module.



rate H . However, when the promoter is not bound, then the gene is inactive and the basal transcription rate is neglected. If mRNA transcript is present in the cell, protein translation proceeds at a rate K . The mRNA transcript degradation rate is set to 1 to eliminate a nuisance parameter, and therefore the protein degradation rate r reflects the ratio of the corresponding protein and mRNA degradation rates. Such normalization implies that the time course as well as other parameters describing the system are rescaled by the degradation rate of the mRNA transcript. The model can be summarized in the following reactions:



where A and I denote possible conformations of the promoter with the transcription factor bound and the active gene in the state A , and the empty promoter and inactive gene in the state I . Function G is a binary variable describing the status of a gene, i.e., the promoter conformation, $G(A)=1$ and $G(I)=0$. Degradation of gene products is depicted with symbol ϕ .

If x and y denote the number of mRNA and protein molecules respectively, then the state of the system at any instant of time is given by the triple (x, y, G) , where x and y are nonnegative integer valued random variables and G is a binary random variable. Their joint distribution can be represented as a pair of probability mass functions:

$$f_{xy}(t) = P[\# \text{ mRNA} = x, \# \text{ protein} = y, G = 0], \quad (3.4)$$

$$g_{xy}(t) = P[\# \text{ mRNA} = x, \# \text{ protein} = y, G = 1], \quad (3.5)$$

which correspond to the probability of having x and y amounts of mRNA and protein molecules in the single cell, in the gene inactive and active state, respectively. The marginal distribution $\rho_{xy} := f_{xy} + g_{xy}$ describes the level of gene products regardless of gene status. Because the process is independent from cell to cell, the functions f_{xy} and g_{xy} also describe a proportion of cells in the population with a given amount of gene products, and thus can be related to measurements provided by experiments at the single cell level, such as flow-cytometry.

The process described by Eqs. (3.1-3.3) is a jump process, because the number of molecules in the system changes only at discrete instants of time. In fact the number of molecules changes only by one, and such Markov process is known as a *one step process* ([67], p. 34). The time evolution of distribution function defined by (3.4-3.5) is given by the following master equation [67]:

$$\begin{aligned} \frac{df_{xy}}{dt} = & \quad bg_{xy} - cf_{xy} + G(I)Hf_{x-1,y} + (x+1)f_{x+1,y} - (HG(I) + x)f_{xy} + \\ & \quad + Kxf_{x,y-1} + r(y+1)f_{x,y+1} - (Kx + ry)f_{xy}, \end{aligned} \quad (3.6)$$

$$\begin{aligned} \frac{dg_{xy}}{dt} = & \quad -bg_{xy} + cf_{xy} + G(A)Hg_{x-1,y} + (x+1)g_{x+1,y} - (HG(A) + x)g_{xy} + \\ & \quad + Kxg_{x,y-1} + r(y+1)g_{x,y+1} - (Kx + ry)g_{xy}, \end{aligned} \quad (3.7)$$

where $x, y=0,1,2,3\dots$

The system (3.6-3.7) is equivalent to the master equation (2.20). Note that in the considered case, the mass function $P(\mathbf{n}, t|\mathbf{n}', t')$ from the Eq. (2.20) describes the triple $\mathbf{n} = (x, y, G)$, where G is a binary random variable. Therefore, given that $G \in \{0, 1\}$, the distribution $P(x, y, G, t)$ is represented as a pair of probability mass functions, $P(x, y, 0, t) = f_{xy}(t)$ and $P(x, y, 1, t) = g_{xy}(t)$, and analogically, the master equation (2.20) can be represented by a pair of equations coupled by transitions between $P(x, y, 0, t)$ and $P(x, y, 1, t)$, as in the system (3.6-3.7). The first two terms in Eqs. (3.6)-(3.7) correspond to the probability flow due to intermittent gene activity or, in other words, to the transition between f_{xy} and g_{xy} . The next three terms depict the flow of the probability due to the synthesis and degradation of mRNA molecules, and the last three terms describe the synthesis and degradation of protein molecules. Note that since $G(I)=0$, the mRNA synthesis terms are absent in Eq. (3.6).

The solution to the master equation (3.6)-(3.7) is the primary interest. However, the analytical solution is unknown even for $\frac{df_{xy}}{dt} = \frac{dg_{xy}}{dt} = 0$. A numerical solution is possible to obtain when the number of molecules involved remains small. At the steady state, Eqs. (3.6)-(3.7) can be represented as a set of infinite simultaneous linear algebraic equations. The resulting system of equations is infinite since the probability functions f_{xy} and g_{xy} have infinite supports. In this case, the tails of f_{xy} and g_{xy} can be disregarded since the probability assigned for arbitrarily large x and y is negligible. By this approximation, the steady state distribution is described by a finite system of linear algebraic equations. The number of equations describing the solution grows quadratically with the number of considered molecules. For eukaryotes, where the number of mRNA molecules might be at the order of hundreds and the number of protein molecules at the order of hundreds of thousands

(e.g. molecules involved in NF- κ B regulatory pathway [36], [37]), one would need to solve a system at the order of 10^7 simultaneous linear equations, which is beyond our computational reach. Due to these limitations, the corresponding Chapman-Kolmogorov equation can be effectively solved only for prokaryotic systems, where the number of molecules considered is relatively small. An alternative, although less rigorous approach is to estimate the underlying distribution function by means of Monte Carlo simulations of system (3.1-3.3) using the Gillespie algorithm [18], [19] or other stochastic simulation schemes. Similarly, the latter approach becomes inefficient when a large number of reacting molecules is involved, especially when estimating multidimensional distributions.

3.2 Variance decomposition

Note that despite the fact that equations (3.6)-(3.7) cannot be solved analytically, they can be used to derive the moments of the underlying distribution function [17], [67]. Generating function techniques allow deriving arbitrarily high moments at the steady state as well as their time evolution in some cases (see the Appendix A for details), but for the purpose of this work only the steady state expected value and the variance are presented.

The expected number of mRNA and protein molecules at the steady state in the system (3.1-3.3) is given by

$$E[X] = \frac{c}{c+b}H, \quad (3.8)$$

$$E[Y] = \frac{K}{r}E[X], \quad (3.9)$$

respectively, whereas the variance of the number of mRNA and protein molecules decom-

poses into additive terms resulting from different sources of stochasticity:

$$Var_E[X] = \frac{b}{c(1+c+b)} E^2[X] + E[X], \quad (3.10)$$

$$Var_E[Y] = \frac{rb(1+c+b+r)}{c(1+r)(1+c+b)(r+c+b)} E^2[Y] + \frac{r}{(1+r)} \frac{E^2[Y]}{E[X]} + E[Y]. \quad (3.11)$$

It will be shown in the following sections that the first term in mRNA variance [expression (3.10)] results entirely from intermittent gene activity, while the second is due to the mRNA production/decay noise. Additionally, it will be shown that the first term in protein variance [expression (3.11)] represents the variation due to intermittent gene activity, the second term depicts the mRNA production/decay noise integrated through translation mechanisms and the third term corresponds to the protein production/decay noise.

The following section proposes two approximations to the exact stochastic process (3.1-3.3), which allow assessing the variance decomposition (3.10-3.11): First, the continuous approximation which accounts only for the stochastic effects due to intermittent gene activity. Second, the mixed approximation which accounts for the stochasticity corresponding to intermittent gene activity and mRNA production/decay noise, while the protein production/decay noise is neglected.

3.3 Approximations to the exact description

3.3.1 Continuous model

Taking into account that in eukaryotes the stochastic effects in gene regulation are primarily attributable to intermittent gene activity [15], [30], [5], [53] rather than to the mRNA/protein

production/decay process, one can simplify the system (3.1-3.3). This leads to the exact stochastic description of molecules present in a small number of copies (in this case gene copies) with the ordinary differential equation (ODE) description for processes involving molecules present at larger levels (in this case mRNA and protein molecules). For a single haploid gene without feedback regulation, the continuous approximation of the system (3.1-3.3) yields:



$$\frac{dx}{dt} = -x + HG(t), \quad (3.13)$$

$$\frac{dy}{dt} = Kx - ry, \quad (3.14)$$

where $G(t)$, as in the exact description, is the binary state of a gene promoter, with $G=1$ whenever activating protein occupies the promoter region, and $G=0$, otherwise. Variables x and y , as before, denote the mRNA and protein levels, respectively.

Eq. (3.13), which describes the mRNA transcription and degradation process, is a stochastic differential equation driven by a binary random variable $G(t)$. It is similar to the Langevin equation [67], but it reflects the intrinsic stochasticity connected with gene activation process, rather than an additive white noise. The Eq. (3.14) is an ordinary differential equation describing the protein production and degradation process. As a result, the number of mRNA as well as protein molecules are a continuous random variables, bounded by the steady state solutions to Eqs. (3.13)-(3.14) at the active gene state ($G=1$): $x \in [0, H]$, $y \in [0, \frac{KH}{r}]$.

The state of the system at any instant of time is described by a triple of random variables

(x, y, G) , two continuous (x and y are no longer integer valued) and one binary. Similarly to the exact description, their joint distribution can be represented as a pair of probability density functions $f(x, y, t)$ and $g(x, y, t)$, defined as follows:

$$f(x, y, t)\Delta x\Delta y = P[x(t) \in (x, x + \Delta x), y(t) \in (y, y + \Delta y), G = 0], \quad (3.15)$$

$$g(x, y, t)\Delta x\Delta y = P[x(t) \in (x, x + \Delta x), y(t) \in (y, y + \Delta y), G = 1]. \quad (3.16)$$

By fluid dynamics analogy one can find a system of partial differential equations (PDEs) describing evolution of densities f and g [38]:

$$\frac{df}{dt} + \text{div} \left[\left(\frac{dx}{dt} \Big|_{G=0}, \frac{dy}{dt} \right) f \right] = bg - cf, \quad (3.17)$$

$$\frac{dg}{dt} + \text{div} \left[\left(\frac{dx}{dt} \Big|_{G=1}, \frac{dy}{dt} \right) g \right] = -bg + cf. \quad (3.18)$$

The system (3.17)-(3.18) is obtained from the continuity equations with source terms resulting from the change of gene status (transformation between f and g), Eq. (3.12), while the velocity fields $(dx/dt, dy/dt)|_{G=0}$ and $(dx/dt, dy/dt)|_{G=1}$ are given by Eqs. (3.13) and (3.14), respectively. Eqs. (3.17-3.18) can be further expressed as

$$\frac{df}{dt} - \frac{\partial}{\partial x}(xf) + \frac{\partial}{\partial y}((Kx - ry)f) = bg - cf, \quad (3.19)$$

$$\frac{dg}{dt} + \frac{\partial}{\partial x}((H - x)g) + \frac{\partial}{\partial y}((Kx - ry)g) = -bg + cf. \quad (3.20)$$

The above system of first-order PDEs is analogous to the general form of the differential Chapman-Kolmogorov equation given by Eq. (2.15). In the considered case, the distribution $p(\mathbf{z}, t | \mathbf{y}, t')$ described by the Eq. (2.15) corresponds to the triple $\mathbf{z} = (x, y, G)$, where G is a binary random variable while x and y are continuous. Therefore, given that $G \in \{0, 1\}$, the distribution $p(x, y, G, t)$ is represented as a pair of probability mass functions, $p(x, y, 0, t) = f(t)$ and $p(x, y, 1, t) = g(t)$, and analogically, the Chapman-Kolmogorov equation (2.15) can be represented by a pair of PDEs coupled by transitions between $p(x, y, 0, t)$ and $p(x, y, 1, t)$ as in the system (3.6-3.7). This coupling between f and g , corresponds to the jump process caused by a change of gene status and is described by the right hand sides of Eqs. (3.19-3.20). In addition, the partial derivatives with respect to x and y constitute the drift vector given by $(dx/dt, dy/dt)|_{G=0}$ and $(dx/dt, dy/dt)|_{G=1}$ for f and g , respectively. In this case there is no diffusion in the process, but rather the process is composed of a nonzero drift onto which a jump process is superimposed.

Independently of the exact stochastic description, the first two steady state moments of the gene product marginal distributions given by the system (3.19)-(3.20) are derived (see Appendix A for details). It was found that the expected number of mRNA and protein molecules is given by

$$\begin{aligned} E[X] &= \frac{c}{c+b}H, \\ E[Y] &= \frac{K}{r}E[X], \end{aligned}$$

respectively, whereas the mRNA and protein variance is equal to:

$$Var_C[X] = \frac{b}{c(1+c+b)} E^2[X], \quad (3.21)$$

$$Var_C[Y] = \frac{rb(1+c+b+r)}{c(1+r)(1+c+b)(r+c+b)} E^2[Y]. \quad (3.22)$$

Comparison with the exact stochastic description shows that disregarding the mRNA/protein production/decay noise does not affect the expected number of mRNA and protein molecules. In addition, the protein [as well as the mRNA] variance accounted for in the continuous approximation is equal to the first term of the total variance given by (3.11) [(3.10)]. This shows that the first term in the protein [mRNA] variance in the exact stochastic description, Eq. (3.11) [Eq. (3.10)] is entirely due to the stochastic effects introduced by intermittent gene activity. Furthermore, since the protein noise does not effect the mRNA level, the second term in the total mRNA variance, Eq. (3.11), contributes the variation entirely due to the mRNA transcription/decay noise. In fact, at the steady state, the mRNA production/decay process in the active gene state has a Poisson distribution with parameter (expected value) given by Eq. (3.8) [this can be shown by analytically solving simplified version of the system (3.6-3.7) restricted only to the mRNA transcript with $G=1$].

3.3.2 Mixed model

In the continuous model, the number of mRNA as well as protein molecules is approximated with ODEs. Such treatment is well justified in eukaryotes for the amount of protein molecules, since their number in the cell may be at the order of 10^5 molecules of a given species. A similar description for the mRNA transcript may not be valid since the mRNA is typically much less abundant. Such motivation leads to the model which attributes the

stochasticity to intermittent gene activity and mRNA production/decay process, while the protein translation/degradation noise is neglected:



$$\frac{dy}{dt} = Kx - ry. \quad (3.25)$$

The mixed (discrete in the mRNA and continuous in the protein number) model provides an exact treatment of mRNA transcript, while the amount of the protein is modeled using ODE description as in the continuous approximation.

The state of the system in any instant of time is given by the triple of the random variables (x, y, G) , but in this case, x and G are discrete, and y is continuous. Their joint distribution can be represented by the pair of the probability density functions:

$$f_x(y, t)\Delta y = P[\# mRNA = x, y(t) \in (y, y + \Delta y), G = 0], \quad (3.26)$$

$$g_x(y, t)\Delta y = P[\# mRNA = x, y(t) \in (y, y + \Delta y), G = 1]. \quad (3.27)$$

Similarly to the continuous model, one can write the PDEs for densities f_x and g_x :

$$\frac{df_x}{dt} + \frac{\partial}{\partial y}(f_x(Kx - ry)) = bg_x - cf_x + (x + 1)f_{x+1} - xf_x, \quad (3.28)$$

$$\begin{aligned} \frac{dg_x}{dt} + \frac{\partial}{\partial y}(g_x(Kx - ry)) &= -bg_x + cf_x \\ &+ Hg_{x-1} + (x + 1)g_{x+1} - (H + x)g_x. \end{aligned} \quad (3.29)$$

The right-hand sides of equations (3.28)-(3.29) account for two flows of probability. The first corresponds to the change of gene activity, while the second depicts the probability flow due to the discrete process of mRNA transcription and degradation as in the Chapman-Kolmogorov equation (3.6)-(3.7).

The system (3.28-3.29) is an infinite system of partial differential equations since the densities $f_x(y)$ and $g_x(y)$ have infinitely long tails ($x=0,1,2\dots$), as in the exact stochastic description. Nevertheless, it is analogous to the differential Chapman-Kolmogorov equation (2.15) and can be obtained by explicitly expressing all conditional densities $p(\mathbf{z},t|\mathbf{y},t')$ for $x = 0, 1, 2, \dots$ and $G \in \{0, 1\}$. Similarly to the continuous approximation, Eqs. (3.28)-(3.29) describe a jump process with drift. In this case jumps are generated by a discrete process of mRNA production and change of gene activity, while drift corresponds to the protein production (x is discrete and y is continuous).

Eqs. (3.28)-(3.29) can be approximated by the finite system of ODEs (or the finite system of linear simultaneous algebraic equations for the steady state) by disregarding the infinite tail of the distribution, and then solved using developed discretization techniques (see Appendix C for details).

By employing generating function techniques (see Appendix A for details), the first two steady state moments of the distribution captured by the system (3.28)-(3.29) were derived.

The expected number of mRNA and protein molecules remains the same as in the case of the exact description and continuous approximation, given in expressions (3.8)-(3.9). The mRNA and protein variance is found to account for two sources of stochasticity: intermittent gene activity and mRNA production/decay noise:

$$Var_M[X] = \frac{b}{c(1+c+b)} E^2[X] + E[X], \quad (3.30)$$

$$Var_M[Y] = \frac{rb(1+c+b+r)}{c(1+r)(1+c+b)(r+c+b)} E^2[Y] + \frac{r}{(1+r)} \frac{E^2[Y]}{E[X]}. \quad (3.31)$$

The mRNA variance accounted for in the mixed model is equal to that given by the exact description, Eq. (3.10), since the mRNA treatment is generically the same as in the former by construction. The protein variance accounts for the first two terms in the exact description [Eq. (3.11)], therefore taking into account the continuous approximation, the second term in the protein variance is attributed to the mRNA transcription/decay noise. Finally, the last term in Eq. (3.11), not accounted for in the mixed model, is due to the protein translation/decay noise. Specifically, assuming that mRNA transcript is present at a constant level n in the cell, the protein production/decay noise has a Poisson distribution with mean equal to $\frac{Kn}{r}$ at the steady state.

In addition, it can be shown that if a gene is continuously active ($G=1$), the mRNA and protein variance are given by:

$$Var_{G=1}[X] = E[X], \quad (3.32)$$

$$Var_{G=1}[Y] = \frac{r}{(1+r)} \frac{E^2[Y]}{E[X]} + E[Y], \quad (3.33)$$

respectively, which follows from disregarding the variation due to intermittent gene activity. This result is in agreement with previous investigations, which disregarded intermittent gene activity as a potential source of stochasticity in gene regulation [63].

3.4 Applicability of introduced models

Introduced approximations allow much faster single cell simulations than the exact stochastic description. In the considered case of regulation without feedback mechanism the continuous model (3.12-3.14) requires only stochastic simulations of gene activity, while the mRNA and protein levels can be obtained separately by solving ODEs (3.13-3.14) between the times when the gene activity changes. Such separation follows from the assumption that rates at which gene activity changes, b and c , are constant and do not depend on the produced protein. Similarly in the mixed model, single cell trajectories can be simulated from Eqs. (3.23)-(3.24) according to the Gillespie algorithm, while the protein number can be obtained by solving Eq. (3.25). The mixed approach is still much more efficient than the exact stochastic description (3.1-3.3) since simulations of the protein number constitute the most time consuming part when the Gillespie algorithm is applied.

Introduced approximations not only allow much faster Monte Carlo simulations of single cells than the Gillespie algorithm, but also transformations into equations for probability distribution function. In the continuous approximation, PDEs (3.19)-(3.20) can be numerically solved using developed discretization techniques (see the Appendix C for details). The steady state solution can be obtained by numerically solving a system of simultaneous linear equations, while its time evolution is captured by the system of ODEs. As opposed to the Chapman-Kolmogorov equation (3.6)-(3.7), the distribution function described by the continuous approximation can be solved for an arbitrarily large number of molecules considered.

The mixed approximation (3.28-3.29) is not as efficient as the continuous model, but it has a computational advantage over the Chapman-Kolmogorov equation that allows the consideration of an arbitrarily large number of protein molecules, whereas the mRNA transcript is treated exactly.

Introduced approximations are particularly useful in the case of eukaryotes when the numbers of molecules involved are large and the Chapman-Kolmogorov equation (3.6-3.7) can not be solved. Therefore to quantify their performance, the errors introduced by each of the approximations are derived (see Appendix B for details). In the derivations it is assumed that the protein half-life time is much greater than that of the mRNA transcript, i.e. $r \ll 1$.

The continuous approximation neglects ε_p fraction of the total protein variance when

$$\varepsilon_p = \frac{c(c+b)K}{rb} \frac{1}{E[Y]}. \quad (3.34)$$

whereas, the mixed approximation neglects ε_p fraction of the total protein variance when

$$\varepsilon_p = \frac{c(c+b)}{rb} \frac{1}{E[Y]}. \quad (3.35)$$

Note that as the expected number of protein molecules increases, resulting errors decrease. However, if c and b increase, and thus the gene activity changes more frequently, the errors introduced by approximations increase. The error introduced by the continuous model in the protein variance is approximately K times greater than that of the mixed model, where K is the transcription rate.

The mixed approximation accounts for the total mRNA variance, while the continuous approximation neglects ε_m fraction of the mRNA variance when, assuming that $r \ll 1$,

$$\varepsilon_m = \frac{c(1+b+c)}{b} \frac{1}{E[X]}. \quad (3.36)$$

Therefore the error of the continuous mRNA approximation decreases when then the expected number of mRNA molecules increases.

3.4.1 Significance of various noise sources

Expressions describing mRNA and protein variances allow quantifying the significance of considered sources of stochasticity in the process. Note that the first term in the total mRNA variance, Eq. (3.10), corresponding to the intermittent gene activity is at the order of $E^2[X]$ (the rates at which gene activity changes, c and b , are at the order of 1), while the second, corresponding to the transcription/decay noise, is at the order of $E[X]$. Therefore, in the case of eukaryotes, where $E[X]$, the expected number of mRNA molecules, can be at the order of hundreds of molecules (e.g. species involved in early immune response [36], [37]) the former dominates the latter. In prokaryotes, where $E[X]$ is at the order of one molecule, the transcription/decay noise has a significant contribution to the total variance. Similarly in Eq. (3.11), which describes the protein variance: The first term corresponding to the intermittent gene activity is at the order of $rE^2[Y]$. The second term due to the mRNA transcription/decay noise is at the order of $rE^2[Y]/E[X]$. Finally, the third term resulting from the protein translation/decay noise is at the order of $E[Y]$ (c and b , are at the order of 1, while $r \ll 1$, since the protein molecules are typically much more stable than mRNA molecules). In eukaryotes, $E[Y]$, the expected protein number, can be at the order of hundreds of thousands of molecules. Therefore, the intermittent gene activity contributes most of the total variability in the process, while the protein/decay noise is of least significance. In prokaryotes, where $E[Y]$ may be relatively small (at the order of tens), there is a

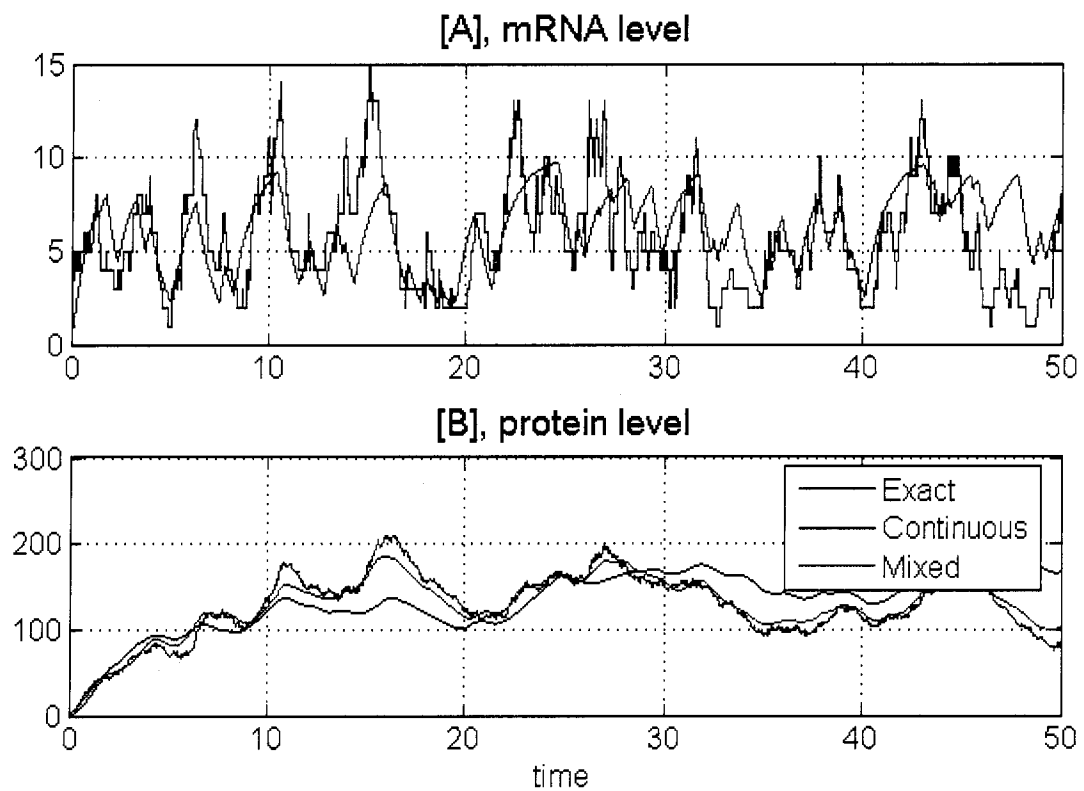
competition between stochastic effects due to intermittent gene activity and mRNA/protein production/decay noise.

The following part of the chapter includes comparison of the introduced models in the terms of the single cell trajectories and the probability distribution functions. Hypothetical cells are parametrized to reflect levels of mRNA and protein molecules of a given species on average in the cell. For "prokaryotic cell" it is assumed that $H=10$, $K=6$, $r=0.25$, which results in a relatively small number of mRNA and protein molecules in the cell (about 10 mRNA molecules and about 200 protein molecules). For "eukaryotes" it is assumed that $H=100$, $K=250$, $r=0.25$, which yields that mRNA number is at the order of 100 molecules on average, while protein number is at the order of 10^5 (which is the case of molecular species involved in NF- κ B regulatory pathway [36], [37]). Two regimes are considered: For $c>1$ and $b>1$ the gene activity changes frequently with respect to the mRNA half-life time and this regime is called to be close to the statistical equilibrium. For $c<1$ and $b<1$, the transcriptional activity changes slowly on the time scale of mRNA half-life time, and this regime is referred as being far from the statistical equilibrium.

3.4.2 Single cell simulations

Introduced approximations are validated by means of conditional trajectories. First, a single trajectory of the system (3.1-3.3) is generated using the Gillespie algorithm [18], [19]. Then, conditioning on the gene activity switching times, the corresponding path for the continuous model can be obtained by solving ODEs (3.13)-(3.14) in the contiguous intervals given by the switching times (setting $G=1$ or $G=0$ depending on the current gene state). The initial conditions for the ODEs, the amount of the mRNA and protein, are passed from the end of the previous time interval to the next. Similarly, for the mixed model, conditioning on the

Figure 3.2 : Single cell conditional trajectories for "prokaryotes".



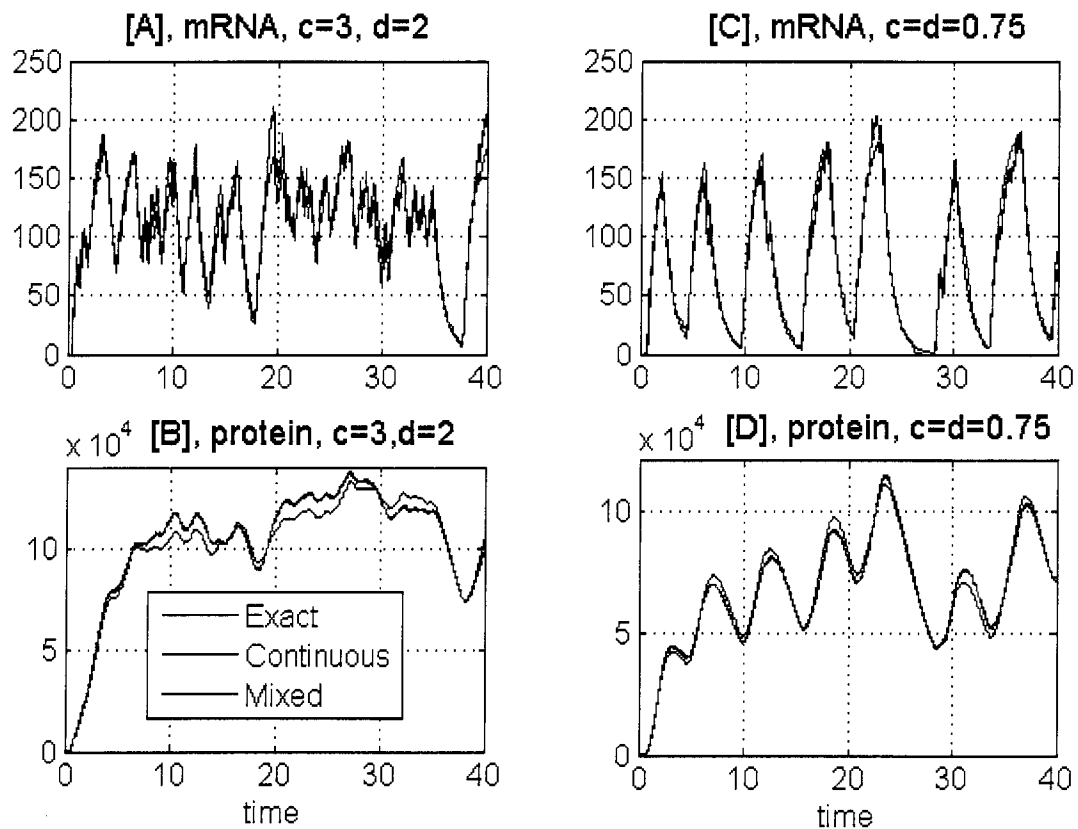
Single cell conditional trajectories for "prokaryotes" in the regime close to the statistical equilibrium ($H=10$, $K=6$, $r=0.25$, $c=3$, $b=2$). Shown here is a single realization of the exact model (red color) augmented with conditional trajectories for continuous (black) and mixed model (blue). Panel A presents mRNA trajectories, while panel B protein trajectories. Note, that the mRNA trajectory for the mixed model and exact description are the same by construction, so the former is not presented as a separate curve (panel A).

time instants when the mRNA amount changes, the protein trajectory can be obtained by solving Eq. (3.25) in the contiguous intervals given by the switching times.

