UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE FÍSICA

Physical Modeling of Cellular Processes: Stochastic Gene Regulation in Embryonic Stem Cells

Tomás de Campos Aquino

Mestrado em Física

(Área de Especialização Física Estatística e Não Linear)

2011

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE FÍSICA

Physical Modeling of Cellular Processes: Stochastic Gene Regulation in Embryonic Stem Cells

Tomás de Campos Aquino

Mestrado em Física
(Área de Especialização Física Estatística e Não Linear)

Orientação: Professora Doutora Ana Maria Ribeiro Ferreira Nunes

2011

Contents

List of Figures

Acknowledgements

A number of people were very important to the writing of this thesis, directly or indirectly. I would like to express my thanks to all of them, whether their contribution was scientific, supportive, or otherwise. Some of them who I feel were most important I address here directly.

I would like to express my sincere gratitude to my advisor, Prof. Ana Nunes, Ph.D., without whom this thesis would certainly not have been possible. I would like to thank her specifically for shaping my approach to physics, throughout both my Degree and Master's; for unending discussions on physics and other related and unrelated subjects; and for her unfailing dedication to the projects in which I collaborated, and indispensable suggestions and contributions.

I thank the people at Unidade de Biologia do Desenvolvimento at Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa. In particular, I thank Elsa Abranches and Domingos Henrique, with whom I have collaborated in the project that led to this thesis. Their critical spirit, hard questions and forthcoming answers and discussions were essential to the work presented here.

I would also like to thank all the professors at Faculdade de Ciências da Universidade de Lisboa who have contributed to my physics formation and enthusiasm for the subject. Special thanks is due to Prof. Jorge Pacheco, Ph.D., for supporting this project.

I thank all my colleagues throughout my Degree and Master's. In particular, I would like to thank João Moreira and Flávio Pinheiro, for constant companionship and many helpful discussions.

Financial support from Fundação para a Ciência e a Tecnologia (FCT) under grant PTDC-FIS-70973-2006 is thankfully acknowledged.

I dedicate this thesis to Telma de Freitas Morna, for everything we have shared; to my closest friends, João Cabral, Martim Martins and André Sobral, who made me live a little more; and to my family, in particular my parents and grandparents, for their unfailing support and the moments I missed. Thank you.

Abstract

In the last few decades, with the advent of single-cell measurement techniques in experimental biology, a growing interest in the role of noise in cellular processes is apparent. In particular, cellular decision processes, based on (intrinsically noisy) regulated gene expression, are of paramount importance, as they can be found across all life, allowing cells to react to the internal and external media.

The tools of statistical physics have proved ideal for the development of a theoretical understanding of the underlying mechanics in noisy cellular processes. In this thesis, we make use of these tools to build a bottom-up model for single-gene auto-regulation, and present an application of this model to a concrete biological system, namely the regulation of the expression of the Nanog protein in mouse embryonic stem (ES) cells. Nanog has been identified as one of the core factors associated with the pluripotent state of ES cells: the concentration of Nanog protein present in a cell is known to be related to differentiation decisions. The structured population distributions of Nanog protein observed in ES cell populations call for mathematical modeling as an important tool to unravel the mechanisms underlying Nanog heterogeneity.

In Chapter 1, we begin with a brief overview of the main biological concepts involved in the following Chapters. After discussing the origin and role of stochasticity in cellular processes, we then present some of the basic modeling procedures and assumptions that have been adopted in the context of these processes.

In Chapter 2, we consider (transcriptional) single-gene auto-regulation specifically. This is the simplest motif in transcriptional gene regulation, and is thus of paramount importance. We adopt a Master Equation approach to describe stochastic effects in an intrinsic, fully dynamic fashion. We discuss the general formulation of the model, and obtain analytical descriptions of protein and mRNA equilibrium distributions using approximations that hold in regimes of strong biological relevance.

In Chapter 3, we discuss the application of our model for gene auto-regulation to the particular case of Nanog in mouse ES cells. To this end, we analyze experimental measurements of Nanog mRNA and protein distributions in populations of ES cells cultured in two different media.

We finish in Chapter 4 with some final remarks regarding the work laid out in this thesis.

Keywords: Physical Biology, Stochastic Processes, Master Equation, Gene Regulation, Nanog, Embryonic Stem Cells.

Resumo

Nas últimas décadas, tem-se verificado um desenvolvimento crescente na Biologia, quer a nível experimental, quer a nível teórico. As técnicas de medição ao nível de células individuais são cada vez mais eficazes e comuns, o que tem levado a um interesse crescente no papel da estocasticidade, ou ruído, nos processos celulares. Em particular, os processos de decisão celular, que se baseiam em expressão genética regulada e intrinsecamente ruidosa, são de fundamental importância, visto que se estendem a todos os organismos vivos, permitindo às células responder ao meio interno e externo.

A existência de uma grande quantidade e variedade de dados experimentes relativos a expressão genética tornam relevante a intervenção da Física na sua interpretação, contribuindo para uma componente formal da Biologia cada vez mais bem definida e desenvolvida. As técnicas da Física Estatística têm-se provado ideais para o desenvolvimento de um entendimento teórico dos mecanismos subjacentes à expressão genética, devido ao carácter estocástico desta última. Nesta dissertação, fazemos uso destas técnicas para construir um modelo *bottom-up* para a auto-regulação de um só gene, e apresentamos uma aplicação deste modelo a um sistema biológico concreto, nomeadamente a regulação da expressão da proteína Nanog em células estaminais embrionárias (EE) de rato. O Nanog foi identificado como um dos elementos principais associados ao estado pluripotente de células EE: sabe-se que a concentração de proteína Nanog presente num célula está relacionada com decisões de diferenciação. Sabe-se também que as proteínas Oct4 e Sox2 são essenciais na manutenção deste estado pluripotente. As distribuições estruturadas de Nanog observadas em populações de células EE apontam para a modelação matemática como uma ferramenta importante para a identificação dos mecanismos subjacentes à heterogeneidade na expressão de Nanog. Nesta dissertação, utilizamos o modelo construído para explorar a hipótese de que o comportamento do Nanog pode ser explicado no contexto da auto-regulação, sem necessidade de intervenção dinâmica de outras proteínas como proposto na literatura.

O trabalho apresentado nesta dissertação decorreu de uma estreita colaboração entre um grupo de Física e um grupo de Biologia, permitindo não só o acesso a dados experimentais, como também a discussão do sistema em causa de dois pontos de vista complementares, levando à construção e exploração de um modelo baseado em princípios matemáticos sólidos e em hipóteses biológicas razoáveis e relevantes.

No Capítulo 1, abrimos com uma breve apresentação dos principais conceitos biológicos envolvidos nos Capítulos seguintes. Em particular, fazemos uma breve descrição dos processos de produção de RNA a partir do DNA (transcrição), e de proteína a partir do RNA (tradução). Discutimos também a origem e o papel da estocacidade nos processos celulares: em primeiro lugar, visto que os mecanismos celulares assentam frequentemente em reacções químicas, e consequentemente em encontros aleatórios entre moléculas ou compostos moleculares, torna-se claro que a aleatoriedade tem um papel fundamental e inegável nestes processos; por outro lado, é também cada vez mais evidente que a estocacidade pode jogar um papel benéfico, por exemplo na medida em que promove a variabilidade e a capacidade de resposta a variações do meio. O carácter intrínseco e dinâmico do "ruído"aqui discutido aponta de imediato para a necessidade de formalismo adequado. Terminamos este Capítulo com uma apresentação de alguns procedimentos e hipóteses básicos que têm sido adoptados no contexto dos processos celulares, começando por formulações deterministas (utilizando *rate equations*) e terminando em formulações estocásticas (onde discutimos excursões estocásticas e apresentamos o formalismo da Master Equation para a evolução de uma distribuição de probabilidade).

No Capítulo 2, consideramos especificamente a auto-regulação (transcricional) de um só gene. Este é o motivo mais simples que pode ser encontrado na regulação genética transcricional, e a sua compreensão detalhada é portanto de fundamental importância. Adoptamos um formalismo baseado em Master Equations para descrever os efeitos estocásticos de uma forma intrínseca e totalmente dinâmica. Em particular, o modelo descreve os processos que têm lugar ao nível do DNA, mRNA e proteína, e incorpora a produção de mRNA e proteína em *bursts* estocásticos (produzindo um número de moléculas segundo uma distribuição geométrica), para os quais existem diversas evidencias experimentais. Discutimos a formulação geral do modelo, considerando em detalhe o caso em que a regulação é mediada por dimerização da proteína seguida de ligação ao respectivo promotor, e obtemos descrições analíticas para as distribuições de equilíbrio de proteína e mRNA recorrendo a aproximações válidas em regimes de forte relevância biológica. A descrição destes fenómenos é feita ao nível da célula individual, mas as distribuições de equilíbrio de mRNA e proteína obtidas neste contexto são também válidas para as distribuições destas espécies em populações de células idênticas e sem interacção entre si, como é o caso para as culturas de células EE aqui referidas.

No Capítulo 3, discutimos a aplicação do modelo desenvolvido ao caso particular do Nanog em células EE de rato. Para este fim, analisamos medições experimentais de distribuições de mRNA e proteína Nanog em po-

pulações de células EE em dois meios de cultura distintos. Nestes dois meios de cultura observam-se distribuições de mRNA e proteína Nanog distintas, permitindo testar em mais detalhe o modelo proposto. Com base nos detalhes das medições experimentais, e numa exploração do modelo recorrendo a métodos analíticos e numéricos, concluímos que estas não apontam para a intervenção de outras proteínas na regulação do Nanog, e discutimos a possibilidade de a regulação do Nanog ter por base um mecanismo ao nível da tradução e não ao nível da transcrição. Fazemos também uma breve análise deste mecanismo como uma extensão do modelo proposto.

Terminamos no Capítulo 4 com alguns comentários finais relativos ao trabalho apresentado na dissertação. Em particular, referimos a necessidade da utilização de um formalismo adequado aos processos em questão, nomeadamente no que diz respeito ao seu carácter estocástico, visto que os aspectos fundamentais que podem ser observados nas medições experimentais apenas ganham sentido à luz deste tratamento. A análise dos dados experimentais no contexto de um modelo em que a estocacidade é formulada dinamicamente permitiu não só uma discussão crítica dos mecanismos propostos na literatura para o sistema em causa, como também a apresentação e discussão de hipóteses originais, num tema de grande actualidade e relevância científica.

Palavras-chave: Física de Sistemas Biológicos, Processos Estocásticos, Regulação Genética, Master Equation, Nanog, Células Estaminais Embrionárias.

Chapter 1

Introduction

1.1 The Genetic Revolution

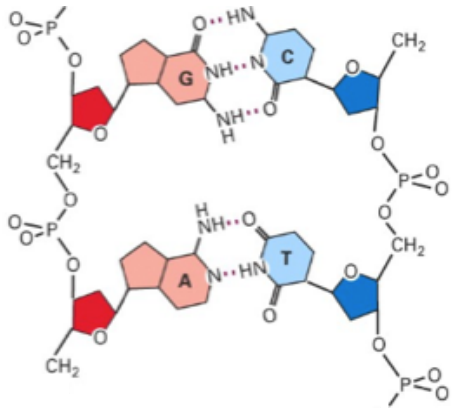

Figure 1.1: Standard DNA base-pairing. The double-helix structure of the DNA molecule is bound together by bonds between the nitrogenous bases adenine (A), thymine (T), guanine (G) and cytosine (C) of each chain. Adenine binds to thymine and guanine to cytosine. This scheme holds for RNA-DNA strand binding, except thymine is substituted by uracil in RNA. Adapted from [27].

When we consider high-level, everyday aspects of life on Earth, we find outstanding variability, from morphological to behavioral traits. As a measure of this diversity, the number of catalogued species on the planet currently ranks at around 2 million, with estimates for the total number of species ranging from around 5 to 100 million.

However, a closer inspection reveals the staggering universality of life's fundamental building blocks. The discovery of cells is usually ascribed to

Robert Hooke, who identified the cellular structure of cork using a microscope, as early as the 17th century. Later, in 1838, botanist Matthias Schleiden stated that all plants are made of cells; just one year later, physiologist Theodor Schwann proposed the same for animals. It is now common knowledge that life universally relies on cells, organized structures with a characteristic length on the order of $1 - 100$ μm, where the fundamental chemical processes that sustain life take place. Inside each cell there resides a DNA molecule, usually confined to a characteristic length of $1 - 10$ μm.

In 1953, James Watson and Francis Crick identified the double-helix structure of the DNA molecule, a landmark of molecular biology. However, Walter Sutton proposed the existence of "genes", responsible for heredity and residing in the chromosomes, as early as 1902, see the 1903 paper [41]. It is now well known that the DNA molecule features sequences of base pairs corresponding to specific proteins according to a universal "Genetic Code". These proteins carry out the bulk of cellular functions, from being the building blocks of the physical/mechanical structure of the cell to acting as catalyzers of the essential biochemical reactions. They are also involved in higher level organization through cell-cell signaling.

In 1958, Crick formulated the so called "Central Dogma" of molecular biology, which is again transversal across all life. While the original idea of the Central Dogma focuses on the irreversibility of information transfer from DNA into proteins, it is of more importance here to briefly discuss each step of the process of obtaining proteins from their coding in the DNA, to give some insight on the complexity of the phenomena at the molecular level. There are essentially two steps: transcription, where a gene in the DNA is used as a template for an RNA molecule, and translation, where the information stored in an RNA molecule is used to produce a protein.

Transcription starts when an enzyme called RNA polymerase binds the DNA molecule at a promoter region, where a start site for the copying of the template strand into RNA is located. For the reading of the template strand to proceed, the double-helix structure of the DNA molecule is unraveled, and a transcription bubble is formed. After this, the polymerase catalyses phosphodiester linkage of the first two ribonucleotides, the building blocks of an RNA strand, to the nucleotides in the template DNA strand, according to a well-defined complimentary base-pairing code (see Figure 1.1). The polymerase then advances along the DNA molecule, and the transcription bubble advances with it as the transcribed DNA reforms the helix structure, the transcribed RNA unbinds and progressively exits the polymerase, and the polymerase melts the upcoming region (see Figure 1.2). Eventually a specific sequence in the DNA, called a stop site, is reached; the polymerase then unbinds from the DNA and transcription ends. It should also be noted that,

in eukaryotes, the RNA molecules thus produced are sometimes called pre-RNAs, because they must be processed by additional molecular machinery in order to produce functional RNA; most strikingly, many RNA molecules have functional sequences (exons) interspersed by non-functional sequences (introns), and the introns must be cut off – this is called splicing.

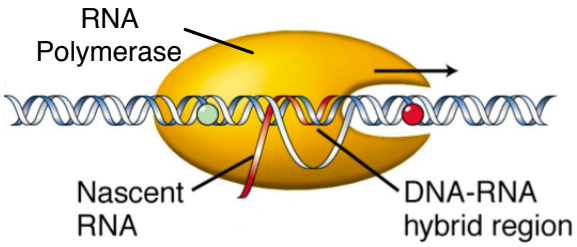

Figure 1.2: Basic elements of transcription (see main text). Adapted from [27].

Translation occurs in the cytoplasm of the cell. The fundamental machines of translation are the ribosomes, which are built from proteins and a particular type of RNA termed rRNA (r for ribosomal). The RNAs that code for proteins are called mRNAs (m for messenger). When a ribosome binds an mRNA molecule, its ribonucleotides are read in triplets; each triplet codes for a particular amino-acid, a building block of protein, according to the universal genetic code (see for example [27]. Note that triplets in the mRNA are complementary to the triplets in the DNA; the genetic code conventionally refers to the latter). Another type of RNA, called tRNA (t for transfer), exists in the cytoplasm; each sports a particular base-pair triplet, and binds the corresponding amino-acid. When a ribosome reads a triplet in the mRNA strand, it binds the complimentary tRNA and uses its amino acid to progressively assemble a polypeptide chain, which will eventually fold into a protein (see Figure 1.3).

For more details on the molecular complexes and reactions involved in transcription and translation, we refer the reader to [27]. We note that standard quantitative models of gene expression existing in the literature reduce transcription and translation to events characterized by some rate of occurrence, eventually subject to regulation. This procedure is essentially justified *a posteriori* by the explanatory and predictive power of the model, and the possible effects of lower-level details are still by and large unknown. A simple example of such an effect, which has been proposed in the literature as a key ingredient for observed protein expression oscillations in somitogenesis, is the delay between transcription/translation initiation and the formation of a functional RNA/protein, see [25].

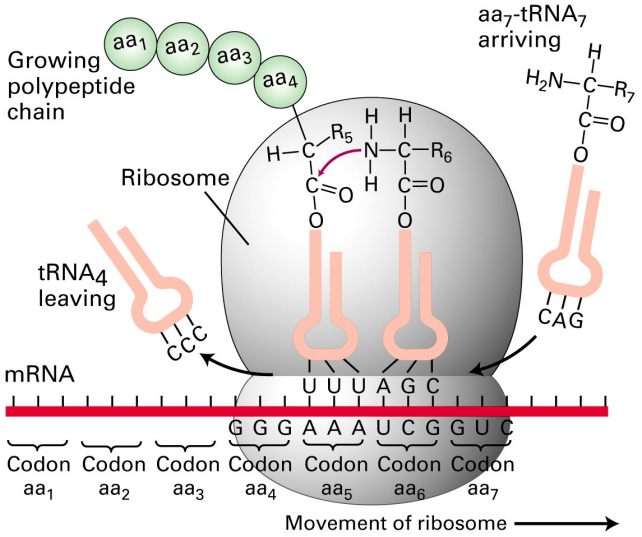

Figure 1.3: Basic elements of translation (see main text). Adapted from [27].

1.2 Dynamic Gene Expression

Far from its taxonomical origins, biology was now unveiling the molecular basis of life. The idea that the process of development obeyed a deterministic "genetic program", written in the DNA and uniquely determining cell fate, gained popularity following the successful cracking of the genetic code. However, differences in phenotype for cells with the same genetic content defy the simplistic interpretation of the idea of a genetic program. In fact, most cells in a single organism contain the same genetic material, namely its species' whole genome!

