Département de Mathématiques
Université de Fribourg

Stochastic models of a self regulated gene network

Thèse

présentée à la Faculté des Sciences de l'Université de Fribourg (Suisse)
pour l'obtention du grade de *Doctor scientiarum mathematicarum*

Thomas Fournier

de

Nendaz (Valais)

Thèse n° 1620
Mécanographie, Université de Fribourg
2008

Acceptée par la Faculté des Sciences de l'Université de Fribourg (Suisse) sur la proposition du jury :

Prof. Ruth KELLERHALS, Université de Fribourg, présidente du jury,
Prof. Christian MAZZA, Université de Fribourg, directeur de thèse,
Prof. Stephan MORGENTHALER, Ecole Polytechnique Fédérale de Lausanne,
Prof. Jean-Pierre GABRIEL, Université de Fribourg.

Fribourg, le 21 novembre 2008.

Le Directeur de thèse :

.....

Prof. Christian MAZZA

Le Doyen :

.....

Prof. Titus JENNY

CONTENTS

RÉSUMÉ

Les modèles stochastiques jouent un rôle déterminant dans la compréhension des mécanismes génétiques conduisant à la production de protéines. De part la complexité de ces systèmes, très peu de méthodes exactes sont connues et l'on se voit contraint d'approcher les valeurs d'équilibre par simulation stochastique.

Dans cette thèse, après avoir brièvement introduit les méthodes stochastiques usuelles pour modéliser les réactions biochimiques, on étudie en profondeur un module génétique composé d'un promoteur et d'un gène reponsable pour la synthèse d'une protéine. L'intérêt principal de ce modèle réside dans le fait que le promoteur est autorégulé, soit positivement, soit négativement. Ce module génétique a été largement étudié dans la littérature, voir par exemple Kepler & Elston (2001). La première approche par un modèle Markovien remonte à Peccoud & Ycart (1995), qui donnent un résultat exact dans le cas le plus simple où le promoteur n'est pas régulé, alors que Hornos *et al.* (2005) calcule la mesure invariante dans le cas d'une rétroaction négative linéaire.

On développe ici une méthode générale pour calculer la mesure invariante du processus, applicable à toute sorte de rétroactions positives ou négatives, avec comme seule restriction l'hypothèse d'un espace d'états fini. Cette méthode n'étant pas directement adaptée au calcul concret, on donne un algorithme de renormalisation qui le rend possible, et on explore plus particulièrement le cas où le module est autocatalytique, c'est-à-dire positivement autorégulé. On montre qu'il existe des régimes de paramètres dans lesquels la distribution invariante est bimodale. Cette propriété est en accord avec certains résultats expérimentaux, mais ne peut pas être prédite par des approches plus simples connues dans la littérature.

On adapte finalement cette approche pour un réseau génétique plus compliqué étudié expérimentalement par Imhof *et al.* (2000). Ce réseau est composé de trois gènes, un répresseur et deux modules similaires à celui décrit précédemment, avec un mécanisme d'autorégulation contrôlé par un élément extérieur. On montre qu'avec notre modèle, on peut reproduire le comportement qualitatif du système, et que la régulation autocatalytique peut produire un commutateur génétique puissant.

ABSTRACT

Stochastic models play a crucial role in the understanding of genetic mechanisms leading to protein synthesis. Due to the complexity of such systems, very few exact methods are known and one has to rely on stochastic simulation to estimate the equilibrium values.

In this thesis, after briefly describing the usual stochastic methods for biochemical reactions, we study in-depth a simple genetic module comprising a promoter and a gene responsible for protein synthesis. The main interest of this model is the possibility for the promoter to be self-regulated, either positively or negatively. This genetic module has been widely studied in the literature, see for example Kepler & Elston (2001). The first Markovian approach to the model goes back to Peccoud & Ycart (1995), where the authors provide an exact result in the simplest case where the promoter is not regulated, while Hornos *et al.* (2005) compute the invariant measure in the case of a negative linear feedback.

We develop here a general method to compute the invariant measure of the process, applicable to all kind of regulation, whether positive or negative, with the sole restriction that the state space has to be finite. This method is not directly adapted to concrete computation, hence we give a renormalization algorithm allowing it, and explore in more details the case of an autocatalytic module, i. e. positively self-regulated. We show that in some parameter regime, the invariant distribution is bimodal. This feature is in agreement with some experimental results, but can not be predicted using simpler approaches known in the literature.

We finally adapt this approach for a more complicated gene network studied experimentally by Imhof *et al.* (2000). This network comprises three genes, a repressor and two modules similar to the one described above, with an autoregulation mechanism controled by an external element. We show that our model can reproduce qualitatively the behaviour of the system, and that autocatalytic regulation can produce a potent genetic toggle switch.

Let's play the music, not the background.
Ornette Coleman

Part I

INTRODUCTION

1

HISTORICAL BACKGROUND

In the description of chemical reactions, the theory of Markov processes was first introduced by Delbruck (1940) to model the irreversible first order chemical reaction $\mathcal{A} \to \mathcal{B}$. For this reaction, the stochastic mean value agree with the solution of the deterministic system based on the law of mass action. This is not the case for higher order chemical reactions, the deterministic solution does not account for variations. Informally, one can explain this difference by the fact that for similar modelization, the stochastic equation involves moments of order k where the deterministic one has the mean to the power k. A broad discussion is given in Darvey *et al.* (1966) for second order reactions, including the dimerization process $2\mathcal{A} \rightleftharpoons \mathcal{B}$ we will discuss later on. One of the goals of these approaches was to point out the fact that in the limit where the number of each reactant molecules is large, the usual laws of chemistry are valid, with the usual square root standard error.

With the emergence of genetics, systems with a very low number of each reactant molecules became popular. Several experiments showed that the role of stochasticity in these systems is crucial, see for example Kærn *et al.* (2005) for an interesting review on the subject. Due to the complexity of such systems, very few among them could be solved exactly. In the seventies, Gillespie took advantage of the growing computing capacity and proposed in Gillespie (1976) an easy to implement stochastic simulation algorithm that allows to tackle harder problems where closed formulas are unattainable. This algorithm, today known as the Gillespie algorithm, is simply the simulation of the regular jump Markov process that solves the Kolmogorov forward equation given by the transition rates of the chemical reaction, also called chemical master equation.

In recent years, advances in biotechnology has allowed the construction of synthetic gene network, enabling to experiment in vivo simple gene networks and compare them with qualitative and quantitative models. The basic gene network modeling transcriptional regulation consists of an upstream regulatory DNA site, called operator, a nucleotide sequence to which RNA polymerase bind to begin transcription, called promoter, and an activator. A more detailed description of the biology of this gene network is given in Appendix A. The Markov process describing the system is the bivariate process $(N(t), Y(t))_{t\geq 0}$, where $N(t)$ is an integer representing the number of gene product monomer and $Y(t)$ is a binary variable corresponding to the two possible state of the operator, 1 for ON or 0 for OFF, depending on whether the activator is bound to the promoter or not. The stochastic model was presented in Peccoud & Ycart (1995) and compared

to in vivo experiments in Guido *et al.* (2006). A particularly interesting special case of this gene network occurs when the transitions between ON and OFF states are governed by positive or negative feedback, that is the gene product is itself an activating or repressing regulatory protein for its gene.