Shown in the Fig. 3.2 are stochastic conditional trajectories for a single prokaryotic cell in the regime close to statistical equilibrium ($H=10$, $K=6$, $r=0.25$, $c=3$, $b=2$). One can observe discrepancies between the exact description and its continuous and mixed approximations. The mRNA trajectory simulated by means of Gillespie algorithm (Fig. 3.2A- red curve) exhibits much larger variability than the trajectory resulting from the continuous approximation. The jumps corresponding to the production or degradation of a single mRNA molecule are clearly visible in the exact description, while the trajectory for the continuous approximation is smooth, except for the kinks reflecting the changes of gene activity. Moreover, the number of mRNA molecules in the exact trajectory elevates to as much as 15 molecules, while for the continuous approximation it is bounded by 10, which is the number given by the solution of Eq. (3.13) at the gene active state ($G=1$). Protein trajectories for the approximating models are much more like these given by the Gillespie algorithm (Fig. 3.2B). Variability due to the protein production/decay noise in the exact description is much smaller than that for the mRNA molecules, although still visible. Overall, the protein approximation seems to be quite good, especially for the mixed model.

Conditional trajectories for the hypothetical eukaryotic cell ($H=100$, $K=250$, $r=0.25$) are depicted in Fig. 3.3. For eukaryotes, where the large number of both mRNA and protein is involved, the approximations are very good. The number of mRNA molecules is approximated quite well by the continuous model (Fig. 3.3A, C -black vs. red curve), but still the large variability in Gillespie based trajectory, especially in the gene active state, is not captured by the continuous model. Nevertheless, these discrepancies are much less visible for the amount of the protein molecules, and the approximation given by the continuous model

Figure 3.3 : Single cell conditional trajectories for "eukaryotes".



Single cell conditional trajectories for "eukaryotes" ($H=200$, $K=250$, $r=0.25$) in two regimes: Panel A, B -close to the statistical equilibrium ($c=3$, $b=2$), panel C, D -far from the equilibrium ($c=0.75$, $b=0.75$). Shown here a single realization (in each case) of the exact model (red color) augmented with conditional trajectories for continuous (black) and mixed model (blue). Note that the mRNA trajectory from the mixed model and exact model are the same by construction, so the former is not presented as a separate curve (panel A, C).

appears to be very good (Fig. 3.3B, D). The mixed model, discrete in mRNA and continuous in protein number, seems to approximate the exact description almost perfectly (Fig. 3.3. blue vs. red curve). The mRNA trajectory is generically the same by construction, and the protein trajectory overlaps with that of the exact description.

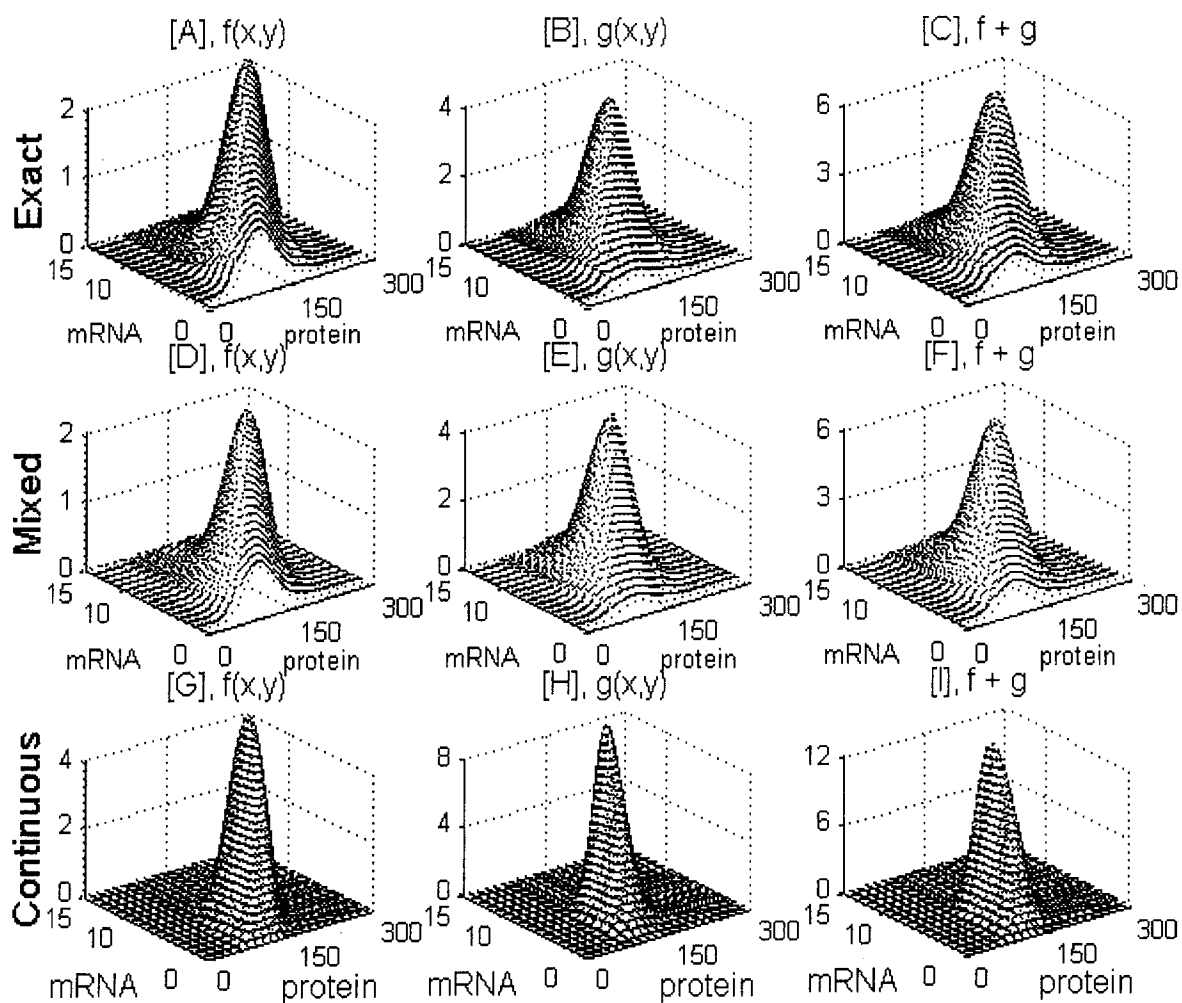
3.4.3 Distribution functions

In this section the steady state solutions to the equation for the probability distribution functions in considered models are presented. The Chapman-Kolmogorov equation (3.6)-(3.7) is solved when the number of molecules involved remains small (for the hypothetical prokaryotic cell). PDEs resulting from the continuous and mixed model, Eqs. (3.19)-(3.20) and Eqs. (3.28)-(3.29), are solved using developed discretization techniques (see the Appendix C for details).

Fig. 3.4 presents two-dimensional mRNA-protein distributions for "prokaryotes" in the regime close to statistical equilibrium ($H=10$, $K=6$, $r=0.25$, $c=3$, $b=2$). Distributions are consistent with the single cell trajectories presented in Fig. 3.2. While the distributions given by the exact description seem to be approximated quite well by the mixed model (Fig. 3.4A, B, C vs. C, D, E), the distributions given by continuous model (Fig. 3.4G, H, I) are much more narrow and peak much higher. In fact, the continuous approximation accounts only for 40% of the total variance in the mRNA number and 39% in the protein number in this case. The mixed approximation explains 100% of the mRNA variance (by construction) and 89% of the protein variance (Table 3.1).

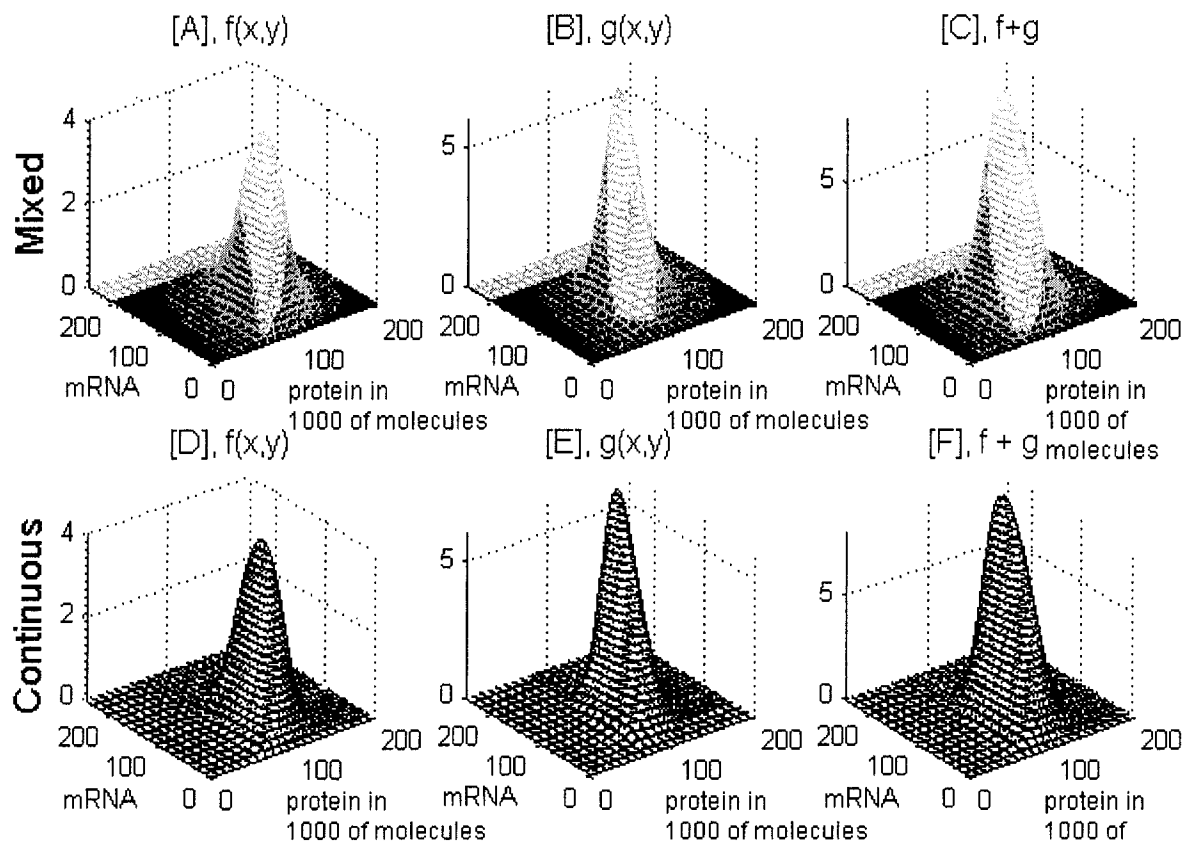
For eukaryotes, the Chapman-Kolmogorov equation (3.6)-(3.7) cannot be solved anymore due to its computational limitations. However, as indicated by the single cell trajectories (Fig. 3.3) the mixed model provides almost exact approximation in this case. Shown in Fig.

Figure 3.4 : Two-dimensional mRNA-protein distributions for "prokaryotes".



Two-dimensional mRNA-protein distributions for "prokaryotes" in the regime close to the statistical equilibrium ($H=10$, $K=6$, $r=0.25$, $c=3$, $b=2$): Panel A, B, C- exact distributions calculated on a grid $20 \text{ mRNA} \times 300 \text{ protein}$ molecules, panel D, E, F- mixed model on a grid 20×240 and panel G, H, I continuous model calculated on a grid 100×100 . Distributions shown in the same scale ($15 \text{ mRNA} \times 300 \text{ protein}$ molecules) and normalized to the exact description. In the left column, $f(x,y)$ - distributions for the gene in the inactive state ($G=0$), in the middle column, $g(x,y)$ - distributions for the gene in the active state ($G=1$) and in the right - the marginal distribution $\rho(x,y)=f(x,y)+g(x,y)$.

Figure 3.5 : 2D distributions for "eukaryotes" close to the statistical equilibrium.



Two dimensional mRNA-protein distributions for "eukaryotes" in the regime close to the statistical equilibrium ($H=200$, $K=250$, $r=0.25$, $c=3$, $b=2$): Panel A, B, C -distributions for mixed model calculated on a grid 300×300 , panel D, E, F -continuous model on a grid 200×200 . Distributions shown in the same scale ($250 \text{ mRNA} \times 2 \cdot 10^5 \text{ protein molecules}$) and normalized to the mixed model. In the left column, $f(x,y)$ - distributions for the gene in the inactive state ($G=0$), in the middle column, $g(x,y)$ - distributions for the gene in the active state ($G=1$) and in the right - the marginal distribution $\rho(x,y)=f(x,y)+g(x,y)$.

Table 3.1 : Variance explained for "prokaryotes".

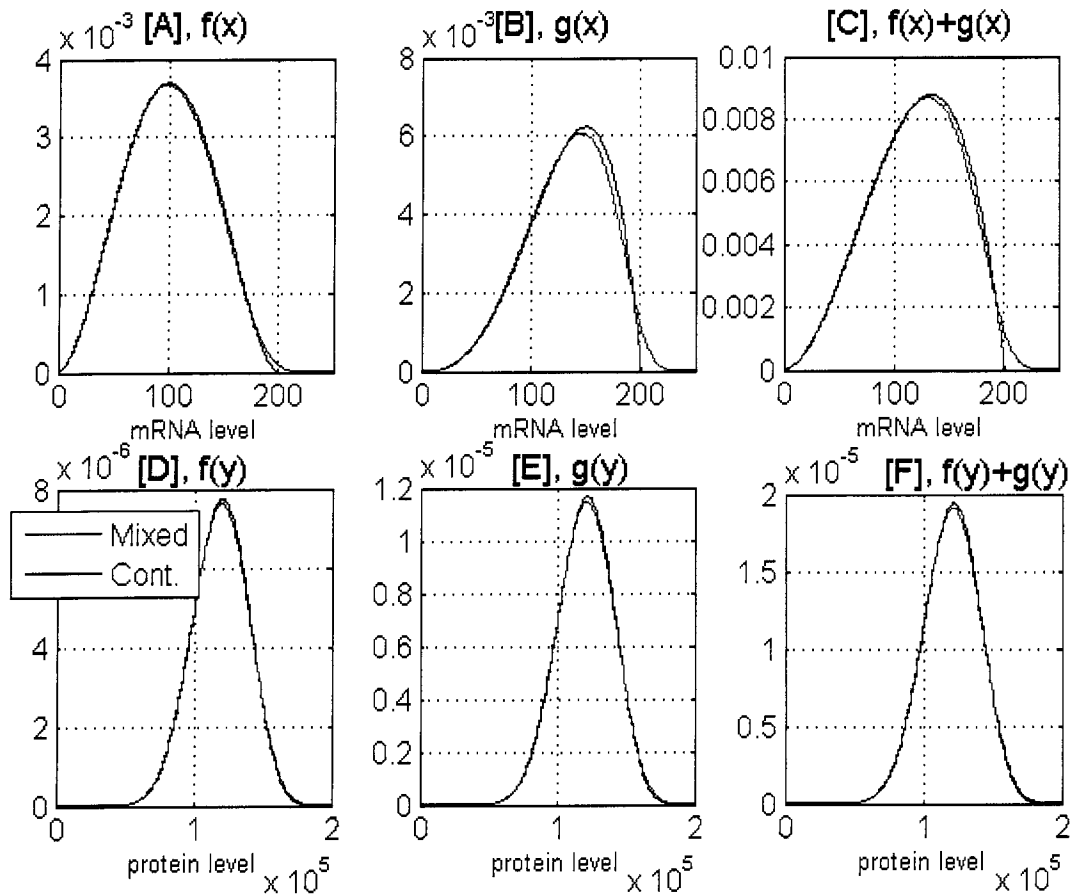
<i>model</i>	<i>mRNA</i>	<i>protein</i>
Mixed	1	0.89
Continuous	0.4	0.39

Percentage of the total variance given by the exact description explained by the mixed and continuous model for "prokaryotes" in the regime close to the equilibrium ($H=10$, $K=6$, $r=0.25$, $c=3$, $d=2$). Shown here are the ratios of mRNA and protein variance for the introduced continuous and mixed models, Eqs. (3.21)-(3.22) and Eqs. (3.30)-(3.31), respectively, and total variance given in Eqs. (3.10)-(3.11).

3.5 are the joint mRNA-protein distributions calculated for the mixed and continuous model for "eukaryotes" in the regime close to statistical equilibrium ($H=200$, $K=250$, $r=0.25$, $c=3$, $b=2$). The distributions are consistent with the single cell trajectories presented in Fig. 3.3. There is almost no visible difference between the continuous (panels E, F, G) and mixed approximation (panels A, B, C). Their marginal distributions (Fig. 3.6) reveal some discrepancies between them. The continuous approximation introduces artifacts in the mRNA marginal distributions (Fig. 3.6A, B, C -black curve): The distributions are bounded by 200 molecules, while the mixed model approximation (blue curve) exceeds that number. These minor differences in mRNA marginal distribution seem not to affect the protein marginal distribution; the continuous and mixed protein marginal distributions almost overlap each other. As shown in the Table 2, the mixed model in this case explains 100% of the total variance in mRNA and 99.9% of the protein variance. The continuous approximation accounts for 93% and 94% of the total mRNA and protein variance, respectively.

Shown in Fig. 3.7 are the joint mRNA-protein distributions calculated for the mixed and continuous model for "eukaryotes" in the regime far from the statistical equilibrium ($H=200$, $K=250$, $r=0.25$, $c=0.75$, $b=0.75$). The artifacts introduced by the continuous model are

Figure 3.6 : Marginal distributions for "eukaryotes" close to the statistical equilibrium.



Marginal distributions for "eukaryotes" ($H=200$, $K=250$, $r=0.25$, $c=3$, $b=2$) calculated from joint distributions presented in Fig. 3.5. Panel A, B, C: marginal mRNA distributions for mixed (blue) and continuous model (black), panel C, D, F marginal protein distribution for mixed (blue) and continuous model (black). The exact model The mRNA marginal distributions given by the mixed model are the same as for the exact description by construction.

Table 3.2 : Variance explained for "eukaryotes".

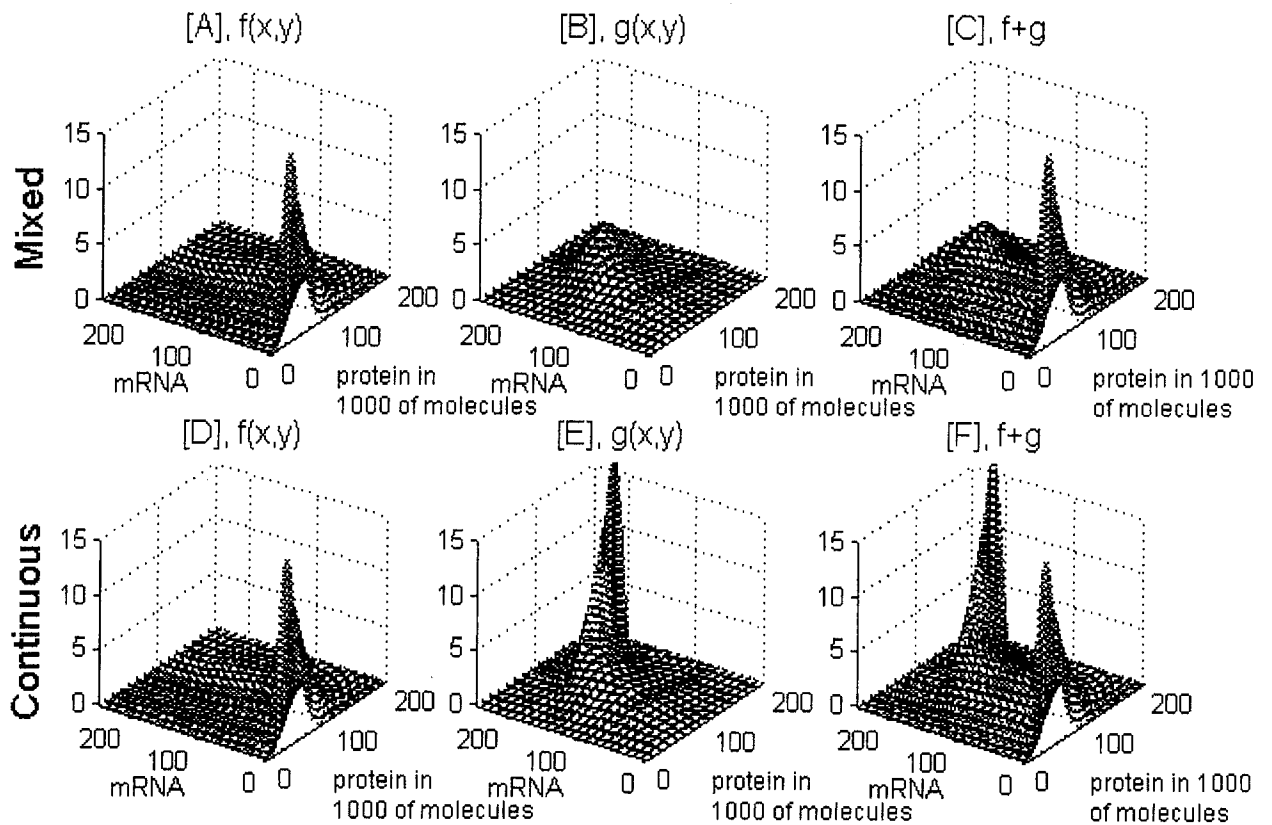
regime	c=3, d=2		c=0.75, d=0.75	
<i>Model</i>	<i>mRNA</i>	<i>protein</i>	<i>mRNA</i>	<i>protein</i>
Mixed	1	0.999	1	0.9999
Continuous	0.93	0.94	0.97	0.98
Kepler-Elston	-	1.12	-	1.11

Percentage of the total variance given by the exact description explained by the mixed, continuous and Kepler-Elston model for "eukaryotes" ($H=200$, $K=250$, $r=0.25$) in two regimes.

clearly visible: The boundary induced by ODEs (3.13)-(3.14) creates an artificially high peak at the gene active state (Fig. 3.7E, F), whereas the corresponding peak in the mixed approximation (Fig. 3.7B, C) is much smaller since the distribution can freely disperse. The distributions at the inactive gene state (Fig. 3.7A vs. D) are very much alike. The marginal distributions depicted in Fig. 3.8 reveal that the artifacts are present only for the amount of the mRNA transcript at the gene active state: The continuous approximation (black curve) is bounded by 200 mRNA, the mixed approximation (blue curve) exceeds this number (Fig. 3.8B, C). Nevertheless, the protein marginal distributions for both of the models overlap each other almost exactly (Fig. 3.8D, E, F). In this case, see Table 3.2, the mixed model explains 100% of the total variance in mRNA and 99.99% of the protein variance, while the continuous approximation accounts for 97% and 98% of mRNA and protein variation, respectively.

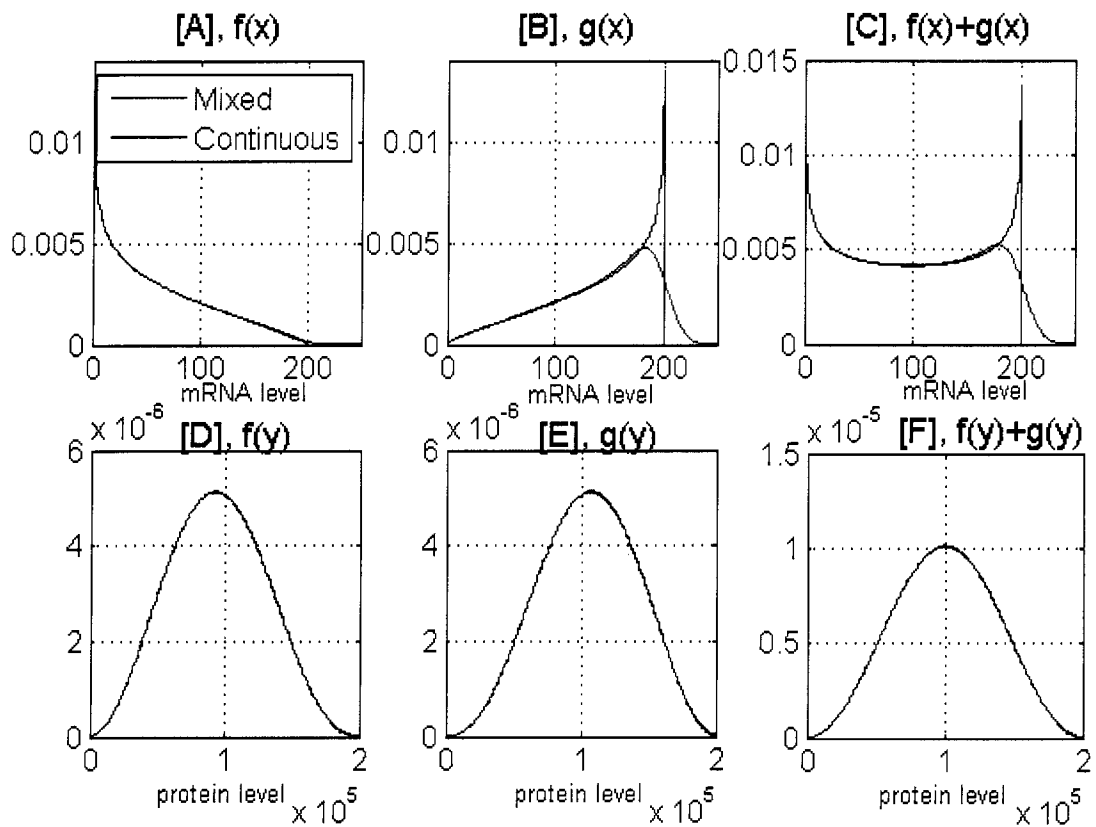
When the protein half-life time is smaller than the mRNA half-life time, i.e., $r > 1$, the artifacts introduced by the continuous approximation at the mRNA level (Fig. 3.8) are also exhibited at the protein level. It is due to the fact that the protein level follows the mRNA level very closely (data not shown). This case is depicted in Fig. 3.9, which includes marginal distributions for eukaryotes in the regime far from the statistical equilibrium ($H=200$, $K=250$,

Figure 3.7 : 2D distributions for "eukaryotes" far from the statistical equilibrium.



Two dimensional mRNA-protein distributions for "eukaryotes" in the regime far from the statistical equilibrium ($H=200$, $K=250$, $r=0.25$, $c=0.75$, $b=0.75$): Panel A, B, C -distributions for mixed model calculated on a grid 300×300 , panel D, E, F -continuous model on a grid 200×200 . Distributions shown in the same scale ($250 \text{ mRNA} \times 2 \cdot 10^5 \text{ protein molecules}$) and normalized to the mixed model. In the left column, $f(x,y)$ - distributions for the gene in the inactive state ($G=0$), in the middle column, $g(x,y)$ - distributions for the gene in the active state ($G=1$) and in the right- the marginal distribution $\rho(x,y)=f(x,y)+g(x,y)$.

Figure 3.8 : Marginal distributions for "eukaryotes" far from the statistical equilibrium.



Marginal distributions for "eukaryotes" ($H=200$, $K=250$, $r=0.25$, $c=0.75$, $b=0.75$) calculated from joint distributions presented in Fig. 3.7. Panel A, B, C: marginal mRNA distributions for mixed (blue) and continuous model (black), panel C, D, F marginal protein distribution for mixed (blue) and continuous model (black).

$r=2$, $c=0.75$, $b=0.75$). The protein degradation rate is assumed to be twice greater than that of the mRNA, i.e., $r=2$. Although the continuous model accounts for 98.1% of the total protein variance (mixed model account for 99.98%), the former predicts the marginal protein distributions with three maxima, while the correct distribution given by the mixed model has only two, Fig. 3.9F. Considered case of $r > 1$, i.e., the mRNA more stable than the protein can be encountered in signaling pathways when the rapid signal propagation is achieved through active protein degradation. This is the case of the NF- κ B pathway [36], [37], where the NF- κ B inhibitor I κ B α is catalytically degraded with a half-life time of about 10 min, while its mRNA has a half-life of about 20 min. In this case, the two-dimensional I κ B α mRNA-protein distribution is important for understanding the underlying dynamics of NF- κ B transcription factor, and moreover, only the mixed model provides a good approximation for that system.