Part of the answer to this apparent contradiction lies in taking into account the dynamic nature of gene expression. A well-known and early statement of the dynamic character of cell decision-making is Conrad Waddington's "epigenetic landscape", dating back to 1957. Waddington intuitively pictured the decision-making process by a movement in a "potential", defined by internal and external conditions, with specific cell states being represented by local minima (see Figure 1.4). Despite the presence of the whole genome, at each time only parts are being transcribed. In fact, which proteins are being coded for at a specific time may depend on a cell's surrounding medium, or on the behavior of its neighbors through cell-cell signaling; because proteins are degraded in the cell, and diluted due to cell division, this conditions protein abundances, and may lead cells to different activities, or even to

irreversible commitment to different roles at key moments in development. Furthermore, the presence or absence of some proteins may affect their own or other genes' expression, the best-known mechanism for this interaction being the binding of proteins to specific sequences in the DNA called promoter sites. This phenomenon is termed gene regulation, and was discussed as early as 1961 by Jacques Monod and François Jacob for the *lac* operon in *E. coli*. Note that so-called post-transcriptional regulation, affecting various processes such as RNA splicing or translation, is less well understood but seems also to play a prominent role, see [17, 37, 40].

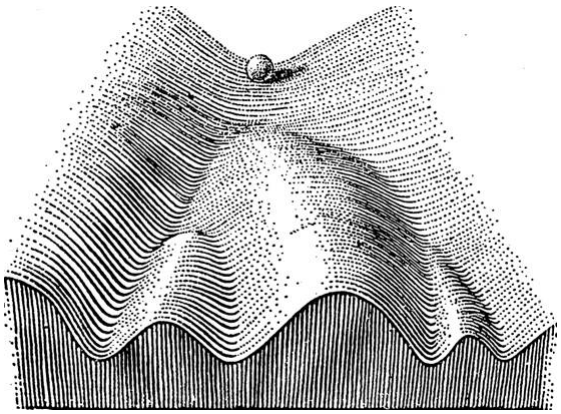

Figure 1.4: Conrad Waddington's epigenetic landscape. From [46].

The Drosophila fly is a famous example of an organism whose development is well understood in terms of gene expression and regulation (see for example [9] for the role of the bicoid gene in morphogenesis). Within this framework, the development of the Drosophila has in fact been found to obey a strict genetic program. Indeed, mutants with reproducible phenotypes result from specific changes in the fly's DNA, see [16] for an example.

However, not all differences in cellular behavior can be explained in the light of medium or neighbor influences. From early on in the history of genetics, a debate between determinism and stochasticity in development has been present. In various systems (such as stem cells in the blastocyst, neural tube formation, leukemia, etc.), cells have been found to make different decisions when in apparently identical conditions, see for example [28] for a review. To address this question, it becomes essential to understand in more detail the dynamics of gene expression and regulation, and the possible roles of stochasticity in cellular processes.

1.2.1 Sources of Randomness

The role of stochasticity in cells and microorganisms has been discussed theoretically since the 1970s, see the classical papers [1, 2]. The evolution of experimental molecular biology techniques has made single-cell measurements possible, and brought numerous confirmations of the presence of stochastic effects in gene expression, see [11] for a classical example. The new experimental possibilities have accordingly brought on a renewed interest in the mechanisms underlying gene expression and regulation in general, and specifically on the sources of randomness affecting them.

Because cellular processes often rely on chemical reactions, and correspondingly on chance encounters between molecules or molecular complexes, stochastic effects due to small numbers and rare events will undoubtedly play some role. The fact that genes coding for specific proteins are often present in single copies provides a striking example; gene activation and regulation, often depending on random association and dissociation events, may introduce considerable noise. Furthermore, transcription is typically rare, with mRNAs being commonly present in low copy numbers, from a few to a few hundred molecules, and many proteins also exist in low number. Because transcription, translation and degradation events are stochastic, finite size fluctuations in mRNA and protein numbers become important. If cellular division plays a role, the partitioning of the mother cell's contents between the daughter cells also introduces an additional source of stochasticity.

Measuring of population distribution of proteins and mRNA often shows heavy-tailed distributions that are still difficult to explain in terms of the previous ingredients. More recently, the development of single-molecule techniques has led to the experimental identification of another, more specific source of variability in gene expression that accounts for heavy-tailed distributions. Both transcription and translation have been found, in many cases, to occur in time-localized bursts resulting in a geometrically distributed number of molecules, see [7, 23, 31, 42]. While the concept of bursts and the mechanisms underlying them are still open to discussion (see for example [32]), some simple ideas clarify this process. First, if transcription/translation events are widely spaced compared to their duration, it is reasonable to speak of burst events. Second, the geometric distribution relates to the number of successive "heads" in the throwing of a (in general, biased) coin; thus, if during a burst event there is a fixed probability that *another* molecule will be produced, a geometrically distributed number of molecules results. A major achievement of the burst description is that the resulting predicted form of unregulated protein expression distributions (see [14, 38], Chapter 2 and Appendices D and E of this thesis for theoretical descriptions) is remarkably

simple and fits an impressive number of experimental distributions measured for *E. coli* populations, see [42]. Bursty gene expression is also fertile ground for current theoretical work, see for example [10, 33].

When considering specific models or experiments, another form of randomness, relating to fluctuations of the biological parameters of the system under consideration, becomes apparent. For example, we may characterize an active gene by a constant effective transcription rate, while in fact this rate depends on the cell's transcription machinery, and eventually on the presence of transcription factors in approximately constant concentrations. Fluctuations of these concentrations or other biochemical parameters will result in fluctuations over time of the effective transcription rate. Additionally, we also expect variations of typical parameters across the individuals of a population of cells or microorganisms.

1.2.2 Noise as a Functional Element

Because life evolved in naturally noisy conditions, many organisms and intracellular systems have developed strategies and structural architectures that make use of stochastic variability rather than reducing it. Whereas some aspects of development, for example, require strictly regulated behaviors, usually mediated by strictly regulated expression of certain genes, other phenomena may in fact benefit from increased variability.

Generically, the basic role of randomness in gene expression is to provide a natural means of generating variability across a population. For example, a bacterial population in a fast-changing medium benefits from the presence of distinct phenotypes, since they allow the colony to quickly adapt to a wide range of conditions. These ideas apply more generally to populations of microorganisms or cells where a swift response to a wide variety of stimuli is desirable. A striking example is presented in [6], where a less noisy synthetic version of a natural system is built and found to be less robust (i.e., the native system is functional in a wider range of conditions).

Another possibility opened by random fluctuations are so-called stochastic excursions. Consider for example a gene with a stable expression level, which would be maintained in the absence of fluctuations. Under certain conditions, discussed in more detail in the next section, finite-size stochastic effects may suffice to drive long excursions of the gene's expression to higher or lower values, producing well-defined pulses and/or bistable expression distributions in a population. Examples of theoretical approaches to these ideas can be found in [14, 22, 36].

While many works focus on the limits imposed by stochasticity and the evolution and properties of noise-minimization strategies in cells and microor-

ganisms, see for example [1, 3, 4, 13], a growing interest in possible functional roles of noise is reflected in recent literature, as discussed above. In this thesis we follow this trend: in Chapter 2 we develop a stochastic model for a particular instance of gene regulation, and in Chapter 3 we discuss in more detail a system where variability plays an important role.

1.3 Biophysical Modeling

The central aspect of this thesis is the physical modeling of biological systems, in particular at the cellular level. The uncovering of unifying fundamental principles, as well as the quality and availability of experimental methods, is transforming biology ever more into a fully quantitative science. As this trend of quantification progresses, measurements of many biological quantities become available in the literature, see for example [37] for general quantitative data on gene expression.

Parallel to the experimental development, quantitative theoretical understanding becomes imperative. Physics, with its history of applying mathematical formalism to concrete problems, is the natural candidate to help bring biology closer to quantitative rigor and to explore the insights available in this new framework. In particular, the stochastic character inherent to many biological problems makes many techniques developed in statistical physics ideal to be adapted and applied in this context.

Throughout this section we provide a brief introduction to some of the fundamental concepts of biophysical modeling, from a deterministic to a stochastic description. For a more comprehensive discussion of the deterministic approach, we refer the reader to [12]; for a more detailed formal approach to stochastic gene regulation, see [47]. In the next Chapter, a concrete illustration of the application of these ideas is provided.

1.3.1 Rate Equations

Classically, the available experimental methods dictated that biological quantities, such as the expression of a protein, were described by their cell-population averages. Deterministic models of biological processes were thus popular, because in most cases they provide a good description at this level of detail, and because of their theoretical simplicity.

When the quantities involved can be modeled as continuous, and stochastic effects can be neglected, rate equations are the standard approach to model time evolution and equilibrium values. Note that the actual quantities to be described are usually discrete, such as the number of molecules of

a certain species; the corresponding continuous variables described by this approach result from multiplication by a small parameter. The simplest and most intuitive example is the description of reacting chemical species in a large fixed volume, where the natural continuous variables to be used are each species' concentration (the corresponding small parameter being the inverse of the volume). More generally, the key idea for the validity of the rate equation approach is that the relative size of the stochastic fluctuations due to small numbers should go down with system size; the van Kampen Expansion formalizes this idea, and shows how the systematic expansion of a Master Equation in a small parameter measuring the inverse of system size leads in lowest order to a rate equation describing a macroscopic variable, see [45]. For these reasons, we generically term the continuous variables in this approach "concentrations".

In the rate equation framework, the evolution of the system is described by a set of coupled ordinary differential equations (ODEs) for the relevant concentration variables, where each concentration-changing process is characterized by their mean effective rate of occurrence: if $x = \lambda n$ is the concentration associated with the discrete variable n, and if a process that increases n to $n + \mu$ takes place at rate γ, then the effective rate characterizing this process will be $\gamma\tilde{\mu}$, with $\tilde{\mu} = \lambda\mu$.

The flexibility of this modeling strategy relies on the possibility of nonlinear rates of occurrence, i.e. rates with arbitrary dependence on the concentrations. In biology, the assumptions made about this dependence are usually termed the choice of kinetics. Again the simplest example are so called mass-action kinetics, which correspond to standard chemical reactions happening with a certain probability on chance encounters between the intervening species, and with the additional ingredient of homogeneity in a fixed volume V. Consider that the species X_i, $i \in I$ react with each other simultaneously, originating the species X_i,$i \in J$ (with I and J arbitrary sets of species). If s_i and r_i are, respectively, the number of intervening and resulting copies of species X_i, we may write this process in chemical reaction notation as:

$$\sum_{i\in I} s_i X_i \xrightarrow{\gamma} \sum_{i\in J} r_i X_i \ . \tag{1.1}$$

If we adopt mass-action kinetics, this reaction translates into the system of coupled ODEs for the concentrations x_j:

$$\dot{x_j} = \gamma V^{-1}(r_j - s_j)\prod_{i\in I} x_i^{s_i} \ , \tag{1.2}$$

where j spans all involved species. Note that $\gamma \prod_{i \in I} x_i^{s_i}$ is the number of collisions per unit time that result in reaction (1.1) taking place, and the whole right-hand side of equation (1.2) is the effective rate characterizing this reaction. Because each reaction is completely characterized by its effective rate given the copy number of each species present at a certain time, the terms corresponding to additional reactions are simply summed to the appropriate equations. For more details on chemical reaction systems, see for example [45].

On the other hand, biological processes routinely involve large numbers of intervenients in complicated combinations. Rather than model each individual reaction using mass-action kinetics, it is sometimes more relevant to describe a composite process by effective kinetics, which may be the result of a large number of subprocesses. Examples are Michaelis-Menten kinetics and Hill kinetics, which take into account saturation in metabolic (e.g. enzymatic) processes, yielding sigmoid-like functions (see [12]); in general, custom functions of the concentrations of involved species may be used to describe a particular process. Remember also that in order to study a population (e.g. of cells or microorganisms), reactions may model processes for a single individual if they are independent and identical, or they may include communication between cells if it is important. If spatial localization or mass flow are important, the same kind of ODE formalism can be applied to study the evolution of concentrations at each point in space, as long as concentration flow is taken into account. A famous example of a successful application of deterministic rate equations to biological systems is detailed in [44], a classical work by Alan Turing dating back to 1951, where pattern formation is described in terms of reaction-diffusion.

Many works in metabolic- and gene regulatory networks focus on studying bifurcation diagrams in the context of deterministic models, see [12] for a review and [29] for an introduction to bifurcation theory. In the context of cellular processes, changes in the number and properties of equilibria as a function of biological parameters are very important because they provide a means for the cell to function in different modes according to media stimuli. For an exploration of bifurcation properties as cellular decision switches using rate equations see [20]. Classically, different modes of cellular functioning, regarding for example gene expression, are associated with different equilibria of the deterministic dynamics. Because rate equation models describe the system in terms of coupled ODES, a large number of intervening species is often necessary to describe non-trivial equilibrium dynamics; in fact, for a single molecular species, corresponding to a single ODE, the only possible equilibria are fixed points. More recently, a deterministic model for somitogenesis involving cell-cell communication and delay was proposed in [25].

In the context of delay differential equations, stable limit cycles may exist even for a single intervening species, permitting oscillations of a single component in a deterministic setting (for further works relating to delay in gene expression, see for example [5, 15]).

1.3.2 Stochastic Processes

As cell-level variability became experimentally accessible, and single-cell measurement techniques became available, the need for theoretical descriptions where stochasticity was taken into account became evident. In their pioneering paper of 1977, see [1], Berg and Purcell considered the effects of diffusion in nutrient search by single-celled organisms, using the tools of equilibrium statistical physics. The variability due to stochastic effects in gene expression was also tackled in 1978, see [2], but many open problems still consist in explaining observed distributions of protein and mRNA in cell or microorganism populations.

As mentioned in the previous subsection, a limitation of simple deterministic, rate equation models is the necessity of high dimensionality to explain non-trivial equilibrium behavior. One of the simplest effects of stochasticity, due for example to small copy numbers, is the introduction of fluctuations that can change the behavior around equilibria: if the deterministic model exhibits a stable fixed point for a certain species, in the corresponding stochastic description we may find a nontrivial distribution. An important example, as mentioned above, are so-called stochastic excursions, in which noise allows a system to move away from the deterministic equilibrium. The basic role of stochasticity and the concept of stochastic excursions are easily understood in the Langevin framework, which we address very briefly here (for an in-depth description of Langevin equations, we refer the reader to [45]; for an informal introduction to stochastic differential equations, see [24]). For simplicity, consider a one-dimensional system; if a deterministic equation describing concentration x in the limit of no fluctuations is known, a noise term $M(x)L(t)$ may be added, yielding:

$$\dot{x} = F(x) + M(x)L(t) \ , \tag{1.3}$$

where $F(x)$ describes the deterministic behavior. The stochastic Langevin term $L(t)$ is characterized by its its moments and multiplied by a magnitude M(x), both of which must be identified from the physical properties of the system.

As an example, consider the simplest stochastic term, additive Gaussian white noise: the magnitude $M(x)$ is a constant which we take here to be

unity, and $L(t)$ is fully determined by having zero average, $\langle L(t)\rangle = 0$ for all t, and the auto-correlation function:

$$\langle L(t)L(t')\rangle = \delta_D(t-t') \ , \tag{1.4}$$

where the averages refer to an ensemble of identical systems. Higher moments are defined as for the Gaussian distribution, and the Dirac delta form of the autocorrelation function characterizes white noise, the infinitely sharp peak corresponding to zero correlation time.[1] In this particular case, equation (1.3) can be shown to be equivalent to a Fokker-Planck equation for the probability distribution $p(x,t)$ of finding concentration x at time t (see [45]):

$$\dot{p}(x,t) = -\partial_x(F(x)p(x,t)) + \frac{1}{2}\partial_x^2 p(x,t) \ . \tag{1.5}$$

In equilibrium ($\dot{p} = 0$) the solution is readily found to be:

$$p^{eq}(x) = p^{eq}(0)e^{2\int_0^x F(u)du} \ , \tag{1.6}$$

with $p^{eq}(0)$ determined by normalization.

To illustrate the effects of noise with a concrete example, let F be a polynomial of degree 3, and write $F(x)$ as:

$$F(x) = \alpha(x-x_0)(x-x_1)(x-x_2) \ . \tag{1.7}$$

In deterministic equilibrium ($\dot{x} = 0$), the deterministic system will rest at a fixed concentration x^* obeying $F(x^*) = 0$. In the presence of noise, however, the fluctuations render other values of concentration accessible to the system. If we look at the equilibrium distribution of concentration, the simplest effect of randomness is a broadening around a single deterministic equilibrium x_0, as illustrated in Figure 1.5-A. If $F(x) = 0$ has more than one solution, corresponding to multistability of the deterministic system, stochastic-induced transitions from one equilibrium to the other may occur, producing multimodal distributions, as in Figure 1.5-B.

Finally, consider that F is asymmetrical around a single equilibrium x_0 and that $|F|$ is small in a region to one side of x_0 (compared to the magnitude of the Langevin term), as exemplified in Figure 1.6. Because in this region

[1] The idea of Gaussian white noise is embodied by the effect of a random force due to collisions in the speed of a Brownian particle: successive collisions of a small particle with the more massive particles of a surrounding fluid can be thought of, at the timescale of diffusion, as being instantaneous (very low correlation time) and arriving randomly and very frequently. Note that if the correlation time of the Langevin term is non-negligible, i.e. if noise in non-white, memory effects are important and the process is non-Markovian, see [45].