Most real networks are much more complicated and often a large number of chemical species are involved in the reaction. Nevertheless, Gillespie's algorithm can still be easily implemented but it is difficult to decide how long the algorithm should be run to reach equilibrium, and thus simulation can be extremely time consuming. Some acceleration techniques has been investigated in different settings, as for example the $\tau-$leaping method introduced in Gillespie (2001), the multiscale algorithm presented in Cao *et al.* (2005b) or the coarse-grained equation-free approach to multiscale computation proposed in Erban *et al.* (2006). Since chemical reactions rates are not known precisely, a screening of the possible parameter values can be helpful, and would imply to begin the whole simulation anew for each set of rates.

In his book, p. 168, Wilkinson (2006) claims the following:

> "*It is important to bear in mind that although analytic analysis of simple processes is intellectually attractive and can sometimes give insight into more complex problems, the class of models where analytic approaches are possible is very restricted and does not cover any models of serious interest in the context of systems biology (where we are typically interested in the complex interactions between several intricate mechanisms). Therefore, computationally intensive study based on stochastic simulation and analysis is the only realistic way to gain insight into system dynamics in general.*"

One of the aim of this thesis is to invalidate this sentence, at least partially, by providing an exact method for analytical computation in the case of the simple genetic module described above and adapting the technique to a more complicated gene network. The method allows for fast computation instead of simulation, parameter screening and possibly simple statistical estimation through parameter fitting. Despite its simplicity, this genetic module is of wide use in bioengineering and has been studied extensively, see for example Kepler & Elston (2001), and has proven to be a *model of serious interest in the context of systems biology*. In some very particular cases, an exact formula was already known in the literature, as for the case without feedback in Peccoud & Ycart (1995) or the case of negative linear feedback in Hornos *et al.* (2005). We provide a closed formula for arbitrary negative or positive feedbacks.

2

ORGANISATION

In Part II of this work, we briefly present the stochastic modeling of chemical reactions, from the usual derivation of the chemical master equation and the implementation of the Gillespie algorithm to more sophisticated equations for the moments and generating functions.

In Part III, we focus on the basic gene network described above. To emphasize the method of generating functions and its limitations, we describe two simple cases in depth, one whithout feedback and one with linear positive feedback. With the restriction that the state space has to be finite, we give a closed and exact formula for the invariant measure for arbitrary feedbacks based on a mathematical technique adapted from Bolthausen & Goldsheid (2000) and similar to the method presented in Fournier *et al.* (2007). Self-regulation is usually modelled through feedbacks that are polynomial functions of the number of monomers, dimers, trimers, quadrimers or higher order polymers for a given gene product. We give a detailed discussion of the case of positive feedback through dimers in the limit when the time scale of the dimerization is much more rapid than the other reactions. We further introduce time delay in a semi-stochastic setting and discuss some condition ensuring convergence of the process.

In Part IV, we adapt this modelization to a more complicated self regulated gene network that was engineered and experimented in Imhof *et al.* (2000) and investigate various situations. This gene network consists of a repressor, an activator and a transgene and is designed to act as a potent genetic switch. Regulated by an external factor, the doxycycline, known as a safe drug with a long history of use in humans, the network is designed to act as a potent genetic switch, completely silent in the absence of doxycycline and able to reach rapidly its maximal production with the adjunction of the harmless antibiotic.

Part V deals with the dimerization process. We present an alternative way to compute the moments of the number of dimers based on recurrences leading to a continuous fraction. The method is especially suited for the fast dimerization discussed in Part III and IV.

Part II

MATHEMATICAL MODELS OF CHEMICAL REACTIONS

3

CHEMICAL REACTIONS AND DETERMINISTIC MODEL

3.1 CHEMICAL REACTIONS

Mathematical models in biology divide roughly in two categories, the deterministic and the stochastic models. In this section we briefly outline the deterministic and stochastic models describing chemical reactions as well as the notations we will use throughout the thesis and emphasize it by mean of a simple example. The reference mathematical book for exact stochastic methods in natural sciences stays Gardiner (1983), although new topics emerged in genetics since then and can be found in some more recent books as Wilkinson (2006), which focus mainly on stochastic simulation. In their review article, El Samad *et al.* (2005) also provide a good introduction to the topic and a concise description of the usual modern notation. Another book of interest is Bromberg & Dill (2002) who describe the mathematical models commonly used in biochemistry.

A simple reversible reaction between the chemical species $\mathcal{S}_1$ and $\mathcal{S}_2$ where a molecule of $\mathcal{S}_1$ binds to two molecules of species $\mathcal{S}_2$ to form a species $\mathcal{S}_3$ is written as

$$\mathcal{S}_1 + 2\,\mathcal{S}_2 \underset{c_2}{\overset{c_1}{\rightleftharpoons}} \mathcal{S}_3,$$

where c_1 and c_2 are the chemical rate constants corresponding to the binding of $\mathcal{S}_1$ and $2\,\mathcal{S}_2$ respectively the splitting of $\mathcal{S}_3$. Both deterministic and stochastic approaches rely on the law of mass action.

3.2 DETERMINISTIC MODEL

In the deterministic setting, if $\phi_{\mathcal{P}}$ denotes the concentration of species $\mathcal{P}$ in the solution, the evolution of the three species is given by

$$\begin{aligned}
\frac{d\phi_{\mathcal{S}_1}(t)}{dt} &= c_2\,\phi_{\mathcal{S}_3}(t) - c_1\,\phi_{\mathcal{S}_1}(t)\,(\phi_{\mathcal{S}_2}(t))^2 \\
\frac{d\phi_{\mathcal{S}_2}(t)}{dt} &= 2\,c_2\,\phi_{\mathcal{S}_3}(t) - 2\,c_1\,\phi_{\mathcal{S}_1}(t)\,(\phi_{\mathcal{S}_2}(t))^2 \\
\frac{d\phi_{\mathcal{S}_3}(t)}{dt} &= c_1\,\phi_{\mathcal{S}_1}(t)\,(\phi_{\mathcal{S}_2}(t))^2 - c_2\,\phi_{\mathcal{S}_3}(t).
\end{aligned}$$

Summing these equations, one can notice the conservation condition

$$\phi_{\mathcal{S}_1}(t) + \phi_{\mathcal{S}_2}(t) + 3\,\phi_{\mathcal{S}_3}(t) = \text{ constant.}$$

The equilibrium constant is $K = \frac{c_1}{c_2} = \frac{\phi_{\mathcal{S}_3}^{(eq)}}{\phi_{\mathcal{S}_1}^{(eq)}(\phi_{\mathcal{S}_2}^{(eq)})^2}$, where $\phi_{\mathcal{P}}^{(eq)}$ denotes the concentration of species $\mathcal{P}$ at equilibrium.

4

STOCHASTIC MODEL

In the stochastic case, one models the evolution in time of the three species as a continuous-time Markov Process with values in a finite or countable set. The main asumption or mathematical simplification here is that the evolution of the process only depends on the past through its present state, or equivalently that the times between two reactions are exponentially distributed. An interesting discussion about the asumption of exponential holding times can be found in Feller (1968), Chapter XVII.6. In the finite state space case, an alternative equivalent condition is that the probability for a change to occur in a small time interval is roughly the length of the interval over the mean time the process stays in the current state. Chapter 2.8 of Norris (1997) provides a detailed discussion of the equivalent characterizations of a continuous-time Markov Process on a finite state space. In the infinite state space case, we need the stronger asumption that the probability for two reactions in the system to occur in a time interval Δt is $o(\Delta t)$, unconditioned on the current space, see Gillespie (1992) for a rigorous analysis.