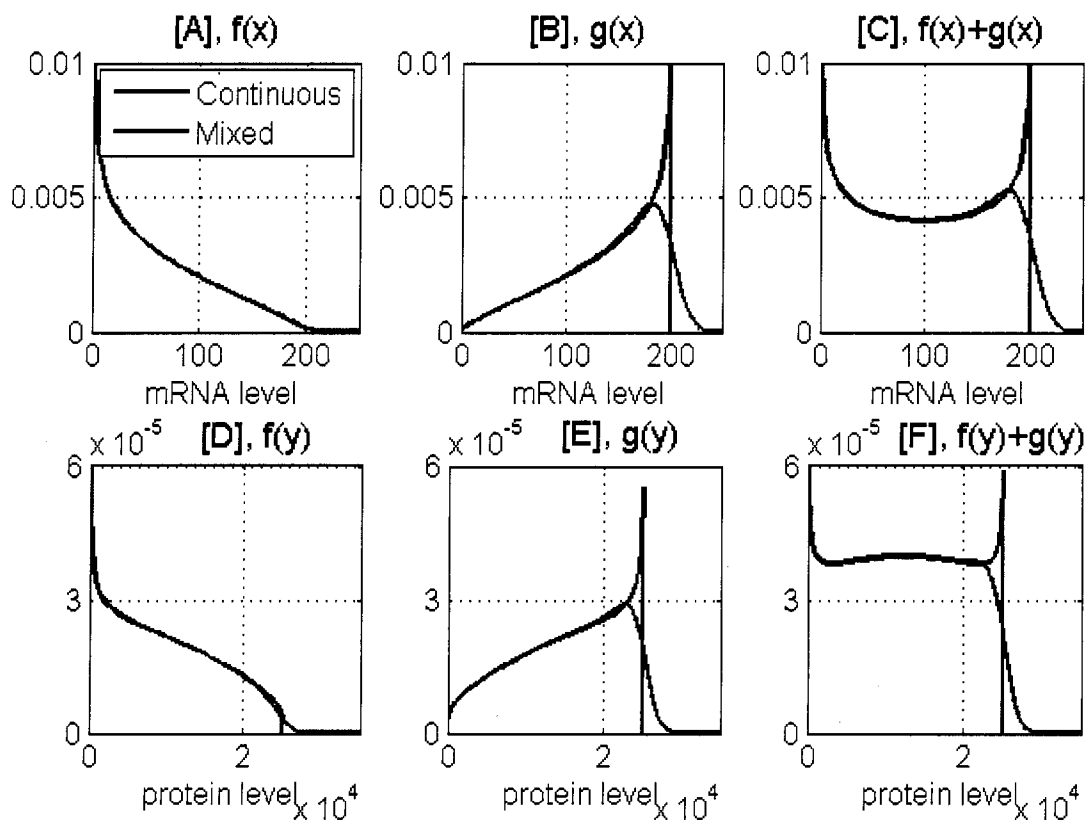
3.5 Model extensions

The following sections extend the proposed models, first, to the case of single haploid gene with feedback regulation, and second, to the case of n -allelic gene without feedback regulation.

3.5.1 Single haploid gene with feedback regulation

Consider a single haploid gene whose activity is regulated by the synthesized protein. Namely, assume that the gene activity rates $c = c(y)$ and $b = b(y)$ depend on the amount of the produced protein y [30]. Then, the stochastic process governing intermittent gene activity is given by the reactions analogous to the Eq. (3.1):

Figure 3.9 : Marginal distributions for "eukaryotes" when mRNA is more stable than protein.



Marginal distributions for "eukaryotes" in the regime far from the statistical equilibrium ($H=200$, $K=250$, $r=2$, $c=0.75$, $b=0.75$). The protein degradation rate is assumed to be twice greater than that of the mRNA, i.e., $r=2$. As a result the artifacts introduced by the continuous approximation at the mRNA level, Fig. 3.8, are now exhibited also at the protein level.



while the reactions (equations) describing the mRNA/protein production/decay noise remain the same for a given model, i.e., the exact stochastic description as well as the continuous or mixed approximation. This implies that the underlying distribution function for a given model is described by the system of equations analogous to the case without a feedback regulation, but with $c = c(y)$ and $b = b(y)$. In this case steady state distribution functions can be obtained with the same computational expenses as in the former description (Appendix C).

Note, that the extension to the feedback regulation do not allow deriving the moments of the underlying distribution. Consider an example of a haploid gene activated by produced protein monomers with a constitutive activation rates c_0 and b_0 , i.e., $c(y) = c_0 + c_1y$ and $b(y) = b_0$. When applying generation function techniques one finds that due to the linear term in $c(y)$, the n^{th} partial moments depend on the moments of the $n + 1$ order. Therefore, unlike the case without feedback regulation (Appendix A), the ODEs for the partial moments do not factorize into independent pairs of linear equations, but rather create a infinite system of ODEs which cannot be solved analytically. Therefore, when the feedback regulation is introduced into the model, the moments can be obtained only numerically based on the marginal mRNA/protein distribution.

3.5.2 n -allelic gene without feedback regulation

Previous sections were dedicated to the analysis of stochastic regulation of a single haploid gene. Such mode of regulation is natural for all prokaryotes, however higher organisms can possibly have more gene copies (alleles). In general eukaryotes are diploid, which means that

each gene has two homologous copies distributed among the chromosomes. In some cases one of these copies can become transcriptionally inactive. Occasionally, due to the gene or chromosomal duplications, the number of alleles per gene can be substantially larger.

Consider a single n -allelic gene without feedback regulation. Since the transition rates c and b are constant and do not depend on the amount of the produced protein y , the gene products resulting from different alleles are independent. Moreover, each double of random variables (X_j, Y_j) , $j = 1, \dots, n$, describing the number of mRNA and protein molecules produced by the j^{th} gene copy is identically distributed with a joint mRNA-protein distribution given for a single haploid gene considered previously.

Therefore, the moments of the gene products for a single n -allelic gene, $\tilde{X} = \sum_{j=1}^n X_j$ and $\tilde{Y} = \sum_{j=1}^n Y_j$, can be obtained based on the previous results:

$$E[\tilde{X}] = nE[X], \quad (3.38)$$

$$E[\tilde{Y}] = nE[Y], \quad (3.39)$$

$$\text{Var}[\tilde{X}] = n\text{Var}[X], \quad (3.40)$$

$$\text{Var}[\tilde{Y}] = n\text{Var}[Y], \quad (3.41)$$

where (X, Y) is the amount of mRNA/protein for a single haploid gene. Note, that the above is true for all of the considered models, i.e., the exact stochastic description as well as the continuous and mixed approximation.

In addition, the marginal mRNA-protein distribution function $\rho(\tilde{x}, \tilde{y})$ can be calculated by n^{th} order convolution of the distribution obtained for a single copy. In this case, the convolution has to be carried numerically.

The following chapter introduces further simplifications to the model, which considers systems of multiple interacting genes.

Chapter 4

Kepler-Elston approximation as a special case of continuous model

4.1 Derrivations of the Kepler-Elston approximation

While considering the continuous approximation (3.13-3.14) note, that for $r \ll 1$ (protein molecules much more stable than the mRNA transcript) the mRNA transcript reaches equilibrium much faster than the protein number. In this case, the dynamics of the mRNA production, Eq. (3.13), can be neglected in favor of its steady state $x=HG$. This assumption yields the following approximation in the case of a single haploid gene:



$$\frac{dy}{dt} = HKG - ry. \quad (4.2)$$

The system (4.1-4.2) is equivalent to the model introduced by Kepler and Elston (2001) [30], which assumed direct protein production from DNA.

The joint distribution of the double of random variables (y, G) describing the state of the system in any given time is given by the pair of probability density functions:

$$f(y, t)\Delta y = P[y(t) \in (y, y + \Delta y), G = 0], \quad (4.3)$$

$$g(y, t)\Delta y = P[y(t) \in (y, y + \Delta y), G = 1], \quad (4.4)$$

Analogically to the system (3.19-3.20), PDEs for probability density functions f and g yield:

$$\frac{df}{dt} - r \frac{\partial}{\partial y}(yf) = bg - cf, \quad (4.5)$$

$$\frac{dg}{dt} + \frac{\partial}{\partial y}((HK - ry)g) = -bg + cf. \quad (4.6)$$

The densities described by (4.5-4.6) have an analytical steady state solution for $y \in [0, \frac{KH}{r}]$. At the steady state Eqs. (4.5-4.6) for $f(y)$ and $g(y)$ yield

$$-\frac{d}{dy}(yf) = b_r g - c_r f, \quad (4.7)$$

$$\frac{d}{dy}\left(\left(\frac{KH}{r} - y\right)g\right) = -b_r g + c_r f, \quad (4.8)$$

where $c_r = \frac{c}{r}$, $b_r = \frac{b}{r}$. Adding Eqs. (4.7) and (4.8) gives the integral

$$\frac{d}{dy} \left[-yf + \left(\frac{KH}{r} - y\right)g \right] = 0, \quad (4.9)$$

which implies that

$$-yf + \left(\frac{KH}{r} - y\right)g = -f\left(\frac{KH}{r}\right) = g(0). \quad (4.10)$$

Since $f(y)$ and $g(y)$ are nonnegative, the condition $-f\left(\frac{KH}{r}\right) = g(0)$ implies that $f\left(\frac{KH}{r}\right) = g(0) = 0$. Hence, from Eq. (4.10), $g = yf/\left(\frac{KH}{r} - y\right)$. Then, the Eq. (4.7) yields

$$-\frac{d}{dy}(yf) = \frac{b_r y}{\frac{KH}{r} - y} f - c_r f, \quad (4.11)$$

or equivalently

$$\frac{f'}{f} = \frac{c_r - 1}{y} - \frac{b_r}{\frac{KH}{r} - y}. \quad (4.12)$$

This implies that

$$f(y) = A \cdot y^{c_r-1} \left(\frac{KH}{r} - y\right)^{b_r}, \quad (4.13)$$

$$g(y) = A \cdot y^{c_r} \left(\frac{KH}{r} - y\right)^{b_r-1}, \quad (4.14)$$

where

$$A = \frac{\Gamma(c_r + b_r)}{\Gamma(c_r)\Gamma(b_r)} \left(\frac{r}{KH}\right)^{c_r+b_r}, \quad (4.15)$$

is a normalizing constant. For $c_r < 1$, $\lim_{y \rightarrow 0} f(y) = \infty$ and for $b_r < 1$, $\lim_{y \rightarrow 1} g(y) = \infty$, while for $c_r > 1, b_r > 1$ one has that $f(0) = g(0) = f\left(\frac{KH}{r}\right) = g\left(\frac{KH}{r}\right) = 0$. The marginal distribution $\rho(y) := f(y) + g(y)$,

$$\rho(y) = f(y) + g(y) = A \cdot \frac{KH}{r} y^{c_r-1} \left(\frac{KH}{r} - y \right)^{b_r-1}, \quad (4.16)$$

describes the protein level regardless of the gene status. For $c_r < 1$ and $b_r < 1$, function $\rho(y)$ has a minimum between 0 and $\frac{KH}{r}$, whereas $\lim_{y \rightarrow 0} \rho(y) = \infty$ and $\lim_{y \rightarrow 1} \rho(y) = \infty$. For $c_r > 1$ and $b_r > 1$, $\rho(y)$ has one maximum. The larger c_r and b_r are, the distribution $\rho(y)$ is more concentrated since the gene activity changes more frequently.

Note that the steady state densities $f(y)$, $g(y)$ and $\rho(y)$ are in fact rescaled beta densities, therefore the partial protein moments joint with gene activity as well as the marginal moments can be directly derived. Alternatively, the moments can be derived by analysis of the system (4.5-4.6) with generating functions. The second approach allows derivation not only the moments at the steady state but also their time evolution. The expected value $E_{KE}[Y]$ and the variance $Var_{KE}[Y]$ of the stationary marginal distribution of protein, $\rho(y) = f(y) + g(y)$, are given by (the partial moments are presented in the Appendix A):

$$E_{KE}[Y] = \frac{c}{c+b} \frac{KH}{r}, \quad (4.17)$$

$$Var_{KE}[Y] = \frac{br}{c(r+c+b)} E_{KE}^2[Y]. \quad (4.18)$$

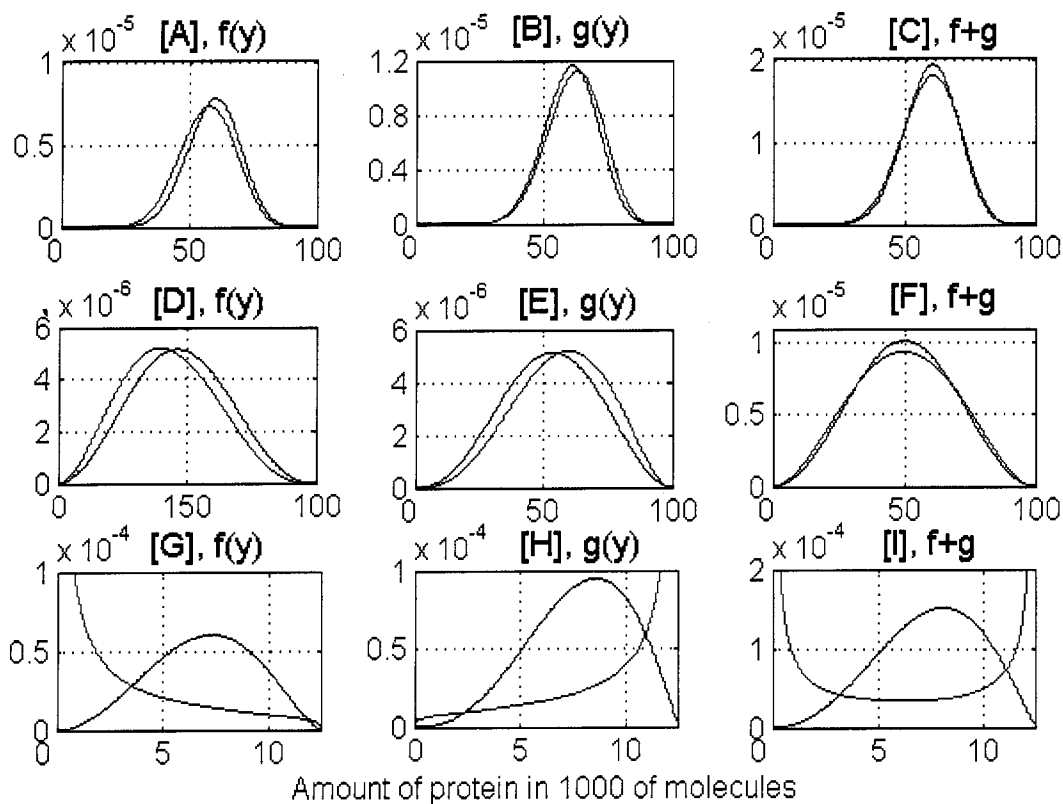
Please note that the expected value of the protein is the same as in the exact stochastic description, while the variance $Var_{KE}[Y]$ can be obtained from $Var_C[Y]$, i.e., the continuous approximation by assuming that $r \ll 1$.

4.2 Applicability of the Kepler-Elston approximation

To verify the Kepler-Elston approximation, stationary densities $f(y)$, $g(y)$ and $\rho(y)$ given by Eqs. (4.13), (4.14) and (4.16) are compared against the marginal protein distributions $\int f(x, y)dx$, $\int g(x, y)dx$ and $\int \rho(x, y)dx$ obtained numerically for the continuous approximation, Eqs. (3.28-3.29). As expected, the Kepler-Elston approximation is quite accurate when $r < 1$ (Fig. 4.1A-F, where $r=0.25$), especially for large c and b . However, even for a small r it introduces more variation than the continuous approximation: The marginal protein distributions, Fig. 4.1C, E are broader and have lower maxima. In fact, for $r=0.25$, $c=3$, $b=2$ Kepler-Elston model overapproximates the exact protein variance by 12%, and for $r=0.25$, $c=0.75$, $b=0.75$ by 11% (see Table 3.2). Fig. 4.1A, B, D, E reveals minor discrepancies between the Kepler-Elston and continuous protein distributions joint with gene activity, $f(y)$, $g(y)$. In fact, as given in Eq. (A.27), the expected values of the protein number joint with gene activity, $E_{KE}[Y, G = 0]$ and $E_{KE}[Y, G = 1]$, differ from these given by the continuous model. Note that the continuous as well as the mixed model gives the same expectations joint with gene activity as the exact stochastic description, Eq. (A.26).

The Kepler-Elston approximation fails when $r > 1$. The stationary protein distribution $\rho(y)$ for $r=4$, $c=3$, $b=2$ presented in Fig. 4.1I (magenta curve) is bimodal, while the marginal distribution $\int \rho(x, y)dx$ calculated based on the continuous approximation has only one maximum. This is due to the fact that although $c > 1$, $b > 1$, but $c_r < 1$, $b_r < 1$. Condition $r > 1$ corresponds to the case when the protein molecules are degraded faster than the mRNA transcript. Such situation may be encountered in the molecular pathways where the rapid signal propagation is achieved throughout catalytic protein degradation, e.g., NF- κ B pathway [36], [37]. In such systems the two-dimensional mRNA-protein distributions are important to understand the underlying dynamics.

Figure 4.1 : Protein distributions in the Kepler-Elston approximation.



Marginal protein distributions for a Kepler-Elston approximation (magenta curve), Eqs. (4.13), (4.14) and (4.16), compared against the marginal protein distribution calculated numerically based on the two-dimensional mRNA-protein distributions for a continuous model (black curve). Hypothetical "eukaryotic" cells are considered ($H=200$, $K=250$): Panels A, B, C correspond to $r=0.25$, $c=3$, $b=2$; Panels D, E, F correspond to $r=0.25$, $c=0.75$, $b=0.75$; Panels G, H, I correspond to $r=4$, $c=3$, $b=2$. In the left column, $f(y)$ -distributions for the gene in the inactive state, in the middle column, $g(y)$ -distributions for the gene in the active state and in the right - the marginal distribution regardless of gene status $\rho(y)=f(y)+g(y)$.

The Kepler-Elston approximation proves to be very accurate and well justified when $r \ll 1$ and provides a great simplification in the analysis. It can be used to analyze systems of two, and possibly more interacting genes.

4.3 Two gene systems in the Kepler-Elston approximation

Consider a system of two interacting genes in the Kepler-Elston approximation. Let y_1 and y_2 denote the amounts of protein related to the first and second gene, respectively. By analogy with Eq. (4.2) the ODE description of the model reads

$$\frac{dy_1(t)}{dt} = -r_1 y_1 + H_1 K_1 G_1(t), \quad (4.19)$$

$$\frac{dy_2(t)}{dt} = -r_2 y_2 + H_2 K_2 G_2(t), \quad (4.20)$$

where G_1 and G_2 are the binary random variables describing the state of each of the genes. Eqs. (4.19-4.20) are parametrized analogically to the system (3.1-3.3) with subscripts corresponding to the respective gene. Assume general transition rules between the gene activity states:

$$I_1 \xrightarrow{c'_1(y_1, y_2)} A_1, A_1 \xrightarrow{b'_1(y_1, y_2)} I_1, G_1(I_1) = 0, G_1(A_1) = 1, \quad (4.21)$$

$$I_2 \xrightarrow{c'_2(y_1, y_2)} A_2, A_2 \xrightarrow{b'_2(y_1, y_2)} I_2, G_2(I_2) = 0, G_2(A_2) = 1, \quad (4.22)$$

where the transition rates are, in general, functions of produced protein y_1 and y_2 . More specifically, assume that

$$c'_1(y_1, y_2) = c'_{10} + c'_{11}y_1 + c'_{12}y_2, \quad (4.23)$$

$$b'_1(y_1, y_2) = b'_{10} + b'_{11}y_1 + b'_{12}y_2, \quad (4.24)$$

$$c'_2(y_1, y_2) = c'_{20} + c'_{21}y_1 + c'_{22}y_2, \quad (4.25)$$

$$b'_2(y_1, y_2) = b'_{20} + b'_{21}y_1 + b'_{22}y_2, \quad (4.26)$$

where free terms correspond to the basal (constitutive) activation rates, while linear (inducible) terms correspond to the regulation due to the protein monomer binding.

One can introduce rescaled variables for problem (4.19-4.26):

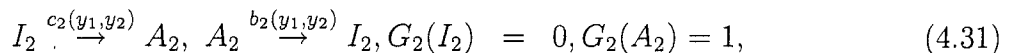
$$y_1^* = \frac{r_1}{H_1 K_1} y_1, y_2^* = \frac{r_2}{H_2 K_2} y_2, t^* = t \cdot r_1. \quad (4.27)$$

Substituting new variables and dropping the asterisks yields the following system:

$$\frac{dy_1(t)}{dt} = -y_1 + G_1, \quad (4.28)$$

$$\frac{dy_2(t)}{dt} = r(-y_2 + G_2), \quad (4.29)$$

where $r = \frac{r_2}{r_1}$ relates the half-life times of involved proteins (recall that r_i is the ratio of the protein and mRNA degradation rates for the i^{th} gene). The transition rules are given by



where

$$c_1(y_1, y_2) = c_{10} + c_{11}y_1 + c_{12}y_2, \quad (4.32)$$

$$b_1(y_1, y_2) = b_{10} + b_{11}y_1 + b_{12}y_2, \quad (4.33)$$

$$c_2(y_1, y_2) = c_{20} + c_{21}y_1 + c_{22}y_2, \quad (4.34)$$

$$b_2(y_1, y_2) = b_{20} + b_{21}y_1 + b_{22}y_2, \quad (4.35)$$

and $c_{i0} = c'_{i0}/r_1$, $b_{i0} = b'_{i0}/r_1$, $c_{i1} = c'_{i1}/K_1/H_1$, $b_{i1} = b'_{i1}/K_1/H_1$, $c_{i2} = c'_{i2} \cdot r/K_2/H_2$, $b_{i2} = b'_{i2} \cdot r/K_2/H_2$ for $i = 1, 2$ respectively.

The state of the system in any instant of time is given by the four random variables (y_1, y_2, G_1, G_2) . Define the probability density function

$$f_{ij}(y_1, y_2, t)\Delta y_1\Delta y_2 = P[y_1(t) \in (y_1, y_1 + \Delta y_1), y_2(t) \in (y_2, y_2 + \Delta y_2), G_1 = i, G_2 = j], \quad (4.36)$$

where $i, j = 0, 1$. By analogy with the continuous approximation, the time evolution of f_{ij} 's is given by the following system of PDEs:

$$\frac{df_{00}}{dt} + \text{div} \left[\left(\frac{dy_1}{dt} \Big|_{G_1=0}, \frac{dy_2}{dt} \Big|_{G_2=0} \right) f_{00} \right] = -(c_1 + c_2)f_{00} + b_1f_{10} + b_2f_{01}, \quad (4.37)$$

$$\frac{df_{10}}{dt} + \text{div} \left[\left(\frac{dy_1}{dt} \Big|_{G_1=1}, \frac{dy_2}{dt} \Big|_{G_2=0} \right) f_{10} \right] = -(c_2 + b_1)f_{10} + c_1f_{00} + b_2f_{11}, \quad (4.38)$$

$$\frac{df_{01}}{dt} + \text{div} \left[\left(\frac{dy_1}{dt} \Big|_{G_1=0}, \frac{dy_2}{dt} \Big|_{G_2=1} \right) f_{01} \right] = -(c_1 + b_2)f_{01} + c_2f_{00} + b_1f_{11}, \quad (4.39)$$

$$\frac{df_{11}}{dt} + \text{div} \left[\left(\frac{dy_1}{dt} \Big|_{G_1=1}, \frac{dy_2}{dt} \Big|_{G_2=1} \right) f_{11} \right] = -(b_1 + b_2)f_{11} + c_1f_{01} + c_2f_{10}. \quad (4.40)$$

At the steady state Eqs. (4.37-4.40) yield:

$$\frac{\partial}{\partial y_1}(-y_1 f_{00}) + \frac{\partial}{\partial y_2}(-r y_2 f_{00}) = -(c_1 + c_2) f_{00} + b_1 f_{10} + b_2 f_{01}, \quad (4.41)$$

$$\frac{\partial}{\partial y_1}[(1 - y_1) f_{10}] + \frac{\partial}{\partial y_2}(-r y_2 f_{10}) = -(c_2 + b_1) f_{10} + c_1 f_{00} + b_2 f_{11}, \quad (4.42)$$

$$\frac{\partial}{\partial y_1}(-y_1 f_{01}) + \frac{\partial}{\partial y_2}[r(1 - y_2) f_{01}] = -(c_1 + b_2) f_{01} + c_2 f_{00} + b_1 f_{11}, \quad (4.43)$$

$$\frac{\partial}{\partial y_1}[(1 - y_1) f_{11}] + \frac{\partial}{\partial y_2}[r(1 - y_2) f_{11}] = -(b_1 + b_2) f_{11} + c_1 f_{01} + c_2 f_{10}. \quad (4.44)$$

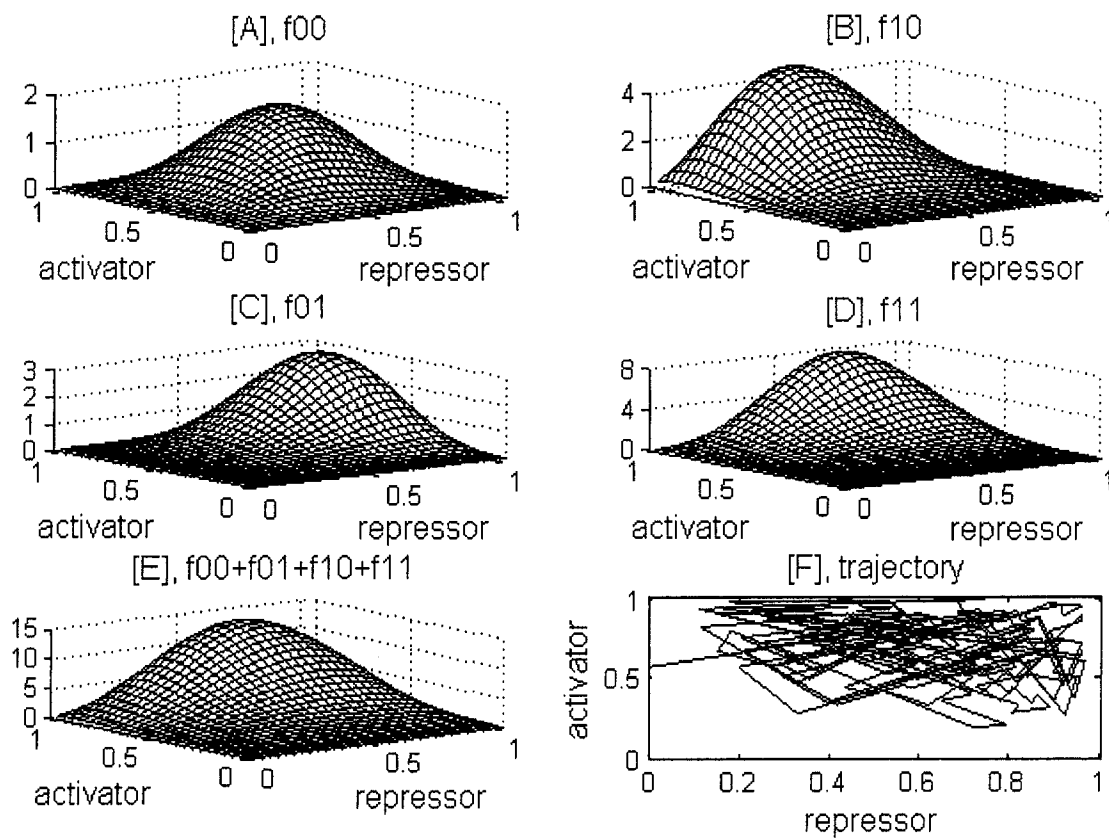
The system (4.41-4.44) can be numerically solved using developed numerical techniques. Appropriate discretization for the general case is included in the Appendix C. Note, that without the Kepler-Elston approximation, one would need to analyze four-dimensional distributions.

The following sections considers examples of two-gene regulatory networks: the activator-repressor and the repressor-repressor system.

4.3.1 Activator-repressor system

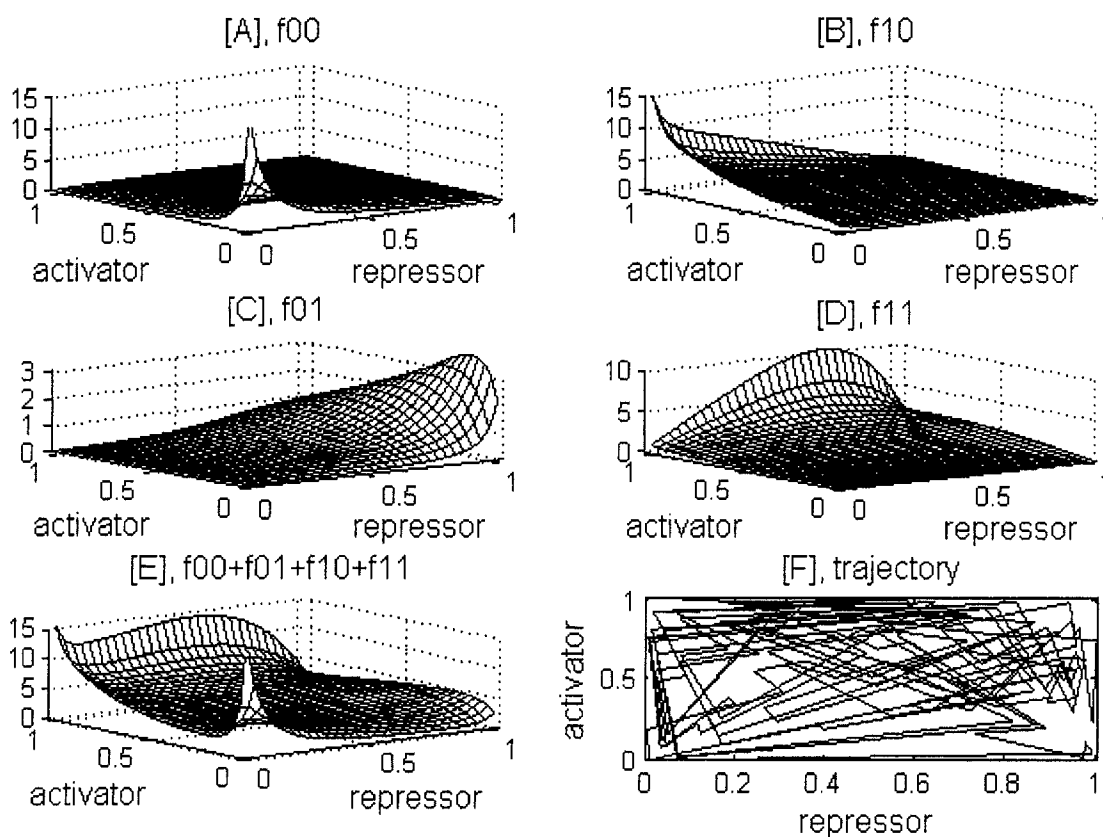
In the case of the activator y_1 - repressor y_2 system, assume the following transition intensities:

Figure 4.2 : 2D distributions for the activator-repressor system, case one.