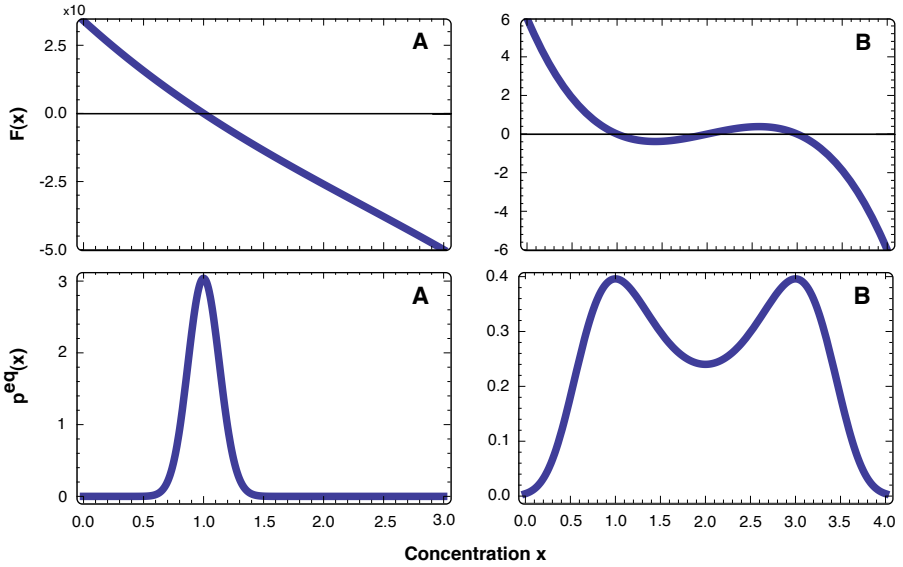

Figure 1.5: Effects of Gaussian white noise on the equilibrium distribution of concentration. The dynamics correspond to equation (1.3) with deterministic term given by (1.7) and Gaussian white noise with unit magnitude (see main text).
(**A**) $\alpha = 1$, $x_0 = 1$, $x_1 = 3 + 5i$, $x_2 = 3 - 5i$.
(**B**) $\alpha = 1$, $x_0 = 1$, $x_1 = 2$, $x_2 = 3$.

deterministic effects are weak, the system is dominated by stochastic effects, and is said to perform stochastic excursions away from the deterministic equilibrium. As illustrated in Figure 1.6, nontrivial equilibrium distributions may result from this effect. While this is not present for the simple example discussed here, stochastic effects may even give rise to bimodality when bistability does not exist in the deterministic description, as reported in [43].

The fundamental difference between a deterministic and a stochastic description of the time evolution of a quantity X is reflected in the mathematical objects used to describe them. In a deterministic setting, X takes a definite value x at each time t, and the dynamics are thus described by a function $x(t)$ (which is usually the solution to a differential equation, as discussed above). On the other hand, if stochastic effects are present, fundamental unpredictability is introduced in the evolution of x. The dynamics now constrain only the probability of $X = x$ at each time t, and the function $x(t)$ is replaced by a family of probability distributions $p(x, t)$ describing the time evolution of the probability distribution of x values. The aim of a stochastic description of the system is thus to obtain $p(x, t)$, or to identify some of its properties. The determination of the properties of $L(t)$ from physical principles may however present problems, and if the stochastic term

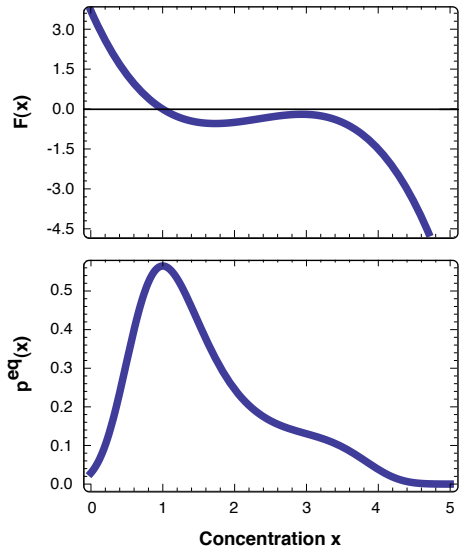

Figure 1.6: Effects of Gaussian white noise on the equilibrium distribution of concentration. Dynamics are as in Figure 1.5, but with:
$\alpha = 0.4$, $x_0 = 1$, $x_1 = 3 + 0.5i$, $x_2 = 3 - 0.5i$.

is not Gaussian and of constant magnitude, additional conceptual difficulties arise, see [45]. The determination of a macroscopic equation to which noise with given moments is added may also present problems when stochastic effects are intrinsic to the process under consideration. For these reasons a Master Equation approach, in which the assumptions of a theoretical model are translated directly into equations for the the evolution of a probability distribution, is often preferable, especially when noise is not due to an external effect such as in Brownian motion.

A Master Equation is a description of a Markov process, and is essentially a balance equation for the flow of probability between system states. Following [45], to construct the Master Equation for a homogeneous Markov process, let $T_\tau(x_2 \mid x_1)$ be the corresponding transition probability from $X = x_1$ to $X = x_2$ in the time interval τ, where X may refer to a set of quantities. The Chapman-Kolmogorov equation applies, and reads:

$$T_{\tau+\tau'}(x_3 \mid x_1) = \int T_{\tau'}(x_3 \mid x_2) T_\tau(x_2 \mid x_1)\, dx_2 \,, \tag{1.8}$$

for all $\tau, \tau' \geqslant 0$. This is closely related to the Markov property, and states that the probability of a transition from x_1 to x_3 arises from all possible intermediate transitions through intermediate values x_2. In the limit of small times, to first order in τ we have:

$$T_\tau(x_2 \mid x_1) = (1 - \alpha\tau)\delta_D(x_2 - x_1) + \tau W(x_2 \mid x_1) \ , \tag{1.9}$$

where δ_D is the Dirac delta function. The second term represents the probability of a single transition (in the considered time interval τ), with $W(x_2 \mid x_1)$ the transition rate (i.e. probability of transition per unit time) from x_1 to x_2; the first term corresponds to the probability of no transitions, with:

$$\alpha = \int W(x_2 \mid x_1)\, dx_2 \ , \tag{1.10}$$

so that $\alpha\tau$ is the probability of any transition. The occurrence of more than one transition is of higher order in τ and is thus omitted. Substituting this form for $T_{\tau'}$ in equation (1.8), dividing by τ' and taking the limit $\tau' \to 0$ we find:

$$\partial_\tau T_\tau(x_3 \mid x_1) = \int \left[W(x_3 \mid x_2)T_\tau(x_2 \mid x_1) - W(x_2 \mid x_3)T_\tau(x_3 \mid x_1)\right]\, dx_2 \ . \tag{1.11}$$

This is the Master Equation for the transition probability, which can be seen to be a differential form of the Chapman-Kolmogorov equation given by (1.8). The first term under the integral accounts for probability flow to $X = x$, and the second term for probability flow away from $X = x$. If we consider that at time t_0 the system state is known to be $X = x_0$, then the master equation may be written in its usual form:

$$\partial_t p(x,t) = \int \left[W(x \mid x')p(x',t) - W(x' \mid x)p(x,t)\right]\, dx' \ , \tag{1.12}$$

where the initial condition $p(x, t_0) = \delta_D(x - x_0)$ is implicitly assumed. In this way, the probability distribution $p(x,t)$ becomes identified with a transition probability from some known initial value.

Finally, the problem of calculating the transition rates $W(x \mid x')$ remains, and of course depends on the particular system. As an illustration, consider again reaction (1.1). Note that, because the Master Equation description is fully stochastic, a continuous approximation is not necessary, and we may work with actual copy numbers n_i. For concreteness consider N species, and let $s_i = 0$ if $i \notin I$ and $r_i = 0$ if $i \notin J$. The stochastic version of mass-action kinetics leads to the transition rate:

$$W_{(m_i)_i|(n_i)_i} = \gamma \prod_{j \in I} \frac{n_j!}{V^{s_j}(n_j - s_j)!} \delta_{(m_i)_i,(n_i + r_i - s_i)_i} \ , \tag{1.13}$$

where the notation $(n_i)_i$ is shorthand for $(n_1, \ldots, n_N)$, and the Kronecker Delta symbol $\delta_{(m_i)_i,(n_i+r_i-s_i)_i}$ accounts for the only possible transition allowed by the specified reaction. The expression

$$\frac{n_j!}{(n_j - s_j)!} = n_j(n_j - 1)\ldots(n_j - s_j + 1) \tag{1.14}$$

replaces $n_j^{s_j}$ so that (1.13) remains valid for small copy numbers. According to (1.12), the Master Equation for this process reads (note that the integral reduces to a sum because of discreteness):

$$\begin{aligned}\dot{p}_{(n_i)_i}(t) = \gamma \Bigg[& \prod_{j\in I} \frac{(n_j - r_j + s_j)!}{V^{s_j}(n_j - r_j)!} p_{(n_i+s_i-r_i)_i}(t) + \\ & - \prod_{j\in I} \frac{n_j!}{V^{s_j}(n_j - s_j)!} p_{(n_i)_i}(t) \Bigg] \ .\end{aligned} \tag{1.15}$$

As in the rate equation description, the terms corresponding to other reactions are simply added to the Master Equation.

The Master Equation formalism thus takes stochastic effects into account intrinsically, and describes the evolution of a probability distribution for the state of the system given information about transition rates, which are determined from physical principles. In the following Chapter we exploit this formalism to build a model of single-gene auto-regulation.

Chapter 2

A Model for Single-Gene Auto-Regulation

In this Chapter we study the cell-level dynamics, and corresponding population distributions, of a single protein capable of transcriptional auto-regulation and its mRNA. The population is assumed to be non-interacting, and regulation to take place through protein dimerization and subsequent binding to the promoter. Since this regulation takes place at the DNA level, we need to model processes at three different levels: the promoter's (DNA), the mRNA's, and the protein's. As is common in nature, the timescale of promoter reactions ($\sim$ seconds) is assumed much shorter than that of the mRNA ($\sim$ minutes to hours) and protein (usually $\sim$ hours), see for example [3, 37] for discussions.

The scheme in Figure 2.1 illustrates the basic structure of the system, and in the rest of this section we build a detailed model for each process.

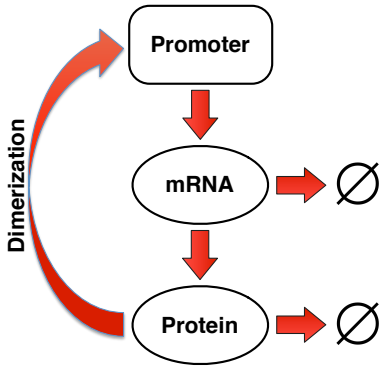

Figure 2.1: Basic structure of the dynamics of a single protein that transcriptionally auto-regulates through dimerization.

2.1 DNA Level

Let us then assume that auto-regulation takes place by dimerization of the protein followed by binding to a single specific promoter site in the DNA molecule. The promoter site is also assumed to bind only one dimer molecule at a time.

For promoter dynamics, we essentially follow [20], adapted to a fully stochastic description. Denote by P_f the free promoter state and by P_b the bound state of the promoter and a dimer. Then this setup can be written in chemical reaction notation as:

$$P_f \underset{k^-}{\overset{k^+}{\rightleftharpoons}} P_b \ . \tag{2.1}$$

Here k^+ and k^- stand for promoter site binding and unbinding rates, respectively. It should be noted, however, that these are not proper chemical reactions, in the sense that the concentration of each promoter certainly cannot be assumed homogeneous in the cell (there is only one promoter!). We use the chemical notation for convenience, and deal with this fact by translating it into an appropriate Master Equation.

Let us then specify the meaning of (2.1). Unbinding dynamics are straightforward: unbinding of a dimer from the promoter occurs at the rate k^- whenever we have the bound state P_b. Binding is somewhat more subtle. Start by assuming a characteristic length l for the promoter, and denote also $V_P \equiv l^3$. If at a certain time there are n_2^P dimer molecules in a volume $\sim V_P$ near the promoter, binding occurs at the rate $k^+ n_2^P$ if the promoter is free.

Denote now, for each instant t, $p(P_f, t)$ as the probability of the promoter being free, and $p(P_b, t)$ as the probability of it being bound to a dimer. If we have n_2^P dimers near the promoter, the above recipe for the evolution of the probability of the bound state is spelled out as the Master Equation:

$$\dot{p}(P_b, t \mid n_2^P) = k^+ n_2^P p(P_f, t \mid n_2^P) - k^- p(P_b, t \mid n_2^P) \ . \tag{2.2}$$

Notice also that at all times the promoter is either free or bound to a dimer. This leads to the conservation equation:

$$p(P_f, t \mid n_2^P) + p(P_b, t \mid n_2^P) = 1 \ . \tag{2.3}$$

To avoid unnecessarily heavy notation, in what follows we assume a given value of protein copy number n in the cell; probabilities should accordingly be taken as conditional probabilities given n. We also use the same symbol for different probability distributions when no confusion is possible.

Let us first find the probability distribution for the number of dimers in volume V_P. Consider two volumes V_1 and V_2 with a total of N particles, which can communicate through diffusion. Recall from standard statistical physics that, in the limit $V_2 \to +\infty$ with finite density N/V_2, the number n of particles in V_1 in equilibrium obeys the Poisson distribution with mean NV_1/V_2. If V is the cell volume, we have $V \gg V_P$. Furthermore, for typical values of protein (and protein dimer) diffusion coefficients and dimerization rates, see [3, 26, 49], it is much more likely for a dimer to wander into the small volume V_P then for a protein to dimerize inside before leaving. If we let $\lambda \equiv V_P/V$, we can thus write (for fixed protein number n in the cell) the equilibrium distribution for the number j of dimers in V_P as:

$$p_j = P_j(\lambda n_2(n)) \,, \tag{2.4}$$

where $P_j(\theta)$ is the Poisson distribution of mean θ (evaluated at j), and $n_2(n)$ is the number of dimers in the cell as a function of protein copy number (in a rate equation description), given by (see Appendix A):

$$n_2(n) = \frac{n}{2} + a^2 - a\sqrt{n + a^2} \,, \tag{2.5}$$

where a is determined by protein dimerization properties. Note that V_P/V is typically very small, since promoters have linear dimensions on the nanometer range, as discussed for example in [3]. In order for auto-regulation to work with a small number of proteins in the cell, as is typical for many transcription factors, active transport is then necessary for the promoter to gauge the actual number of molecules in the cell. However, the previous assumptions leading to the Poisson distribution can be relaxed. Assuming transport does not distinguish between dimers, and that the number of dimers does not influence the transport of a single dimer (essentially, that dimers are independent regarding transport, as is the case for diffusion), the distribution of dimers in V_P is binomial in general, with an "effective rate of volumes" parameter λ. In the relevant limit $\lambda \ll 1$ we regain the Poisson distribution above. Note that a lower bound to the timescale of the transport process is set by diffusion and it is, for typical diffusion coefficients, much faster than the protein timescale.

We will now explicitly take into account that the promoter timescale is much shorter than the protein timescale by assuming that the reactions (2.1) have time to reach equilibrium for each fixed value of the number of proteins. Note that we are interested in the distribution of total protein copy number (regardless of wether it is bound in dimers or free). Note also that if a dimer is bound to the promoter only unbinding can occur and the rate does not depend on the number of free molecules.

Our goal is now to find the equilibrium solutions for the probability distributions $p(P_f, t)$, $p(P_b, t)$, with fixed protein copy number n in the cell. Equation (2.4) allows us to write, in equilibrium:

$$p^{eq}(P) = \sum_{j \geqslant 0} p^{eq}(P \mid j) P_j(\lambda n_2(n)) \ , \tag{2.6}$$

with $P \in \{P_f, P_b\}$. Solving equation (2.2) in equilibrium ($\dot{p} = 0$), and defining the dimensionless parameter $k \equiv k^+/k^-$ leads to:

$$p^{eq}(P_b \mid n_2^P) = k \, n_2^P p^{eq}(P_f \mid n_2^P) \ . \tag{2.7}$$

This gives the probability of finding the promoter bound state as a function of the probability of finding its free state. Substituting in equation (2.3) yields:

$$p^{eq}(P_f \mid n_2^P) = \frac{1}{1 + k n_2^P} \ , \tag{2.8a}$$

$$p^{eq}(P_b \mid n_2^P) = \frac{k n_2^P}{1 + k n_2^P} \ . \tag{2.8b}$$

Substitution into (2.6) finally gives us the desired result, where we now emphasize copy number n in the cell:

$$p^{eq}(P_f \mid n) = \sum_{j \geqslant 0} \frac{1}{1 + kj} P_j(\lambda n_2(n)) \ , \tag{2.9a}$$

$$p^{eq}(P_b \mid n) = \sum_{j \geqslant 0} \frac{kj}{1 + kj} P_j(\lambda n_2(n)) \ . \tag{2.9b}$$

2.2 mRNA and Protein Levels

Because transcription and translation have been found, in many cases, to occur in sharp geometrical bursts (see Chapter 1), we adopt here an approach along the lines of [14] and [38], in which bursts are formulated in a stochastic framework for these two processes.

Let us first discuss mRNA dynamics. Recall that our goal is to study mRNA and protein dynamics and distributions. Owing to the timescale separation between promoter and mRNA/protein dynamics, it is appropriate to write, again in chemical reaction notation:

$$\begin{cases} \varnothing \xrightarrow{\beta_m f(n)} & \mu_m m \ , \\ m \xrightarrow{\delta_m} & \varnothing \ , \\ m \xrightarrow{\beta} & m + \mu p \ , \\ p \xrightarrow{\delta} & \varnothing \ . \end{cases} \tag{2.10}$$

Here m is the mRNA and p is the protein, while n stands for protein copy number. f is the regulation function, such that:

$$\begin{aligned} f(n) &= p^{eq}(P_f \mid n) + \rho\, p^{eq}(P_b \mid n) \ , \\ &= \sum_{j \geqslant 0} \frac{1 + \rho k j}{1 + k j} P_j(\lambda n_2(n)) \ . \end{aligned} \tag{2.11}$$

Thus, β_m is the transcription rate when the promoter is free, and $\rho\beta_m$ is the transcription rate when the promoter is bound to a dimer; the protein exhibits negative auto-regulation (auto-inhibition) if $\rho < 1$, and positive auto-regulation (auto-activation) if $\rho > 1$; μ_m is the mean transcriptional burst size (see below for details). With the burst scenario in mind, the transcription rates above are to be interpreted as the mean rates at which a transcription event takes place; this event is modeled as the instantaneous transcription of a certain number (drawn from a geometric distribution) of mRNA molecules. We assume here that regulation affects only the base transcription rate, and not burst size. Finally, δ_m is the mRNA degradation rate.

Similar definitions stand for the protein parameters (with β the translation rate, interpreted as the rate at which a single mRNA molecule initiates an instantaneous translational burst, μ the mean translational burst size, and δ the protein degradation rate).