4.1 SHORT-TIME EVOLUTION

In our simple example, if $X(t) = (X_1(t), X_2(t), X_3(t))$ where $X_1(t)$, $X_2(t)$ and $X_3(t)$ denotes the number of molecules of species $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ at time t, the short-time evolution of the process is modeled by

$$
\begin{aligned}
&P(X(t+\Delta t) = (x_1 - 1, x_2 - 2, x_3 + 1) \mid X(t) = x) = c_1 x_1 x_2 (x_2 - 1)\Delta t + o(\Delta t),\\
&P(X(t+\Delta t) = (x_1 + 1, x_2 + 2, x_3 - 1) \mid X(t) = x) = c_2 x_3 \Delta t + o(\Delta t),\\
&P(X(t+\Delta t) = (x_1, x_2, x_3) \mid X(t) = x) = 1 - (c_1 x_1 x_2 (x_2 - 1) + c_2 x_3)\Delta t + o(\Delta t),\\
&P(X(t+\Delta t) = \text{ anything else } \mid X(t) = x) = o(\Delta t),
\end{aligned}
$$

where $x = (x_1, x_2, x_3)$, Δt is small and $o(\Delta t)$ means that $\lim_{\Delta t \downarrow 0} \dfrac{o(\Delta t)}{\Delta t} = 0$.

Let us introduce the general setting emphasized on the preceding example, mainly paraphrasing El Samad *et al.* (2005) Section 4 and other articles on the subject.

We consider a well-stired system consisting of molecules of N chemical species $\{\mathcal{S}_1, \ldots, \mathcal{S}_N\}$ interacting through M chemical reaction channels $\{R_1, \ldots, R_M\}$, and denote by $X(t) = \{X_1(t), \ldots, X_N(t)\}$ the vector with $X_i(t)$ the number of molecules of species S_i at time t.

Each reaction channel R_j can be characterized by two mathematical quantities,

the state-change vector $v_j = (v_{1j}, v_{2j}, \ldots, v_{Nj})$ and the propensity function a_j. If the system is in state $x = (x_1, \ldots, x_N)$ and reaction j occurs, the system instantaneously jump in state $x + v_j$.

In our simple example $N = 3$, $M = 2$, $x = (x_1, x_2, x_3)$, and the other quantities are

Reaction R_j	Propensity function a_j	State-change vector v_j
$R_1 : \mathcal{S}_1 + 2\,\mathcal{S}_2 \xrightarrow{c_1} \mathcal{S}_3$	$a_1(x) = c_1 x_1 x_2 (x_2 - 1)$	$v_1 = (-1, -2, 1)$
$R_2 : \mathcal{S}_3 \xrightarrow{c_2} \mathcal{S}_1 + 2\,\mathcal{S}_2$	$a_2(x) = c_2 x_3$	$v_2 = (1, 2, -1)$

To model the process in the general setting, we will make the following asumptions:

A0: With Δt and $o(\Delta t)$ as above, the short-time evolution of the process is governed by the $M + 2$ equations

$$\begin{aligned} &P(X(t+\Delta t) = x + v_j \mid X(t) = x) = a_j(x)\,\Delta t + o(\Delta t), \quad \text{for } 1 \le j \le M, \\ &P(X(t+\Delta t) = x \mid X(t) = x) = 1 - \sum_{j=1}^{M} a_j(x)\,\Delta t + o(\Delta t), \\ &P(X(t+\Delta t) = y \mid X(t) = x) = o(\Delta t), \quad \text{for } y \neq x + v_j,\ 1 \le j \le M. \end{aligned} \tag{4.1}$$

A1: The probability that more than one reaction occur in the time interval between t and $t + \Delta t$ is $o(\Delta t)$.

When the state space E is finite, asumption A1 follows from A0. If the state space is finite and we asume A0 or the state space is infinite and we asume A0 and A1, the process $\{X(t)\}_{t \ge 0}$ is a time-continuous Markov process on E, see Chapter 2.8 of Norris (1997) for the finite case and Gillespie (1992) in the infinite case.

In the thesis, we will always asume A0 when the state space is finite and A0 and A1 when the state space is infinite.

4.2 CHEMICAL MASTER EQUATION (CME)

One of the key concepts in this model is the so-called Chemical Master Equation or CME, which denotes in fact Kolmogorov's forward equation in this special setting. We recall that the state space E can either be finite or infinite. A finite state space allows to avoid some technical difficulties and can be motivated by biological considerations like for example the finite size of a cell, and an infinite state space is sometimes more convenient to avoid boundary conditions.

The idea of the CME is simply to look at the possible evolution of the process from time t to time $t + \Delta t$ using the Chapman-Kolmogorov identity, more

precisely for an $x \in E$,

$$P(X(t+\Delta t)=x)=\sum_{y\in E} P(X(t+\Delta t)=x \mid X(t)=y)P(X(t)=y)$$

To avoid notational difficulties on the boundary, let us enlarge the set E by defining the set $\tilde{E}$ to be

$$\tilde{E}=E\cup E_- \quad \text{with} \quad E_-=\bigcup_{j=1}^{M}\{x-v_j; x\in E\},$$

and extend P on $\tilde{E}$ setting $P(X(t)=y)\equiv 0$ and $a_j(y)=0$ for $y\in \tilde{E}\backslash E$, $1\le j\le M$. In the finite case, $\tilde{E}$ is also finite.
Since all terms in $\tilde{E}\backslash E$ have probability 0, the above partition can also be written as

$$P(X(t+\Delta t)=x)=\sum_{y\in \tilde{E}} P(X(t+\Delta t)=x \mid X(t)=y)P(X(t)=y).$$

Theorem 1 *The temporal evolution of the probability $P(X(t)=x)$ is given by the Chemical Master Equation (CME)*

$$\frac{d}{dt}P(X(t)=x)=\sum_{j=1}^{M} a_j(x-v_j)P(X(t)=x-v_j)-\sum_{j=1}^{M} a_j(x)P(X(t)=x), \quad (4.2)$$

where $\frac{d}{dt}$ denotes the right derivative with respect to t.

Proof: Using the evolution dynamics described in equations (4.1) and noting that $\tilde{E}$ can be written as the disjoint union $\tilde{E}=E_1\cup E_2\cup E_3$, with $E_1=\{x-v_1,\dots,x-v_M\}$, $E_2=\{x\}$ and $E_3=\tilde{E}\backslash\{x,x-v_1,\dots,x-v_M\}$, the sum can be separated in three parts and the right hand side is simply

$$\begin{aligned}
P(X(t+\Delta t)=x) &= \sum_{y\in E_1} P(X(t+\Delta t)=x \mid X(t)=y)P(X(t)=y)\\
&\quad+\sum_{y\in E_2} P(X(t+\Delta t)=x \mid X(t)=y)P(X(t)=y)\\
&\quad+\sum_{y\in E_3} P(X(t+\Delta t)=x \mid X(t)=y)P(X(t)=y)\\
&=\sum_{j=1}^{M} P(X(t+\Delta t)=x \mid X(t)=x-v_j)P(X(t)=x-v_j)\\
&\quad+P(X(t+\Delta t)=x \mid X(t)=x)P(X(t)=x)\\
&\quad+\sum_{y\in E_3} P(X(t+\Delta t)=x \mid X(t)=y)P(X(t)=y)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{M} (a_j(x - v_j)\,\Delta t + o(\Delta t)) P(X(t) = x - v_j) \\
&\quad + \left(1 - \sum_{j=1}^{M} (a_j(x)\,\Delta t + o(\Delta t))\right) P(X(t) = x) \\
&\quad + \sum_{y \in E_3} o(\Delta t)\, P(X(t) = y)
\end{aligned}
$$