Two-dimensional protein-protein distributions for $c_{10} = 4$, $b_{12} = 3$, $c_{21} = 4$, $b_{20} = 2$, $r = 1$. Panels A, B, C, D and E show functions f_{00} , f_{10} , f_{01} , f_{11} and the marginal distribution $f = f_{00} + f_{10} + f_{01} + f_{11}$. Panel F shows the example single trajectory for the same set of parameters.

Figure 4.3 : 2D distributions for the activator-repressor system, case two.



Two-dimensional protein-protein distributions for $c_{10} = 1$, $b_{12} = 1.5$, $c_{21} = 1.6$, $b_{20} = 1$, $r = 1$. Panels A, B, C, D and E show functions f_{00} , f_{10} , f_{01} , f_{11} and the marginal distribution $f = f_{00} + f_{10} + f_{01} + f_{11}$. Panel F shows the example single trajectory for the same set of parameters.

$$c_1(y_1, y_2) = c_{10}, \quad (4.45)$$

$$b_1(y_1, y_2) = b_{12}y_2(t), \quad (4.46)$$

$$c_2(y_1, y_2) = c_{21}y_1(t), \quad (4.47)$$

$$b_2(y_1, y_2) = b_{20}, \quad (4.48)$$

which implies that the inactivation of the activator y_1 is proportional to the amount of the repressor y_2 , while the activation of the repressor is proportional to the amount of the activator. In Figs. 4.2 and 4.3 the steady state solutions of the activator-repressor system for two sets of parameters are presented. When the transition intensities c_{10} , b_{12} , c_{21} , b_{20} are relatively large, the resulting protein-protein distribution $f = f_{00} + f_{10} + f_{01} + f_{11}$ has a single mode (Fig. 4.2). However, when transition intensities are smaller, the resulting distribution $f = f_{00} + f_{10} + f_{01} + f_{11}$ has three maxima and a complicated profile (Fig. 4.3).

4.3.2 Repressor-repressor system

In the case of the repressor y_1 - repressor y_2 system, assume the following transition intensities:

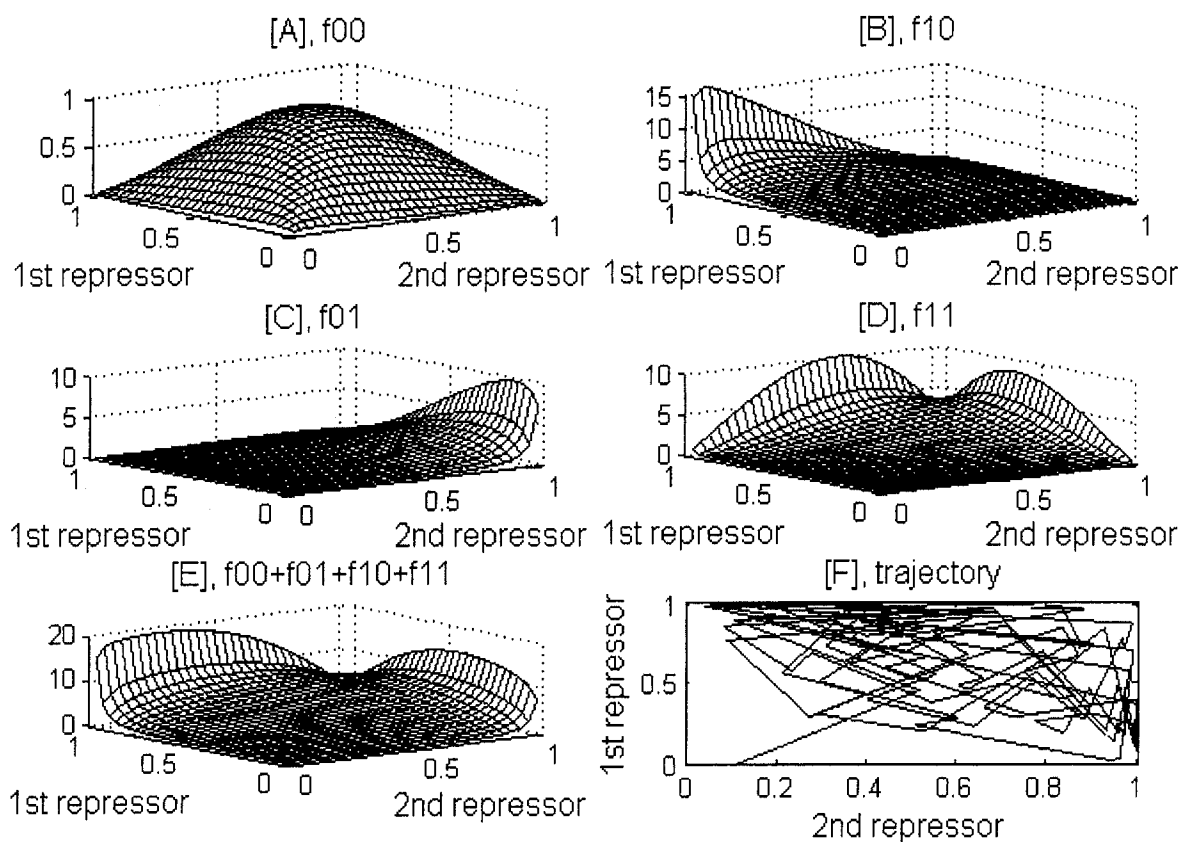
$$c_1(y_1, y_2) = c_{10}, \quad (4.49)$$

$$b_1(y_1, y_2) = b_{12}y_2(t), \quad (4.50)$$

$$c_2(y_1, y_2) = c_{20}, \quad (4.51)$$

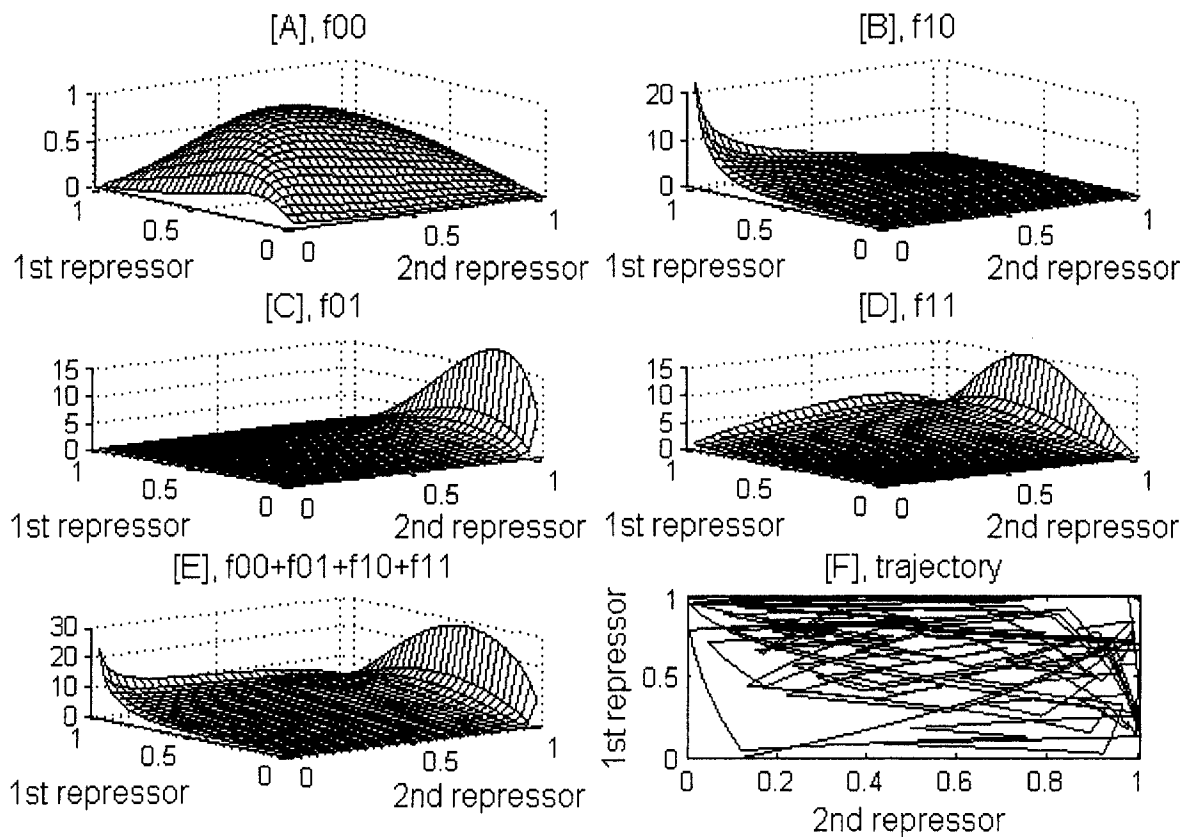
$$b_2(y_1, y_2) = b_{21}y_1(t), \quad (4.52)$$

Figure 4.4 : 2D distributions for the repressor-repressor system, case one.



Two-dimensional protein-protein distributions for $c_{10} = 1.7$, $b_{12} = 1.5$, $c_{20} = 1.5$, $b_{21} = 1.5$, $r = 1$. Panels A, B, C, D and E show functions f_{00} , f_{10} , f_{01} , f_{11} and the marginal distribution $f = f_{00} + f_{10} + f_{01} + f_{11}$. Panel F shows the example single trajectory for the same set of parameters.

Figure 4.5 : 2D distributions for the repressor-repressor system, case two.



Two-dimensional protein-protein distributions for $c_{10} = 2$, $b_{12} = 2$, $c_{20} = 2$, $b_{21} = 2$, $r = 2$. Panels A, B, C, D and E show functions f_{00} , f_{10} , f_{01} , f_{11} and the marginal distribution $f = f_{00} + f_{10} + f_{01} + f_{11}$. Panel F shows the example single trajectory for the same set of parameters.

which implies that the inactivation of the 1st repressor y_1 is proportional to the amount of the 2nd repressor y_2 and vice versa. In Figs. 4.4 and 4.5 the steady state solutions of the activator-repressor system for two sets of parameters are presented. One can observe that the repressor-repressor system is relatively unstable. Relatively small differences in constitutive production rates ($c_{10} = 1.7$, $c_{20} = 1.5$) lead to the substantial asymmetry in the resulting protein-protein distribution (Fig. 4.4). This asymmetry can be also a result of different half-life times of involved proteins as shown in Fig. 4.5, where the activation intensities are the same for both genes, while $r = 2$.

Chapter 5

Collective actions of multiple activators

For the sake of simplicity it is usually assumed that the transcriptional gene activity is due to the actions of single *trans*-acting regulatory molecule (transcription factor) and single *cis*-acting regulatory element, i.e., operator in bacteria or promoter in eukaryotes [32], [30], [29], [63], [7], [61], [66]. In fact, the specific patterns of gene expression may be governed by combinatorial interactions of series of transcription factors that bind to various regulatory sites within gene promoters and enhancers ([68], p. 72).

Such regulatory mechanism is hypothesized for the NF- κ B dependent genes in HeLa cells, where activity of the primary transcription factor (NF- κ B) cannot explain the behavior of induced genes [51].

5.1 Biological motivation

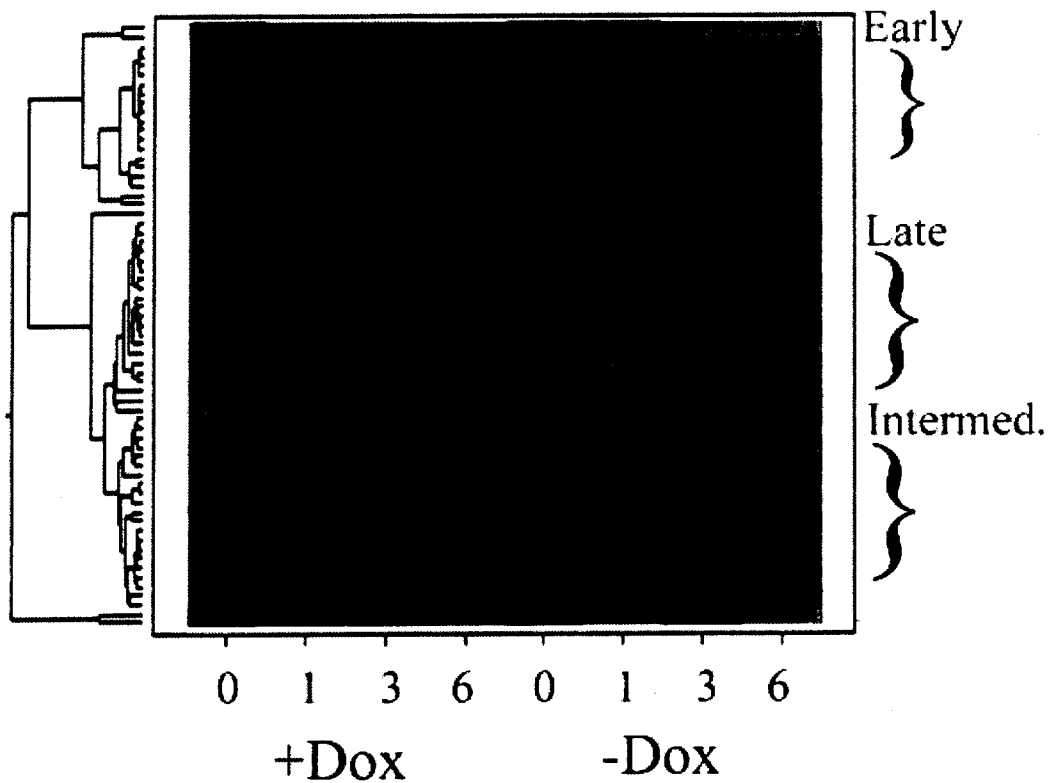
The NF- κ B family of transcription factors plays an important role in pathogen or cytokine inflammation, immune response, cell proliferation and survival [65]. In mammals, the NF- κ B family contains five members but its RelA subunit is responsible for the most common NF- κ B binding activity. In resting cells NF- κ B is sequestered in the cytoplasm by association with the members of another family of inhibitory proteins called I κ B. In response to extracellular signals such as the tumor necrosis factor- α (TNF), I κ B inhibitory proteins are degraded, which allows NF- κ B to translocate into the nucleus, bind to κ B motifs present in promoters

of numerous genes and upregulate their transcription.

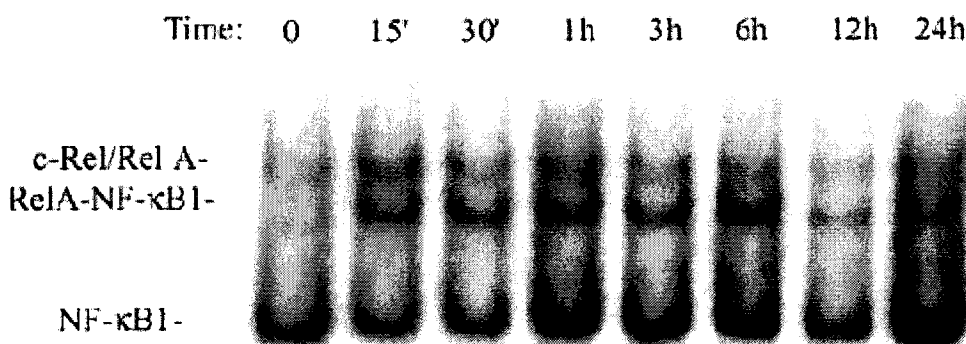
Recently, cells engineered to have NF- κ B activity controlled by exogenous doxycycline have been used to empirically identify the members of the NF- κ B dependent gene network by high density microarrays [64], [65]. In this system, the pattern of gene expression in wild type cells in response to a stimulus is compared against the pattern produced by the same stimulus in the absence of NF- κ B. In response to the TNF, 91 genes were identified to be NF- κ B dependent by analysis of variance. Hierarchical clustering was used to stratify these genes into common expression profiles. As shown in Fig. 5.1, the NF- κ B responsive genes can be grouped into 3 characteristic classes (regulons): early (such as $I\kappa B\alpha$, A20, $Gro\beta$ or IL8), for which the amount of mRNA transcript has its maximum at about 1 hour, intermediate (such as NF- κ B1 or TNFAIP2) with the maximum at 3 hours, and late (such as NAF1, NF- κ B2 or TRAF1) with the maximum at about 6 hours.

For previously studied human fibroblast, the nuclear activity of NF- κ B is terminated by the newly synthesized $I\kappa B\alpha$, which enters the nucleus, binds to NF- κ B and takes it out into the cytoplasm [64], [44], [36]. The experiments show that in HeLa cells, in contrast, NF- κ B is not effectively lead out of the nucleus by the $I\kappa B\alpha$, but rather, after entering the nucleus at 15 min from the beginning of TNF stimulation, it remains there for at least 6 hours at a steady level (Fig. 5.2A). Analysis of RelA, i.e., NF- κ B's functional subunit, association with promoters of early and late inducible genes reveals similar binding kinetics: Within 30 min from TNF treatment, transcriptionally active RelA binds to gene promoters and persists bound for at least 6 hours at a steady level (Fig. 5.2B).

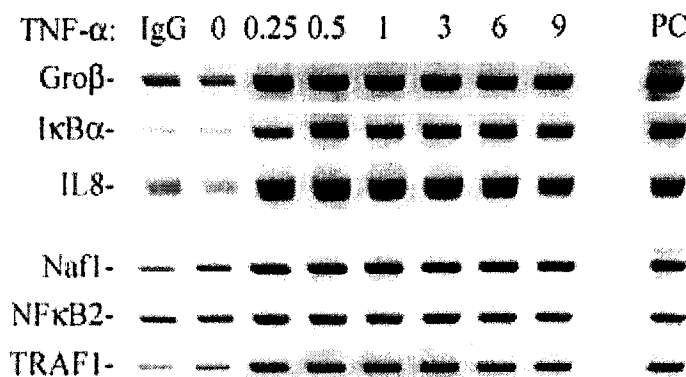
The finding that NF- κ B remains associated with early promoters throughout the 6 hours time course, even though expression of corresponding genes is actively being terminated (Fig. 5.1) strongly suggests the presence of a "repressor" that silences NF- κ B activity. Additionally,

Figure 5.1 : Three classes of NF- κ B dependent genes.

Kinetics of NF- κ B-dependent gene expression in HeLa cells (high expression depicted with red color). Data represent the mean of three independent time course experiments analyzed by high-density microarrays. In this experiment the NF- κ B nuclear translocation is enabled by culturing cells in the presence of doxycycline (Dox). The expression profiles for selected genes of each group were confirmed by Northern blots. The data reveal three characteristic classes of genes: early, intermediate and late, stratified by the time of 1, 3 and 6 hours at which the level of the mRNA transcript reaches the maximum value.

Figure 5.2 : Kinetics of NF- κ B transcription factor in HeLa cells.[A]. NF- κ B nuclear migration by EMSA

[B]. ChIP analysis of RelA association with specific promoters



[A] The Electrophoretic Mobility Shift Assay (EMSA) of the NF- κ B binding activity. HeLa S3 cells were stimulated with TNF α (30 ng/ml) for the indicated times prior to nuclear extraction and analysis of NF- κ B binding by EMSA. Shown is an autoradiogram of the protein-DNA complexes. The relative migration of the specific NF- κ B heterodimers is labeled. Rel A/NF- κ B1 and c-Rel-RelA complexes are rapidly induced by TNF treatment within 15 min. and persist in the nucleus for 6 h. A later peak at 24 h is also seen. [B] ChIP analysis of RelA association with promoters of early (I κ B α , Gro β and IL8) and late (NAF1, NF- κ B2 and TRAF1) genes. HeLa cells were stimulated for various times with TNF α prior to formaldehyde fixation. Rel A was used as the immuno-precipitating antibody. Far right lane is genomic DNA control. Similar RelA association with early and late promoters is observed: Binding occurs 30 min after TNF stimulation and persist for 9h at the steady level.

the fact that the same NF- κ B binding kinetics results in three different transcription profiles among the dependent genes, proves that the dynamics of NF- κ B transcription factor is not able to explain the observed phenomena. It is hypothesized that some other factors (co-activators) are required to initiate and terminate gene expression.

5.2 Model and derivations of expected expression profiles

To explain the expression profiles among 3 classes of genes, the model involving 3 activators and 1 repressor is proposed. Without an attempt to identify these regulatory factors, it is hypothesized that they might be activating (e.g. histone acetylation) or repressing factors, not necessarily connected with the DNA/protein binding.

The model relies on the introduced continuous approximation, however the analysis is limited to the mRNA level, since the protein abundance is not measured.

It is assumed that each gene has n potentially active homologous copies, and the activation and repression of these copies proceeds independently. The amount of mRNA transcript in a cell is a sum over amounts of transcript produced by each of homologous gene copies. Amount of mRNA transcript produced by a single gene copy j of a gene from the i^{th} class, denoted with $x_i^j(t)$, is given by

$$\frac{dx_i^j(t)}{dt} = -r_i \cdot x_i^j(t) + H \cdot G_i^j(t), \quad (5.1)$$

where $i = 1, 2, 3$ corresponds to 1st (early), 2nd (intermediate) and 3rd (late) class of genes, respectively, and $j = 1, \dots, n$ denotes the homologous gene copy. As previously, H depicts the transcription rate, while r_i corresponds to the mRNA degradation rate, which in general may differ between the genes from different classes. The function $G_i^j(t)$ is a binary

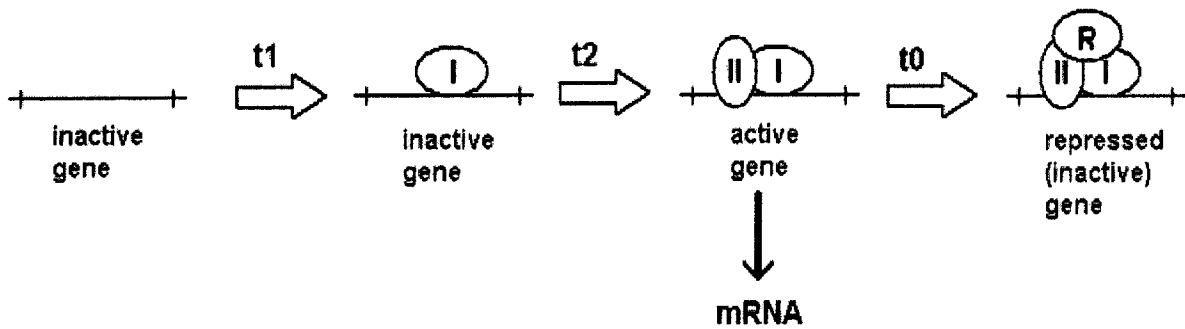
random variable describing the status of a gene.

The differences between considered gene classes are modeled through the stochastic process governing the gene activity, i.e., the function $G_i^j(t)$. It is assumed that an activation of a gene from the 1st class (early regulon) requires only binding of the first activator, while to activate a gene from the 2nd class (intermediate regulon) one additional co-activator is required, finally to activate a gene from the 3rd class (late regulons) two co-activators are required in addition of the first one. Activity of all genes is eventually terminated by a repressor. Such description implies that at a given instant of time a gene can be either in the inactive, active or repressed state. Analogous to the previous considerations, it is assumed that $G_i^j(t) = 1$ whenever all required activators occupy promotory region, but the repressor is not bound, and $G_i^j(t) = 0$ otherwise. Binding of the activators and of the repressor occurs with intensities for the activators equal to λ_1 , λ_2 and λ_3 , respectively, and with the intensity equal to λ_0 for the repressor. It is assumed that λ_i 's, $i = 0, 1, 2, 3$ are constant, which means that the corresponding regulatory factors are present at the constant amount in the cell.

It is assumed that co-activators act according to a conditional mechanism, so there exists a certain order of events. Let t_1 be the time of the first activator binding from the beginning of stimulation. Then, let t_2 be the time of the second activator binding, counted from the time of binding of the first activator. Similarly, let t_3 be equal to the duration of the period between second and third activator binding. Finally, let t_0 be the time between binding of the last activator and the repression event. Since the binding intensities are assumed to be constant, the t_i 's, $i = 0, 1, 2, 3$ are independent random variables exponentially distributed with parameters λ_i , $i = 0, 1, 2, 3$, respectively. The schematic representation of the model is depicted in Fig. 5.3 for the case of a gene from the second (intermediate) class.

In addition, let τ_i and ς_i be activation and repression times of genes from the i^{th} class.

Figure 5.3 : Schematic representation of the model for intermediate genes.



Schematic representation of the model for the genes from the second (intermediate) class. Two activators, I and II , required for gene to start mRNA transcription bind at the exponentially distributed times, t_1 and t_2 , respectively. Gene activity is terminated by repressor, R , binding at exponentially distributed time t_0 following activation. The representation for other gene classes would differ only in number of activators needed to initiate gene expression.

Specifically, for the first class, the activation time is equal to $\tau_1 = t_1$, for the second $\tau_2 = t_1 + t_2$, and for the third $\tau_3 = t_1 + t_2 + t_3$. The repression time among genes from the i^{th} class is equal to $\varsigma_i = \tau_i + t_0$. Each of these characteristic times is a sum the corresponding exponentially distributed random variables and its distributions can be analytically derived.

The solution of the Eq. (5.1) for a given initial conditions depends on the function $G_i^j(t)$, which is determined by the underlying stochastic process. The status of each homologous gene copy j from the i^{th} class $G_i^j(t) = 1$ if $t \in (\tau_i, \tau_i + t_0)$ and zero otherwise. Therefore to obtain the amount of transcript in the single cell produced by the gene from the i^{th} class, first, the function $G_i^j(t)$ is simulated by drawing τ_i and t_0 from the underlying probability densities for each gene copy j , second, the Eq. (5.1) is solved, and third, the amount of transcript is summed over all gene copies.

It is possible to describe the solution of Eq. (5.1) in the terms of its expected value, which corresponds to the gene expression measurements at the population level. Please note that from the Eq. (5.1), the expected number of mRNA transcript over time, $E[x_i^j(t)]$, (the average expression profile of a gene copy j from the i^{th} class) is described by the following differential equation:

$$\frac{dE[x_i^j(t)]}{dt} = g_i^j(t) \cdot H - r_i \cdot E[x_i^j(t)], \quad (5.2)$$

where $g_i^j(t) = E[G_i^j(t)]$ is the expected gene activity. The expected gene activity $g_i^j(t)$ can be evaluated as:

$$g_i^j(t) = \int_0^\infty \int_0^\infty 1_{[\tau_i, \tau_i + t_0)}(t) f_{\tau_i}(\tau_i) f_0(t_0) dt_0 d\tau_i = \int_0^t \int_{t-\tau_i}^\infty f_{\tau_i}(\tau_i) f_0(t_0) dt_0 d\tau_i, \quad (5.3)$$

where $1_A(t)$ is an indicator function equal to 1 if $t \in A$ and zero otherwise, $f_0(t)$ is the probability density function of repressor binding and $f_{\tau_i}(t)$, $i = 1, 2, 3$ are the activation distributions for consecutive gene classes.