It is interesting to see that the timescale separation for promoter dynamics allows all details of regulation to be condensed in the regulation function. Different regulatory dynamics affecting only the transcription rate and obeying the same timescale separation may be modeled in this framework simply by considering a different form of $f(n)$. Note also that a useful approximation to the promoter occupation function as defined by (2.11) exists if $k \ll 1$. If $\lambda n_2(n)$ is small, the low j terms of the sum will dominate; taylor expansion of the denominator to lowest order in kj (for $kj \ll 1$) and explicit calculation of the sum leads to:

$$f(n) \approx \frac{1 + \rho k \lambda n_2(n)}{1 + k \lambda n_2(n)} \ . \tag{2.12}$$

If $\lambda n_2(n)$ is large, the large j terms dominate, and the approximation given by (2.12) remains valid because $(1+\rho k\alpha)/(1+k\alpha) \approx \rho$ for large α. Direct numerical calculation reveals that (2.12) is a good approximation overall, even for moderate values of $k < 1$, see Figure 2.2.

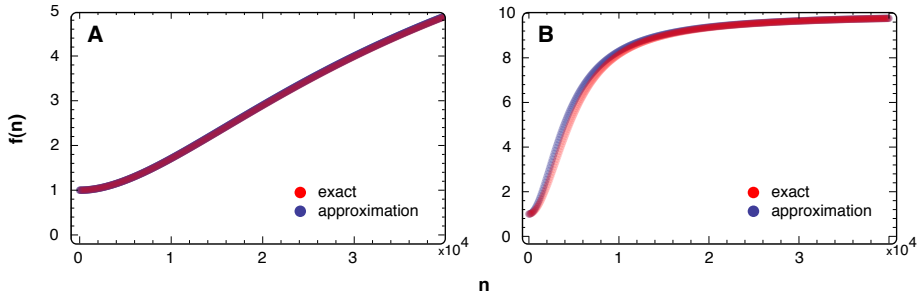

Figure 2.2: Aproximation (2.12) for the regulation function. We fixed $\lambda = 0.01$, $\rho = 10$, $a = 100$ for a typical example.
(**A**) $k = 10^{-2}$; (**B**) $k = 0.5$.

A meaningful interpretation of reactions (2.10) in terms of a Master Equation is straightforward, taking into account the description above. Let:

$$E_i(\theta) \equiv \frac{(\theta-1)^{i-1}}{\theta^i} \; ; \tag{2.13}$$

this is the geometric distribution of mean θ (evaluated at i), conditioned to non-zero values ($i \geqslant 1$) because a burst of zero molecules has no physical meaning.[1] Let also $p_{j,n}(t)$ be the joint probability distribution of protein and mRNA copy numbers (evaluated at mRNA copy number j and protein copy number n) at time t. Then the Master Equation describing the process above reads:

$$\begin{aligned} \dot{p}_{j,n}(t) = \Bigg[& \beta_m f(n) \sum_{i\geqslant 1} E_i(\mu_m)(\mathbb{E}_m^{-i} - 1) + \delta_m(\mathbb{E}_m - 1)j + \\ & + \beta j \sum_{i\geqslant 1} E_i(\mu)(\mathbb{E}^{-i} - 1) + \delta(\mathbb{E} - 1)n \Bigg] p_{j,n}(t) \; , \end{aligned} \tag{2.14}$$

[1] Note that, if $E^0(\theta)$ is the non-conditioned geometrical distribution of mean θ, the relation $E_i(\theta) = \frac{\theta}{\theta-1} E_i^0(\theta - 1)$ holds for all $i \geqslant 1$. Thus, "conditioned bursting" with frequency α and mean θ is equivalent to "non-conditioned bursting" with frequency $\frac{\theta}{\theta-1}\alpha$ and mean $\theta - 1$, and the difference becomes relevant only for $\theta \sim 1$. In this text, all biological burst frequencies and mean sizes refer to the conditioned distribution.

where we have made use of the "step operators" $\mathbb{E}_m$, $\mathbb{E}$ defined by:

$$\begin{aligned} \mathbb{E}_m^i g_{j,n}(t) &= g_{j+i,n}(t) \ , \\ \mathbb{E}^i g_{j,n}(t) &= g_{j,n+i}(t) \ , \end{aligned} \tag{2.15}$$

for any function g depending on mRNA copy number j, protein copy number n, and time t.

2.3 Finding Expression Distributions

Studying the general system described above calls for direct numerical simulations of the dynamics or numerical integration techniques. However, further timescale separations between mRNA and protein dynamics are common. In this section we study the model proposed above for the case of fast mRNA compared to protein dynamics, and vice-versa. We explore both the discrete scenario and a continuous approximation.

2.3.1 Fast mRNA Dynamics

mRNA dynamics

It is convenient in this case to consider fixed protein copy number n, since fast mRNA dynamics should allow mRNA copy number to equilibrate for each fixed protein copy number. This means we are considering the reactions:

$$\begin{cases} \varnothing \xrightarrow{\beta_m f(n)} \mu_m m \ , \\ m \xrightarrow{\delta_m} \varnothing \ , \end{cases} \tag{2.16}$$

at fixed n, with the interpretation provided in Section 2.2. Let $q_{j|n}(t)$ be the distribution of mRNA copy number (evaluated at j) at time t, given n protein molecules in the cell. The Master Equation for this process has the simple form:

$$\dot{q}_{j|n}(t) = \left[\beta_m f(n) \sum_{i \geqslant 1} E_i(\mu_m)(\mathbb{E}_m^{-i} - 1) + \delta_m(\mathbb{E}_m - 1)j \right] q_{j|n}(t) \ . \tag{2.17}$$

Let $q_{|n}^{eq}$ be the equilibrium distribution of mRNA copy number, for each protein copy number n. The mean value of mRNA corresponding to this distribution can be found (see Appendix B), giving:

$$\langle id \rangle_{q^{eq}_{|n}} = \mu_m \gamma_m f(n) \ , \tag{2.18}$$

where $\gamma_m = \beta_m/\delta_m$, and id is the identity function[2].

Protein dynamics

In this scale, we have the reactions:

$$\begin{cases} m & \xrightarrow{\beta} & m + \mu p \ , \\ p & \xrightarrow{\delta} & \varnothing \ . \end{cases} \tag{2.19}$$

Since protein translation has well-defined rates for a certain number of corresponding mRNA molecules, and since we assume the mRNA distribution to quickly reach equilibrium for fixed values of protein, these reactions now have a straightforward interpretation. If we let $p_n(t)$ be the distribution of protein copy number (evaluated at n) at time t, the Master Equation reads:

$$\begin{aligned} \dot{p}_n(t) &= \left[\sum_{j \geqslant 0} \sum_{i \geqslant 1} \beta E_i(\mu) (\mathbb{E}^{-i} - 1) j q^{eq}_{j|n} + \delta(\mathbb{E} - 1) n \right] p_n(t) \ , \\ &= \left[\beta \sum_{i \geqslant 1} E_i(\mu) (\mathbb{E}^{-i} - 1) \langle id \rangle_{q^{eq}_{|n}} + \delta(\mathbb{E} - 1) n \right] p_n(t) \ , \\ &= \left[\beta \mu_m \gamma_m \sum_{i \geqslant 1} E_i(\mu) (\mathbb{E}^{-i} - 1) f(n) + \delta(\mathbb{E} - 1) n \right] p_n(t) \ . \end{aligned} \tag{2.20}$$

We see that, when mRNA is fast, protein dynamics depends at each time only on the average mRNA corresponding to the available protein number n. Specifically, the translation rate becomes proportional to $\langle id \rangle_{q^{eq}_{|n}}$, which is in turn proportional to $f(n)$. Through this mechanism, promoter-level regulation yields a measure of the number of molecules present in the cell at a certain time that is available at the level of translation. Note also that further details of mRNA dynamics, including burst-like production, are lost at the level of protein.

[2]Note that, in physics, if p is the probability distribution for some random variable X, $\langle id \rangle_p$ is usually written $\langle X \rangle$. Here we chose this rather more abstract notation involving the identity function for greater clarity when dealing with multiple variables and distributions throughout the text. In general, for some function f, we write $\langle f \rangle_p$ instead of $\langle f(X) \rangle$.

Let us consider as well a continuous approximation of the dynamics. For this we take $x \equiv \lambda n$ as an "approximately continuous" variable (recall that $\lambda \ll 1$). A continuous Master Equation for the distribution $p(x,t)$ of protein "concentration" x reads (see Appendix C):

$$\begin{aligned}\dot{p}(x,t) =& \beta\mu_m\gamma_m \int_0^x f(x')\left[E(x-x',\tilde{\mu}) - \delta_D(x-x')\right] p(x',t)\, dx' + \\ &+ \delta\, \partial_x \left[xp(x,t)\right] \ .\end{aligned} \tag{2.21}$$

Here we write:

$$E(x,\theta) \equiv (1/\ \theta) e^{-x/\theta} \tag{2.22}$$

for the exponential probability distribution of mean θ evaluated at x, and δ_D is the Dirac Delta. For simplicity we have chosen to keep the symbol f, such that $f(x) = f(n)$ for $x = \lambda n$. The exponential distribution term accounts for the contribution to $p(x)$ due to bursts leading to concentration x, and the Dirac delta term accounts for bursts away from x; $\tilde{\mu}$ is the rescaled burst size, $\tilde{\mu} \equiv \lambda\mu$. The last term is due to protein degradation.

Protein distribution

The equilibrium solution of (2.21) can be found analytically in terms of a primitive (see Appendix D), leading to an explicit continuous approximation for the protein distribution:

$$p^{eq}(x) = A_c\, x^{-1} e^{-x/\tilde{\mu}} e^{\gamma \int_c^x du f(u)/u} \ , \tag{2.23}$$

where $\gamma \equiv \mu_m\gamma_m\beta/\delta$ (note that, for each copy number n, $\gamma f(n)$ is the effective rate of translation burst events, due to all mRNA molecules, scaled by the degradation rate of the protein). The constant A_c is determined by normalization (depending on the arbitrary integration limit c).

If we solve equation (2.20) directly in the discrete setting (see Appendix E), we find the solution:

$$p_n^{eq} = \frac{\gamma\, p_0^{eq}}{n} \prod_{i=1}^{n-1} \left(\gamma \frac{f(i)}{i} + \frac{\mu - 1}{\mu} \right) \ , \tag{2.24}$$

for $n \geqslant 1$, with p_0^{eq} determined by normalization.

The performance of the continuous approximation within its range of validity is exemplified in Figure 2.3. Generically, the continuous approximations presented throughout this section are very accurate for burst sizes of

order 10 and higher. It should be noted, however, that very sharp peaks (with a width of the order of a single molecule) that arise for zero protein or mRNA in some parameter ranges are not well captured, since the continuity approximation breaks down.

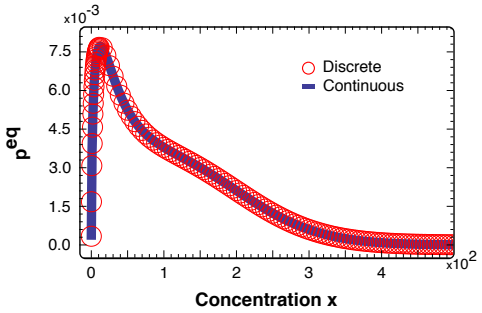

Figure 2.3: Illustration of the the performance of the continuous approximation for protein with fast mRNA. Example parameters are $\gamma = 1.5$, $\mu = 2 \cdot 10^3$, $\rho = 10$, $k = 5 \cdot 10^{-2}$, $a = 10^2$, $\lambda = 10^{-2}$.

The fast mRNA scenario is common in nature, and is particularly suited to transcriptional regulation due to the properties of protein dynamics discussed above (see also [37]). The problem of finding protein distributions in this regime is thus of particular importance; the role of the biological parameters in the qualitative features of the protein distribution is also particularly clear, especially in the continuous setting. To study some of these features, consider the derivative of the probability distribution given by (2.23); concentrations x where probability peaks correspond to $\partial_x p^{eq}(x) = 0$, leading to (see Appendix F):

$$\gamma\tilde{\mu} f(x) = x + \tilde{\mu} \ . \tag{2.25}$$

Let us consider the regulation function as given by the approximation described by (2.12). In the continuous description we write:

$$f(x) \approx \frac{1 + \rho k x_2(x)}{1 + k x_2(x)} \ , \tag{2.26}$$

with:

$$x_2(x) \equiv \lambda n_2(n) = \frac{x}{2} + \tilde{a}^2 - \tilde{a}\sqrt{x + \tilde{a}^2} \tag{2.27}$$

and $\tilde{a} = a/\sqrt{\lambda}$. By noting that equation (2.25) is equivalent to a quartic equation in $z = \sqrt{x + \tilde{a}^2}$, we conclude that it has at most four real solutions, of which only two may correspond to maxima of p^{eq}. Note also that p^{eq} peaks

at zero if and only if $\partial_x p^{eq}(0) < 0$, which yields $\gamma < 1$. However, it is possible to prove that p^{eq} is at most bimodal, see Appendix F.

In the case of negative auto-regulation ($\rho < 1$), p^{eq} is always unimodal because the regulation function is monotonically decreasing. Positive auto-regulation ($\rho > 1$) is necessary for more structured distributions, and bimodal distributions do in fact arise for some parameter sets. It is interesting to note that in the limit of weak dimerization (large $\tilde{a}$) p^{eq} is always unimodal, while in the limit of strong dimerization (small $\tilde{a}$) it is unimodal if $\gamma > 1$ and bimodal with a peak at zero if $\gamma < 1$; bimodal distributions that do not peak at zero are present only for intermediate dimerization (see Figure 2.4). Near parameter regions allowing for bimodality, promoter affinity also strongly affects the shape of p^{eq}, as exemplified in Figure 2.5. The effect of varying γ and ρ is similar, but has a stronger effect on peak positions. Although the burst size parameter $\tilde{\mu}$ also affects the position and relative size of peaks in p^{eq}, its essential role is to produce the heavy tailed distributions commonly observed experimentally.

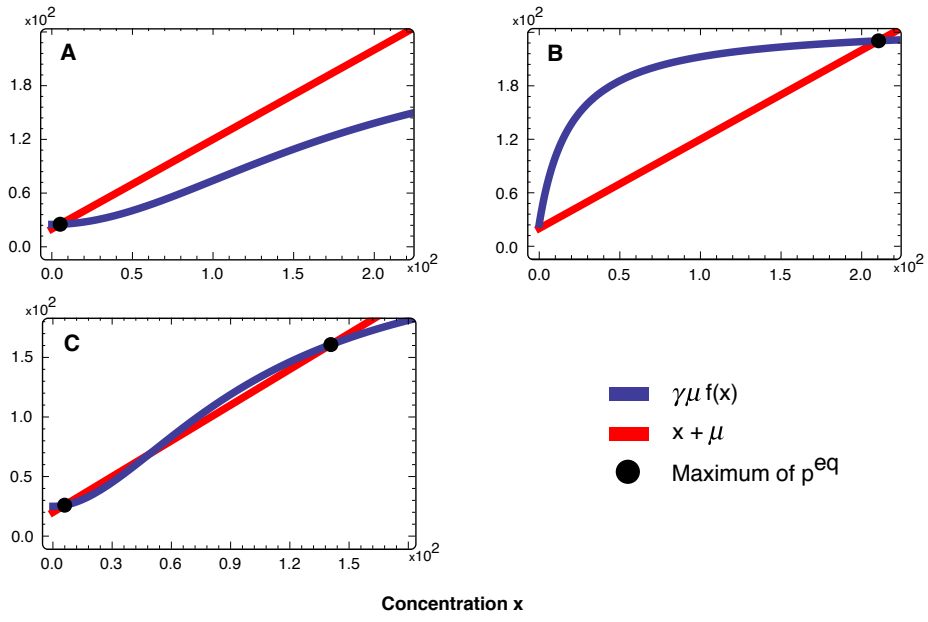

Figure 2.4: Illustration of the effect of varying the dimerization parameter $\tilde{a}$ when bimodality is possible. For low dimerization (**A**), there is only a low-concentration peak. For high dimerization (**B**), there is only a high-concentration peak. Bimodal distributions with both peaks at non-zero concentrations arise only for intermediate dimerization (**C**). Parameters are $\gamma = 1.25$, $\tilde{\mu} = 20$, $\rho = 10$, $k = 10^{-1}$, and: (**A**) $\tilde{a} = 20$; (**B**) $\tilde{a} = 0$; (**C**) $\tilde{a} = 10$.