If E is finite, $\sum_{y \in \tilde{E}} o(\Delta t)\, P(X(t) = y) \leq \sum_{y \in \tilde{E}} o(\Delta t)$ is again $o(\Delta t)$. If E is not finite, $\sum_{y \in \tilde{E}} o(\Delta t)\, P(X(t) = y)$ is roughly the probability that more than one reaction occur in the time interval between t and $t + \Delta t$, hence it is $o(\Delta t)$ according to asumption A1. After deducting $P(X(t) = x)$ on both sides, we can rewrite the above equation as

$$
\begin{aligned}
P(X(t + \Delta t) = x) - P(X(t) = x) = &\sum_{j=1}^{M} a_j(x - v_j)\,\Delta t\, P(X(t) = x - v_j) \\
&- \sum_{j=1}^{M} a_j(x)\,\Delta t\, P(X(t) = x) + o(\Delta t).
\end{aligned}
$$

The CME is obtained from the above equation after dividing by Δt and letting Δt go to 0. □

Notice that the term $P(X(t) = x)$ on the right side does not depend on j and could be taken before the summation sign, however we prefer to write it that way to emphasize the following interpretation of the CME:

> *the change in the probability to be in state x at time t is the difference between the weighted averages of the incoming rates in x and the outgoing rates from x.*

In the simple example we consider, the CME is hence

$$
\begin{aligned}
\frac{d}{dt} P(X(t) = x) = &c_1(x_1 + 1)(x_2 + 2)(x_2 + 1) P(X(t) = (x_1 + 1, x_2 + 2, x_3 - 1)) \\
&+ c_2(x_3 + 1) P(X(t) = (x_1 - 1, x_2 - 2, x_3 + 1)) \\
&- (c_1 x_1 x_2(x_2 - 1) + c_2 x_3) P(X(t) = x).
\end{aligned}
$$

4.3 THE GILLESPIE ALGORITHM

The Gillespie stochastic simulation algorithm, often abbreviated SSA, is simply the simulation of the time-continuous Markov process whose short-time evolution correspond to asumption A0 in Section 4.1. According to the usual theory of Markov processes, the process can be separated in two simpler processes, the embedded jump chain and the time process.

The embedded jump chain is a simple discrete time Markov chain $\{J(k)\}_{k\in\mathbb{N}}$ with values in E, initial condition $J(0) = X(0)$ and transition matrix $\mathbb{P}$ satisfying

$$\mathbb{P}(x, x+v_j) = P(J(k+1) = x+v_j \mid J(k) = x) = \frac{a_j(x)}{\sum_{i=1}^{M} a_i(x)}.$$

It can be viewed as the skeleton of the process, only recording the successive states of the process whitout accounting for the time spend in each state. The transition matrix gives the probability to go from state x to a state $x+v_j$ when the next reaction occur, in other words the probability that the next reaction to occur is the j-th reaction, and this probability is proportional to the propensity function $a_j(x)$.

The time process $\{T(k)\}_{k\in\mathbb{N}}$ is a partial sum of independent exponential variables T_l, $l = 1, 2, 3, \ldots$, such that the parameter of the distribution of T_l is $\sum_{i=1}^{M} a_i(J(l))$. More precisely, $T(0) = 0$ and $T(k+1) = T(k) + T_k$, where T_k is exponential with parameter $\sum_{i=1}^{M} a_i(J(k))$ and independent of $T(k)$. The time process records the time the process spend in the successive states of the jump process. If the jump process is in state x, the time continuous process stays in x until one reaction occur. The time for reaction i to occur is an exponentially distributed random variable Y_i with parameter $a_i(x)$, independent of the other reactions. Hence the time that the process stays in x is the minimum of M independent exponential random variables with parameter $a_i(x)$, $1 \le i \le M$, whose distribution is again exponential with parameter $\sum_{i=1}^{M} a_i(J(k))$ since

$$P(\min(Y_1, \ldots, Y_M) > t) = \prod_{i=1}^{M} P(Y_i > t) = e^{-\sum_{i=1}^{M} a_i(x)}.$$

The time-continuous process $\{X(t)\}_{t\ge 0}$ is equal in distribution to the process $\{J(\tau(t))\}_{t\ge 0}$, where $\tau(t)$ is the renewal time

$$\tau(t) = \inf_{k\in\mathbb{N}}\{t \ge T(k)\}.$$

The Gillespie stochastic simulation algorithm can be implemented as follows:

STEP 0: Specify the initial condition $J(0)$ and the time t_{end} to end the simulation. Set $k = 0$, $J(k) = J(0)$ and $T(k) = 0$.

STEP 1: Update $k \to k+1$.
Draw two pseudo random numbers U_1 and U_2 uniform on $[0, 1]$.
Choose the next reaction to occur to be the $j-$th one for j satisfying

$$\sum_{i=1}^{j-1} a_i(J(k-1)) \le U_1 \cdot \sum_{i=1}^{M} a_i(J(k-1)) < \sum_{i=1}^{j+1} a_i(J(k-1)),$$

with the convention that an empty sum is 0.
Set $J(k) = J(k-1) + v_j$ and $T(k) = T(k-1) - \dfrac{\ln(U_2)}{\sum_{i=1}^{M} a_i(J(k-1))}$.

STEP 2: If $T(k) < t_{\text{end}}$, go to STEP 1, else stop.

In the last part of STEP 1, we use the fact that an exponential random variable with parameter λ is equal in distribution to $-\dfrac{\ln(U)}{\lambda}$ for U uniform on $[0,1]$. The form of the algorithm given here returns a single trajectory of the process $\{X(t)\}_{t\geq 0}$ from the time 0 to $T(k)$. With slight modifications it can be performed in parallel for several trajectories at small computational time cost.

One is usually interested in the asymptotic behaviour of the system and the SSA is in this case a Monte Carlo method for approximating the invariant measure $\pi(x),\ x \in E$. Two approaches are possible, the first one is simply based on the fact that for t large enough, $X(t)$ is approximately in the invariant measure, while the second one rely on the ergodic theorem for time-continuous Markov processes.

The first method is the most simple one but requires far more trajectories to simulate. The idea is to look at the value $X(t_{\text{end}})$ for a large number of trajectories and approximate $\pi(x)$ as the empirical frequence of state x. It is worth noticing the $X(t_{\text{end}})$ is not the last state in the skeleton trajectory but the next to last. The empirical distribution of the last value in the skeleton of the trajectory does not approach π but $\pi \cdot \mathbb{P}$, where $\mathbb{P}$ is the transition matrix of the jump chain, and the two quantities can be very different.