The probability density functions $f_0(t)$ and $f_{\tau_1}(t)$ are given by assumption:

$$f_0(t) = \lambda_0 e^{-\lambda_0 t}, \quad (5.4)$$

$$f_{\tau_1}(t) = \lambda_1 e^{-\lambda_1 t}, \quad (5.5)$$

while to calculate $f_{\tau_2}(t)$, note that $\tau_2 = t_1 + t_2$ is a sum of two exponentially distributed independent random variables, thus its density is given by the convolution formula:

$$f_{\tau_2}(t) = \int_0^t \lambda_1 \lambda_2 e^{-\lambda_1 \tau_2} e^{-\lambda_2 (t - \tau_2)} d\tau_2 = \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} (e^{-\lambda_1 t} - e^{-\lambda_2 t}). \quad (5.6)$$

Analogically, the distribution $f_{\tau_3}(t)$ follows by noticing that $\tau_3 = \tau_2 + t_3$, i.e.,

$$f_{\tau_3}(t) = \frac{\lambda_1 \lambda_2 \lambda_3}{\lambda_2 - \lambda_1} \left[\frac{1}{\lambda_3 - \lambda_1} (e^{-\lambda_1 t} - e^{-\lambda_3 t}) - \frac{1}{\lambda_3 - \lambda_2} (e^{-\lambda_2 t} - e^{-\lambda_3 t}) \right]. \quad (5.7)$$

Given that, $g_i^j(t)$, the expected gene activity of a single copy j from the i^{th} class, $i =$

1, 2, 3, can be obtained by solving Eq. (5.3). One has the following:

$$g_1^j(t) = \frac{\lambda_1}{\lambda_1 - \lambda_0} e^{-\lambda_0 t} + \frac{\lambda_1}{\lambda_0 - \lambda_1} e^{-\lambda_1 t}, \quad (5.8)$$

$$g_2^j(t) = \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_0)(\lambda_2 - \lambda_0)} e^{-\lambda_0 t} + \frac{\lambda_1 \lambda_2}{(\lambda_0 - \lambda_1)(\lambda_2 - \lambda_1)} e^{-\lambda_1 t} \quad (5.9)$$

$$+ \frac{\lambda_1 \lambda_2}{(\lambda_0 - \lambda_2)(\lambda_1 - \lambda_2)} e^{-\lambda_2 t}, \quad (5.10)$$

$$g_3^j(t) = \frac{\lambda_1 \lambda_2 \lambda_3}{(\lambda_1 - \lambda_0)(\lambda_2 - \lambda_0)(\lambda_3 - \lambda_0)} e^{-\lambda_0 t} + \frac{\lambda_1 \lambda_2 \lambda_3}{(\lambda_0 - \lambda_1)(\lambda_2 - \lambda_1)(\lambda_3 - \lambda_1)} e^{-\lambda_1 t} \quad (5.11)$$

$$+ \frac{\lambda_1 \lambda_2 \lambda_3}{(\lambda_0 - \lambda_2)(\lambda_1 - \lambda_2)(\lambda_3 - \lambda_2)} e^{-\lambda_2 t} + \frac{\lambda_1 \lambda_2 \lambda_3}{(\lambda_0 - \lambda_3)(\lambda_1 - \lambda_3)(\lambda_2 - \lambda_3)} e^{-\lambda_3 t}.$$

In general, in the case of N gene classes the expected gene activity $g_i^j(t)$, $1 \leq i \leq N$, can be expressed as:

$$g_i^j(t) = \sum_{k=0}^i \left[\left(\frac{\prod_{l=1}^i \lambda_l}{\prod_{l=0, l \neq k}^i (\lambda_l - \lambda_k)} \right) e^{-\lambda_k t} \right]. \quad (5.12)$$

Then, the average expression profile of a gene copy j from the i^{th} class, can be derived by substituting Eq. (5.12) into Eq. (5.2) and solving it for $E[x_i^j(t)]$. Assuming zero initial conditions, i.e., $x_i^j(0) = 0$,

$$E[x_i^j(t)] = \sum_{k=0}^i \left[\frac{H}{r_i - \lambda_k} \left(\frac{\prod_{l=1}^i \lambda_l}{\prod_{l=0, l \neq k}^i (\lambda_l - \lambda_k)} \right) (e^{-\lambda_k t} - e^{-r_i t}) \right]. \quad (5.13)$$

More specifically, in the case of considered three gene classes, $i = 1, 2, 3$, the Eq. (5.13) gives the following expected expression profiles:

$$E[x_1^j(t)] = \frac{\lambda_1 H}{(\lambda_1 - \lambda_0)(r_1 - \lambda_0)} (e^{-\lambda_0 t} - e^{-r_1 t}) + \frac{\lambda_1 H}{(\lambda_0 - \lambda_1)(r_1 - \lambda_1)} (e^{-\lambda_1 t} - e^{-r_1 t}), \quad (5.14)$$

$$\begin{aligned}
E [x_2^j(t)] &= \frac{\lambda_1 \lambda_2 H}{(\lambda_1 - \lambda_0)(\lambda_2 - \lambda_0)(r_2 - \lambda_0)} (e^{-\lambda_0 t} - e^{-r_2 t}) \\
&+ \frac{\lambda_1 \lambda_2 H}{(\lambda_0 - \lambda_1)(\lambda_2 - \lambda_1)(r_2 - \lambda_1)} (e^{-\lambda_1 t} - e^{-r_2 t}) \\
&+ \frac{\lambda_1 \lambda_2 H}{(\lambda_0 - \lambda_2)(\lambda_1 - \lambda_2)(r_2 - \lambda_2)} (e^{-\lambda_2 t} - e^{-r_2 t}),
\end{aligned} \tag{5.15}$$

$$\begin{aligned}
E [x_3^j(t)] &= \frac{\lambda_1 \lambda_2 \lambda_3 H}{(\lambda_1 - \lambda_0)(\lambda_2 - \lambda_0)(\lambda_3 - \lambda_0)(r_3 - \lambda_0)} (e^{-\lambda_0 t} - e^{-r_3 t}) \\
&+ \frac{\lambda_1 \lambda_2 \lambda_3 H}{(\lambda_0 - \lambda_1)(\lambda_2 - \lambda_1)(\lambda_3 - \lambda_1)(r_3 - \lambda_1)} (e^{-\lambda_1 t} - e^{-r_3 t}) \\
&+ \frac{\lambda_1 \lambda_2 \lambda_3 H}{(\lambda_0 - \lambda_2)(\lambda_1 - \lambda_2)(\lambda_3 - \lambda_2)(r_3 - \lambda_2)} (e^{-\lambda_2 t} - e^{-r_3 t}) \\
&+ \frac{\lambda_1 \lambda_2 \lambda_3 H}{(\lambda_0 - \lambda_3)(\lambda_1 - \lambda_3)(\lambda_2 - \lambda_3)(r_3 - \lambda_3)} (e^{-\lambda_3 t} - e^{-r_3 t}).
\end{aligned} \tag{5.16}$$

In addition, define a gene from the i^{th} class be active at the time t if the mRNA transcript is produced by at least one of its copies. Assuming that n copies of a gene act independently, the proportion of active cells (i.e. cells with an active gene) in the population at time t can be evaluated using Eq. (5.12):

$$E [g_i(t)] = 1 - (1 - g_i^j(t))^n. \tag{5.17}$$

$E [g_i(t)]$ constitutes a measure of transcriptional variability among genes from different classes.

5.3 Fit at the population level

The proposed model was fit to reproduce 3 different expression profiles of NF- κ B dependent genes presented in Fig. 5.1 (i.e. the characteristic times of the maximum mRNA transcript abundance given at 1, 3 and 6 hours after the stimulation for early, intermediate and late genes, respectively). It is assumed that every gene has four potentially active homologous copies (in fact HeLa cell are almost tetraploidy, since their modal number is 82 [2]). In addition, it is assumed that all genes have the same mRNA degradation half-life time equal to 20 min, which corresponds to degradation rate $r=0.00057 \text{ s}^{-1}$. This is in agreement with the experimental data from Blattner et al. (2000) [6], who estimated the degradation half-life time for early gene I κ B α to be within 15 to 30 min range. In fact, the two fold increase of the degradation half-life time results in relatively small change in expression profiles [51]. A common transcription rate among genes is assumed, equal to 4 mRNA molecules per minute per gene copy ($H=0.0667 \text{ s}^{-1}$), which was confirmed for β -actin by single RNA transcript visualization [15]. Given this, the profiles are determined by the set of four parameters λ_i , $i = 0, 1, 2, 3$ describing binding rates of three activators and one repressor.

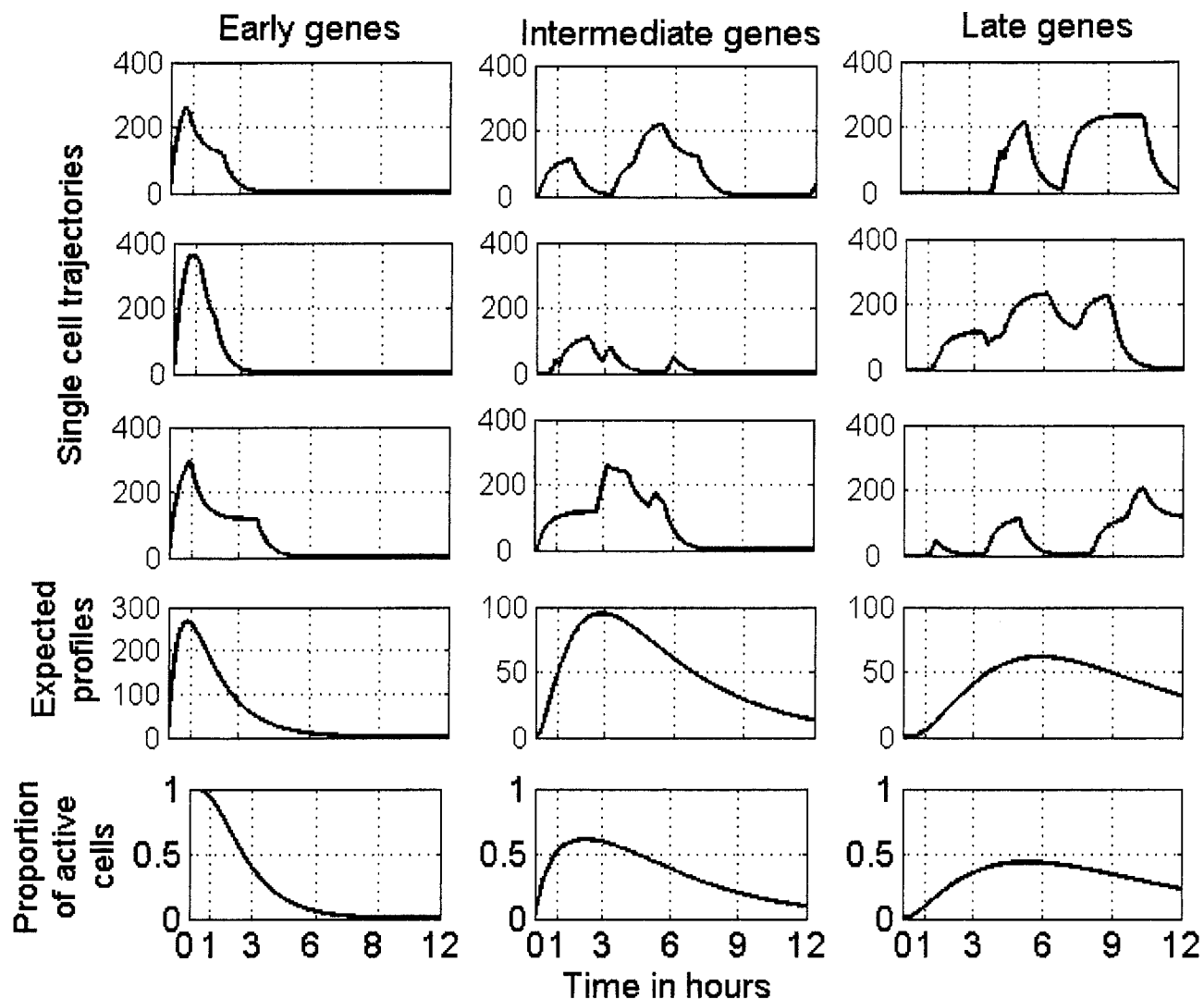
The fitting procedure was carried out analytically by taking the time derivative of the profiles given by Eq. (5.14)-(5.16) and equating it to zero. Given the 3 characteristic times of peak transcription at 1, 3, and 6 hour the parameters can be derived by solving the equations so obtained. Unfortunately there are four unknown parameters and only three expressions describing them. Thus, taking into account the fast dynamics of NF- κ B, the expected binding time of the first activator is assumed to be 10 min, which corresponds to the binding rate $\lambda_1=1.667 \times 10^{-3} \text{ s}^{-1}$. Then, the remaining parameters are determined by numerically solving the given equations. As a result, the following binding rates were fitted: $\lambda_2=7.4 \times 10^{-5} \text{ s}^{-1}$, $\lambda_3=8.1 \times 10^{-5} \text{ s}^{-1}$ and $\lambda_0=1.96 \times 10^{-4} \text{ s}^{-1}$ which corresponds to

the expected binding times of 225, 205 and 85 min for the two remaining activators and one repressor, respectively. Note that the obtained fit is not unique, the different assumption about the degradation rate and expected binding of the first activator would result in different estimates for other parameters.

The fit for three hypothetical genes belonging to the early, intermediate and late class, respectively, is depicted in Fig. 5.4. First three rows correspond to the single cell simulations. These trajectories have characteristic kinks, which reflect initiation or termination of expression in any of four homologous copies of the gene. The fourth row includes expected mRNA trajectories analytically derived in Eq. (5.14)-(5.16), corresponding to expression profiles averaged over the population of cells. These should be compared to experimental data in Fig. 5.1. Finally, in the last row, the gene activity over time is shown, defined in Eq. (5.17) as the proportion of the cells in the culture with at least one of the four gene copies active. Single cell expression trajectories are significantly different from the profiles obtained for the population of cells. In fact, no individual cell behaves like an “average” one. This is especially visible for the late genes, where the variability among single cell profiles is much larger than for early and intermediate genes. Another interesting observation is that not all cells (genes) are active in the population even at their peak transcriptional activity. The proportion of active cells (with at least one active copy) at their maximum activity is 1 for the early class, but significantly decreases to 0.61 and 0.43 for intermediate and late genes, respectively. The same statistics for a haploid cells yields proportions of 0.75, 0.21 and 0.13 for early, intermediate and late genes in the active state in the population at their peak transcriptional activity.

The average expression profiles obtained fit the microarray observations on NF- κ B dependent genes very well. The fact that variability among single cell profiles is larger for

Figure 5.4 : Expected expression profiles.



The mRNA profiles for early, intermediate and late genes. First 3 rows show single cell mRNA profiles, while the profiles in fourth row depict the expected expression in the cell culture derived in Eq. (5.14)-(5.16). These latter should be compared to experimental data in Fig. 5.1. Finally, in the last row, gene activity over time is shown, defined in Eq. (5.17) as the proportion of cells in the culture with at least one of four gene copies active. The kinks visible on single cell profiles correspond to initiation or termination of expression in any of the homologous copies of the gene. The difference among single cell profiles is larger for the late genes and, as a result, the averaged expression profile is broader, what is well confirmed by Northern blot data [47].

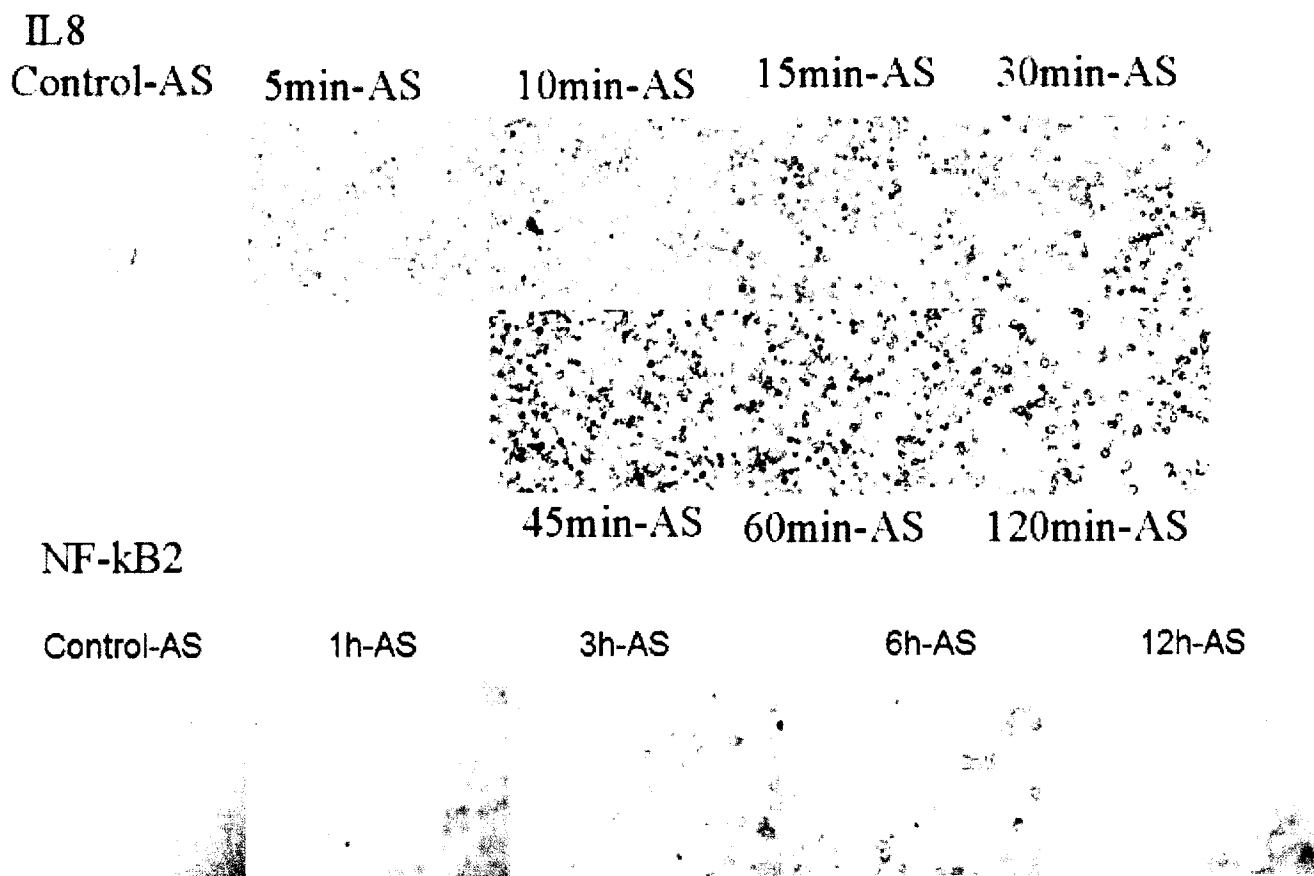
late genes and, as a result, the averaged expression profile is broader, was well confirmed by Northern blot data [47]. From the other hand, derived expression profiles cannot be fitted assuming just one or two activators including NF- κ B, unless biologically unjustified parametrization is assumed [51].

5.4 Experimental distribution functions

In addition to the data on the expected expression profiles, another set of experiments provided a time dependent mRNA distributions for two specific NF- κ B regulated genes. The experiments were conducted for IL8, which is an early gene, and NF- κ B2, which is a late gene. The method of in situ hybridization was used: TNF stimulated HeLa cells were harvested at multiple time instants and fixed, and then the targeted mRNA was hybridized to the probes, which were then visualized by dye. Each of the individual cell can be characterized by the amount of the dark dye, which is proportional to the level of the mRNA transcript in the cell at a given time instant. Measurements (images) with multiple replicates were taken at 0, 5, 10, 15, 30, 45, 60 and 120 min after TNF stimulation for the IL8 gene, and at 0, 1, 3, 6, 12 hours for the NF- κ B2. Acquired images (only one replicate per time is shown) depicting large colonies of HeLa cells are presented in Fig. 5.5. Further image quantification yielded time dependent distributions (histograms) of the dye intensity in the population of cells (Figs. 5.6 and 5.7) [16]. Each cell from the population captured at consecutive time instants was represented by their average dye intensity in the CMY space. Two directions in the the color space were estimated for each population: First, corresponding to dye intensity, second corresponding to the background staining. Since the obtained vectors were not perpendicular, the background staining component was removed by subtraction.

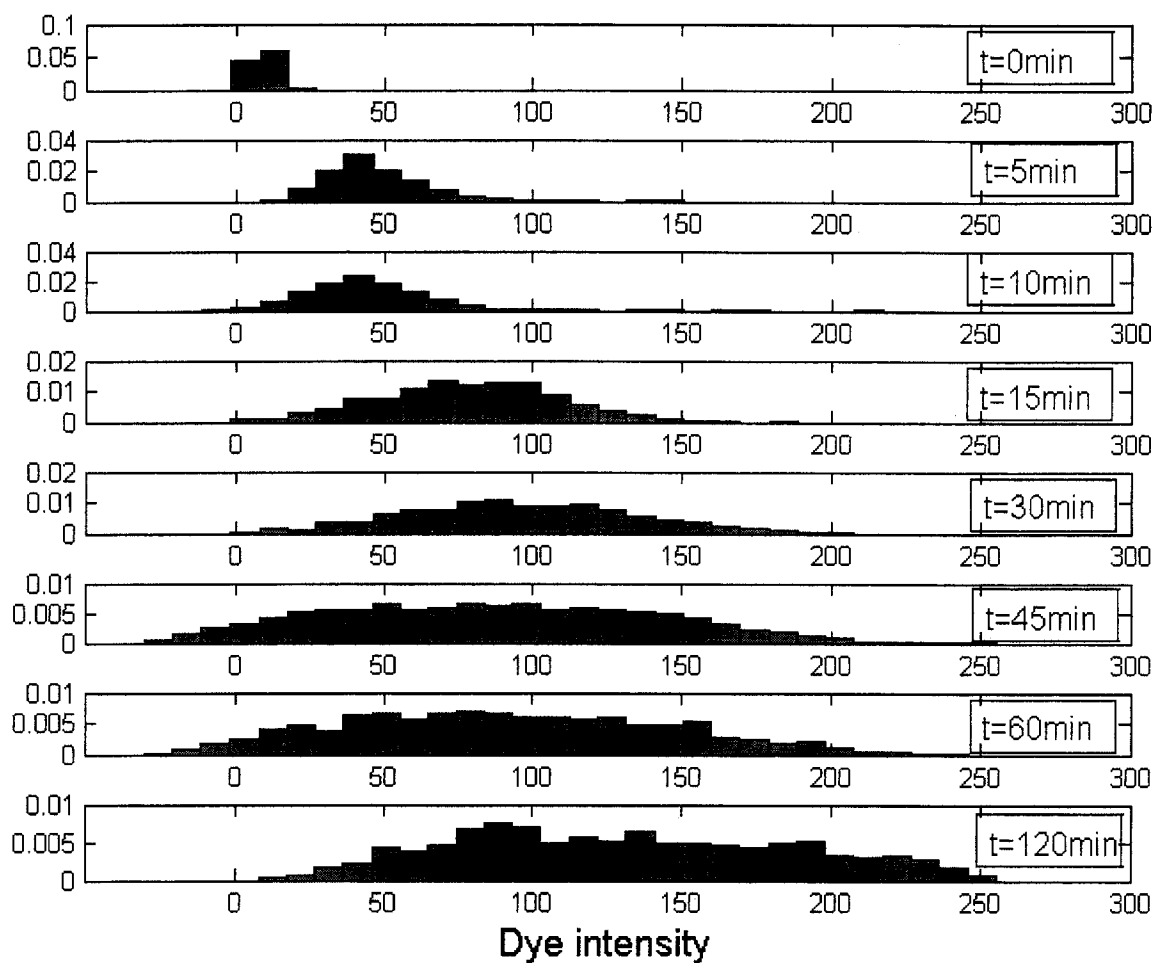
As shown in the Fig. 5.6, the initial dye intensity distribution for the IL8 gene spreads

Figure 5.5 : In situ hybridization measurements.



TNF stimulated HeLa cells are harvested and the target mRNA transcript is hybridized to a dyed probes. Shown here are the acquired images of cell colonies: Top panel, IL8 mRNA transcript at 0, 5, 10, 15, 30, 45, 60 and 120 min after TNF stimulation; Bottom panel, NF- κ B2 mRNA transcript at 0, 1, 3, 6, 12 hours after TNF stimulation. Each individual cell can be characterized by the amount of the dark dye, which is proportional to the level of the corresponding mRNA transcript.

Figure 5.6 : Quantified dye intensity distributions for IL8 gene.



Quantification of the in situ hybridization data on the IL8 mRNA. Shown are histograms of dye intensity at 0, 5, 10, 15, 30, 45, 60 and 120 min after TNF stimulation. Note the negative dye intensity at 10, 45 and 60 min, which is attributable to the different background colors.

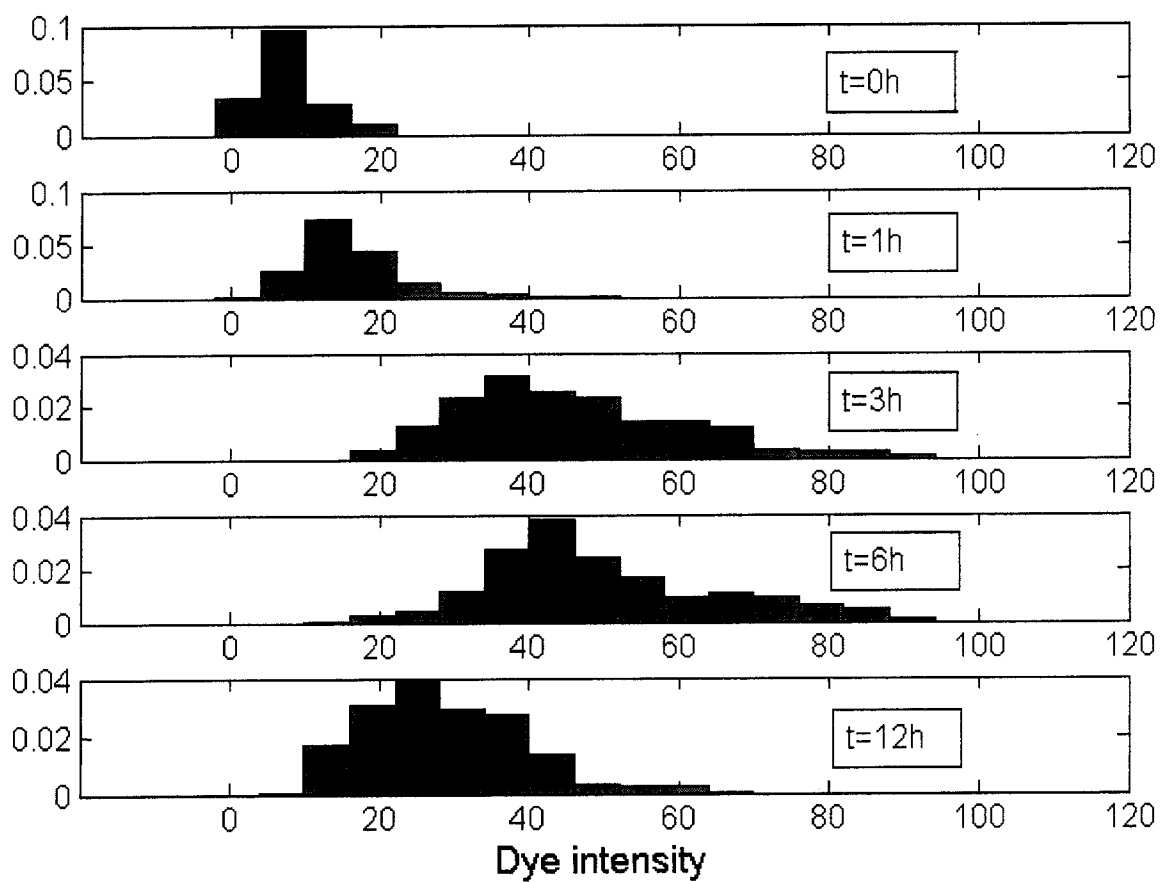
right in time as more mRNA transcript is being produced. However, some discrepancies are observed: At 10, 45 and 60 min after stimulation more dye than expected is concentrated around zero intensity. Similarly, by the time of 120 min one still cannot observe effective repression, which was apparent at this time from the expected expression profiles depicted in Fig. 5.4 or microarray data, Fig. 5.1. Instead, the dye distribution continues to spread right comparing with the measurement at 60 min. In the case of the NF- κ B2 (Fig. 5.7), where the time horizon is larger, the initial distribution spreads right with time again, to be effectively repressed at 12 hours as the maximum dye intensity decreases.

Shown in Fig. 5.8 are the average dye intensities calculated based on quantified in situ measurements. What was already apparent, the peak abundance of IL8 transcript occurs at 2 hours after the stimulation, Fig. 5.8A, rather than at 1 hour as suggested by microarray experiments, Fig. 5.1. In addition, the measurements at 10, 45 and 60 min are not consistent with the expected expression profiles. This is not the case of average dye intensities for NF- κ B2 gene, Fig. 5.8B, which are in agreement with the previous data (Fig. 5.1).

Nevertheless, the in situ measurements support proposed herein mechanism of collective activation of NF- κ B dependent genes in HeLa cells. Although there are discrepancies between the in situ and microarray data, please note that the former are collected for two specific genes, while the latter correspond to a group of genes with some heterogeneity among individuals. In addition, the discrepancies observed for IL8 transcript at 10, 45 and 60 min are also likely due to the fact that the corresponding images have different colors and backgrounds than the others, Fig. 5.5, which may have affected the quantification [16].