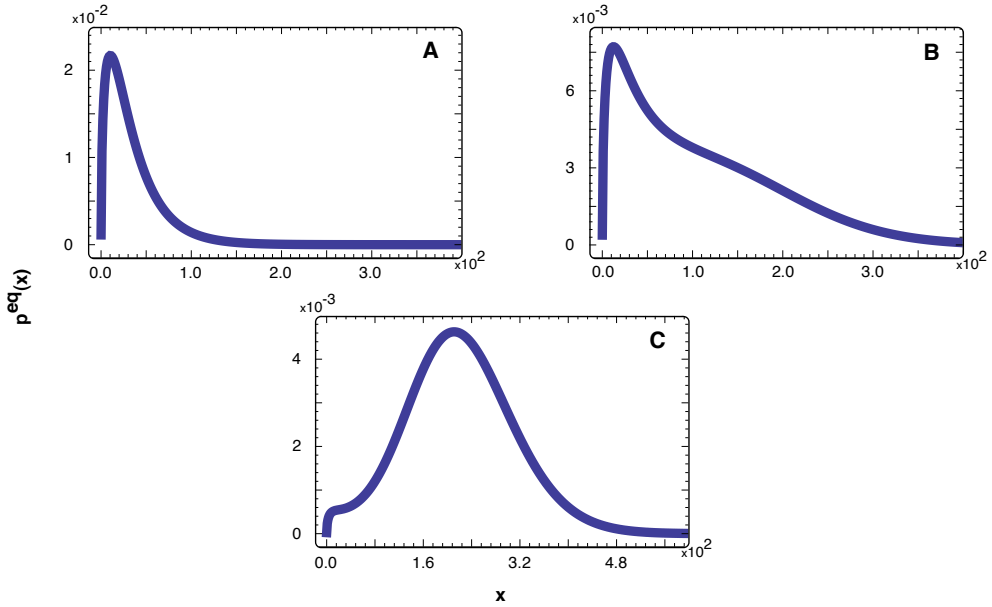

Figure 2.5: Illustration of the effect of varying promoter affinity k when bimodality is possible. Example parameters are $\gamma = 1.5$, $\tilde{\mu} = 20$, $\rho = 10$, $\tilde{a} = 10$, and: (**A**) $k = 10^{-2}$; (**B**) $k = 5 \cdot 10^{-2}$; (**C**) $k = 10^{-1}$.

mRNA distribution

It is now easy to obtain the distribution of mRNA expression. For the continuous approximation, taking into account the Master Equation (2.17), and following the recipe above, we write:

$$\begin{aligned} \dot{q}(z,t \mid x) =& \beta_m f(x) \int_0^z \left[E(z-z',\tilde{\mu}_m) - \delta(z-z')\right] q(z',t \mid x)\, dz' + \\ &+ \delta_m\, \partial_z \left[z q(z,t \mid x)\right] \ . \end{aligned} \tag{2.28}$$

This is an evolution equation for the distribution of a "continuous" mRNA concentration variable $z \equiv \lambda j$, given a fixed protein concentration $x = \lambda n$ (with $\tilde{\mu}_m$ again a rescaled burst size). It is, naturally, of the same type as (2.21). Since f depends on protein but not mRNA concentration, the "no regulation" recipe (see Appendix D) applies, and we find:

$$q^{eq}(z \mid x) = G(z, \gamma_m f(x), \tilde{\mu}_m) \ . \tag{2.29}$$

This is a Gamma distribution (again, see Appendix D for details). To find the equilibrium distribution of mRNA, we take the integral over all values

of protein concentration, weighted by the respective probabilities given by (2.23):

$$\begin{aligned} q^{eq}(z) &= \int_0^\infty q^{eq}(z \mid x) p^{eq}(x)\, dx \ , \\ &= \langle G(z, \gamma_m f, \tilde{\mu}_m) \rangle_{p^{eq}} \ . \end{aligned} \tag{2.30}$$

Similarly, the solution for the discrete dynamics, corresponding to equation (2.17), is (c.f. Appendix E):

$$q^{eq}_{j|n} = N\left(j, \frac{\mu_m}{\mu_m - 1}\gamma_m f(n), \frac{1}{\mu_m}\right) \ . \tag{2.31}$$

This is a Negative Binomial distribution, as defined in Appendix E. The discrete equilibrium distribution for mRNA is found in this case by summing over all protein copy numbers n, weighing with the discrete protein distribution given by (2.24):

$$\begin{aligned} q^{eq}_j &= \sum_{n \geqslant 0} q^{eq}_{j|n} p^{eq}_n \ , \\ &= \left\langle N\left(j, \frac{\mu_m}{\mu_m - 1}\gamma_m f, \frac{1}{\mu_m}\right) \right\rangle_{p^{eq}} \ . \end{aligned} \tag{2.32}$$

An illustration of the continuous approximation within its validity range is shown in Figure 2.6.

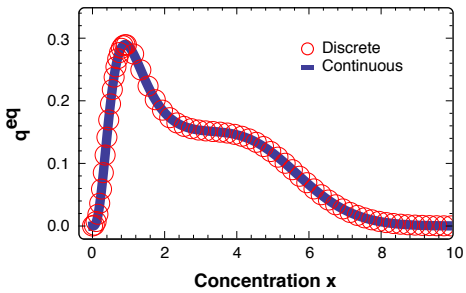

Figure 2.6: Illustration of the the performance of the continuous approximation for mRNA with fast mRNA. The example corresponds to the protein distributions from Figure 2.3, with $\gamma_m = 4$, $\mu_m = 20$.

2.3.2 Fast Protein Dynamics

Protein dynamics

It is now convenient to consider fixed mRNA copy number j, since in this case protein dynamics is much faster and should equilibrate. Let $p_{n|j}(t)$ be the distribution of protein copy number (evaluated at n) at time t, given j mRNA molecules in the cell. We have again reactions (2.19), but in this case we write the Master Equation for fixed mRNA copy number j:

$$\dot{p}_{n|j}(t) = \left[\beta j \sum_{i\geqslant 1} E_i(\mu)(\mathbb{E}^{-i} - 1) + \delta(\mathbb{E} - 1)n\right] p_{n|j}(t) \ . \tag{2.33}$$

In the continuous approximation, we find:

$$\begin{aligned} \dot{p}(x,t \mid z) &= \lambda^{-1}\beta z \int_0^x \left[E(x-x',\tilde{\mu}) - \delta_D(x-x')\right] p(x',t)\, dx' + \\ &\quad + \delta\, \partial_x \left[xp(x,t)\right] \ . \\ &= \frac{\delta\gamma}{\tilde{\mu}_m\gamma_m} z \int_0^x \left[E(x-x',\tilde{\mu}) - \delta_D(x-x')\right] p(x',t)\, dx' + \\ &\quad + \delta\, \partial_x \left[xp(x,t)\right] \ . \end{aligned} \tag{2.34}$$

This equation can be solved for the equilibrium distribution in exactly the same way as equation (2.28), yielding:

$$p^{eq}(x \mid z) = G(x, \tilde{\gamma}_2 z, \tilde{\mu}) \ , \tag{2.35}$$

where $\tilde{\gamma}_2 \equiv \beta/(\lambda\delta) = \gamma/(\tilde{\mu}_m\gamma_m)$. Note that this solution is only valid for $z \neq 0$. If we consider equation (2.34) with $z = 0$, it represents a process where only protein degradation is present (since there is no mRNA). Accordingly, the (normalizable) equilibrium solution is, in the sense of distributions:

$$p^{eq}(x \mid 0) = \delta_D(x) \ . \tag{2.36}$$

Since $G(x, 0, \tilde{\mu})$ is undefined, we may define $G(x, 0, \alpha) \equiv \delta_D(x)$ (for any α), so that (2.35) remains valid.

Similarly, the discrete solution (equation (2.33)) is:

$$p^{eq}_{n|j} = N\left(n, \frac{\mu}{\mu - 1}\gamma_2 j, \frac{1}{\mu}\right) \ , \tag{2.37}$$

where $\gamma_2 \equiv \beta/\delta = \gamma/(\mu_m\gamma_m)$ and $N(n, 0, \alpha) \equiv \delta_{n,0}$, with $\delta_{n,0}$ a Kronecker Delta symbol.

mRNA dynamics

For each protein copy number n, we have again the reactions (2.16). Following arguments similar to those leading to equation (2.20), the Master Equation for mRNA reads in this case:

$$\begin{aligned}\dot{q}_j(t) &= \left[\sum_{n\geqslant 0}\sum_{i\geqslant 1}\beta_m E_i(\mu_m)(\mathbb{E}_m^{-i}-1)f(n)p^{eq}_{n|j} + \delta_m(\mathbb{E}_m-1)j\right] q_j(t)\ , \\ &= \left[\beta_m\sum_{i\geqslant 1}E_i(\mu_m)(\mathbb{E}_m^{-i}-1)\langle f\rangle_{p^{eq}_{|j}} + \delta_m(\mathbb{E}_m-1)j\right] q_j(t)\ .\end{aligned} \tag{2.38}$$

The corresponding continuous Master Equation reads:

$$\begin{aligned}\dot{q}(z,t) = \beta_m\int_0^z \langle f\rangle_{p^{eq}(|z')}\left[E(z-z',\tilde{\mu}_m)-\delta_D(z-z')\right]q(z',t)\,dz' + \\ +\,\delta_m\,\partial_z\left[zq(z,t)\right]\ .\end{aligned} \tag{2.39}$$

mRNA distribution

The equilibrium solution of equation (2.39) can be found through the same method as the one used for equation (2.21), yielding:

$$q^{eq}(z) = A_c z^{-1}e^{-z/\tilde{\mu}}e^{\gamma_m\int_c^z du\langle f\rangle_{p^{eq}(|u)}/u}\ , \tag{2.40}$$

where A_c is a (different) normalization constant, depending on the (again arbitrary) lower integration limit c.

The discrete solution, for equation (2.38), is:

$$q^{eq}_j = \frac{\gamma_m\, q^{eq}_0}{j}\prod_{i=1}^{j-1}\left(\gamma_m\frac{\langle f\rangle_{p^{eq}_{|i}}}{i}+\frac{\mu_m-1}{\mu_m}\right)\ , \tag{2.41}$$

for $j\geqslant 1$, with q^{eq}_0 determined by normalization (note $\langle f\rangle_{p^{eq}_{|0}} = f(0) = 1$).

An illustration of the continuous approximation for favorable parameters, as discussed above, is shown in Figure 2.7.

protein distribution

In the continuous approximation, the distribution of protein concentration follows immediately from the integration of the conditional distribution given by equation (2.35):

$$\begin{aligned} p^{eq}(x) &= \int_0^\infty p^{eq}(x \mid z) q^{eq}(z)\, dz \ , \\ &= \langle G(x, \tilde{\gamma}_2\, id, \tilde{\mu}) \rangle_{q^{eq}} \ . \end{aligned} \tag{2.42}$$

The corresponding discrete distribution is:

$$\begin{aligned} p_n^{eq} &= \sum_{j \geqslant 0} p_{n|j}^{eq} q_j^{eq} \ , \\ &= \left\langle N\left(n, \frac{\mu}{\mu - 1}\gamma_2\, id, \frac{1}{\mu}\right) \right\rangle_{q^{eq}} \ . \end{aligned} \tag{2.43}$$

As expected, in this timescale regime the role of the regulation function is confined to the level of mRNA. The protein distribution depends only on the mRNA distribution, plus translation rate and protein burst size.

The performance of the continuous approximation for favorable parameters is illustrated in Figure 2.8.

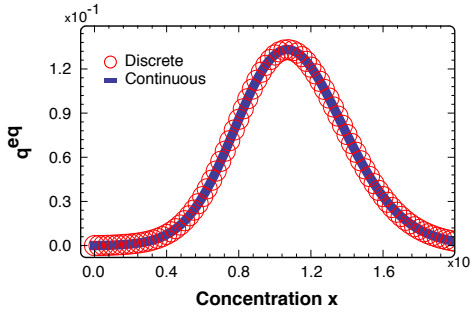

Figure 2.7: Illustration of the the performance of the continuous approximation for mRNA with fast protein. Example parameters are $\gamma = 1.5$, $\mu = 2 \cdot 10^3$, $\rho = 10$, $k = 5 \cdot 10^{-2}$, $a = 10^2$, $\lambda = 10^{-2}$, $\gamma_m = 3$, $\mu_m = 50$.

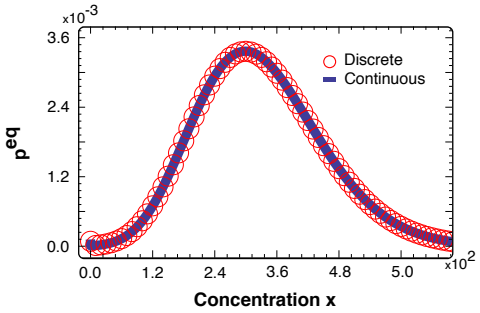

Figure 2.8: Illustration of the the performance of the continuous approximation for protein with fast protein. The example corresponds to the mRNA distributions from Figure 2.7.

Chapter 3

Applying the Model

In this chapter we explore a particular system, namely a population of mouse embryonic stem (ES) cells, cultured under two different conditions. We discuss how the model developed in Chapter 2 can be used to account for the observed heterogeneity and structure of the expression of the protein Nanog at the population level in terms of auto-regulation.

3.1 Biological System

3.1.1 Mouse Embryonic Stem Cells

In mammals, the blastocyst is a structure formed early in the development of the embryo. It consists of the inner cell mass (ICM), which will form the embryo proper, and the trophoblast, which is a layer of cells surrounding the ICM and will form the placenta. In mice, blastocyst formation occurs at approximately 4 to 5 days after fertilization, and the ICM comprises on the order of 100 cells. The cells of the ICM are undifferentiated and pluripotent, as they will eventually give rise to the new organism. They are thus called embryonic stem cells, and are the object of this study.

Here we analyze the population distribution of Nanog protein and mRNA. For a deeper understanding of the underlying dynamics, we study ES cells in two different culture conditions, GMEM and iStem. GMEM is the standard culture medium for ES cells; some inhibitors are added to capture most cells in the undifferentiated state. In iStem, differentiation is inhibited more efficiently by blocking all the genetic pathways known to be responsible for differentiation.

3.1.2 The NOS Network

Three genes have been identified as central to the pluripotent state of ES cells and their ability to regulate differentiation decisions, namely Nanog, Oct4 and Sox2 (NOS). The expression of Nanog itself is related to a cell's decision to differentiate: cells with low Nanog expression are much more likely to exit the pluripotent state and differentiate. As shown by the experimental data of the following section, the population distribution of Nanog is highly structured. Experimental evidence for regulatory effects of Oct4, Sox2 and Nanog itself on Nanog's expression have been reported in the literature, see for example [50]. Because Nanog expression is related to differentiation decisions, see [8], in GMEM culture a much larger fraction of the cell population expresses low Nanog protein levels, while in iStem most cells express high Nanog protein levels, see Figures 3.3-A and 3.4-A. While we do not deal here with the detailed effect of the differentiation signals involved, the large difference between the protein distributions observed in the two culture media provides an opportunity to test our model's ability to describe different system behaviors and discuss possible mechanisms of Nanog regulation.

3.2 Experimental Data

Experimental measurements of Nanog mRNA and protein across the ES cell population were carried out through Fluorescence In Situ Hybridization (FISH), and immunocytochemical staining measured with a Fluorescence-Activated Cell Sorting (FACS) machine, respectively.

With the FISH technique, individual molecules of mRNA can be identified at the single cell level. The resulting experimental measurements consist of histograms of mRNA copy number. The mRNA histograms presented in this Chapter were obtained in this manner. They have been normalized to unit sum and correspond to averages over three experiments, totaling $\sim 10^3$ cells.

The FACS device allows for measurement of different cellular characteristics, as well as physical cell sorting. In particular, the expression of multiple proteins in a cell can be measured through fluorescence, but careful processing of experimental measurements must be carried out to obtain histograms that relate directly to protein number. We turn to a brief discussion of this procedure here.

Because during culture some cells in the population die or differentiate, the bulk population corresponding to living ES cells must be identified. The FACS machine measures the intensity of light that is forward-scattered (FSC) and side-scattered (SSC) by each cell. The first measure correlates with cell

size, and the second with "complexity", or structure, measured in terms of granularity; these morphological measures allow different cellular populations to be identified in a SSC vs FSC plot, and the dead cell population to be excluded. A low FSC population can be identified in our experimental measurements; to ascertain that this corresponds to cellular debris, a fluorescent molecule, propidium iodide (PI), that stains dead cells was added to the population, and its presence in this subpopulation was verified (see Figure 3.1 for a typical illustration).

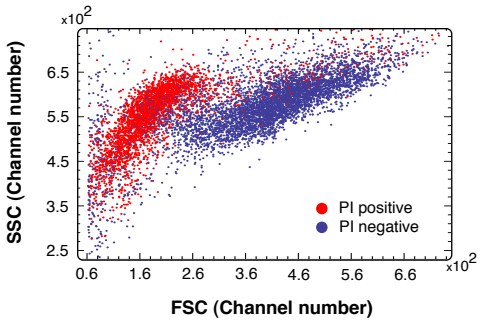

Figure 3.1: Example of Side Scatter (SSC) vs Forward Scatter (FSC) plot for ES cells in GMEM culture. Note the strong overlap between PI positive (dead) cells and the low FSC cluster. FSC- and SSC-provenient light are captured by different channels of the FACS device according to their intensity, as discussed in the text for fluorescence.

To directly measure the expression of Nanog protein, cells were fixed and stained with an appropriate fluorescent antibody, which binds to Nanog protein. Cell fluorescence in the corresponding wavelength range is then measured using the FACS machine. Different fluorescence intensities are detected by different channels of the device, allowing for quantitative measurements of concentration. To permit reading in a wide range of intensities, fluorescence acquisition is usually logarithmic. In terms of measured channel number n, a fluorescence measure linear in the number of proteins present in a cell is given by:

$$f = 10^{nR/N} \,, \tag{3.1}$$

where R and N are characteristic of the FACS machine (respectively the linear range and number of decades).

Finally, so-called cellular auto-fluorescence, due to native proteins that fluoresce in the same wavelength range as the target, must be taken into account. The fluorescence measured at the population level may be described

by a random variable X_t with an appropriate distribution. X_t is then the sum of two random variables, X_N and X_a, corresponding to Nanog fluorescence and auto-fluorescence respectively. Thus, according to Appendix G, the probability distribution of X_t is the convolution of the probability distributions of X_N and X_a. By considering cells without Nanog antibody staining, we obtain a measurement of the distribution of X_a, and extract the distribution of X_N through deconvolution. An example is shown in Figure 3.2.