The second approach rely on the ergodic theorem for time-continuous Markov processes, see for example Norris (1997), Chapter 3.8. The ergodic theorem states that for large t, $\pi(x)$ is approximately the fraction of time the process $\{X(s)\}_{0\leq s\leq t}$ stays in state x, more precisely

$$\pi(x) = \lim_{t\to\infty} \frac{\int_0^t 1_{\{X(s)=x\}} ds}{t}.$$

Hence for a trajectory $\{J(m)\}_{0\leq m\leq k}$ and $\{T(m)\}_{0\leq m\leq k}$,

$$\pi(x) \approx \frac{\sum_{m=0}^{k-1} (T(m+1) - T(m)) \cdot 1_{\{J(m)=x\}}}{T(k)}.$$

In practice, we deal with a big number N_{traj} of trajectories $\left\{J^{(i)}(m)\right\}_{0\leq m\leq k_i}$ and $\left\{T^{(i)}(m)\right\}_{0\leq m\leq k_i}$, $1 \leq i \leq N_{\text{traj}}$, and the above formula is replaced by

$$\pi(x) \approx \frac{\sum_{i=1}^{N_{\text{traj}}} \sum_{m=0}^{k_i-1} (T^{(i)}(m+1) - T^{(i)}(m)) \cdot 1_{\{J^{(i)}(m)=x\}}}{\sum_{i=1}^{N_{\text{traj}}} T^{(i)}(k_i)}.$$

To minimize the effects of the transient phase in the system, one can also proceed in a similar way over a part of the trajectories.

In large systems, the algorithm can be extremely slow, especially when there are several time scales and the most interesting reactions are the slower ones. Most of the reactions simulated are those of the most rapid time scale and one needs much more steps in the algorithm to reach the end of the simulation. Accelerated methods have been developed to handle such problems, as the slow-scale stochastic simulation algorithm (ssSSA) or the τ-leaping method, for a review of these methods see El Samad *et al.* (2005).

4.4 EVOLUTION OF THE MOMENTS

We will need the following lemma, stating that the weighted propensity functions are the same in E and in $E_- = \bigcup_{j=1}^{M} \{x - v_j; x \in E\}$.

Lemma 1 *For any function f defined on $\tilde{E}$,*

$$\sum_{y \in E_-} f(y) a_j(y) P(X(t) = y) = \sum_{x \in E} f(x) a_j(x) P(X(t) = x).$$

Proof: We recall that $\tilde{E} = E \cup E_-$. Taking the difference eliminates the common terms and yields

$$\sum_{y \in \tilde{E} \backslash E} f(y) a_j(y) P(X(t) = y) - \sum_{x \in \tilde{E} \backslash E_-} f(x) a_j(x) P(X(t) = x).$$

The first sum is zero since for $y \in \tilde{E} \backslash E$, $P(X(t) = y) \equiv 0$. For the second sum, suppose that $a_j(x) \neq 0$, it means that $x + v_j$ is a possible state, i. e. $x + v_j \in E$ and $x = (x + v_j) - v_j \in E_-$. Hence, for each element of $\tilde{E} \backslash E_-$, the propensity function a_j vanishes and the second sum is also 0. □

Using this Lemma and simple manipulations of the CME (4.2), one can easily derive equations for a handful of quantities of interest, but these equations are usually unwieldy or impossible to solve except in very special simple cases.

For example, the evolution in time of the kth-moment of the number of molecules of the chemical species $\mathcal{S}_i$ can be obtained as follows:

Theorem 2 *The kth-moment of the number of molecules of the chemical species $\mathcal{S}_i$ follows the differential equation*

$$\frac{d}{dt} \mathbb{E}\left(X_i(t)^k\right) = \sum_{j=1}^{M} \mathbb{E}\left((X_i(t) + v_{ij})^k a_j(X(t))\right) - \sum_{j=1}^{M} \mathbb{E}\left(X_i(t)^k a_j(X(t))\right). \quad (4.3)$$

Developing the right hand side, this equation can be reformulated as

$$\frac{d}{dt} \mathbb{E}\left(X_i(t)^k\right) = \sum_{j=1}^{M} \sum_{l=0}^{k-1} \binom{k}{l} v_{ij}^{k-l} \mathbb{E}\left(X_i(t)^l a_j(X(t))\right).$$

Proof: Since all terms are non negative,

$$\frac{d}{dt}\mathbb{E}\left(X_i(t)^k\right) = \frac{d}{dt}\sum_{x\in E} x_i^k P(X(t)=x) = \sum_{x\in E} x_i^k \frac{d}{dt}P(X(t)=x),$$

or with the CME (4.2)

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\left(X_i(t)^k\right) =& \sum_{x\in E} x_i^k \sum_{j=1}^{M} a_j(x-v_j)P(X(t)=x-v_j)\\ &- \sum_{x\in E} x_i^k \sum_{j=1}^{M} a_j(x)P(X(t)=x)\\ =& \sum_{j=1}^{M}\sum_{y\in E_-} (y_i+v_{ij})^k a_j(y)P(X(t)=y)\\ &- \sum_{j=1}^{M}\sum_{x\in E} x_i^k a_j(x)P(X(t)=x).\end{aligned}$$

Using Lemma 1,

$$\sum_{y\in E_-} (y_i+v_{ij})^k a_j(y)P(X(t)=y) = \sum_{x\in E}(x_i+v_{ij})^k a_j(x)P(X(t)=x)$$

and hence

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\left(X_i(t)^k\right) =& \sum_{j=1}^{M}\sum_{x\in E}(x_i+v_{ij})^k a_j(x)P(X(t)=x)\\ &- \sum_{j=1}^{M}\sum_{x\in E} x_i^k a_j(x)P(X(t)=x)\\ =& \sum_{j=1}^{M}\mathbb{E}\left((X_i(t)+v_{ij})^k a_j(X(t))\right) - \sum_{j=1}^{M}\mathbb{E}\left(X_i(t)^k a_j(X(t))\right).\end{aligned}$$

A simple binomial expansion of $(X_i(t)+v_{ij})^k$ allows to eliminate the term $\sum_{j=1}^{M}\mathbb{E}\left(X_i(t)^k a_j(X(t))\right)$ and yields the second equation. □

As a simple case of Theorem 2, the evolution of the number of molecules of the chemical species i reads

$$\frac{d}{dt}\mathbb{E}\left(X_i(t)\right) = \sum_{j=1}^{M} v_{ij}\mathbb{E}\left(a_j(X(t))\right). \tag{4.4}$$

Since the propensity functions $a_j(x)$, $x=(x_1,\ldots,x_n)$ are usually polynomials, the term $\mathbb{E}\left(X_i(t)^l a_j(X(t))\right)$ is very likely to contain moments of $X_i(t)$ of order higher than k and mixed terms $\mathbb{E}\left(X_i(t)^m X(t)^n\right)$ for some m, $n\in\mathbb{N}$. Hence to solve the equation for the moment of order k, one usually has to solve an infinite system of equations for the moments of $X_i(t)$ and mixed terms $\mathbb{E}\left(X_i(t)^m X(t)^n\right)$. To derive equations for the mixed moments, the formalism of moment generating functions discussed below is very useful.

4.5 EVOLUTION OF THE GENERATING FUNCTION

Concerning the moment generating function, things become more involved. With the convention

$$s^k = s_1^{k_1} s_2^{k_2} \cdots s_N^{k_N} \quad \text{for } s = (s_1, \ldots, s_N) \text{ and } k = (k_1, \ldots, k_N),$$

let us consider the function $G : (0,1]^N \times \mathbb{R}_+ \to [0,1]$, defined as

$$G(s,t) = \mathbb{E}(s^{X(t)}) = \sum_{x \in E} P(X(t) = x) s^x$$

and G evolves according to

$$\begin{aligned}
\frac{\partial}{\partial t} G(s,t) &= \sum_{x \in E} s^x \sum_{j=1}^{M} a_j(x - v_j) P(X(t) = x - v_j) \\
&\quad - \sum_{x \in E} s^x \sum_{j=1}^{M} a_j(x) P(X(t) = x) \\
&= \sum_{j=1}^{M} \sum_{x \in E} s^{x+v_j} a_j(x) P(X(t) = x) \\
&\quad - \sum_{j=1}^{M} \sum_{x \in E} s^x a_j(x) P(X(t) = x) \\
&= \sum_{j=1}^{M} (s^{v_j} - 1) \mathbb{E}(s^{X(t)} a_j(X(t))). \qquad (4.5)
\end{aligned}$$

Some terms in the v_j may be negative and the last equation only make sense for s having nonzero components, this is why we excluded this case in the definition of G.