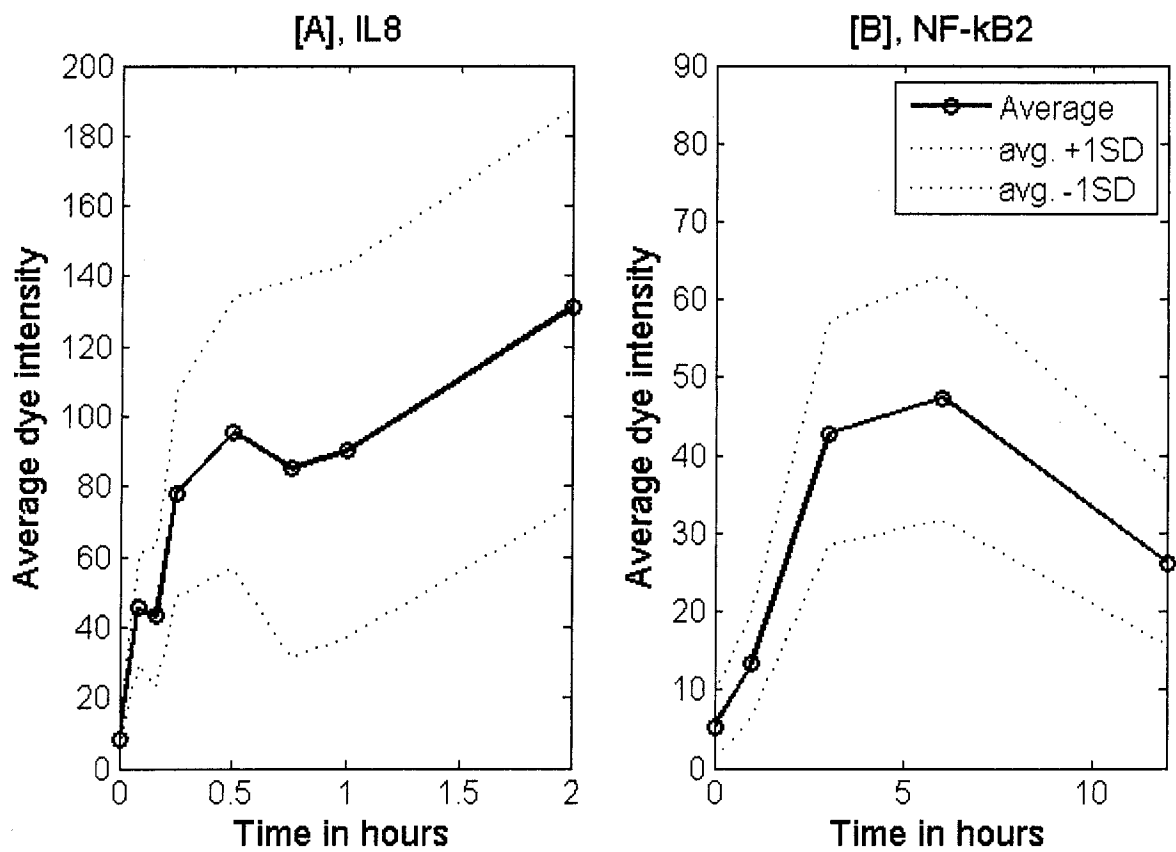
The following sections provide analysis of the proposed model in the terms of the time dependent mRNA distributions. The model is revisited to accommodate the in situ hybridization data on two specific genes.

Figure 5.7 : Quantified dye intensity distributions for NF- κ B2 gene.



Quantification of the in situ hybridization data on NF- κ B2 mRNA. Shown are histograms of dye intensity at 0, 1, 3, 6 and 12 hours after TNF stimulation.

Figure 5.8 : Average dye intensity.



Average dye intensity (circles connected with solid lines) evaluated based on the quantified in situ data for IL8 (Panel A) and NF- κ B2 (Panel B). Confidence bounds of one standard deviation (average \pm 1SD) depicted with dotted lines.

5.5 Refined model and the distribution functions

By analogy with the exact stochastic description (3.1-3.3) the proposed model can be analyzed in the terms of its underlying distribution function. To avoid artifacts introduced by the continuous approximation (3.12-3.14), the analysis is based on the discrete analog (exact stochastic description) of the system given by the Eq. (5.1). Previously it was assumed that the mRNA transcript is produced only when a gene promoter is bound by all of the required activators, and there is no production otherwise. As a result, every mRNA transcript molecule produced after TNF stimulation would be eventually degraded. To the contrary, the quantifications of the in situ experiments reveal an initial mRNA transcript distributions (Fig. 5.6 and 5.7 at $t=0$ min) prior to the stimulation. This new finding is incorporated into the refined model by allowing the mRNA production with a erratic transcription rate from an inactive or repressed gene. This is equivalent to assuming that $G_i^j(A, t) = 1$ and $G_i^j(I, t) = G_i^j(R, t) = \varepsilon_i$, where $\varepsilon_i \ll 1$ and A denotes the active gene state, while I and R denote the inactive and repressed states, respectively.

Please note that the expected expression profiles for the refined model can be obtained based on the Eq. (5.13). One can split the amount of the mRNA transcript produced according to the refined model into two pools: First, produced by a single, continuously active gene with a transcription rate equal to $H \cdot \varepsilon_i$. Second, produced by a transiently active gene with a production rate equal to $H \cdot (1 - \varepsilon_i)$ at the gene active state, and zero otherwise. The former yields constant amount of mRNA transcript on average equal to $\frac{H \cdot \varepsilon_i}{r_i}$, while the latter yields the expected mRNA level as in Eq. (5.13), but with the transcription rate equal to $H \cdot (1 - \varepsilon_i)$. Therefore, the expected expression profiles in the refined model are given by:

$$E[x_i(t)] = \frac{H\varepsilon_i}{r_i} + \sum_{k=0}^i \left[\frac{H(1-\varepsilon_i)}{r_i - \lambda_k} \left(\frac{\prod_{l=1}^i \lambda_l}{\prod_{l=0, l \neq k}^i (\lambda_l - \lambda_k)} \right) (e^{-\lambda_k t} - e^{-r_i t}) \right], \quad (5.18)$$

where $i = 1, 2, 3$ corresponds to early, intermediate and late class of genes, respectively.

Please note, that Eq. (5.18) is valid for the continuous as well as the discrete analog of the model, since as shown previously, the first moments are preserved regardless of the approximation.

5.5.1 Early genes

The exact stochastic description of the system (5.1) for early genes in the refined model reads:



$$G_1(A) = 1, G_1(I) = G_1(R) = \varepsilon_1, \quad (5.20)$$



where I denotes inactive, A active and R repressed gene state and $G_1(A) = 1$, and $G_1(I) = G_1(R) = \varepsilon_1$. While gene is in the active state, the mRNA transcript is produced with the rate $HG_1(A) = H$, but while not in the active state the transcription proceeds with the rate $H\varepsilon_1$, where $\varepsilon_1 \ll 1$. Please note, that in this case, the state of a gene is equivalent

to the state of its promoter: Empty promoter corresponds to the inactive gene, promoter bound by the first activator corresponds to the active gene, and promoter bound by repressor corresponds to the repressed gene.

The state of the system (5.19-5.23) can be described by double random variables (x_1, S_1) , where x_1 denotes the number of mRNA transcript molecules, while S_1 denotes the state of gene promoter, $S_1 \in \{I, A, R\}$. The joint distribution of (x_1, S_1) can be captured by the triple of probability mass functions:

$$w_x(t) = P[\# \text{ of } mRNA = x_1, S_1 = I], \quad (5.24)$$

$$u_x(t) = P[\# \text{ of } mRNA = x_1, S_1 = A], \quad (5.25)$$

$$v_x(t) = P[\# \text{ of } mRNA = x_1, S_1 = R], \quad (5.26)$$

where $w_x(t)$ denotes the mRNA distribution in the gene inactive state, while $u_x(t)$ and $v_x(t)$ in the active and repressed state respectively. Because the process is independent from cell to cell the marginal distribution

$$\rho_x(t) = w_x(t) + u_x(t) + v_x(t) \quad (5.27)$$

describes a time dependent distribution of cells in the population with a given amount of mRNA transcript regardless of the gene status, and thus can be related to the in situ experiments.

By the analogy with the master equation (3.6-3.7), the time evolution of densities (5.24-5.26) is given by:

$$\frac{dw_x(t)}{dt} = H\varepsilon_1 w_{x-1} + r_1(x+1)w_{x+1} - (H\varepsilon_1 + r_1x)w_x - \lambda_1 w_x, \quad (5.28)$$

$$\frac{du_x(t)}{dt} = H u_{x-1} + r_1(x+1)u_{x+1} - (H + r_1x)u_x - \lambda_0 u_x + \lambda_1 w_x, \quad (5.29)$$

$$\frac{dv_x(t)}{dt} = H\varepsilon_1 v_{x-1} + r_1(x+1)v_{x+1} - (H\varepsilon_1 + r_1x)v_x + \lambda_0 u_x, \quad (5.30)$$

where $x = 0, 1, 2, \dots$. The right hand sides of Eqs. (5.28-5.30) account for two flows of probability: The first corresponds to the discrete process of mRNA production and degradation and it is depicted with first three terms in each of the equations. The second flow of probability corresponds to the process of gene activation, for example, the last term in Eq. (5.28) depicts the gene activation event, which is equivalent to the binding of first activator with intensity λ_1 .

The system (5.28-5.30) has a nontrivial marginal steady state distribution, which is a *Poisson* $(\frac{H\varepsilon_1}{r_1})$ [see the discussion following Eqs. (3.21-3.22)]. The steady state results from the erratic transcription rate at the repressed gene state and provides the initial mRNA transcript distribution at the beginning of the simulation.

$$w_x(0) = \text{Poisson}\left(\frac{H\varepsilon_1}{r_1}\right), \quad (5.31)$$

$$u_x(0) = 0, \quad (5.32)$$

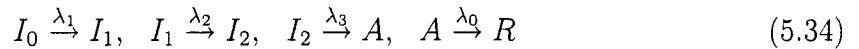
$$v_x(0) = 0, \quad (5.33)$$

The system (5.28-5.30) is in fact an infinite system of linear ODEs, since the support of the distributions w_x , u_x and v_x is infinite. However, similarly to the master equation (3.6-

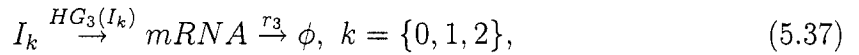
3.7), the tails of w_x , u_x and v_x can be disregarded since the probability assigned for arbitrarily large x is negligible. By this approximation, the underlying time dependent distribution is described by a finite system of linear ODEs, which can be numerically solved.

5.5.2 Late genes

Note, that for the late genes there are three different promoter conformation in the inactive gene state: First, denoted with I_0 , corresponds to the empty promoter; second, denoted with I_1 , corresponds to the promoter bound by the first activator; and third, denoted with I_2 , corresponds to the promoter bound by first and second activator. Binding of the third activator activates a gene, and this state is denoted with A , as previously. When in addition binding of the repressor occurs, the gene becomes repressed and its state is again denoted with R . Therefore, the discrete haploid analog of the system (5.1) for late genes in the refined model reads:



$$G_3(A) = 1, G_3(I_0) = G_3(I_1) = G_3(I_2) = G_3(R) = \varepsilon_3, \quad (5.35)$$



While a gene is in the active state (A), the mRNA transcript is produced with the rate $HG_3(A) = H$, but while in the inactive (I_0 , I_1 or I_2) or repressed state (R), the transcription proceeds with the rate $H\varepsilon_3$, where $\varepsilon_3 \ll 1$.

The state of the system (5.34-5.34) can be described by double random variables (x_3, S_3) , where x_3 denotes the number of mRNA transcript molecules, while S_3 denotes conformation of the promoter, $S_3 \in \{I_0, I_1, I_2, A, R\}$. The joint distribution of (x_3, S_3) can be captured by the following probability mass functions:

$$w_x^0(t) = P[\# \text{ of } mRNA = x_3, S_3 = I_0], \quad (5.39)$$

$$w_x^1(t) = P[\# \text{ of } mRNA = x_3, S_3 = I_1], \quad (5.40)$$

$$w_x^2(t) = P[\# \text{ of } mRNA = x_3, S_3 = I_2], \quad (5.41)$$

$$u_x(t) = P[\# \text{ of } mRNA = x_3, S_3 = A], \quad (5.42)$$

$$v_x(t) = P[\# \text{ of } mRNA = x_3, S_3 = R], \quad (5.43)$$

where $w_x^k(t)$, $k = \{0, 1, 2\}$ describes the mRNA distribution in the respective conformation of the promoter in the inactive gene state, while $u_x(t)$ and $v_x(t)$ describe the distribution in the active and repressed state, respectively. The marginal distribution

$$\rho_x(t) = w_x^0(t) + w_x^1(t) + w_x^2(t) + u_x(t) + v_x(t) \quad (5.44)$$

describes the time dependent mRNA distribution regardless of the gene status.

The time evolution of densities (5.39-5.43) is given by the following master equation:

$$\frac{dw_x^0(t)}{dt} = H\varepsilon_3 w_{x-1}^0 + r_3(x+1)w_{x+1}^0 - (H\varepsilon_3 + r_3x)w_x^0 - \lambda_1 w_x^0, \quad (5.45)$$

$$\frac{dw_x^1(t)}{dt} = H\varepsilon_3 w_{x-1}^1 + r_3(x+1)w_{x+1}^1 - (H\varepsilon_3 + r_3x)w_x^1 + \lambda_1 w_x^0 - \lambda_2 w_x^1, \quad (5.46)$$

$$\frac{dw_x^2(t)}{dt} = H\varepsilon_3 w_{x-1}^2 + r_3(x+1)w_{x+1}^2 - (H\varepsilon_3 + r_3x)w_x^2 + \lambda_2 w_x^1 - \lambda_3 w_x^2, \quad (5.47)$$

$$\frac{du_x(t)}{dt} = Hw_{x-1} + r_3(x+1)u_{x+1} - (H + r_3x)u_x + \lambda_3 w_x^2 - \lambda_0 u_x, \quad (5.48)$$

$$\frac{dv_x(t)}{dt} = H\varepsilon_3 v_{x-1} + r_3(x+1)v_{x+1} - (H\varepsilon_3 + r_3x)v_x + \lambda_0 u_x, \quad (5.49)$$

with initial conditions

$$w_x^0(0) = \text{Poisson}\left(\frac{H\varepsilon_3}{r_3}\right), \quad (5.50)$$

$$w_x^1(0) = w_x^2(0) = u_x(0) = v_x(0) = 0. \quad (5.51)$$

The system (5.45-5.49) is again an infinite system of linear ODEs, however the tails of w_x^k , $k = \{0, 1, 2\}$, u_x and v_x can be disregarded since the probability assigned for arbitrarily large x is negligible. By this approximation, the underlying time dependent distribution is described by a finite system of linear ODEs, which can be numerically solved.

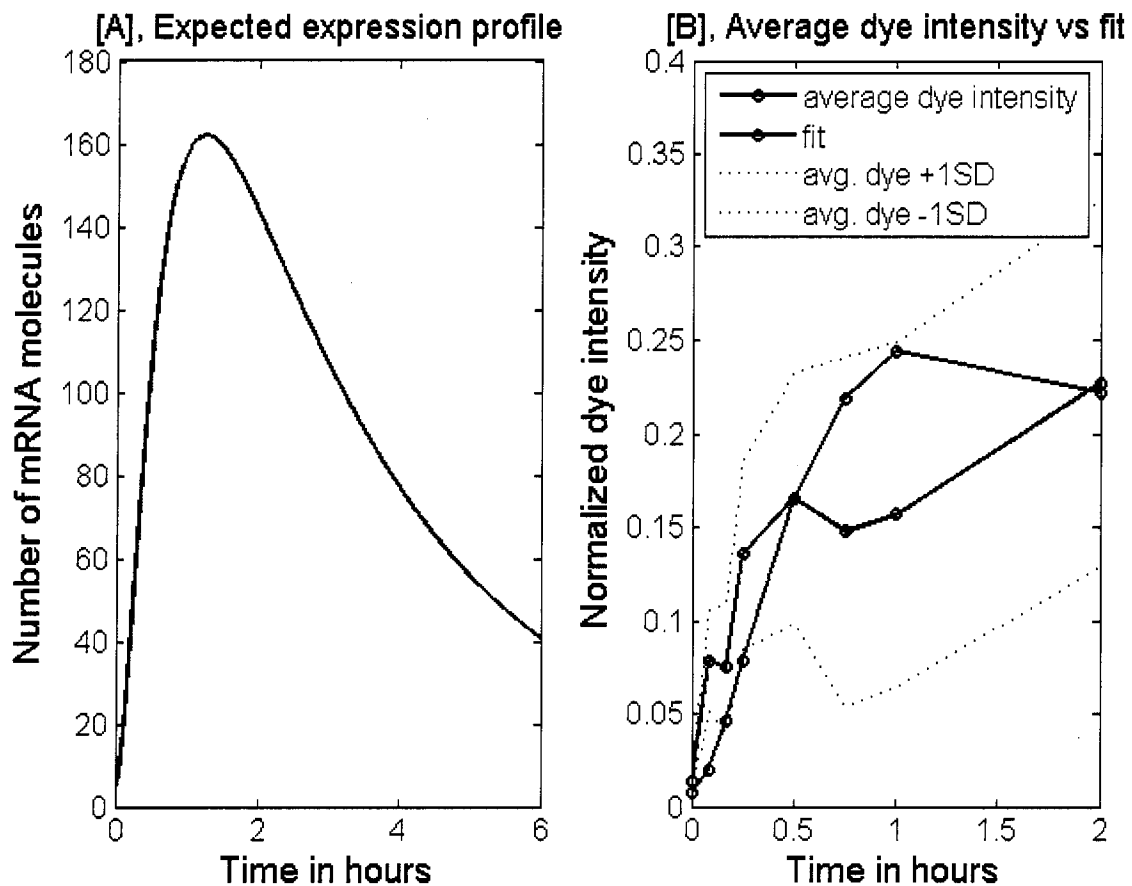
5.6 Comparison with the in situ data

The refined model is fit to reproduce measured time dependent mRNA distributions for the IL8 and NF- κ B2 gene. First, the expected expression profiles derived in Eq. (5.18) are matched against the averaged dye intensities (Fig. 5.8). Then, obtained parametrization is used to predict transient mRNA distributions from the model.

The measured dye intensity is proportional up to some extent to the number of the mRNA transcript in the system, however the corresponding proportionality constant is not estimated, instead the fitting procedure is carried out for the average intensities normalized by the area they span. Therefore neither the transcription rate nor the number of considered copies per gene affect the fitting procedure. However, the best match in the terms of mRNA distribution functions was obtained when assuming that every gene has four potentially active homologous copies. In fact the HeLa cells are almost tetraploid since their modal number is 82 [2](the modal number of a regular diploid human cell is 46). Because the data include rather small number of noisy observations, the fitting procedure is carried out manually based on the parameters estimated previously to explain microarray experiments. The fit was carried separately for the IL8 and NF- κ B2 gene to allow heterogeneity in decay between corresponding mRNA transcripts.

The fit of the expected expression profile for the IL8 gene is depicted in Fig. 5.9. In this case, 20 min mRNA half-life time is assumed ($r_1 = 0.00057s^{-1}$). In addition, the transcription rate of 2 mRNA transcript per minute per gene copy is assumed, which is equal to production rate $H = 0.033s^{-1}$. Notice the two fold decrease comparing to the microarray fit, but it allows obtaining distribution functions relatively less separated at measured time instants, Fig. 5.10. While the gene copy is in the active state, the transcription proceeds with a rate H , but when it is in the inactive or repressed state transcription proceeds with rate $H\varepsilon_1$, where $\varepsilon_1 = 0.02$. The expected binding time of the first activator is assumed to be 10 min, which corresponds to the binding rate $\lambda_1 = 1.667 \times 10^{-3}s^{-1}$. The binding rate of the repressor is fit to be equal to $\lambda_0 = 9.7 \times 10^{-5}s^{-1}$, which corresponds to the expected binding time of 170 min. A two-fold decrease in the repressor binding time allows obtaining the maximum mRNA abundance at 75min after the stimulation (Fig. 5.9A). The initial fast

Figure 5.9 : Fitted expected expression profile for IL8 gene.



Expected expression profile for IL8 mRNA against the measured averaged dye intensity depicted in Fig. 5.8A: Panel A - expected expression profile over 6 hours after TNF stimulation calculated based on the Eq. (5.18) assuming four homologous gene copies. Panel B - comparison between the fit and the average dye intensity, both normalized by the area under the curve.

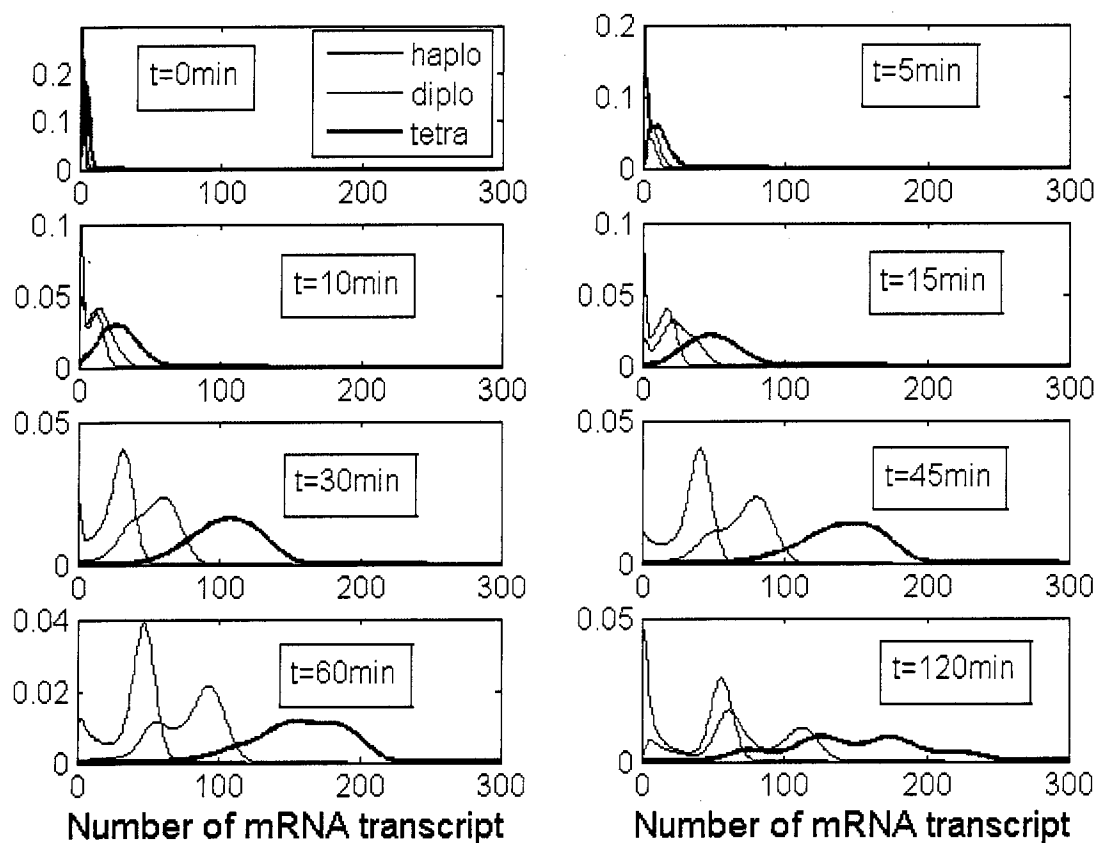
increase in the average mRNA number is well depicted by the model, while measurements at 45 and 60 min are treated as outliers (Fig. 5.9B).

Model predictions for transient distributions of IL8 mRNA corresponding to the obtained parametrization is presented in Fig. 5.10. In addition to the marginal mRNA distribution $\rho_x(t)$ corresponding to a haploid gene (black curve) given by the solution to Eqs. (5.28-5.30), Fig. 5.10 presents distributions for a diploid (red curve) and tetraploid gene (blue curve), obtained from the former by convolutions. One can observe that with time, the initial mRNA distribution resulting from the erratic transcription spreads right as more mRNA is produced in the system. The best match with the quantified in situ data presented in Fig. 5.10 is obtained for a tetraploid gene. Some discrepancies are observed at 10 min after the stimulation, where the measured distribution remains almost unchanged comparing to the 5 min time point. In addition, experimental distributions at 45 and 60 min have more dye concentrated at zero intensity than these predicted from the model, however these two time points were treated as outliers during the fitting procedure.

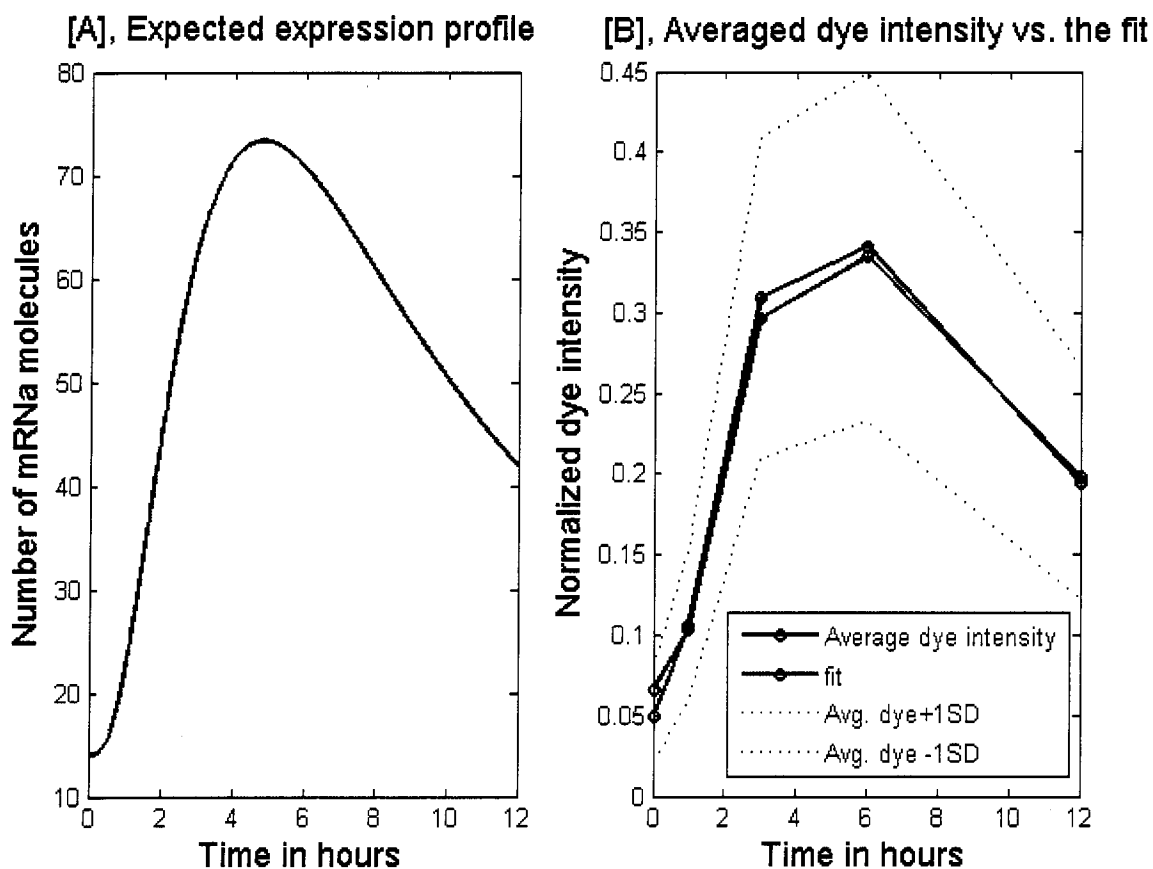
As already mentioned the fit is not unique, however the predictions of mRNA distributions remain in a good agreement with the measurements even after three-fold increase of the IL8 mRNA half-life time and simultaneous fit of other parameters (data not shown). This is not the case of the NF- κ B2 gene. It was found that the model predictions improve when assuming two- and three-fold increase of mRNA half-life time comparing with IL8 gene.

For the NF- κ B2 gene the best fit for the expected expression profile is depicted in Fig. 5.11. It is assumed that NF- κ B2 mRNA half-life time is equal to 60 min, which corresponds to the degradation rate $r_3 = 0.00019s^{-1}$. Analogous with the IL8 gene, the transcription rate of 2 mRNA transcripts per minute per gene copy is assumed, which corresponds to the production rate $H = 0.033s^{-1}$ in the gene active state, and production rate $H\varepsilon_3$, where

Figure 5.10 : Model predictions of transient IL8 mRNA distributions.



Model predictions in the terms of the transient IL8 mRNA distributions for the parametrization as in Fig. 5.9. These should be compared against quantified dye intensities presented in Fig. 5.6.

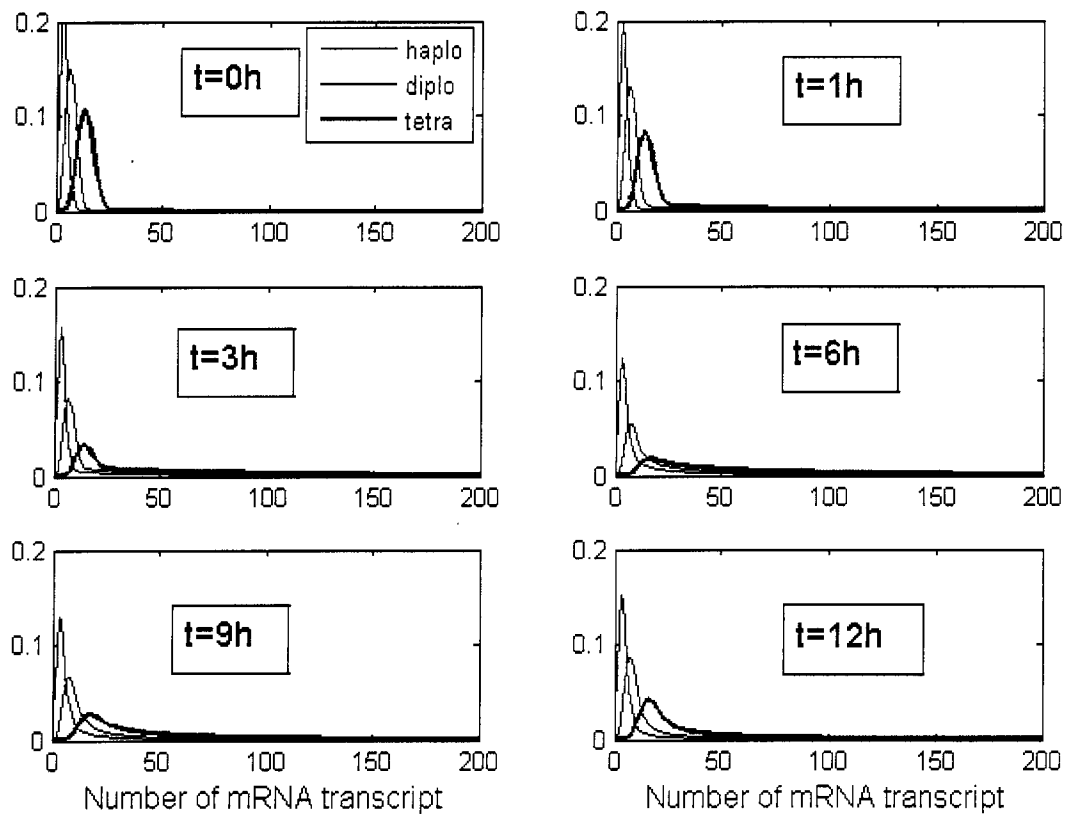
Figure 5.11 : Fitted expected expression profile for NF- κ B2 gene.