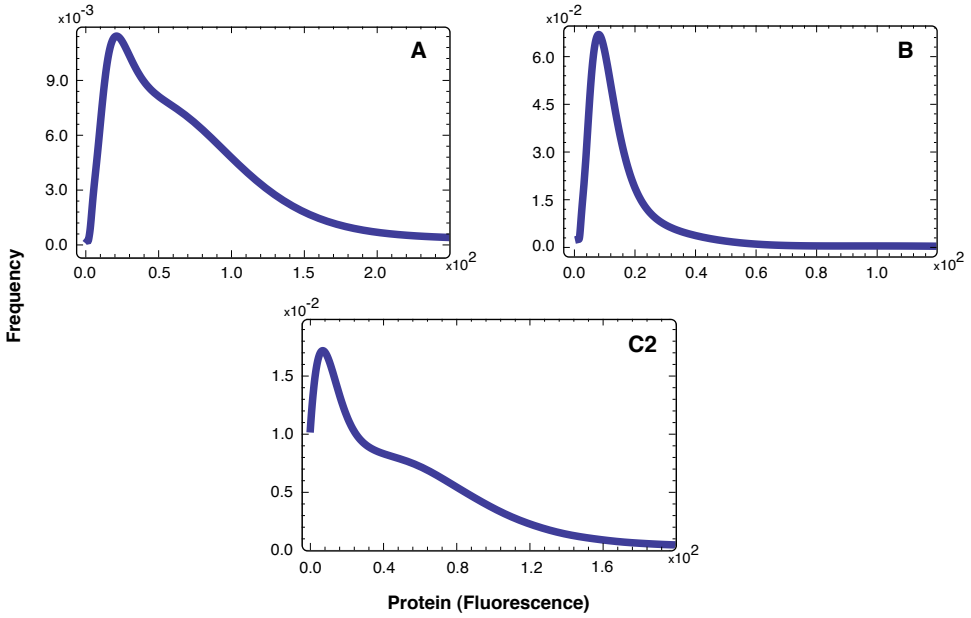

Figure 3.2: Example deconvolution of FACS histogram data (see text for details). The distribution of interest (**C**) is obtained from deconvolution of the autofluorescence data (**B**) from the original data (**A**). Note that smoothing procedures are involved in order to perform deconvolution.

The protein histograms presented in this Chapter were obtained through FACS measurements over $\sim 10^4$ cells and underwent these procedures.

3.3 Modeling the System

The culture media we discussed above allow us to artificially lock the cells in the pluripotent state, and study the corresponding gene expression patterns. Because the cells in these cultures do not communicate with each other, and there is no underlying cell organization, this system is particularly amenable to the study of dynamic gene expression at the single cell level

– since we expect each cell to follow its own program, all cells being independent and identical, population distributions should be a measure of the equilibrium distributions of the corresponding dynamics in a single cell. It has also been shown experimentally that separate culturing of cells with only high or low Nanog expression reestablish the original distribution after some time, see [22], giving evidence for the dynamic origin of the observed Nanog distribution. In fact, this dynamic nature provides a mechanism to regulate the fraction of cells that differentiate without precommitting particular cells.

In the literature, various models for the distribution of Nanog have been proposed, explicitly involving dynamic regulatory effects of Oct4 and Sox2 (see [19, 20, 22]). Although Oct4 and Sox2 are known to be necessary to Nanog expression in ES cells, through binding to a specific site in the Nanog promoter, see [35], it is conceivable that they merely set the proper promoter configuration to allow Nanog transcription to take place. We apply our model for a single auto-regulated gene to test the hypothesis that Nanog auto-regulation, when described in a fully stochastic framework, is enough to account for the observed distributions. We note that, to the best of our knowledge, no study of Nanog expression in ES cells where stochastic effects are formulated in a fully dynamic context exists in the literature. We also point out that the role of Nanog dimers in the regulatory mechanism is speculative; however, it is common for protein dimers to play a role in gene regulation, and the importance of Nanog dimerization for normal ES cell functioning has been confirmed in [48].

3.3.1 Analysis of Population Distributions

In this subsection we analyze the experimental measurements of Nanog protein and mRNA obtained through the techniques discussed above. It should be noted that some quantitative variations are common in the measurements of protein and mRNA distributions pertaining to different experiments (due to small variations of the typical biological parameters, because of cell and media variations), and that quantitative FACS measures and deconvolution techniques have some inherent accuracy limitations. Although our model is in principle amenable to parameter optimization techniques, and good fits exist for specific distributions, intensive parameter fitting to a few experimental results would not say much about the model's generic behavior. We believe that it is more significant that the model is able to reproduce the overall features of the experimental data, including their variability, in terms of biologically reasonable parameters, and that it does so in a parameter region sufficiently large for these features to be found by manual (vs. automatic) inspection of the system's behavior. Therefore, this is the strategy

that we shall adopt here to illustrate the performance of the model.

GMEM

Experimental measurements of Nanog protein in GMEM culture are shown in Figures 3.3-A.[1] While the frequency of zero fluorescence (meaning no protein) is very sensitive to small differences in the data, the presence of a non-zero protein peak is robustly present for all independent cultures measured, and Figures 3.3-A show qualitatively representative data.

Consider also the FISH measurements of Nanog mRNA population distribution in GMEM (see Figure 3.3-C). The histogram clearly shows a pronounced peak corresponding to no mRNA molecules in the population, which translates in our single-cell description into a large fraction of time in which the cell has no mRNA. Intuitively, if Nanog protein had fast dynamics compared to its mRNA, we would expect a pronounced peak corresponding to no protein, because the protein distribution quickly equilibrates to a given mRNA value, and no translation can occur if no mRNA is present. Note that this is independent of the particular form of transcriptional regulation, or the number of regulatory species involved, because these affect the protein distribution only through the mRNA distribution, and not directly. Thus, we assume here that mRNA is fast compared to protein (see subsection 3.3.2 for a full discussion).

In this timescale regime, the analytical solutions for the equilibrium distributions of protein and mRNA are given by (2.24) and (2.32), respectively. Because the low protein peaks in Figures 3.3-A are above but close to zero, we expect γ slightly above 1. On the other hand, the heavy tails suggest large protein bursts, that is, $\mu \gg 1$. For fixed γ and $\tilde{\mu}$, the qualitative behavior of protein distributions is reproduced for different combinations of ρ, k and $\tilde{a}$ (the shape of the protein distributions is set by γ, $\tilde{\mu}$, ρ, k and $\tilde{a}$, and system size, in terms of number of molecules, is then fixed by λ. We adopt this parametrization for this reason). Importantly, positive auto-regulation ($\rho > 1$) is necessary to reproduce the observed features. Note also that the two qualitatively different distributions measured experimentally can easily be explained in terms of the sensitivity of the system to changes in k and γ. For a given protein distribution, an mRNA distribution similar to the experimental data is then easy to obtain for different combinations of $\gamma_m < 1$ and μ_m. Illustrative model results for protein and mRNA distributions are shown in Figures 3.3-B and 3.3-D, respectively.

[1] All experimental and model figures can be found at the end of the Chapter.

iStem

Experimental measurements of Nanog protein and mRNA in iStem culture are shown in Figures 3.4-A and 3.4-C, respectively. The results shown are qualitatively representative. As argued above, the data again support the fast mRNA hypothesis, so the same analytic distributions are used. We shall now seek the general conditions for these analytic solutions to reproduce the observed protein and mRNA distributions in iStem.

A peak is now always found at, or very close to, zero in the protein experimental distributions, suggesting γ slightly smaller than 1. Furthermore, the presence of two close and very well-defined peaks requires a very strongly switch-like behavior of the promoter occupation function. This can be achieved only for very large effective promoter affinity λk. For reasonable system sizes ($\lambda = 10^{-2}$ was used here), this requires that approximately all dimers be available to the promoter;[2] it requires also an increase in the promoter affinity k itself. The experimental mRNA distribution is again sharply peaked at zero, and suitable parameter combinations can be found following a similar strategy as for GMEM. Illustrative model results are shown in Figures 3.4-B and 3.4-D.

3.3.2 Discussion of the Timescale Relation

Experimental measurement of Nanog protein and mRNA lifetimes yielded values consistent with those existing in the literature (respectively ≈ 2 hours and ≈ 4 hours, see [34, 39]), but that are in contradiction with the fast mRNA hypothesis discussed above. This hypothesis is essential to ensure that the output of the model is well described by the analytic solution used in this section to explore it. When the time scales of Nanog protein and mRNA are comparable, as indicated by experiment, the performance of the model must be analyzed by resorting to numerical techniques.

In the experimentally relevant regime, extensive exploration of the remaining relevant parameters through direct numerical simulation (using Gillespie's algorithm, see Appendix H) shows a very robust unimodal behavior with a peak for zero protein. This is exactly what one would expect from the observed mRNA distributions, all of them sharply peaked at zero, if Nanog protein had fast dynamics compared to its mRNA. This can be seen by considering the fast protein approximation discussed in Chapter 2: for the discrete solution, numerical calculation shows that $p_0^{eq} \gg p_1^{eq}$ is robustly satisfied when the observed mRNA distributions are assumed and the re-

[2]Note that, in this case, approximation (2.12) for the regulation function remains valid with $\lambda \approx 1$. This is easily seen from description (2.8) for promoter state probabilities.

maining parameters γ and μ are varied. The biological reason for this sharp zero-peak behavior is that short protein lifetime allows protein expression to reflect the very frequent presence of no mRNA, which corresponds to no protein translation.

We conclude that in the framework of the model proposed here, the observed experimental distributions correspond to typical outputs in the timescale separation regime, but cannot be explained if the Nanog mRNA lifetime is larger than or comparable to the protein lifetime. Furthermore, a modification of the model so as to include, for instance, the intervention of other species in transcriptional regulation or transcription burst-size regulation, would lead to the same conclusion, because this conclusion depends only on how we modeled Nanog mRNA translation.

While it cannot at present be completely excluded that some mechanism, such as the existence of a short-lived, active mRNA population and a long-lived, inactive population, may be masking the effective mRNA lifetime, this seems unlikely in the face of measurements of protein lifetime when transcription is stopped, which yield the same results and thus indicate that the measured lifetime corresponds to a transcriptionally active mRNA population. Therefore, in the light of the available experimental data, we are lead to conclude that the observed Nanog heterogeneity cannot be understood in terms of transcriptional regulation alone, irrespective of the details.

The biological rationale for this conjecture is that transcriptional regulation is not effective in this scenario, because the accumulation of mRNA due to long life prevents the correct dynamic assessment of the number of protein molecules present in the cell at a certain time. In fact, the lack of structure at the level of mRNA, as shown in Figures 3.3-C and 3.4-C, is surprising for long-lived mRNA, and may easily be explained solely by unregulated transcriptional bursting (see Figure 3.5, and below for a derivation of the analytic solution). The strong structure observed for protein distributions must thus arise later, at the level of translation.

Formally, the study of our model and its good agreement with experimental results in the favorable timescale separation regime also support this conjecture. Indeed, in terms of population distribution, the role of the timescale separation between mRNA and protein is to produce a modulation of the translation rate itself. To see this, consider that mRNA production proceeds through bursts without protein regulation. With the standard symbol meanings adopted here, unregulated mRNA production translates into the Master Equation:

$$\dot{q}_j(t) = \left[\beta_m \sum_{i\geqslant 1} E_i(\mu_m)(\mathbb{E}_m^{-i} - 1) + \delta_m(\mathbb{E}_m - 1)j\right] q_j(t) \ . \tag{3.2}$$

The equilibrium solution for an unregulated process of this type, see Appendix E, is a Negative Binomial:

$$q_j^{eq} = N\left(n, \frac{\mu_m}{\mu_m - 1}\gamma_m, \frac{1}{\mu_m}\right) \ . \tag{3.3}$$

If we assume that translation is modulated by a post-transcriptional regulation function $\tilde{f}$ depending on mRNA and protein copy numbers and describing an interaction (direct or indirect) of the protein with its mRNA, the equilibrium protein distribution would be a solution to the probability balance equation:

$$\begin{aligned} 0 &= \left[\gamma \sum_{j\geqslant 0}\sum_{i\geqslant 1} E_i(\mu)(\mathbb{E}^{-i} - 1)\tilde{f}(j,n)q_j^{eq} + (\mathbb{E} - 1)n\right] p_n^{eq} \ , \\ &= \left[\gamma \sum_{i\geqslant 1} E_i(\mu)(\mathbb{E}^{-i} - 1)f(n) + (\mathbb{E} - 1)n\right] p_n^{eq} \ , \end{aligned} \tag{3.4}$$

where $f(n) \equiv \langle \tilde{f}(\,\cdot\,, n)\rangle_{q^{eq}}$. This is the same equation that describes the equilibrium distribution of protein with fast mRNA dynamics, compare to equation (2.20) in equilibrium. Thus, we see that, under appropriate kinetics for translational regulation (leading to an appropriate regulation function f), the protein equilibrium distribution is the same as for fast mRNA, irrespective of mRNA stability. The regulatory mechanism must remain positive, but the possible molecular mechanisms and detailed dynamics leading to the necessary properties are at this point unknown for this system (although the proven role of Nanog dimers for normal ES cell functioning, as mentioned above, suggests they should maintain a central role). However, the main point here is that a stochastic model with a Nanog post-transcriptional positive auto-regulation loop would also robustly reproduce the observed Nanog levels distributions and their variability.

The idea that the regulatory roles of Nanog are mainly post-transcriptional rather than transcriptional is strongly supported by [30], where the artificial over-expression of Nanog is shown to have very little effect on the transcriptome. In addition, recent works suggest that post-transcriptional regulation may have a central role for many genes, and different molecular mechanisms are being discussed in the literature (see for example [17, 21]). Post-transcriptional regulation as an effective alternative to transcriptional

regulation for proteins with stable mRNAs was also previously discussed in [37]. In the post-transcriptional framework, the model's description of the observed distributions will remain essentially the same as described in Section 3.3.1, since the mathematical description is equivalent. The biological interpretation and exact shape of the regulation function, however, would have to be re-examined in the light of new experiments aimed at verifying the existence of post-transcriptional regulation and clarifying its molecular details.

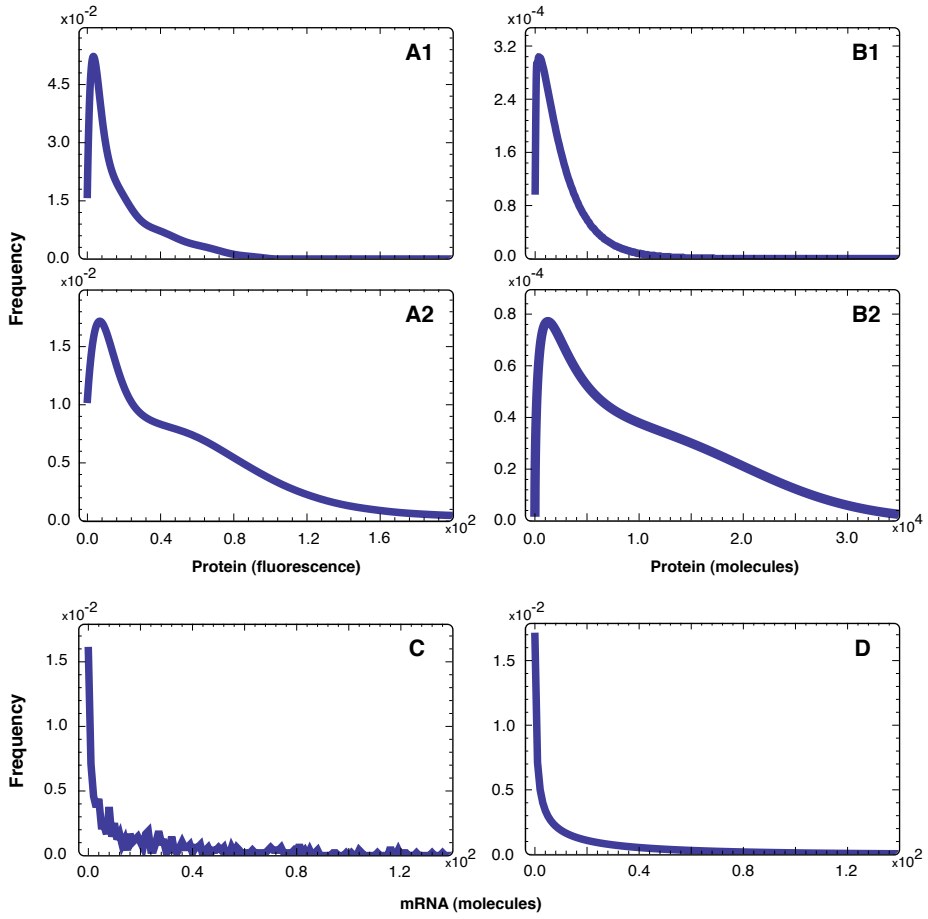

Figure 3.3: Experimental data and model results for Nanog protein and mRNA in GMEM culture. Figures on the left column are experimental data, and adjacent are the corresponding model results (see main text). For the illustrative model results we fixed $\lambda = 10^{-2}$, $\tilde{\mu} = 20$, $\rho = 10$, $\tilde{a} = 10$, and used:
(**B1**) $\gamma = 1.2$, $k = 10^{-2}$.
(**B2**) $\gamma = 1.5$, $k = 5 \cdot 10^{-2}$.
(**D**) Same as (**B1**), plus $\gamma_m = 0.4$, $\mu_m = 60$.

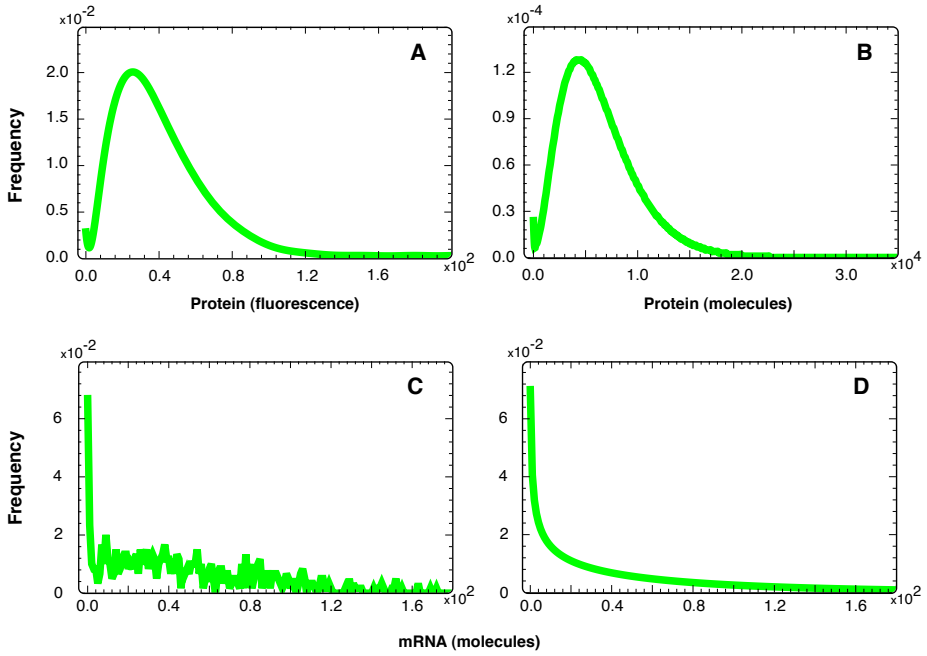

Figure 3.4: Experimental data and model results for Nanog protein and mRNA in iStem culture. Figures on the left column are experimental data, and adjacent are the corresponding model results (see main text). For the illustrative model results, parameters are $\lambda = 10^{-2}$, $\tilde{\mu} = 20$ and $\tilde{a} = 10$, plus:
(**B**) $\gamma = 0.8$, $\rho = 4$, $k = 50$.
(**D**) Same as (**B**), plus $\gamma_m = 0.15$, $\mu_m = 10^2$.