Since a_j is usually a polynomial, equation (4.5) is a partial differential equation for G.

In the following, we suppose that a_j is a polynomial in x and define some useful notations. Let $k = (k_1, k_2, \ldots, k_N)$ denote a multi-index and we write the polynomial a_j as

$$a_j(x) = \sum_{k \in I_j} \alpha_{kj} x^k,$$

where x^k stands as before for $x^k = x_1^{k_1} x_2^{k_2} \cdots x_N^{k_N}$ and I_j is the finite set of multi-indices corresponding to a_j. Since for any i, $1 \le i \le N$,

$$\mathbb{E}(X_i(t) s^{X(t)}) = s_i \frac{\partial}{\partial s_i} G(s,t),$$

it is very convenient to define the differential operator Θ_i as

$$\Theta_i := s_i \frac{\partial}{\partial s_i},$$

and, with the same multi-index k as above and $\Theta = (\Theta_1, \Theta_2, \ldots, \Theta_N)$,

$$\Theta^k := \Theta_1^{k_1}\Theta_2^{k_2}\cdots\Theta_N^{k_N}.$$

With this simple notation and the relation

$$\mathbb{E}\left(X(t)^k s^{X(t)}\right) = \Theta^k G(s,t), \tag{4.6}$$

we can rewrite equation (4.5) and the evolution of the generating function G is the partial differential equation

$$\frac{\partial}{\partial t}G(s,t) = \sum_{j=1}^{M}(s^{v_j}-1)\sum_{k\in I_j}\alpha_{kj}\Theta^k G(s,t). \tag{4.7}$$

From this last equation, one can derive partial differential equations for all kind of mixed moments, since for any multi-index $l = (l_1, l_2, \ldots, l_N)$, if $\mathbb{E}(X^l(t))$ exists, then

$$\mathbb{E}(X^l(t)) = \Theta^l G(s,t)\big|_{s=(1,1,\ldots,1)}.$$

Hence we have

$$\frac{\partial}{\partial t}\mathbb{E}(X^l(t)) = \sum_{j=1}^{M}\sum_{k\in I_j}\alpha_{kj}\Theta^l(s^{v_j}-1)\Theta^k G(s,t)\big|_{s=(1,1,\ldots,1)}.$$

The result given in Theorem 2 is a particular case of this equation with the multi-index k having all components but one equal to zero.

4.6 ASYMPTOTIC BEHAVIOUR OF THE SYSTEM

Even for simple systems, equations (4.3) and (4.7) are very unlikely to be solved analytically. To avoid the difficulty of time dependence, one could only focus on the process at equilibrium. This approach is mathematically much simpler but captures the long term behaviour one is usually interested in.

Suppose the process has an invariant distribution π, i. e. there exists a probability distribution π on E such that

$$\pi(x) = \lim_{t\to\infty} P(X(t) = x) \quad \text{for all } x \in E, \quad \text{and } \sum_{x\in E}\pi(x) = 1,$$

and let $X^{(\infty)} = (X_1^{(\infty)}, \ldots, X_N^{(\infty)})$ be distributed according to π. With the convention $G^{(\infty)}(s) := \mathbb{E}(s^{X^{(\infty)}})$, $\mathbb{E}\left((X_i^{(\infty)})^k\right)$ and $G^{(\infty)}(s)$ have to satisfy equations (4.3), (4.5) respectively (4.7) with the left hand side set to zero, i. e.

$$0 = \sum_{j=1}^{M}\sum_{l=0}^{k-1}\binom{k}{l}v_{ij}^{k-l}\mathbb{E}\left((X_i^{(\infty)})^l a_j(X^{(\infty)})\right),$$

$$0 = \sum_{j=1}^{M}(s^{v_j}-1)\mathbb{E}(s^{X^{(\infty)}}a_j(X^{(\infty)})),$$

$$0 = \sum_{j=1}^{M}(s^{v_j}-1)\sum_{k\in I_j}\alpha_{kj}\Theta^k G^{(\infty)}(s).$$

The formalism developed here seems at first sight very heavy, but with some training its use can indeed avoid pages of tedious calculus.

5

STOCHASTIC VERSUS DETERMINISTIC MODELS OF GENE NETWORKS

The two approaches can yield surprisingly different results, especially when applied to gene expression. Kærn *et al.* (2005) provide a wide review of the biological meaning of stochasticity and the mathematical differences are discussed extensively in El Samad *et al.* (2005) or Goutsias (2007). We give here a short summary of the principal differences between the two approaches.

5.1 CONDITIONS FOR EQUIVALENCE

In Darvey *et al.* (1966), the authors state that for second order chemical reaction, provided the reactions are not irreversible, the difference between the deterministic equilibrium and the stochastic equilibrium mean of one of the chemical species is of order $\frac{1}{N}$, where N is the order of the number of molecules in the system, and the stochastic variance is also of order $\frac{1}{N}$, thus the coefficient of variation, standard deviation over mean, of the usual $\frac{1}{\sqrt{N}}$ order. Furthermore, they argue that "*For chemical systems of practical interest N is generally very large and the term* (... difference between stochastic mean and deterministic equilibrium ...) *is negligible by comparison with* (... the equilibrium mean). *Thus it follows that for such systems the mean value of the stochastic solution agrees with the deterministic solution.*" This also explain why the topic of stochastic modeling is not essential for classical chemistry where the number of molecules of each species is huge.

In biochemistry however, the total number of molecules is not always large. In the setting of gene expression, one has to take into account the particular role of the promoter. The promoter is usually modeled binary as repressed or active, or equivalently OFF or ON, and its state largely influence the behaviour of the system. According to Kærn *et al.* (2005) or Kepler & Elston (2001), the two conditions that need to be satisfied for the two approaches to be similar are

- large system size, i. e. high numbers of expressed mRNA and protein and large cell volumes,

- fast promoter kinetics, i. e. fast switches between active and inactive state.

These two conditions are often not fulfilled and in this case the effects of molecular-level noise can be very large, as emphasized in the following examples.

5.2 MATHEMATICAL EXPRESSION OF MOLECULAR NOISE

In this section, we briefly describe the various levels at wich molecular noise can act and generate different behaviour as the deterministic equilibrium. The exemples are mostly taken from El Samad *et al.* (2005) where a detailed discussion can be found.