Expected expression profile for NF- κ B2 mRNA against the measured averaged dye intensity depicted in Fig. 5.8B: Panel A - expected expression profile over 12 hours after TNF stimulation calculated based on the Eq. (5.18) assuming four homologous gene copies. Panel B - comparison between the fit and the averaged dye intensity, both normalized by the area under the curve.

$\varepsilon_3 = 0.02$, in the inactive and repressed state. In addition, the following binding rates were fitted: $\lambda_1 = 1.667 \times 10^{-3} s^{-1}$, $\lambda_2 = 7 \times 10^{-4} s^{-1}$, $\lambda_3 = 3.88 \times 10^{-5} s^{-1}$ and $\lambda_0 = 2.8 \times 10^{-4} s^{-1}$ which corresponds to the expected binding times of 10, 24, 430 and 60 min for the three activators and one repressor, respectively. The expected expression profile fits very well averaged dye intensity calculated based on the quantified experiments, Fig. 5.11B, however the maximum abundance of NF- κ B2 mRNA transcript is observed at about 5 hours after the stimulation. The parameters fitted are quite different from the previous estimates from the microarray data (except of the first activator binding time), with a very long expected time for the third activator binding, i.e., 430 min, and rather short for the second activator, i.e., 24 min (previously 225 and 205 min, respectively). This indicates that, in fact, only one step required for activation of NF- κ B2 gene is time limiting and causes most of the cell-to-cell variability (due to assumed constant activation rates, the binding events are independent and thus their order is interchangeable). As in the case of IL8, the system is quite robust to changes of parameters (increase of mRNA half-life time followed by simultaneous fit of other parameters), however the disproportion between the fitted expected binding time of the third and second (or first) activator is still observed.

Model predictions in the terms of the transient mRNA distributions are depicted in Fig. 5.12 and these should be compared against the data depicted in Fig. 5.7. The marginal mRNA distributions for a single gene copy (black curve), obtained by solving Eqs. (5.45-5.49) are augmented with the distributions for the diploid (red curve) and tetraploid (blue curve) gene, derived based on the former by convolutions. The initial mRNA distribution spreads right with time, although rather slowly (at 1 hour one can notice only a slightly heavier tail of the mRNA distribution) to arrive at the maximum spread at 6 hours followed by a clear repression at 12 hours after TNF treatment. One problem with the fit can be

Figure 5.12 : Model predictions of transient NF- κ B2 mRNA distributions.



Model predictions in the terms of the transient NF- κ B2 mRNA distributions for the parametrization as in Fig. 5.11. These should be compared against quantified dye intensities in Fig. 5.7.

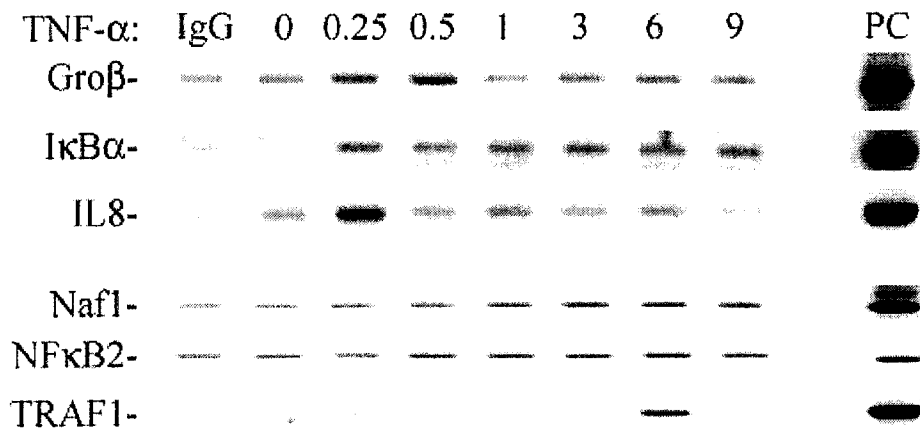
observed: At any instant of time, model distribution functions range from about 5 mRNA, while in the data, the corresponding left extreme of the dye intensity tends to follow the average over the time course of the experiment, i.e., the initial increase is followed by the decrease. This feature cannot be depicted by the model due to its simplicity and probably indicates a very complex dynamics in the original biological system. Besides of that, the model predictions are in good agreement with the experimental data presented in Fig. 5.7.

5.7 The role of phosphorylated NF- κ B

The model presented assumes that the expression of NF- κ B dependent genes in HeLa cells is governed by the collective action of multiple regulatory factors. Without an attempt to identify these factors, it is hypothesized that they might be activating (e.g. histone acetylation) or repressing events, not necessarily connected with the DNA/protein binding.

Recent experimental findings report that a subset of NF- κ B regulated genes are dependent on the formation of specifically phosphorylated functional subunit of the NF- κ B transcription factor (phospho-Serine 276 RelA) [48]. It was shown that the phospho-Serine 276 RelA is required to facilitate expression of most of the early genes including IL8 and coincides with their peak expression (Fig. 5.13 top three panels). Phosphorylated RelA enters the nucleus about 15 minutes after TNF treatment and remains associated with early promoters only for a short period of time. In fact, in the case of IL8 gene there is no measured binding activity at 30 min after stimulation. Recall, that in the same system the bulk (unphosphorylated) RelA remains associated with IL8 promoter for 6 hours after TNF treatment, Fig. 5.2B. To the contrary, experimental evidence that phospho-Serine 276 RelA is responsible for transcriptional activity of late genes is not conclusive. In the case of NAF1 and NF- κ B2, Fig. 5.13, one can observe that the binding activity of phosphorylated RelA

Figure 5.13 : Kinetics of phospho-Serine 276 RelA in HeLa cells.



ChIP analysis of phospho-Serine 276 Rel A association with promoters of early (Gro β , I κ B α , IL8) and late (NAF1, NF- κ B2, TRAF1) genes. HeLa cells were stimulated for various times with TNF α prior to formaldehyde fixation. Phospho-Serine 276 Rel A binding coincides with peak expression of early genes at 1 hour after TNF α treatment.

remains roughly the same during the time course of the experiment. However, there is some suggestion that the peak activity occurs at 6 hour by analogy with the TRAF1 promoter. More importantly, kinetics of phospho-Serine 276 RelA association with the late gene promoters is significantly different that with the early ones, which was not the case of the bulk RelA, Fig. 5.2B.

Nevertheless, above experimental findings [48] are in agreement with the proposed model of regulation of NF- κ B dependent genes. In the case of early genes, the only one activator required in the model to initiate transcription can be identified with the phospho-Serine 276 RelA. The repressor assumed in the model, whose task is to terminate gene expression, can explain heterogeneity of phospho-Serine 276 RelA dissociation from early promoters (Fig. 5.13). As ChIP assay suggests, the termination of phospho-Serine 276 RelA binding activity

can be due to a some unknown process of active repression, e.g., dephosphorylation. In the case of late genes, it is not clear if the kinetics of phospho-Serine 276 RelA can explain corresponding expression profiles by itself. In addition, there is no explanation how the RelA is being phosphorylated (with exception of its initial wave which is assumed to be phosphorylated in the cytoplasm due to the TNF treatment). The model proposed suggests that RelA associates with late promoters prior to the serine 276 phosphorylation event, while the phosphorylation event might be a time limiting step required for late gene activation. This hypothesis has to be further examined experimentally.

Chapter 6

Discussion

Stochasticity in genetic regulatory systems results from a small numbers of involved molecules of DNA, mRNA and protein of a given species. These effects are especially important in prokaryotes, where the abundance of gene products might be as low as one mRNA transcript molecule and several protein molecules on average in the cell [1], [41], [3], [29]. Therefore, the production or degradation of a single mRNA or protein molecule has a significant effect on the cell's behavior [13]. In eukaryotes, and especially in higher eukaryotes, fluctuations in the system are significantly influenced by the process governing intermittent gene activity. Over a decade ago, Ko (1991, 1992) [32], [33] postulated that the interactions between transcription factors and DNA (gene promoters) contribute major stochastic effects in the process. More precisely, Ko postulated that at a given instant of time a gene copy is thought to be either "*switched on*" by having transcription complex bound to its promoter, or "*switched off*" by having transcription complex not bound. Typically, to activate a eukaryotic gene, several regulatory proteins, i.e., transcription factors, are required with prior chromatin remodeling. Therefore, when a gene becomes active for a sufficiently long period of time, it results in production of large bursts of mRNA transcript followed by protein molecules. The resulting cell-to-cell heterogeneity cannot be explained only by effects due to the small number of involved molecules of mRNA and protein [60], [59], since their levels in eukaryotes can be fairly large with up to hundreds of transcript molecules and hundreds of thousand of protein molecules (e.g. species involved in early immune response [36], [37]).

To better understand the phenomenon of gene expression, gene regulatory systems are subject to extensive investigations. Interactions between molecules of DNA and their mRNA and protein products are being modeled as systems of coupled chemical reactions. Under the assumption of spatial homogeneity, the stochastic process governing the reacting molecules is a Markov process and its distribution can be exactly captured by a Chapman-Kolmogorov equation [67]. Unfortunately, this approach is limited by the number of involved molecules and the corresponding Chapman-Kolmogorov equation can be solved only for the bacterial systems. A second, although less rigorous, method relies on the stochastic simulation algorithm [18], [19] (or its approximations [20], [8], [23], [55], [9]), which instead of directly solving the corresponding Chapman-Kolmogorov equation, allows numerical simulations of the underlying Markov process (or its approximations).

The latter, although historically older approach is extensively applied to the analysis of large biological networks [41], [42], [3], [10], [29], [5]. Unfortunately, the specifics of the considered biological systems and the method itself does not allow providing a rigorous description and understanding the details of processes connected with gene regulation. This can be obtained by the former approach, which relies on the rigorous analysis of small (one- or two-) gene regulatory systems in the terms of the corresponding Chapman-Kolmogorov equations [30], [63] [61], [62], [66], [49]. The methods introduced allow to derive analytical expressions for the moments of the marginal mRNA and protein distributions, usually under some simplifying assumptions. Among mentioned work, stochasticity caused by a switching of a gene status, was first rigorously analyzed by Kepler and Elston (2001) [30]. In their influential paper, Kepler and Elston (2001) considered synthesis of protein oligomers in the process, however they assumed a direct protein translation from the DNA. The approach involved a Chapman-Kolmogorov equation for the underlying probability distribution function

approximated by a Fokker-Planck equation. In the case of a single self-activating gene, they further simplified the Fokker-Planck equation by neglecting the diffusion term, which lead to the first order system of PDEs for the underlying protein distribution function. Recently, stochasticity due to the intermittent gene activity was also analyzed by Raser and O'Shea (2004) [49] in a more general model incorporating in addition the mRNA/protein production/decay noise. The authors derived the normalized steady state protein variance, however they did not address the problem of solving the corresponding Chapman-Kolmogorov equation for the underlying distribution function.

This thesis is dedicated to the analysis of stochastic effects in small gene regulatory networks. The existing description of the underlying stochastic process utilizing a Chapman-Kolmogorov equation proves to be limited and inefficient. The present work introduces a much more efficient yet accurate modeling approach. Opposite to other methods, proposed modeling approach allows analyzing stochasticity in the system in the terms of the underlying distribution function in addition to Monte Carlo simulations.

The novel modeling approach is motivated by the analysis of a single gene module without feedback regulation with three major sources of stochasticity: intermittent gene activity, mRNA transcription/decay noise and protein translation/decay noise. Although the corresponding Chapman-Kolmogorov equation cannot be solved when a large number of molecules is considered, the first two moments of the marginal mRNA and protein distributions are derived. The variance of the number of mRNA and protein molecules is found decomposable into terms corresponding to different sources of stochasticity, which allows to quantify their significance in the process. It is shown that in eukaryotes, intermittent gene activity contributes most of the total variability in the process, while the protein/decay noise is of the least significance. However, in prokaryotes, there is a competition between stochastic

effects due to intermittent gene activity and the mRNA/protein production/decay noise. This result extends findings of Thattai and Oudenaarden (2001) [63] and Tao (2004, 2004a) [61], [62] who disregarded the intermittent gene activity as a potential noise source, and work of Kepler and Elston (2001) [30] who neglected the mRNA transcription/decay noise in the system. An analogical model of a single gene regulatory module was analyzed by Raser and O'Shea (2004) [49] who derived analogical expressions for the normalized steady state protein variance, and concluded that the balance between gene promoter activation and transcription determines the magnitudes of different stochastic effects in the process.

Based on the variance decomposition, two approximations to the original stochastic process at the single cell level are proposed: First, the continuous approximation, which accounts only for the stochastic effects due to the intermittent gene activity [38]. Second, the mixed approximation, which accounts for the stochasticity corresponding to the intermittent gene activity and mRNA production/decay noise, while the protein production/decay noise is neglected [52]. Approximations yield systems of linear first-order PDEs for the underlying probability distribution functions, Eqs. (3.19-3.20) and Eqs. (3.28-3.29), for the continuous and mixed model, respectively. Although they were originally derived based on the fluid dynamics analogy, resulting PDEs follow from the differential Chapman-Kolmogorov equation (2.15), first derivations of which are due to Kolmogorov (1931) [35]. Similar equations were used to describe the time evolution of the distribution function in the process governed by the Langevin equation [54], [14], [30], however the latter included diffusion term resulting from the white noise term in the Langevine equation. Such equations has been used in physics to describe noise induced transitions [25] and in theoretical mechanics to describe dynamics of rigid bodies under trains of random impulses [26], [27].

Discretization techniques developed (Appendix C) allow reducing resulting PDEs into

large systems of the linear algebraic equations for the stationary two-dimensional mRNA-protein distribution functions or large systems of ODEs for their time evolution. In addition, the errors introduced by each approximation are evaluated, which allow the choice of appropriate approximation in the specific modeling task. This is especially important because of the trade-off between the continuous and mixed approximation: Mixed model is more accurate, but also more computationally intensive than the continuous approximation, while the continuous approximation is very efficient but it may introduce artifacts at the mRNA and protein level.

The marginal protein distribution resulting from the continuous approximation is compared against the marginal distribution given by the Kepler-Elston model [30], which disregards mRNA and assumes direct translation of protein from the DNA. The approach taken by Kepler and Elston (2001) relies on the Chapman-Kolmogorov equation for the underlying probability distribution function approximated by a Fokker-Planck equation. These equations are further simplified by neglecting the diffusion term, which leads to the first-order system of PDEs, analogous to the system (4.5-4.6). Whereas Kepler and Elston introduced their approximations at the population level, considered herein continuous and mixed models rely on the approximation of the exact stochastic description at the single cell level. The latter approach allows validating approximations not only by comparison in the terms of the distribution functions (Figs 3.4-3.9), as does the Kepler-Elston model, but also in the terms of the single cell conditional trajectories (Fig. 3.2 and 3.3). It is shown that the Kepler-Elston approximation fails when the mRNA transcript is more stable than the corresponding protein (Fig. 4.1). However, in the opposite case, the approximation is satisfactory and provides a great simplification in the analysis. It is used here to analyze the two-gene systems for which the two-dimensional protein-protein distributions are calculated, namely

for the activator-repressor and the repressor-repressor system. Without the Kepler-Elston approximation, this would require derivations of four-dimensional distributions. A simplified system of two repressors was analyzed by Kepler and Elston (2001) [30] by means of Monte-Carlo techniques, but the authors assumed that the genes considered share the same operator and have the same kinetic parameters.

For the sake of simplicity it is usually assumed that the transcriptional gene activity is due to the actions of single a *trans*-acting regulatory molecule (transcription factor) and a single *cis*-acting regulatory elements, i.e., an operator in bacteria or a promoter in eukaryotes [32], [30], [29], [63], [7], [61], [66]. In fact, the specific patterns of gene expression are governed by the combinatorial interactions of series of transcription factors that may bind to various regulatory sites within gene promoters and enhancers ([68] p. 72, [40], [53]). This mode of gene regulation is exploited in last chapter of this thesis. Collective mechanism of multiple regulatory factors is hypothesized to explain the dynamics of NF- κ B dependent genes in HeLa cells important in cell survival and inflammation. The modeling treatment follows the exact stochastic description and its continuous approximation introduced throughout this work, restricted to the amount mRNA transcript measured in microarray and in situ experiments. Expected expression profiles derived, as well as time dependent mRNA distributions, fit well experimental measurements. Although some of assumed regulatory factors remain unknown, it is hypothesized that they might be activating (e.g. histone acetylation) or repressing factors, not necessarily connected with DNA/protein binding. Nevertheless, the model developed confirms that the phospho-Ser276 RelA is responsible for the regulation of early NF- κ B dependent genes, and in addition it suggests that Ser276 phosphorylation of RelA is a time limiting step in activation of late NF- κ B dependent genes.

The approach taken provides interesting insights. Single cell expression profiles are sig-

nificantly different from the profiles constructed by averaging over the population of cells (Fig. 5.4). No individual cell behaves like an “average” one. This is especially visible in the example of the late NF- κ B dependent genes, where the variability among single cell profiles for late genes is much larger than for early and intermediate group. This observation has a strong implication in the terms of understanding the microarray experiments. The time course microarray experiments provide us with measurements of gene expression averaged over the cell population. These measurements have continuous values, but in a single cell, at a given time moment, the targeted gene is either turned “on” or “off”. It is also shown that not all cells (genes) are always active in the culture, in fact, their number may be unexpectedly small. Thus it is misleading to think that every cell in the tissue responds gradually in the terms of the expression level, as the microarray measurements might seem to suggest. One should rather think about a proportion of the transcriptionally active cells at the given time in the population.

Microarray studies rely on the assumption that the abundance of mRNA molecules is tightly correlated with the abundance of the corresponding protein, therefore the measured mRNA expression levels are being often extrapolated to the protein level. This can also be misleading. It is shown (Appendix A) that in the case of the single gene without feedback regulation corresponding correlation coefficient is equal to 1 only when the protein half-life time approaches zero, i.e., when the produced protein is immediately degraded. Moreover, when the protein half-life time increases and approaches infinity, i.e., when the protein never degrades, the correlation coefficient tends to 0 at a rate \sqrt{r} , where r is ratio of mRNA and protein degradation rates. Hence, in the biologically more realistic situation, in which the protein molecules are much more stable than the mRNA, $r \ll 1$, the correlation between the amount of the mRNA and protein molecules can be very small. This is the case of the

NF- κ B pathway [36], [37], where the NF- κ B inhibitor I κ B α is catalytically degraded with a half-life time of about 10 min, while its mRNA has a half-life of about 20 min.

Recent advances in genetic engineering allow synthesis of small artificial genetic networks based on the available well-characterized genetic components that naturally occur in their context [31]. Such relatively simple systems are needed to study in the quantitative way how the genetic structure and connectivity of cellular networks are related to their function including origin and consequences of stochastic gene expression. Currently, a libraries of small synthetic networks are being designed in *Escherichia coli* based on the transcriptional regulators such as *LacI*, *λ cI* and *TetR* [12], [22], [34]. In addition, a variety of artificial gene networks is being reported in *Saccharomyces cerevisiae* based on the eukaryotic components [5], [49]. In these biological systems, the ability to quantitatively and rigorously interpret the underlying dynamics is of crucial importance. In this context, the mathematical tools introduced herein may prove to be especially important by allowing rigorous analysis in the terms of the underlying distributions functions, which is being measured experimentally.

Bibliography

- [1] Ackers, G.K., Johnson, A.D., Shea, M.A., 1982. Quantitative model for gene regulation by phage λ repressor. *Proc. Natl Acad. Sci. USA* 79, 1129-1133.
- [2] Data provided by The Global Bioresource Center at www.atcc.org, April 2006. Search performed for the HeLa cell line ATCC number (CCL-2).
- [3] Arkin, A., Ross, J., McAdams, H.H., 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ -infected *Escherichia coli* cells. *Genetics* 149, 1633-1648.
- [4] Barken, D., Wang, C. J., Kærns, J., Cheong, R., Hoffman, A., Levchenko, A., 2005. Comment on "Oscillations in NF- κ B signaling control the dynamics of gene expression". *Science* 308, 52a.
- [5] Blake, W.J., Kærns, M., Cantor, C.R., Collins, J.J., 2003. Noise in eukaryotic gene expression. *Nature* 422, 633-637.
- [6] Blattner, C., Kannouche, P., Litfin, M., Bender, K., Rahmsdorf, H.J., Angulo, J.F., and Herrlich, P., 2000. UV-Induced Stabilization of c-fos and Other Short-Lived mRNAs. *Molecular and Cellular Biology* 20, 3616-3625.
- [7] Bundschuh, R., Hayot, F., Jayaprakash, C., 2003. The role of dimerization in noise reduction of simple genetic networks. *J. Theor. Biol.* 220, 261-269.

- [8] Cao, Y., Petzold, L.R., Rathinam, M., Gillespie, D.T., 2004. The numerical stability of leaping methods for stochastic simulation of chemically reacting systems. *J. Chem. Phys.* 121, 12169-12178.
- [9] Cao, Y., Gillespie, D.T., Petzold, L.R., 2005. The slow-scale stochastic algorithm. *J. Chem. Phys.* 122, 014116.
- [10] Cook, D.L., Gerber, A.N., Tapscott, S.J., 1998. Modeling stochastic gene expression: Implications for haplosufficiency. *Proc. Natl Acad. Sci. USA* 95, 15641-15646.
- [11] Eberharter, A., Becker, P. B., 2002. Histone Acetylation: a switch between repressive and permissive chromatin. *EMBO reports* vol 3, 224-229.
- [12] Elowitz, M.B., Leibler, S., 2000. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403, 335-338.
- [13] Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S., 2002. Stochastic gene expression in a single cell. *Science* 297, 1183-1186.
- [14] Emch, G.G., Liu, C., 2002. *The logic of thermostatical physics*. Springer, Berlin, p.494.
- [15] Femino, A.M., Fay, F.S., Fogarty, K., Singer, R. H., 1998. Visualization of Single RNA Transcripts in Situ. *Science* 280, 585-590.
- [16] Fujarewicz, K., 2006. Unpublished data.
- [17] Gardiner, C.W., 2003. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, Spronger-Verlag. New York.
- [18] Gillespie, D.T., 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.* 22, 403-434.

- [19] Gillespie, D.T., 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81, 2340-2361.
- [20] Gillespie, D.T., 2001. Approximate accelerated stochastic simulation of chemically reacting system. *J. Phys. Chem.* 115, 1716-1733.
- [21] Gregory, P. D., Wagner, K. and Hörz, W., 2001. Histone acetylation and chromatin remodeling. *Experimental Cell Research* 265, 195-202.
- [22] Guet, C.C., Elowitz, M.B., Hsing, W., Leibler, S., 2002. Combinatorial synthesis of genetic networks. *Science* 296, 1466-1469.
- [23] Hasltine, E.L., Rawlings, J.B., 2002. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *J. Chem. Phys.* 117, 6959-6969.
- [24] Hasty, J., Pradines, J., Dolnik, M., Collins, J.J., 2000. Noise-based switches and amplifiers for gene expression. *Proc. Natl Acad. Sci. USA* 97, 2075-2080
- [25] Horsthemke, W., Lefever, R., 1984. Noise induced transitions. *Theory and applications in Physics, Chemistry and Biology.* Springer, Berlin.
- [26] Iwankiewicz, R., 1995. Dynamical mechanical systems under random impulses. *Series on advances in mathematic for applied sciences-vol. 36.*
- [27] Iwankiewicz, R., Nielsen, S.R.K., 2000. Solution techniques for pulse problems in non-linear stochastic dynamics. *Prob. Eng. Mech.* 15, 25-36.
- [28] Johnson, A.D., Meyer, B.J., Ptashne, M., 1979. Interactions between DNA-bound repressors govern regulation by the phage λ repressor. *Proc. Natl. Acad. Sci. USA* 79, 5061-5065.

- [29] Kierzek, A.M, Zaim, J., Zilenkiewicz, P., 2001. The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression, *J. Biol. Chem.* 276, 8165-8172.
- [30] Kepler, T. B., Elston, T. C., 2001. Stochasticity in Transcriptional Regulation: Origins, Consequences, and Mathematical Representations. *Biophys. J.* 81, 3116-3136.
- [31] Kærn, M., Blake, W.J., Collins, J.J., 2003. The engineering of gene regulatory networks. *Annu. Rev. Biomed. Eng.* 5, 179-206.
- [32] Ko, M.S.H., 1991. A Stochastic Model for Gene Induction. *J. Theor. Biol.* 153, 181-194.
- [33] Ko, M.S.H., 1992. Induction Mechanism of a Single Gene Molecule: Stochastic or Deterministic? *Bioassays* 14, 341-346.
- [34] Kobayashi, H., Kærn, M., Araki, M., Chung, K., Gardner, T.S., Cantor, C.R., Collins, J.J., 2004. Programmable cells: Interfacing natural and engineered gene networks. *Proc. Natl Acad. Sci. USA* 101, 8414-8419.
- [35] Kolmogorov, A.N., 1931. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Math. Ann.* 104, 415-458.
- [36] Lipniacki, T., Paszek, P., Brasier, A. R., Luxon, B., Kimmel, M., 2004. Mathematical model of NF- κ B module. *J. Theor. Biol.* 228, 195-215.
- [37] Lipniacki, T., Paszek, P., Brasier, A. R., Luxon, B., Kimmel, M., 2006. Stochastic regulation in early immune response, *Biophys. J.* 90, 725-742.
- [38] Lipniacki, T., Paszek, P., Marciniak-Czochra, A., Brasier, A.R., Kimmel, M., 2006. Transcriptional stochasticity in gene expression, *J. Theor. Biol.* 238, 348-367.

- [39] Lewin, B., 1997. Genes VI. Oxford University Press, Oxford, UK.
- [40] Louis, M., Holm, L., Sanchez, L., Kaufman, M., 2003. A Theoretical Model for the Regulation of Sex-lethal, a Gene That Controls Sex Determination and Dosage Compensation in *Drosophila melanogaster*. *Genetics* 165, 1355-1384.
- [41] McAdams, H.H., Shapiro, L., 1995. Circuit Simulation of Genetic Networks. *Science* 269, 650-656.
- [42] McAdams, H.H., Arkin, A., 1997. Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA* 94, 814-819.
- [43] McAdams, H.H., Arkin, A., 1998. It's a noisy business! Genetic regulation at the nanomolar scale. *TIG* (15) 65-69.
- [44] Nelson, G., Paraoan, L., Spiller, D. G., Wilde, G. J. C., Browne, A. M., Djali, P. K., Unitt, J. F., Sullivan, E., Floettmann, E. and White, M. R. H., 2002. Multi-parameter analysis of the kinetics of NF- κ B signaling and transcription in single living cells. *J. Cell Science* 115, 1137-1148.
- [45] Nelson, D. E., Ihekweba, A.E.C., Elliot, M., Johnson, J.R., Gibney, C.A., Foreman, B.E., Nelson, G., See, V., Horton, C.A., Spiller, D.G., Edwards, S.W., McDowell, H.P., Unitt, J.F., Sullivan, E., Grimley, R., Benson, N., Broomhead, D., Kell, D.B., White, M.R.H. 2004. Oscillations in NF- κ B signaling control the dynamics of gene expression. *Science* 306, 704-708.
- [46] Nelson, D. E., Horton, C.A. , See, V., Johnson J.R., Nelson , G., Spiller, D.G., Kell, D.B., White M.R.H., 2005. Response to comment on "Oscillations in NF- κ B signaling control the dynamics of gene expression". *Science* 308, 52b.

- [47] Nowak, D., E., 2004. Unpublished data.
- [48] Nowak, D., E., 2005. Mechanistic analysis of the TNF- α induced, NF- κ B regulated gene networks. PhD Thesis, The University of Texas Graduate School of Biomedical Sciences at Galveston.
- [49] Raser, J.M., O'Shea, E.K., 2004. Control of stochasticity in eukaryotic gene expression. *Science* 304, 1811-1814.
- [50] Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D., van Oudenaarden, A., 2002. Regulation of noise in the expression of a single gene. *Nature Genetics* 31, 69-73.
- [51] Paszek, P., Lipniacki, T., Brasier, A.R., Tian, B., Nowak, D.E., Kimmel, M. 2005. Stochastic effects of multiple regulators on expression profiles in eukaryotes. *J. Theor. Biol.* 233, 423-433.
- [52] Paszek. P., 2006. Significance of various noise sources in gene regulation. *Bul. Math. Biol.* submitted.
- [53] Pirone, J. R., Elston T. C., 2004. Fluctuations in transcription factor binding can explain the graded and binary responses observed in inducible gene expression. *J. Theor. Biol.* 226, 111-121.
- [54] Rao, Ch. V., Wolf, D.M., Arkin, A.P., 2002. Control, exploitation and tolerance of intracellular noise. *Nature* 420, 231-237.
- [55] Rao, Ch. V., Arkin, A.P., 2003. Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm. *J. Chem. Phys.* 118, 4999-5010.