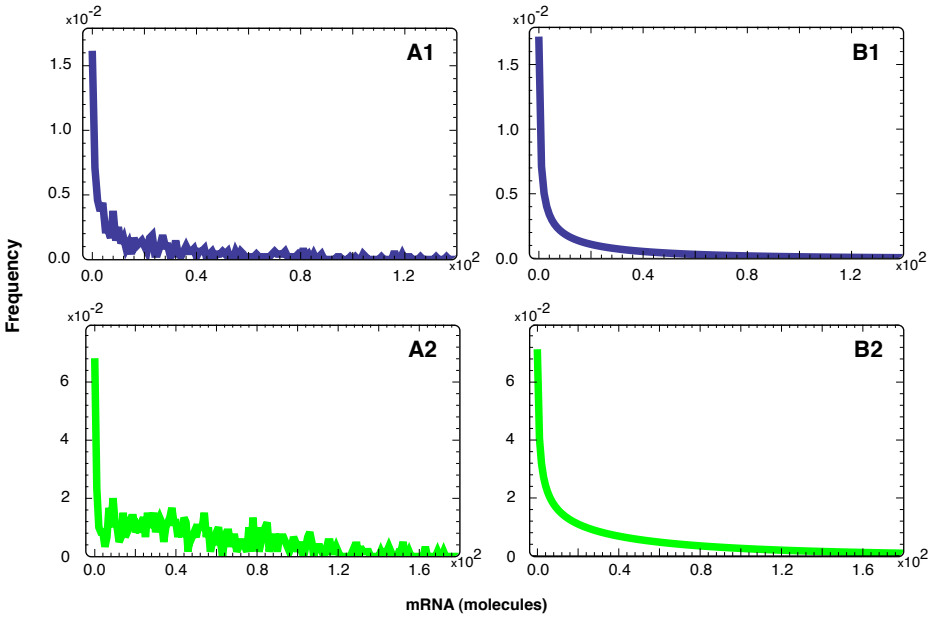

Figure 3.5: If translation is unregulated, the corresponding equilibrium distribution is given by (3.3). The experimental data for mRNA in GMEM (**A1**) and iStem (**A2**) is easily explained for different parameters in this framework. For this example we took:
(**B1**) GMEM: $\gamma_m = 0.4$, $\mu_m = 10^2$.
(**B2**) iStem: $\gamma_m = 0.5$, $\mu_m = 2.5 \cdot 10^2$.

Chapter 4

Final Remarks

In this thesis we have presented and discussed the construction of a model for gene regulation based on solid physical principles. We discussed its general formulation and assumptions, and presented analytical descriptions in important biological limits. The analysis of a concrete biological system led us to a comprehensive exploration of the model in terms of its biological parameters, using both analytic and numeric approaches. We emphasize here the importance of mathematical modeling, based on clear principles and assumptions, for the understanding of the mechanisms involved in biological systems. In particular, it is important to use a framework appropriate to the problem at hand. A deterministic description of cellular processes where noise is added by hand, as opposed to dynamically formulated, often obscures essential biological aspects that are natural in a truly stochastic approach.

The study of experimental data on Nanog expression in ES cells in light of our theoretical model precipitated a number of important discoveries and considerations that, as far as we can tell, have not been discussed in the literature. We have seen that the observed Nanog mRNA and protein distributions in GMEM and in iStem correspond to typical outputs of our model when Nanog mRNA lifetime is much shorter than that of the protein. We have also seen that any model that assumes that translation is simply described, as common, by a burst process occurring at a constant rate followed by exponential protein degradation will be unable to reproduce the observed Nanog mRNA and protein distributions when their lifetimes are comparable, as indicated by experiment. Thus, we have found no evidence that regulatory activity of other genes is implied by the observed Nanog heterogeneity. We discussed Nanog translational auto-regulation as an alternative that is consistent with current experimental data. Further experiments are necessary to ascertain the validity and details of this proposal.

As a closing remark, I believe a note on my personal experience regarding

the work presented in this thesis is in order. In the past two years I had the opportunity to work in an interdisciplinary project comprising a physics team and a biology team. This collaboration allowed me to work on the development of a theoretical model in close contact with experimental data, and this has certainly informed, shaped and matured my approach to physics. The need to relate theoretical descriptions to experimental findings promotes the creation of models that provide clear descriptions, based on physically meaningful quantities. It is my belief that the fundamental role of the physicist in an interdisciplinary collaboration (and indeed in any scientific project) is to provide a fresh point of view, that is on the one hand rooted in firm mathematical formalism, and on the other based on clear physical assumptions and providing new insights and intuitions regarding the system under study.

To finish, I believe the major achievement of the project that led to this thesis was the strong interplay between the biology and physics teams, between theoretical model-building and prediction and experiment design and interpretation. This interplay led to original and relevant scientific discoveries on a major and active subject, as well as strong personal and scientific enrichment for all involved.

Appendices

Appendix A

Dimer Dynamics

Consider a cell of volume V where there are n copies of some molecular species that can be characterized by a dimerization rate k_d^+ (dimensions volume.time^{-1}) and an undimerization rate k_d^- (dimensions time^{-1}). Our goal here is to find the explicit form of $n_2(n)$, the number of dimers as a function of (fixed) total copy number n. The equations governing dimerization dynamics of this species at fixed total density $\phi \equiv n/V$ are:

$$\begin{cases} \dot{\phi_1} = k_d^+ \phi_f^2 - k_d^- \phi_2 \ , & \text{(A.1a)} \\ \phi = \phi_f + 2\phi_2 \ . & \text{(A.1b)} \end{cases}$$

Equation (A.1a) is the rate equation for temporal dynamics, and the conservation equation (A.1b) reflects that molecules are either free ($\phi_f \equiv n_f/V$) or bound in pairs as dimers ($\phi_2 \equiv n_2/V$).

Defining $k_d \equiv k_d^+/k_d^-$, equation (A.1a) yields in equilibrium:

$$\phi_2 = k_d \phi_f^2 \ . \tag{A.2}$$

Using equation (A.1b) for ϕ_f leads, in terms of copy number, to the desired result:

$$n_2(n) = \frac{n}{2} + a^2 - a\sqrt{n + a^2} \ , \tag{A.3}$$

where a is a dimensionless parameter defined by $a \equiv \sqrt{V/(8k_d)}$.

It is also interesting to note that there are two limits in which (A.3) becomes very simple and intuitive. One the one hand, if $a^2 \ll n$, we find:

$$n_2(n) \approx \frac{n}{2} \ . \tag{A.4}$$

In physical terms, this can be understood as follows: for a certain density n/V, if k_d is high enough most proteins will bind in dimers; conversely, for

a certain k_d, if density is high enough most proteins will again be bound because of increased collision probability. On the other hand, if $a^2 \gg n$, we are in the opposite limit where most proteins will be free. Taylor expansion of the square root leads in lowest order to:

$$n_2(n) \approx \frac{k_d}{V} n^2 \ . \tag{A.5}$$

This result can also be found by setting $\phi_f \approx \phi$ in (A.2).

Appendix B

Mean mRNA in Equilibrium (Fast mRNA)

Consider the mRNA Master Equation (2.17). Multiplying both sides by j and summing over j we find an equation for the mean:

$$\begin{aligned}\partial_t \langle id \rangle_{q_{|n}(t)} = \Bigg[& \beta_m f(n) \sum_{i \geqslant 1} E_i(\mu_m) \sum_{j \geqslant 0} j(\mathbb{E}_m^{-i} - 1) + \\ & + \delta_m \sum_{j \geqslant 0} j(\mathbb{E}_m - 1) j \Bigg] q_{j|n}(t) \ .\end{aligned} \tag{B.1}$$

Let us compute (omitting the arguments t, n for simplicity):

$$\begin{aligned}\sum_{i \geqslant 1} E_i(\mu_m) \sum_{j \geqslant 0} j(\mathbb{E}_m^{-i} - 1) q_j &= \sum_{i \geqslant 1} E_i(\mu_m) \sum_{j \geqslant 0} j(q_{j-i} - q_j) \ , \\ &= \sum_{i \geqslant 1} E_i(\mu_m) \sum_{j \geqslant -i} (j+i) q_j + \\ &\quad - \sum_{i \geqslant 1} E_i(\mu_m) \sum_{j \geqslant 0} j q_j \ , \\ &= \sum_{i \geqslant 1} i E_i(\mu_m) \sum_{j \geqslant 0} q_j \ , \\ &= \mu_m \ .\end{aligned} \tag{B.2}$$

In the penultimate step we have made use of the fact that $q_j = 0$ whenever copy number j is negative.

Now let us look at:

$$\begin{aligned}\sum_{j\geqslant 0} j(\mathbb{E}_m - 1)jq_j \ , &= \sum_{j\geqslant 0} j(j+1)q_{j+1} - \sum_{j\geqslant 0} j^2 q_j \ , \\ &= \sum_{j\geqslant 1} (j-1)jq_j - \sum_{j\geqslant 0} j^2 q_j \ , \\ &= -\sum_{j\geqslant 0} jq_j \ , \\ &= -\langle id \rangle_{q_{|n}(t)} \ .\end{aligned} \tag{B.3}$$

Since we are looking for the equilibrium mean we now set the left-hand side of (B.1) to zero, and using results (B.2) and (B.3) we find the desired result:

$$\langle id \rangle_{q^{eq}_{|n}} = \mu_m \gamma_m f(n) \ . \tag{B.4}$$

Appendix C

Continuous Approximation

Here we study a continuous approximation for equations of the form:

$$\dot{p}_n(t) = \left[\delta\gamma \sum_{i\geqslant 1} E_i(\mu)(\mathbb{E}^{-i} - 1)f(n) + \delta(\mathbb{E} - 1)n\right] p_n(t) \ , \tag{C.1}$$

where f is some function of (protein or mRNA) copy number n, $\gamma \neq 0$ and $\delta \neq 0$ are constants, and the step operator raises n. For some time t, let copy number n be fixed, and let $x = \lambda n$ be the corresponding concentration. In accordance with the main text (see Chapter 2), the convention $f(n) = f(x)$ will be used. First, note that a reasonable definition for the continuous distribution obeys:

$$\begin{aligned} p_n(t) \equiv p(x,t)\lambda\left[(n+1/2) - (n-1/2)\right] &\approx \int_{n-1/2}^{n+1/2} p(x,t)dx \\ &= \lambda p(x,t). \end{aligned} \tag{C.2}$$

Now consider the conditioned geometric distribution. We have:

$$\begin{aligned} E_n(\mu) &= \frac{(\mu-1)^{n-1}}{\mu^n} \ , \\ &= \frac{1}{\mu-1} e^{n[\log(\mu-1) - \log(\mu)]} \ , \\ &= \frac{1}{\mu-1} e^{-n\log(1-1/\mu)} \ . \end{aligned} \tag{C.3}$$

If we take $\mu \gg 1$ (which is biologically common, especially for proteins, see for example [23, 42]) and expand $\log(1 - 1/\mu)$ around $1/\mu = 0$ we find to lowest order:

$$\begin{aligned} E_n(\mu) &\approx \frac{1}{\mu} e^{-n/\mu} \ , \\ &= \lambda \frac{1}{\tilde{\mu}} e^{-x/\tilde{\mu}} \ , \\ &= \lambda E(x, \tilde{\mu}), \end{aligned} \tag{C.4}$$

with $\tilde{\mu} = \lambda\mu$.

Now notice that, apart from constant coefficients, the creation term in equation (C.1) may be written:

$$\begin{aligned} \sum_{i \geqslant 1} E_i(\mu)(\mathbb{E}^{-i} - 1) f(n) p_n(t) &= \sum_{i \geqslant 1} E_i(\mu) f(n-i) p_{n-i} - f(n) p_n(t) \ , \\ &= \sum_{i=0}^{n} \left(E_{n-i}(\mu) - \delta_{n,i} \right) f(i) p_i \ , \end{aligned} \tag{C.5}$$

where $\delta_{n,i}$ is a Kronecker Delta symbol. Note that the upper limit of the sum can be extended to infinity by taking $E_j(\mu) = 0$ for $j \leqslant 0$, and the lower limit can be extended to negative infinity since $p_i = 0$ for $i < 0$.

The Kronecker Delta term reads:

$$\begin{aligned} \sum_{i=0}^{n} \delta_{n,i} f(i) p_i &= f(n) p_n(t) \ , \\ &= \lambda \int_0^x \delta_D(x - x') f(x') p(x', t) \, dx' \ , \end{aligned} \tag{C.6}$$

where δ_D is the Dirac Delta. Notice that, for a meaningful conversion to the continuous case, the lower limit of the integral must be strictly included (in order to encompass the contribution of the Delta function). Thus, the upper and lower limits of the integral may be extended to infinity.

For the conditioned geometric distribution term in (C.5) we may write:

$$\begin{aligned} \sum_{i=0}^{n} E_{n-i}(\mu) f(i) p_i &\approx \sum_{i=0}^{n} \lambda E(\lambda(n-i), \tilde{\mu}) f(i) \lambda p(\lambda i, t) \ , \\ &\approx \lambda \int_0^x E(x - x', \tilde{\mu}) f(x') p(x', t) \, dx' \ , \end{aligned} \tag{C.7}$$

where again the upper and lower limits of the integral may be extended do plus and minus infinity by considering, respectively, $E(y, \tilde{\mu}) = 0$ and $p(y, t) = 0$ for negative y. Here, the approximations $\mu \gg 1$ (approximating the conditioned geometric distribution with an exponential distribution) and

$\lambda \ll 1$ (approximating the sum with an integral, i.e. considering x continuous) have been explicitly used.

Finally, the degradation term in equation (C.1) reads, apart from a factor of δ:

$$\begin{aligned}(\mathbb{E}-1)np_n(t) &= \left[(n+1)p_{n+1}(t) - np_n(t)\right] \ , \\ &= \frac{1}{\lambda}\left[(x+\lambda)\lambda p(x+\lambda)(t) - x\lambda p(x,t)\right] \ , \\ &\approx \lambda\partial_x(xp(x,t)) \ ,\end{aligned} \tag{C.8}$$

where we again make use of $\lambda \ll 1$ to approximate a finite difference with a derivative. Noting that $\dot{p}_n(t) = \lambda\dot{p}(x,t)$ and collecting terms we find:

$$\begin{aligned}\dot{p}(x,t) = &\gamma\int_0^x f(x')\left[E(x-x',\tilde{\mu}) - \delta_D(x-x')\right]p(x',t)\,dx' + \\ &+ \delta\,\partial_x\left[xp(x,t)\right] \ .\end{aligned} \tag{C.9}$$

Appendix D

Continuous Equilibrium Distributions

Here we follow [14] to obtain an analytical solution to equation (C.9). As discussed in Appendix C, the upper and lower integration limits may be extended to plus and minus infinity, respectively. Thus, defining:

$$w(x\,,\tilde{\mu}) = E(x,\tilde{\mu}) - \delta_D(x)\;, \tag{D.1}$$

we may write:

$$\dot{p}(x,t) = \delta\gamma(w(\tilde{\mu}) * fp)(x,t) + \delta\,\partial_x\left[xp(x,t)\right]\;, \tag{D.2}$$

where $*$ is a convolution product. In equilibrium we have:

$$-\,\partial_x\left[xp^{eq}(x)\right] = \gamma(w(\tilde{\mu}) * fp^{eq})(x)\;. \tag{D.3}$$

Laplace transformation of this equation leads to:

$$\begin{aligned} s\partial_s\hat{p}(s) &= \gamma\,\hat{w}(s)\,\mathcal{L}(fp^{eq})(s)\;,\\ &= \gamma\,\hat{w}(s)(\hat{f} * \hat{p})(s)\;,\\ &= -\gamma\frac{s}{s+1/\tilde{\mu}}(\hat{f} * \hat{p})(s)\;. \end{aligned} \tag{D.4}$$

Here, $\hat{g}(s) = \mathcal{L}(g)(s) = \int_0^{+\infty} e^{-sx}g(x)\,dx$ is the Laplace transform of function g (evaluated at s), and $\hat{p} = \mathcal{L}(p^{eq})$. Note that the integration limit 0 is strictly included when dealing with Dirac Delta functions. Convolution theorems have been used in the first and second lines, and in the third line the explicit form of $\hat{w}(s)$ was substituted. Rearranging terms we have:

$$(s+1/\tilde{\mu})\partial_s\hat{p}(s) = -\gamma(\hat{f} * \hat{p})(s)\;, \tag{D.5}$$

which inverse-transforms to:

$$\partial_x[xp^{eq}(x)] = (\gamma f(x)/x - 1/\tilde{\mu})\, xp^{eq}(x) \; . \tag{D.6}$$

This equation can easily be solved, leading to:

$$p^{eq}(x) = A_c\, x^{-1} e^{-x/\tilde{\mu}} e^{\gamma \int_c^x du f(u)/u} \; . \tag{D.7}$$

The constant A_c is determined by normalization (depending on the arbitrary integration limit c).

Consider now the case $f(x) = 1$, for all x. Solving the integral in (2.23) and normalizing the probability distribution to integral unity we find:

$$\begin{aligned} p^{eq}(x) &= \frac{x^{\gamma-1} e^{-x/\tilde{\mu}}}{\tilde{\mu}^{\gamma}\Gamma(\gamma)} \; , \\ &= G(x, \gamma, \tilde{\mu}) \; . \end{aligned} \tag{D.8}$$

This is the Gamma distribution of parameters γ and $\tilde{\mu}$ (Γ is the Euler Gamma function). It's interesting to see that, with $\gamma = \mu_m \gamma_m \beta/\delta$ and $\tilde{\mu}$ the mean rescaled protein burst size (with definitions according to Chapter 2 of the main text), this is the equilibrium solution for unregulated protein dynamics with fast mRNA.