When comparing the evolution equations for the deterministic concentration $\phi_{\mathcal{S}_i}$ and the stochastic mean $\mathbb{E}(X_i)$, the terms involving products between species can be very different. Typically terms of the form $\phi_{\mathcal{S}_i} \cdot \left(\phi_{\mathcal{S}_j}\right)^m$ in the deterministic setting do not correspond in the stochastic setting to $\mathbb{E}(X_i) \cdot (\mathbb{E}(X_j))^m$ but to the much more complicated term $\mathbb{E}\left(X_i \cdot (X_j)^m\right)$, taking into account noise and correlations between species. Moreover additional terms appears in the stochastic setting when reactions involve several reactants of the same species.

i) MONOSTABLE SYSTEMS

Even for simple monostable systems, noise can play a crucial role, as shown in the example 3.3.1 in El Samad *et al.* (2005). A system with a unique stable equilibrium point in the positive quadrant is presented. The authors provide two sets of parameters yielding the same deterministic equilibrium but showing very different behaviour when simulated stochastically. The first set of parameters behaves almost like the deterministic trajectory, while the second set shows stochastic excursions reaching up to four-fold the deterministic value.

ii) GENETIC SWITCHES

Multistable systems possess several stable equilibrium points. In a deterministic setting, if the initial conditions of a trajectory lie in the basin of attraction of an equilibrium point, the trajectory will tend asymptotically to this point. Different initial conditions yields a different asymptotic behaviour. When the dynamics is stochastic, the trajectory can jump between the basins of attraction of several stable points, hence a trajectory can stay a very long time near an equilibrium point and suddenly escape from its basin of attraction and stay for a while near another equilibrium point. For obvious reasons, this feature is often called genetic switch. In Figure 5.1, one can clearly see the trajectory oscillating between the neighbourhoods of two almost stable states. The example is taken from a simulation of the genetic module described in Part III when the promoter state is governed by positive feedback and the steady state distribution is bimodal.

iii) GENETIC OSCILLATORS

The circadian rythm is the molecular mechanism that generate oscillations with a period close to 24 hours. This clock permits living organisms to adapt to natural periodicity such as the alternance of night and day, and regulate their behaviour accordingly. In Vilar *et al.* (2002), the author shows that circadian clock is not only noise resistant, but paradoxally even enhanced by noise. In some parameter regime, both deterministic and stochastic model show oscillation, whereas in other the deterministic system has a unique stable equilibrium point while stochasticity permits to maintain oscillations.

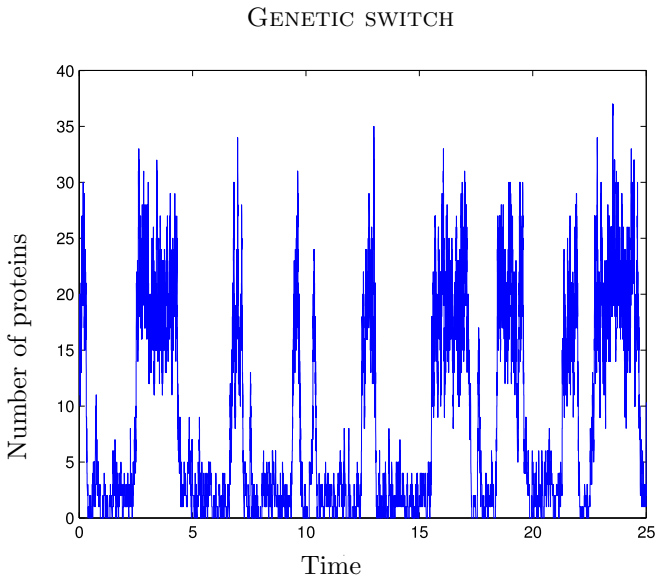

Figure 5.1: Example of a genetic switch between two equilibrium points. The trajectory oscillates between the neighbourhoods of two almost stable states, one at about 2 proteins and the other at about 20. The example is taken from a simulation of the genetic module described in Part III when the promoter state is governed by positive feedback and the steady state distribution is bimodal.

iv) Molecular fluctuation and the role of the promoter

In molecular biology, single events have sometimes a huge effect on the global behaviour of the system. This is the case in the heat shock model presented in El-Samad *et al.* (2005), where a species is almost always absent in the system, but at rare periods it can be present at very low level corresponding to an important event for gene expression that can not be captured when considering only an averaged behaviour.

Similarily, the promoter plays a crucial role in gene expression, yielding a completely different behaviour of the system depending on its state, ON or OFF, as discussed in and Zhou & Davies (2004). Again, some specificities of the system can not be catched considering only an averaged behaviour of the promoter level. For example, Ozbudak *et al.* (2002) propose a mean-field approximation of the promoter state, yielding a poissonian type steady state. When varying the parameters, this poissonian model fails to catch the experimentally observed bimodal distribution and peak of variance that can be reproduced using our method presented in Chapter 10.

5.3 BIOLOGICAL BENEFITS OF STOCHASTICITY

In opposition to the popular wisdom that noise is a nuisance, leading to errors and negatively affecting cell regulation, randomness can also be of biological benefit. Fedoroff & Fontana (2002) and more recently Kærn *et al.* (2005) dis-

cuss the potentially beneficial roles of stochasticity. The first article deals with stochasticity in general biological processes while the second one focuses on gene expression.

In microorganisms, stochasticity in gene expression can result in phenotypic heterogeneity. Let us summarize some of the biological advantages a heterogenous population can gain.

i) RESISTANCE TO ENVIRONMENTAL STRESS

As shown in Thattai & van Oudenaarden (2004), a heterogeneous bacterial population can achieve higher growth rate than an homogenous one in a fluctuating environment, provided the bacterial response time to changing conditions is sufficiently slow. The bacteria in the heterogeneous population are able to anticipate and take advantage of sudden changes in their environment.
Recently, Acar *et al.* (2008) and Bishop *et al.* (2007) have also investigated this subject, and come to the conclusion that hetereogeneity increases the fitness of the population:

> "*The diversity is introduced naturally through the stochastic nature of gene expression, allowing isogenic populations to mitigate the risk by 'not putting all their eggs in one basket'.*"
>
> (Acar *et al.* (2008))

ii) RESISTANCE TO ANTIBIOTICS

Switching between phenotypic states has been proposed as a likely cause of persistent bacterial infections after treatment with antibiotics. While most of the population is rapidly killed by the treatment, the subpopulation of dormant persistor is less affected and can survive an extended period of exposure. When the drug treatment is over, the surviving persistors randomly switch from the dormant to an active phenotype and the infection can reemerge.

iii) AUTOREGULATION AND SUBSTABLE PHENOTYPIC STATES

In self regulated gene networks, it is commonly admitted that

- *negative feedback* provides a noise-reduction mechanism, see for example Becskei & Serrano (2000),
- *positive feedback* amplifies fluctuations, see for example Isaacs *et al.* (2003) or Becskei *et al.* (2001), often yielding bimodal distributions, see Fournier *et al.* (2007).

Tuning the feedbacks can hence decrease or increase the population heterogeneity, and as an extreme case of positive feedback one can have multiple substable states.
We will discuss the effects of self regulation in more details in Chapter 11 for the simple genetic module investigated in this thesis.

Stochasticity also plays a role in the development of higher organisms and in diseases, for more details see Kærn *et al.* (2005).

Part III

MODEL OF A SIMPLE GENETIC MODULE

6

INTRODUCTION

In this part, we describe the model of a simple genetic module consisting of a promoter and a gene that codes for a protein. When the promoter is activated or in ON state, proteins are synthesized at a constant rate whereas when the promoter is OFF, there is no production. Proteins are degradated in both promoter states according to a propensity function increasing in the number of protein, corresponding to the fact that the more proteins are present in the system, the more probable it is for one of them to be degradated. An important feature of the model is the self-regulation, acting on the promoter. Positive feedback means that the more proteins there is, the more likely the promoter is in the ON state, whereas negative feedback increase the probability for the promoter to be OFF. We give a very general description, to allow copies of such simple modules in a suited particular form to be integrated in more complicated networks, as for example the engineered gene network discussed in Part IV.