- [56] Shea, M.A., Ackers, G.K., 1985. The Or control system of bacteriophage lambda: a physical-chemical model for gene regulation. *J. Mol. Biol.* 181, 211-230.
- [57] Snustad, D.P., Simmons, M.J., 2003. Principles of genetics. Wiley & Sons Inc.
- [58] Simpson, M.L., Cox, C.D., Sayler, G.S., 2004. Frequency domain chemical Langevin analysis of stochasticity in gene transcriptional regulation. *J. Theor. Biol.* 229, 383-394.
- [59] Stirland, J.A., Seymour, Z.C., Windeatt, S., Norris, A.J., Stanley, P., Castro, M.G., Loudon, A.S.I., White, M.R.H., Davis, J.R.E., 2003. Real-time imaging of gene promoter activity using an adenoviral reporter construct demonstrates transcriptional dynamics in normal anterior pituitary cells, *J. Endocrinology* 178, 61-69.
- [60] Takasuka, N., White, M.R.H., Wood, C.D., Robertson, W.R., Davis, J.R.E., 1998. Dynamic changes in prolactin promoter activation in individual living lactotrophic cells, *Endocrinology* 139, 1361-1368.
- [61] Tao, Y., 2004. Intrinsic and external noise in an auto-regulatory genetic network. *J. Theor. Biol.* 229, 147-156.
- [62] Tao, Y., 2004. Intrinsic noise, gene regulation and steady-state statistics in a two gene network. *J. Theor. Biol.* 231, 563-568.
- [63] Thattai, M., van Oudenaarden, A., 2001. Intrinsic noise in gene regulatory networks. *Proc. Natl Acad. Sci. USA* 98, 8614-8619.
- [64] Tian, B., Zhang, Y., Luxon, B. A., Garofalo, R. P., Casola, A., Sinha, M., Brasier, A. R., 2002. Identification of NF- κ B-Dependent Gene Networks in Respiratory Syncytial Virus-Infected Cells. *J. Virol.* 76, 6800-6814.

- [65] Tian, B., Brasier, A. R., 2003. Identification of a Nuclear Factor Kappa B-dependent Gene Network. *Recent Progress in Hormone Research* 58, 95-130.
- [66] Tomioka, R., Kimura, H., Kobayashi, T.J., Aihara, K., 2004. Multivariate analysis of noise in genetic regulatory networks. *J. Theor. Biol.* 229, 501-521.
- [67] van Kampen, N.G., 2004. *Stochastic processes in physics and chemistry*. North-Holland.
- [68] White, R.J., 2001. *Gene Transcription: Mechanisms and Control*, Blackwell Science Ltd, Oxford, UK.
- [69] Wolfe, A. P., Pruss, D., 1996. Targeting Chromatin Disruption: Transcription Regulators that Acetylate Histones. *Cell* 84, 817-819.

Appendix A

Moments of the marginal mRNA and protein distribution

A.1 Equations for MGFs

Exact description. Given the probability generating functions (PGFs):

$$F(z, s) = \sum_{x,y} z^x s^y f_{xy}, \quad (\text{A.1})$$

$$G(z, s) = \sum_{x,y} z^x s^y g_{xy}, \quad (\text{A.2})$$

Eqs. (3.6)-(3.7) lead to the following partial differential equations for PGFs :

$$\begin{aligned} \frac{\partial F(z, s)}{\partial t} = & -(z-1) \frac{\partial F(z, s)}{\partial z} + Kz(s-1) \frac{\partial F(z, s)}{\partial z} \\ & -r(s-1) \frac{\partial F(z, s)}{\partial s} + bG(z, s) - cF(z, s), \end{aligned} \quad (\text{A.3})$$

$$\begin{aligned} \frac{\partial G(z, s)}{\partial t} = & H(z-1)G(z, s) - (z-1) \frac{\partial G(z, s)}{\partial z} + Kz(s-1) \frac{\partial G(z, s)}{\partial z} \\ & -r(s-1) \frac{\partial G(z, s)}{\partial s} - bG(z, s) + cF(z, s). \end{aligned} \quad (\text{A.4})$$

Equations (A.3)-(A.4) can be used to derive the moments of the mass functions f and g , both at the steady state as well as their time evolution. The method of generating functions allows deriving moments of arbitrary order, but for the purpose of this work only the expected

value and the variance are shown. By the property of PGFs, the first partial moments joint with the state of gene activity are given by:

$$E[X; G = 0] = \frac{\partial F(z, s)}{\partial z} \Big|_{z=s=1}, \quad (\text{A.5})$$

$$E[X; G = 1] = \frac{\partial G(z, s)}{\partial z} \Big|_{z=s=1}, \quad (\text{A.6})$$

$$E[Y; G = 0] = \frac{\partial F(z, s)}{\partial s} \Big|_{z=s=1}, \quad (\text{A.7})$$

$$E[Y; G = 1] = \frac{\partial G(z, s)}{\partial s} \Big|_{z=s=1}, \quad (\text{A.8})$$

thus by differentiating Eqs. (A.3)-(A.4) with respect to z , and separately, with respect to s , and substituting $z=s=1$, one gets the system of linear ODEs for the evolution of the first partial moments or by setting time derivative to zero, system of algebraic equations for the steady state. Resulting system must be augmented with the ODE describing changes of gene activity:

$$\frac{dG(1, 1)}{dt} = c(1 - G(1, 1)) - bG(1, 1),$$

where $G(1, 1) = G(z, s)|_{z=s=1}$ is the probability that the gene is active. Unique solution of the resulting system of equations yields quantities (A.5)-(A.8), which are then used to derive the expected number of mRNA and protein given by:

$$E[X] = E[X; G = 0] + E[X; G = 1], \quad (\text{A.9})$$

$$E[Y] = E[Y; G = 0] + E[Y; G = 1]. \quad (\text{A.10})$$

To obtain the second moments, it is required to express in terms of F and G the second factorial moments:

$$E[X(X-1); G = 0] = \frac{\partial^2 F(z, s)}{\partial z^2} \Big|_{z=s=1}, \quad (\text{A.11})$$

$$E[X(X-1); G = 1] = \frac{\partial G^2(z, s)}{\partial z^2} \Big|_{z=s=1}, \quad (\text{A.12})$$

$$E[Y(Y-1); G = 0] = \frac{\partial F^2(z, s)}{\partial s^2} \Big|_{z=s=1}, \quad (\text{A.13})$$

$$E[Y(Y-1); G = 1] = \frac{\partial G^2(z, s)}{\partial s^2} \Big|_{z=s=1}, \quad (\text{A.14})$$

as well as the joint mRNA and protein moments:

$$E[XY; G = 0] = \frac{\partial F^2(z, s)}{\partial z \partial s} \Big|_{z=s=1}, \quad (\text{A.15})$$

$$E[XY; G = 1] = \frac{\partial G^2(z, s)}{\partial z \partial s} \Big|_{z=s=1}. \quad (\text{A.16})$$

To write system of linear ODEs for the evolution of the second factorial moments (A.11-A.14) (or by setting time derivative to zero, system of algebraic equations for the steady state) one needs to differentiate Eqs. (A.3)-(A.4) twice with respect to z , and separately with respect to s , and substitute $z=s=1$. In addition, Eqs. (A.3)-(A.4) must be differentiated with respect to both, z and s , followed with substitution $z=s=1$ to close the system with equations for the joint moments (A.15)-(A.16). The resulting system of equations can be uniquely solved for the partial moments $E[X^2; G=0]$, $E[X^2; G=1]$, $E[Y^2; G=0]$, $E[Y^2; G=1]$, and then the second moments, $E[X^2]$, $E[Y^2]$ follow analogically to (A.9)-(A.10). Then, the variance can be derived, e.g., $\text{Var}[X]=E[X^2]-E^2[X]$.

Continuous model. Given the moment generating functions (MGFs):

$$F(z, s) = \int \int e^{zx+sy} f(x, y) dx dy, \quad (\text{A.17})$$

$$G(z, s) = \int \int e^{zx+sy} g(x, y) dx dy, \quad (\text{A.18})$$

Eqs. (3.19)-(3.20) lead to the following partial differential equations

$$\frac{\partial F(z, s)}{\partial t} + (z - sK) \frac{\partial F(z, s)}{\partial z} + sr \frac{\partial F(z, s)}{\partial s} = bG(z, s) - cF(z, s), \quad (\text{A.19})$$

$$\frac{\partial G(z, s)}{\partial t} - zHG + (z - sK) \frac{\partial G(z, s)}{\partial z} + rs \frac{\partial G(z, s)}{\partial s} = -bG(z, s) + cF(z, s). \quad (\text{A.20})$$

Equations for the partial moments (their time evolution as well as at steady state) can be obtained by differentiating Eqs. (A.19)-(A.20) with respect to z and s and setting $z=s=0$, similarly to the derivations for the exact description. Please note, that the second derivatives of MGFs with respect to z or s directly yield the partial moments ($E[X^2/G=0]$, $E[X^2/G=1]$, $E[Y^2/G=0]$, $E[Y^2/G=1]$), instead of the factorial moments in the case of the PGFs (A.11-A.14).

Mixed model. Given the MGFs:

$$F(z, s) = \sum_x \int_y z^x e^{sy} f_x(y), \quad (\text{A.21})$$

$$G(z, s) = \sum_x \int_y z^x e^{sy} g_x(y), \quad (\text{A.22})$$

Eqs. (3.28)-(3.29) lead the following partial differential equations:

$$\begin{aligned} \frac{\partial F(z, s)}{\partial t} - z s K \frac{\partial F(z, s)}{\partial z} + s r \frac{\partial F(z, s)}{\partial s} &= -(z-1) \frac{\partial F(z, s)}{\partial z} \\ &+ bG(z, s) - cF(z, s), \end{aligned} \quad (\text{A.23})$$

$$\begin{aligned} \frac{\partial G(z, s)}{\partial t} - z s K \frac{\partial G(z, s)}{\partial z} + s r \frac{\partial G(z, s)}{\partial s} &= H(z-1)G(z, s) - (z-1) \frac{\partial G(z, s)}{\partial z} \\ &- bG(z, s) + cF(z, s). \end{aligned} \quad (\text{A.24})$$

Eqs. (A.23)-(A.24) can be used to derive the moments (steady state and time evolution) by taking derivatives with respect to z and s and setting $z=1$, $s=0$, similarly to the exact description. Note, that the second derivative with respect to z yields the factorial partial moments, $E[X(X-1)/G=0]$, $E[X(X-1)/G=1]$, while the second derivative in s provides the regular partial moments, $E[Y^2/G=0]$, $E[Y^2/G=1]$.

A.2 Partial moments

The partial moments joint with the gene activity were derived as described in the previous section. The steady state solutions are presented below. To compress the notation, a binary variable α is introduced: $\alpha=1$ at the gene active state and $\alpha=0$ at the gene inactive state.

The first partial moments for the exact, continuous and mixed models are given by the same expressions:

$$E[X; G = \alpha] = Hc \frac{b(1-\alpha) + (1+c)\alpha}{(1+c+b)(c+b)}, \quad (\text{A.25})$$

$$E[Y; G = \alpha] = K \frac{(b+r)(1-\alpha) + c\alpha}{r(r+c+b)} E[X; G = 0] + K \frac{b(1-\alpha) + (r+c)\alpha}{r(r+c+b)} E[X; G = 1]. \quad (\text{A.26})$$

The first protein partial moments for the Kepler-Elston model are different than for the others and equal to

$$E_{KE}[Y; G = \alpha] = KHc \frac{d(1-\alpha) + (r+c)\alpha}{r(r+c+d)(c+d)}. \quad (\text{A.27})$$

The joint mRNA and protein partial moments for the exact, continuous and mixed model are given by

$$\begin{aligned} E[XY; G = \alpha] &= K \frac{(1+b+r)(1-\alpha) + c\alpha}{(1+r)(1+c+b+r)} E[X^2; G = 0] \\ &+ K \frac{b(1-\alpha) + (1+c+r)\alpha}{(1+r)(1+c+b+r)} E[X^2; G = 1] \\ &+ H \frac{b(1-\alpha) + (1+c+r)\alpha}{(1+r)(1+c+b+r)} E[Y; G = 1], \end{aligned} \quad (\text{A.28})$$

where the second moments (i.e. $E[X^2; G=0]$ and $E[X^2; G=1]$) depend on the specific model. The second partial moments are given below:

- Exact description:

$$E_E[X^2; G = \alpha] = H \frac{b(1-\alpha) + (2+c)\alpha}{(2+c+b)} E[X; G = 1] + E[X; G = 0], \quad (\text{A.29})$$

$$\begin{aligned}
E_E[Y^2; G = \alpha] &= K \frac{(b+2r)(1-\alpha) + c\alpha}{r(c+b+2r)} E_E[XY; G = 0] \\
&+ K \frac{b(1-\alpha) + (c+2r)\alpha}{r(c+b+2r)} E_E[XY; G = 1] \\
&+ E[Y; G = 0],
\end{aligned} \tag{A.30}$$

- Continuous model:

$$E_C[X^2; G = \alpha] = H \frac{b(1-\alpha) + (2+c)\alpha}{(2+c+b)} E[X; G = 1], \tag{A.31}$$

$$\begin{aligned}
E_C[Y^2; G = \alpha] &= K \frac{(b+2r)(1-\alpha) + c\alpha}{r(c+b+2r)} E_C[XY; G = 0] \\
&+ K \frac{b(1-\alpha) + (c+2r)\alpha}{r(c+b+2r)} E_C[XY; G = 1],
\end{aligned} \tag{A.32}$$

- Mixed model:

$$E_M[X^2; G = \alpha] = H \frac{b(1-\alpha) + (2+c)\alpha}{(2+c+b)} E[X; G = 1] + E[X; G = 0], \tag{A.33}$$

$$\begin{aligned}
E_M[Y^2; G = \alpha] &= K \frac{(b+2r)(1-\alpha) + c\alpha}{r(c+b+2r)} E_M[XY; G = 0] \\
&+ K \frac{b(1-\alpha) + (c+2r)\alpha}{r(c+b+2r)} E_M[XY; G = 1],
\end{aligned} \tag{A.34}$$

- Kepler-Elston approximation:

$$E_{KE}[Y^2; G = \alpha] = KH \frac{b(1-\alpha) + (2r+c)\alpha}{r(2r+c+b)} E_{KE}[Y; G = 1]. \tag{A.35}$$

A.3 Correlation between mRNA and protein number

Based on the partial moments correlation between the number of mRNA and protein molecules in the case of the exact stochastic description is derived. The correlation coefficient is given by

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var_E(X)Var_E(Y)}}, \quad (\text{A.36})$$

where the covariance can be expressed as $Cov(X, Y) = E[XY] - E[X]E[Y]$. Based on Eq. (A.28) the steady state covariance is equal to

$$Cov(X, Y) = \frac{rb(1+c+b+r)}{c(1+r)(1+c+b)(r+c+b)}E[Y]E[X] + \frac{K}{(1+r)}E[X]. \quad (\text{A.37})$$

It can be shown, that the first term in Eq. (A.37) is due to the intermittent gene activity, while the second results from the mRNA production/decay noise. In addition, it can be shown, that if $r \rightarrow 0$ (i.e. the protein half-life time approaches infinity, or in other words, the protein accumulates and never degrades), $Corr(X, Y) \rightarrow 0$ with a rate proportional to \sqrt{r} . If $r \rightarrow \infty$ (i.e. the protein half-life time approaches zero, or equivalently, the protein is degraded immediately after translation), $Corr(X, Y) \rightarrow 1$.

Appendix B

Error estimation

Continuous model. Based on th Eq. (3.10) the model neglects ε_m fraction of the total mRNA variance when

$$\varepsilon_m = \frac{E[X]}{Var[X]},$$

which yields that

$$\varepsilon_m = \frac{1}{\frac{b}{c(1+b+c)}E[X] + 1} < \frac{c(1+b+c)}{b} \frac{1}{E[X]}, \quad (\text{B.1})$$

when the expected number of mRNA molecules is much greater than 1.

If the consecutive terms in expression (3.11) are denoted with D , M , S , respectively, to have $Var_E[Y] = D + M + S$, then the continuous model neglects ε_p fraction of the total protein variance equal to

$$\varepsilon_p = \frac{M + S}{D + M + S},$$

or equivalently

$$\frac{\varepsilon_p}{1 - \varepsilon_p} = \frac{M + S}{D}.$$

Assuming that $r \ll 1$ (protein degradation rate much smaller than the mRNA degradation

rate) and introducing respective terms from (3.11) gives that

$$\frac{\varepsilon_p}{1 - \varepsilon_p} = \frac{c(c+b)}{b} \left(\frac{1}{E[X]} + \frac{1}{rE[Y]} \right).$$

Then, since $E[x] = \frac{r}{K}E[y]$ [Eq. (3.9)], the former yields that

$$\frac{\varepsilon_p}{1 - \varepsilon_p} = \frac{c(c+b)(K+1)}{rb} \frac{1}{E[Y]} \simeq \frac{c(c+b)K}{rb} \frac{1}{E[Y]},$$

where the transcription rate, which is the average number of protein molecules produced from a single mRNA transcript, is much greater than 1. In addition, when $\varepsilon_p \ll 1$, i.e. the error remains small, one finds that

$$\varepsilon_p = \frac{c(c+b)K}{rb} \frac{1}{E[Y]}. \quad (\text{B.2})$$

Mixed model. The model neglects ε_p fraction of the total protein variance given in Eq. (3.11) equal to

$$\varepsilon_p = \frac{S}{D + M + S},$$

or equivalently

$$\frac{\varepsilon_p}{1 - \varepsilon_p} = \frac{S}{D + M}.$$

Assuming that $r \ll 1$ and introducing respective terms from (3.11) gives that

$$\frac{\varepsilon_p}{1 - \varepsilon_p} = \frac{1}{\frac{rb}{c(c+b)}E[Y] + K} < \frac{c(c+b)}{rb} \frac{1}{E[Y]}.$$

In the above the second term in the denominator, K , is neglected, since the first term equals $\frac{b}{c(c+b)}K \cdot E[X]$ and thus much greater than the former. In addition, when $\varepsilon_p \ll 1$, one finds that

$$\varepsilon_p = \frac{c(c+b)}{rb} \frac{1}{E[Y]}. \quad (\text{B.3})$$

Appendix C

Discretization techniques

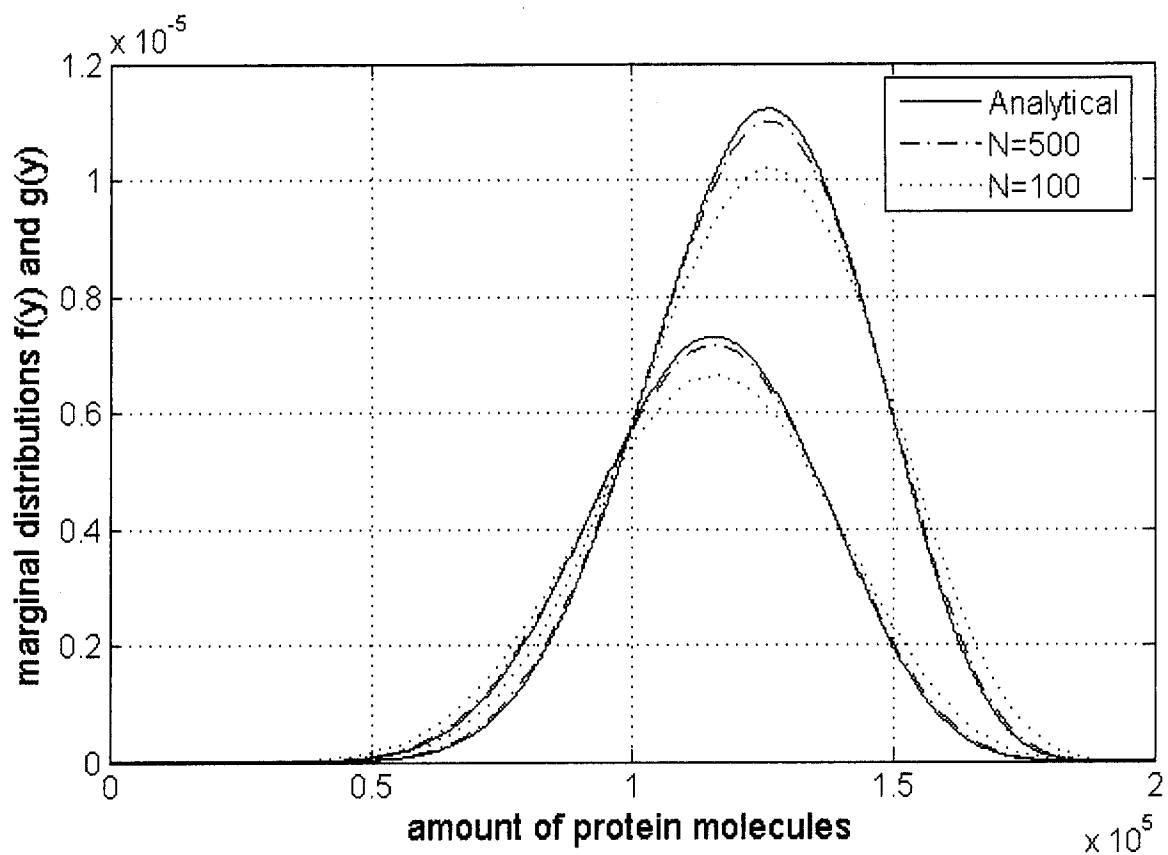
Kepler-Elston approximation. To illustrate the discretization techniques recall the simplified system (4.1-4.2). Consider PDEs (4.7-4.8) on a spatially discretized grid i , $0 \leq i \leq N$. The continuous variable y is replaced by $i \frac{KH}{rN}$, $i = 0, 1, \dots, N$. Let f_i and g_i denote the probability density functions f and g at point i of the grid. The discretized system (4.7-4.8) reads

$$\begin{aligned} \frac{df_i}{dt} &= b_r g_i - c_r f_i - i \frac{HK}{rN} f_i + (i+1) \frac{HK}{rN} f_{i+1} \\ \frac{dg_i}{dt} &= -b_r g_i + c_r f_i - (N-i) \frac{HK}{rN} g_i + (N+1-i) \frac{HK}{rN} g_{i-1} \end{aligned}$$

In each of the equations the first two right-hand side terms correspond to the exchange between densities f and g . The last two terms are due to the transport from and into the grid point i . Note the different directions of transport, from $i+1$ to i in the case of f and from $i-1$ to i in the case of g .

The stationary distributions are calculated by setting $\frac{df_i}{dt} = \frac{dg_i}{dt} = 0$. The resulting system consists of $2 \times (N+1)$ algebraic equations. However, to make the system unique one of its original equations has to be replaced by a normalization equation, i.e. $1/(N+1) \sum_{i=0}^N (f_i + g_i) = 1$. Fig. C.1 presents comparison between the numerical solutions obtained for $N=100$ and $N=500$ and the analytical solution given in Eqs. (4.13)-(4.14). As the size of the grid N

Figure C.1 : Accuracy of the numerical solution.



Numerical solution for $N=100$ (dotted lines) and $N=500$ (dashed-dotted lines) versus analytical solutions (solid lines) given in Eqs. (4.13)-(4.14). Shown are the stationary protein distributions $f(y)$ and $g(y)$ in Kepler-Elston approximation for $H=200$, $K=250$, $r=0.25$, $c=3$, $b=2$, $r=0.25$.

increases, the numerical solution becomes more accurate.

Continuous model. Consider the system (3.19)-(3.20) on a spatially discretized grid i, j , where $0 \leq i \leq N$ and $0 \leq j \leq N$. The continuous variables x and y are replaced by $i \frac{H}{N}$ and $j \frac{KH}{rN}$, respectively. Let $f_{i,j}$ and $g_{i,j}$ denote distributions f and g at point i, j of the grid. The discretized system (3.19)-(3.20) now yields:

$$\begin{aligned} \frac{df_{i,j}}{dt} = & b g_{i,j} - c f_{i,j} - i \frac{H}{N} f_{i,j} + (i+1) \frac{H}{N} f_{i+1,j} - \frac{KH}{N} |i-j| f_{i,j} \\ & + \frac{KH}{N} (i+1-j) f_{i,j-1} L_1 + \frac{KH}{N} (j+1-i) f_{i,j+1} L_2, \end{aligned}$$

$$\begin{aligned} \frac{dg_{i,j}}{dt} = & -b g_{i,j} + c f_{i,j} - (N-i) \frac{H}{N} g_{i,j} + (N+1-i) \frac{H}{N} g_{i-1,j} - \frac{KH}{N} |i-j| g_{i,j} \\ & + L_1 \frac{KH}{N} (i+1-j) g_{i,j-1} + L_2 \frac{KH}{N} (j+1-i) g_{i,j+1}, \end{aligned}$$

where L_1 and L_2 are the logical variables,

$$L_1 = 1 \text{ if } i > j - 1 \text{ and } L_1 = 0 \text{ if } i < j - 1,$$

$$L_2 = 1 \text{ if } i < j + 1 \text{ and } L_2 = 0 \text{ if } i > j + 1.$$

The stationary distributions are calculated by setting $\frac{df_{i,j}}{dt} = \frac{dg_{i,j}}{dt} = 0$. As a result system of $2 \times (N+1)^2$ algebraic linear equations is obtained. To make the solution unique, one of the equations by normalization $\sum_{i,j} (f_{i,j} + g_{i,j}) = (N+1)^2$ is replaced. Note that the matrix of the resulting system is relatively sparse, and the number of nonzero entries grows as N^2 , not as N^4 .

Mixed model. The system (3.28)-(3.29) is considered on the grid i, j , where $0 \leq i \leq N_x$, $0 \leq j \leq N_y$. The first coordinate, i , corresponds to the mRNA count and N_x describes the

maximum number of mRNA molecules considered. The continuous variable y is discretized into units of size $h = \frac{KN_x}{rN_y}$, where N_y is the number of unites considered. The discretized system (3.28)-(3.29) now yields:

$$\begin{aligned} \frac{df_{i,j}}{dt} = & bg_{ij} - cf_{ij} + (i+1)f_{i+1,j} - xf_{ij} \\ & - \frac{1}{h}|Ki - rhj|f_{ij} + L_1 \frac{1}{h}(Ki - rh(j-1))f_{i,j-1} + L_2 \frac{1}{h}(Ki - rh(j+1))f_{i,j+1}, \end{aligned}$$

$$\begin{aligned} \frac{dg_{ij}}{dt} = & -bg_{ij} + cf_{ij} + Hg_{i-1,j} + (i+1)g_{i+1,j} - (i+H)g_{ij} \\ & - \frac{1}{h}|Ki - rhj|g_{ij} + L_1 \frac{1}{h}(Ki - rh(j-1))g_{i,j-1} + \frac{1}{h}L_2(Ki - rh(j+1))g_{i,j+1}, \end{aligned}$$

where L_1 and L_2 are logical variables,

$$L_1 = 1 \text{ if } Ki > rh(j-1) \text{ and } L_1 = 0 \text{ if } Ki < rh(j-1),$$

$$L_2 = 1 \text{ if } Ki < rh(j+1) \text{ and } L_2 = 0 \text{ if } Ki > rh(j+1).$$

The stationary distributions are calculated by setting $\frac{df_{ij}}{dt} = \frac{dg_{ij}}{dt} = 0$. As a result, a system of $2 \times (N_x + 1)(N_y + 1)$ algebraic linear equations is obtained. To make the solution unique the height of the distribution $f(x,y)$ is set to be 1 at its mRNA and protein mean.

Two-gene system in Kepler-Elston approximation. Consider a system (4.37-4.40) on the grid i, j , where $0 \leq i \leq N$, $0 \leq j \leq N$. The continuous variables y_1 and y_2 are

discretized into i/N and j/N , respectively. The discretized system (4.37-4.40) reads

$$\begin{aligned}
\frac{df_{ij}^{00}}{dt} &= -if_{ij}^{00} + (i+1)f_{i+1,j}^{00} - rjf_{ij}^{00} + r(j+1)f_{i,j+1}^{00} \\
&\quad - (c_1^* + c_2^*)f_{ij}^{00} + b_2^*f_{ij}^{01} + b_1^*f_{ij}^{10}, \\
\frac{df_{ij}^{10}}{dt} &= -(N-i)f_{ij}^{10} + (N+1-i)f_{i-1,j}^{10} - rjf_{ij}^{10} + r(j+1)f_{i,j+1}^{10} \\
&\quad - (c_2^* + b_1^*)f_{ij}^{10} + c_1^*f_{ij}^{00} + b_2^*f_{ij}^{11}, \\
\frac{df_{ij}^{01}}{dt} &= -if_{ij}^{01} + (i+1)f_{i+1,j}^{01} - r(N-j)f_{ij}^{01} + r(N+1-j)f_{i,j-1}^{01} \\
&\quad - (c_1^* + b_2^*)f_{ij}^{01} + c_2^*f_{ij}^{00} + b_1^*f_{ij}^{11}, \\
\frac{df_{ij}^{11}}{dt} &= -(N-i)f_{ij}^{11} + (N+1-i)f_{i-1,j}^{11} - r(N-j)f_{ij}^{11} + r(N+1-j)f_{i,j-1}^{11} \\
&\quad - (b_1^* + b_2^*)f_{ij}^{11} + c_2^*f_{ij}^{10} + c_1^*f_{ij}^{01},
\end{aligned}$$

where

$$\begin{aligned}
c_1^* &= c_{10} + c_{11}\frac{i}{N} + c_{12}\frac{j}{N}, \\
b_1^* &= b_{10} + b_{11}\frac{i}{N} + b_{12}\frac{j}{N}, \\
c_2^* &= c_{20} + c_{21}\frac{i}{N} + c_{22}\frac{j}{N}, \\
b_2^* &= b_{20} + b_{21}\frac{i}{N} + b_{22}\frac{j}{N}.
\end{aligned}$$