Appendix E

Discrete Equilibrium Distributions

In this Appendix we analyze, directly in the discrete setting, equation (C.1). Analogously to the continuous case, the discrete Master Equation may be written:

$$\dot{p}_n(t) = \delta\gamma(w(\mu) * fp)(n,t) + \delta\left[(n+1)p_{n+1}(t) - np_n(t)\right] \ , \tag{E.1}$$

where $*$ is now the discrete convolution product, and:

$$w(n\,,\mu) = E_n(\mu) - \delta_{n,0} \ . \tag{E.2}$$

We now follow the procedures of Appendix D using the Z transform instead of the Laplace transform, $\hat{g}(s) = \mathcal{Z}(g)(s) = \sum_{n=0}^{+\infty} s^{-n} g(n)$, $\mathcal{Z}(p^{eq}) = \hat{p}$. The corresponding equation in "momentum space" is:

$$s(s-1)\partial_s\hat{p}(s) + \frac{s}{\mu}\partial_s\hat{p}(s) = -\gamma\left(\hat{f} * \hat{p}\right)(s) \ . \tag{E.3}$$

Inverse-transforming, we get:

$$(n+1)p_{n+1}^{eq} + (1/\mu - 1)np_n^{eq} = \gamma f(n)p_n^{eq} \ , \tag{E.4}$$

leading to the recurrence relation:

$$\begin{cases} p_1^{eq} = \gamma f(0)p_0^{eq}, \\ (n+1)p_{n+1} = \left(\gamma\frac{f(n)}{n} + \frac{\mu-1}{\mu}\right)np_n^{eq}, \qquad n \geqslant 1. \end{cases} \tag{E.5}$$

This is easily solved, yielding:

$$p_n^{eq} = \frac{\gamma f(0) p_0^{eq}}{n} \prod_{i=1}^{n-1} \left(\gamma \frac{f(i)}{i} + \frac{\mu - 1}{\mu} \right) , \tag{E.6}$$

for all $n \geqslant 1$, with p_0^{eq} determined by normalization (and the standard convention that the product equals one when the upper limit is smaller than the lower). Note that if f is a regulation function as per Chapter 2 of the main text we have $f(0) = 1$, since the promoter is necessarily free when no protein is present.

Consider now the case $f(n) = 1$ for all n. Write (E.6) as:

$$\begin{aligned} p_n^{eq} &= \frac{\mu}{\mu - 1} \frac{\gamma f(0) p_0^{eq}}{n} \left(\frac{\mu - 1}{\mu} \right)^n \prod_{i=1}^{n-1} \left(\frac{\mu}{\mu - 1} \gamma \frac{f(i)}{i} + 1 \right) , \\ &= \frac{\gamma' f(0) p_0^{eq}}{n} \left(\frac{\mu - 1}{\mu} \right)^n \prod_{i=1}^{n-1} \left(\gamma' \frac{f(i)}{i} + 1 \right) , \end{aligned} \tag{E.7}$$

with $\gamma' = \gamma\mu/(\mu - 1)$. The product can be solved explicitly in terms of Gamma functions, and normalizing to unit sum we find:

$$\begin{aligned} p_n^{eq} &= \frac{1}{\mu^{\gamma'}} \left(\frac{\mu - 1}{\mu} \right)^n \frac{\Gamma(n + \gamma')}{\Gamma(\gamma')\Gamma(n + 1)} , \\ &= N\left(n, \gamma', \frac{1}{\mu} \right) . \end{aligned} \tag{E.8}$$

This is the Negative Binomial distribution of parameters γ' and $1/\mu$. The parameters are defined such that:

$$N(n, k, p) = p^k (1 - p)^n \binom{n + k - 1}{k - 1} . \tag{E.9}$$

As in the continuous case (Appendix D), with $\gamma = \mu_m \gamma_m \beta/\delta$ and μ the mean protein burst size (definitions according to Chapter 2 of the main text), this is the discrete solution for unregulated protein dynamics with fast mRNA (as reported for example in [38]).

Appendix F

Protein Multimodality in Equilibrium (Fast mRNA)

Consider the continuous equilibrium distribution for protein with fast mRNA, given by (2.23). The derivative of this probability distribution is given by:

$$\partial_x p^{eq}(x) = \left[\gamma\tilde{\mu}f(x) - (x+\tilde{\mu})\right]\frac{p^{eq}(x)}{\tilde{\mu}x} \; . \tag{F.1}$$

If p^{eq} peaks at zero (i.e. if $\partial_x p^{eq}(0) < 0$), the term in brackets in equation (F.1) must be negative at zero. Because $p^{eq}(x) > 0$ for all $x > 0$, other maxima of p^{eq} must satisfy:

$$\gamma\tilde{\mu}f(x) - (x+\tilde{\mu}) = 0 \; . \tag{F.2}$$

Consider $f(x)$ as given by (2.26). As mentioned in the main text (see Chapter 2), a change of variables to $z = \sqrt{x+\tilde{a}^2}$ in equation (F.2) leads to an equivalent quartic equation, $P_4(z) = 0$. Direct algebra shows that, without changing the sign of the left hand side of equation (F.2), the quartic polynomial $P_4(z)$ may be written:

$$P_4(z) = -z^4 + 2\tilde{a}z^3 + \alpha_2 z^2 + \alpha_1 z + \alpha_0 \; , \tag{F.3}$$

where the α_i are real constants determined by the biological parameters.

We now proceed to prove that p^{eq} is at most bimodal. Since zeros of P_4 correspond alternately to maxima and minima of p^{eq}, the presence of more than two maxima requires the presence of a peak for $x = 0$, which is characterized by $P_4(\tilde{a}) < 0$ (note that $x = 0$ corresponds to $z = \tilde{a}$, and $x > 0$ to $z > \tilde{a}$). Furthermore, because $P_4(z) \to -\infty$ when $z \to -\infty$, this condition implies also that trimodality requires P_4 to have four roots for $z > \tilde{a}$. A necessary condition is then that P_4'' has two real roots for $z > \tilde{a}$

(the prime denotes derivation). The equation $P_4''(z) = 0$ is quadratic in z and thus is easily shown to have the two solutions:

$$z = \frac{\tilde{a}}{2} \pm \sqrt{\left(\frac{\tilde{a}}{2}\right)^2 + \frac{\alpha_2}{6}} \,. \tag{F.4}$$

If they are real, one of these solutions necessarily obeys $z < \tilde{a}$. Therefore, trimodality does not arise and p^{eq} is at most bimodal.

Appendix G

Sum of Independent Random Variables

Consider two independent random variables X and Y, with probability distributions p_X and p_Y, respectively. We are interested in the distribution of the sum, p_{X+Y}. For concreteness, let X and Y be continuous random variables, defined in the range $]-\infty, \infty[= \mathbb{R}$.

Note now that, for each $x, z \in \mathbb{R}$, if $X = x$ then $X + Y = z$ if and only if $Y = z - x$. Thus, we may write:

$$p_{X+Y}(z) = \int_{\mathbb{R}} p_X(x \mid Y = z - x)\, dx \ , \tag{G.1}$$

where $p_X(x \mid Y = z - x)$ is the probability that $X = x$ given that $Y = z - x$, and we integrate over every possible value of x. Due to independence of X and Y, we immediately find:

$$\begin{aligned} p_{X+Y}(z) &= \int_{\mathbb{R}} p_X(x) p_Y(z - x)\, dx \ , \\ &= (p_X * p_Y)(z) \ , \end{aligned} \tag{G.2}$$

where $*$ is a convolution product. Thus, the distribution of the sum of two independent random variables is given by the convolution of the distributions of the individual variables.

Finally, this proof holds for discrete random variables defined over the range $\mathbb{Z}$ (of all integers) by substituting the integrals by sums (the convolution becomes a discrete convolution). In both the continuous and discrete cases, if the range is smaller than, respectively, $\mathbb{R}$ or $\mathbb{Z}$, the proof still holds if we consider extensions of the probability distributions that are null outside each variables' definition range.

Appendix H

The Gillespie Algorithm

Here we give a brief description of the rationale behind the Gillespie algorithm (see [18] for the original paper). Consider a system composed of a certain number of components, or "species", each present in a certain copy number at each time. Let the "system state" at some time t be defined as the list of copy numbers of each species. Consider now that at each time one of J state-changing "reactions" can occur. Furthermore, each reaction $j \in \{1, \ldots, J\}$ occurs at a rate r_j that is well-defined given the system state.

If we consider the system at time t, the next reaction will occur at some time $t+\tau$. Note that in the time interval $]t, t+\tau[$ the system state is constant, and consequently so are the rates r_j. Thus, the probability distribution for the next-reaction time τ is the probability distribution for the minimum of the independent variables τ_j, for all $j \in \{1, \ldots, J\}$, such that:

$$p_{\tau_j}(\tau) = r_j e^{-r_j \tau} \ . \tag{H.1}$$

The crucial idea is that, so long as the system state remains fixed, these probability distributions are the same as for shot noise. Since each p_{τ_j} is an exponential distribution, the minimum has the distribution:

$$p(\tau) = r_0 e^{-r_0 \tau} \ , \tag{H.2}$$

where $r_0 = \sum_{i=1}^{J} r_i$. Finally, the probability that the next reaction will be j must be proportional to r_j. We conclude that for each j this probability is given by:

$$p_j = \frac{r_j}{r_0} \ . \tag{H.3}$$

With these results in mind, the algorithm to simulate the system's evolution from time t_0 to time t_{max} proceeds as follows:

1. Set the initial system state at time t_0;
2. Compute reaction rates;
3. Decide the next reaction: generate a (uniformly distributed) random number r in $]0, 1[$; choose reaction j such that j is the smallest integer satisfying $\sum_{i=1}^{j} r_i > r_0\, r$;
4. Decide the time until next reaction: generate a random number τ according to the p_τ distribution;
5. Update system state according to reaction j and add τ to time. Stop if time $> t_{max}$, else go to step 2.

References

[1] H.C. Berg and E.M. Purcell. Physics of chemoreception. *Biophysical Journal*, 20(2):193–219, 1977.

[2] O.G. Berg. A model for the statistical fluctuations of protein numbers in a microbial population. *Journal of Theoretical Biology*, 71(4):587–603, 1978.

[3] W. Bialek and S. Setayeshgar. Physical limits to biochemical signaling. *Proceedings of the National Academy of Sciences of the United States of America*, 102(29):10040, 2005.

[4] W. Bialek and S. Setayeshgar. Cooperativity, sensitivity, and noise in biochemical signaling. *Physical Review Letters*, 100(25):258101, 2008.

[5] D. Bratsun, D. Volfson, L.S. Tsimring, and J. Hasty. Delay-induced stochastic oscillations in gene regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(41):14593, 2005.

[6] T. Çagatay, M. Turcotte, M.B. Elowitz, J. Garcia-Ojalvo, and G.M. Süel. Architecture-dependent noise discriminates functionally analogous differentiation circuits. *Cell*, 139(3):512–522, 2009.

[7] L. Cai, N. Friedman, and X.S. Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–362, 2006.

[8] I. Chambers, J. Silva, D. Colby, J. Nichols, B. Nijmeijer, M. Robertson, J. Vrana, K. Jones, L. Grotewold, and A. Smith. Nanog safeguards pluripotency and mediates germline development. *Nature*, 450(7173):1230–1234, 2007.

[9] O. Crauk and N. Dostatni. Bicoid determines sharp and precise target gene expression in the Drosophila embryo. *Current Biology*, 15(21):1888–1898, 2005.

[10] V. Elgart, T. Jia, A.T. Fenley, and R. Kulkarni. Connecting protein and mRNA burst distributions for stochastic models of gene expression. *Physical Biology*, 8:046001, 2011.

[11] M.B. Elowitz, A.J. Levine, E.D. Siggia, and P.S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183, 2002.

[12] H.W. Engl, C. Flamm, P. Kügler, J. Lu, S. Müller, and P. Schuster. Inverse problems in systems biology. *Inverse Problems*, 25:123014, 2009.

[13] H.B. Fraser, A.E. Hirsh, G. Giaever, J. Kumm, and M.B. Eisen. Noise minimization in eukaryotic gene expression. *PLoS Biology*, 2(6):e137, 2004.

[14] N. Friedman, L. Cai, and X.S. Xie. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Physical Review Letters*, 97(16):168302, 2006.

[15] T. Galla. Intrinsic fluctuations in stochastic delay systems: Theoretical description and application to a simple model of gene regulation. *Physical Review E*, 80(2):021909, 2009.

[16] A. García-Bellido. Genetic control of wing disc development in Drosophila. *Ciba Foundation symposium*, 0(29):161–182, 1975.

[17] F. Gebauer and M.W. Hentze. Molecular mechanisms of translational control. *Nature Reviews Molecular Cell Biology*, 5(10):827–835, 2004.

[18] D.T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115:1716, 2001.

[19] I. Glauche, M. Herberg, and I. Roeder. Nanog variability and pluripotency regulation of embryonic stem cells - insights from a mathematical model analysis. *PloS One*, 5(6):e11238, 2010.

[20] R. Guantes and J.F. Poyatos. Multistable decision switches for flexible control of epigenetic differentiation. *PLoS Computational Biology*, 4(11):e1000235, 2008.

[21] T. Jia and R.V. Kulkarni. Post-transcriptional regulation of noise in protein distributions during gene expression. *Physical Review Letters*, 105(1):18101, 2010.

[22] T. Kalmar, C. Lim, P. Hayward, S. Muñoz-Descalzo, J. Nichols, J. Garcia-Ojalvo, and A. Martinez Arias. Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biology*, 7(7):e1000149, 2009.

[23] B.B. Kaufmann and A. van Oudenaarden. Stochastic gene expression: from single molecules to the proteome. *Current Opinion in Genetics & Development*, 17(2):107–112, 2007.

[24] D.S. Lemons and P. Langevin. *An introduction to stochastic processes in physics*. Johns Hopkins University Press, 2002.

[25] J. Lewis. Autoinhibition with transcriptional delay: a simple mechanism for the zebrafish somitogenesis oscillator. *Current Biology*, 13(16):1398–1408, 2003.

[26] M.K. Liu, P. Li, and J.C. Giddings. Rapid protein separation and diffusion coefficient measurement by frit inlet flow field-flow fractionation. *Protein Science: A Publication of the Protein Society*, 2(9):1520, 1993.

[27] H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell. *Molecular cell biology*. Wiley Online Library, 1995.

[28] R. Losick and C. Desplan. Stochasticity and cell fate. *Science*, 320(5872):65, 2008.

[29] J.D. Meiss. *Differential dynamical systems*. Society for Industrial and Applied Mathematics, 2007.

[30] A. Nishiyama, L. Xin, A.A. Sharov, M. Thomas, G. Mowrer, E. Meyers, Y. Piao, S. Mehta, S. Yee, Y. Nakatake, et al. Uncovering early response of gene regulatory networks in ESCs by systematic induction of transcription factors. *Cell stem cell*, 5(4):420–433, 2009.

[31] E.M. Ozbudak, M. Thattai, I. Kurtser, A.D. Grossman, and A. van Oudenaarden. Regulation of noise in the expression of a single gene. *Nature Genetics*, 31(1):69–73, 2002.

[32] J. Paulsson. Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157–175, 2005.

[33] J.M. Pedraza and J. Paulsson. Effects of molecular memory and bursting on fluctuations in gene expression. *Science*, 319(5861):339, 2008.

[34] S. Ramakrishna, B. Suresh, K.H. Lim, B.H. Cha, S.H. Lee, K.S. Kim, and K.H. Baek. Pest motif sequence regulating human nanog for proteasomal degradation. *Stem Cells and Development*, -Not available-, ahead of print. doi:10.1089/scd.2010.0410.

[35] D.J. Rodda, J.L. Chew, L.H. Lim, Y.H. Loh, B. Wang, H.H. Ng, and P. Robson. Transcriptional regulation of nanog by OCT4 and SOX2. *Journal of Biological Chemistry*, 280(26):24731, 2005.

[36] P. Rué and J. Garcia-Ojalvo. Gene circuit designs for noisy excitable dynamics. *Mathematical Biosciences*, 2011.

[37] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, 2011.

[38] V. Shahrezaei and P.S. Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 105(45):17256, 2008.

[39] L.V. Sharova, A.A. Sharov, T. Nedorezov, Y. Piao, N. Shaik, and M.S.H. Ko. Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Research*, 16(1):45, 2009.

[40] N. Sonenberg and A.G. Hinnebusch. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, 136(4):731–745, 2009.

[41] W.S. Sutton. The chromosomes in heredity. *The Biological Bulletin*, 4(5):231, 1903.

[42] Y. Taniguchi, P.J. Choi, G.W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X.S. Xie. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533, 2010.

[43] T.L. To and N. Maheshri. Noise can induce bimodality in positive transcriptional feedback loops without bistability. *Science*, 327(5969):1142, 2010.

[44] A.M. Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37, 1952.

[45] N.G. van Kampen. *Stochastic processes in physics and chemistry*. North Holland, 1992.

[46] C.H. Waddington. *The strategy of the genes; a discussion of some aspects of theoretical biology*. London: George Allen & Unwin, Ltd., 1957.

[47] A.M. Walczak, A. Mugler, and C.H. Wiggins. Analytic methods for modeling stochastic regulatory networks. *Arxiv preprint arXiv:1005.2648*, 2010.

[48] J. Wang, D.N. Levasseur, and S.H. Orkin. Requirement of Nanog dimerization for stem cell self-renewal and pluripotency. *Proceedings of the National Academy of Sciences of the United States of America*, 105(17):6326, 2008.

[49] D. Xu, C.J. Tsai, and R. Nussinov. Mechanism and evolution of protein dimerization. *Protein Science: A Publication of the Protein Society*, 7(3):533, 1998.

[50] A. Yates and I. Chambers. The homeodomain protein Nanog and pluripotency in mouse embryonic stem cells. *Biochemical Society Transactions*, 33(6):1518–1521, 2005.