This simple genetic module has been studied extensively in the literature. In some very special cases, exact results for the asymptotic behaviour of the system are provided like in Peccoud & Ycart (1995) for the case whithout feedback or in Hornos *et al.* (2005) for the case of linear negative feedback. The method consists in deriving from the chemical master equation asymptotic equations for generating functions and solve them. This method is very particular and the whole derivation has to be done anew for each model, as we emphasize in Chapter 9. For more complicated cases, one usually rely on simulation using the Gillespie stochastic simulation algorithm (SSA) or one of its accelerated variants like τ-leaping or slow-scale SSA (ssSSA), for a detailed review see El Samad *et al.* (2005).

In this Part, we develop a very general and flexible method to compute the invariant distribution of the related time-continuous Markov Process, quite similar to the method presented in Fournier *et al.* (2007). The method is based on the technique of transfer matrices and can be used for all kinds of feedbacks.

6.1 EXACT SOLUTION AND SIMULATION

Compared to simulation, an exact approach entails multiple advantages. In the literature on the subject the use of the word exact is somehow blurry. In constrast to its accelerated approximations, the Gillespie algorithm is often refered to as exact, which is not the case since a computation using Monte Carlo methods is not exact. We can summarize things as follows:

exact methods:	*exact* computation of the invariant distribution of the *exact* system
Gillespie SSA:	Monte Carlo *approximation* of the invariant distribution of the *exact* system
τ-leaping and ssSSA:	Monte Carlo *approximation* of the invariant distribution of an *approximated* system

Another advantage of exact methods is the speed of computation. The knowledge about the model parameters, here the biochemical rates, are usually very vague, and our approach allows to play with the parameters by screening plausible values, what would be far too long using simulations.

7

DESCRIPTION AND MATHEMATICAL MODEL

7.1 BIOLOGICAL DESCRIPTION

Our basic simple genetic module is a piece of DNA consisting of a promoter and a sequence of bases coding for a protein. The promoter has two possible states that we will call ON and OFF, or equivalently 1 and 0.

When the promoter is in the ON state, proteins can be synthesized at a constant rate μ whereas there is no production when the promoter is OFF. In reality, the gene do not directly produce proteins but in a first step called transcription it produces mRNA whose translation synthesize proteins, see Appendix A.2 and A.3 for more details. Here we ignore for simplicity the mRNA level and combine transcription and translation into a single step. In both ON and OFF states, proteins are degradated at rate ν, usually an increasing function of the number of proteins present at that time in the system like for example a constant times the number of proteins. The promoter randomly switches from state ON to OFF at rate κ and from state OFF to ON at rate g. The rates κ and g can be either constants or increasing functions of the number of proteins, corresponding to systems without feedback when both κ and g are constants, negative feedback loop when κ is increasing and g constant, positive feedback loop when κ is constant and g increasing, or both negative and positive feedback loops when κ and g are both increasing.

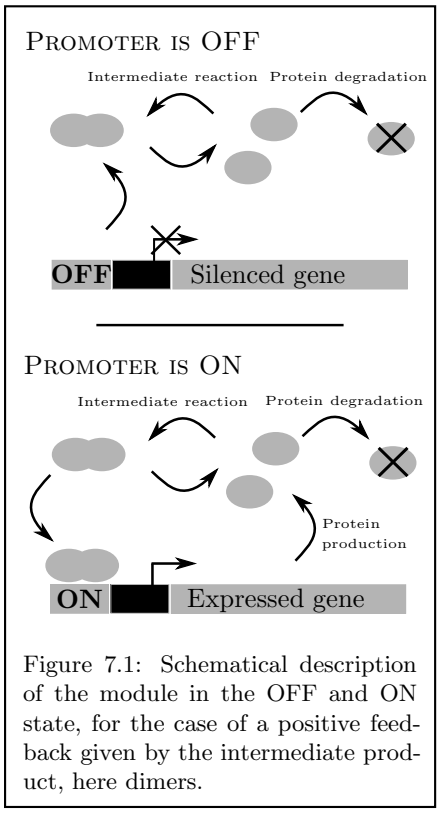

Figure 7.1: Schematical description of the module in the OFF and ON state, for the case of a positive feedback given by the intermediate product, here dimers.

7.2 MATHEMATICAL MODEL

We adapt the general description given in Chapter 3 in this simple particular case. We suppose the infinitesimal transition probabilities to follow asumption A0 from 4.1 and asume moreover A1 if the state space is infinite. Hence $X(t) = (N(t), Y(t))_{t>0}$, where $N(t)$ denotes the number of proteins present in the system at time t and $Y(t)$ the state of the promoter, 1 for ON and 0 for OFF, is a time-continuous regular jump Markov Process. We suppose throughout that $g(0) > 0$, i. e. there is a basal activity and the state $(0, 0)$ is not absorbing. Here the biochemical reactions given the current state $(N(t), Y(t)) = (n, y)$ are

$$\emptyset \underset{\nu(n)}{\overset{y\cdot\mu}{\rightleftharpoons}} \mathcal{P} \qquad \text{protein production / degradation}$$

$$\mathcal{O}_0 \underset{\kappa}{\overset{g(n)}{\rightleftharpoons}} \mathcal{O}_1 \qquad \text{state of the promoter}$$

with $\emptyset$, $\mathcal{P}$, $\mathcal{O}_0$ and $\mathcal{O}_1$ standing for nothing, protein, promoter OFF and promoter ON respectively. In the usual formalism, one can summarize the system for $n \in \mathbb{N}$, $y \in \{0, 1\}$, as

Reaction R_j	Propensity function a_j	State-change vector v_j
$R_1 : \emptyset \xrightarrow{y\cdot\mu} \mathcal{P}$	$a_1(n, y) = y\mu \cdot 1_{\{(n+1,y)\in E\}}$	$v_1 = (1, 0)$
$R_2 : \mathcal{P} \xrightarrow{\nu(n)} \emptyset$	$a_2(n, y) = \nu(n)$	$v_2 = (-1, 0)$
$R_3 : \mathcal{O}_0 \xrightarrow{g(n)} \mathcal{O}_1$	$a_3(n, y) = g(n)(1 - y)$	$v_3 = (0, 1)$
$R_4 : \mathcal{O}_1 \xrightarrow{\kappa(n)} \mathcal{O}_0$	$a_4(n, y) = \kappa(n)y$	$v_4 = (0, -1)$

The propensity function $a_1(n, y) = y\mu \cdot 1_{\{(n+1,y)\in E\}}$ is written in this way to allow for finite or infinite state space. The indicator function $1_{\{x\in A\}}$ is defined as 1 for $x \in A$ and 0 otherwise. If the state space E is $\mathbb{N}\times\{0, 1\}$, $a_1(n, y) = y\mu$ while if the number of proteins is bounded by a maximal number Λ, $a_1(n, y) = y\mu$ if $n < \Lambda$ and 0 otherwise.

The infinitesimal evolution of the system can be written here in compact form since a switch of the promoter state can be written as a transition from y to $1 - y$,

$$\begin{aligned}
&P(X(t + \Delta t) = (n + 1, y) \mid X(t) = (n, y)) = y\mu\Delta t + o(\Delta t),\\
&P(X(t + \Delta t) = (n - 1, y) \mid X(t) = (n, y)) = \nu(n)\Delta t + o(\Delta t),\\
&P(X(t + \Delta t) = (n, 1 - y) \mid X(t) = (n, y)) = (y\kappa(n) + (1 - y)g(n))\Delta t + o(\Delta t),\\
&P(X(t + \Delta t) = (n, y) \mid X(t) = (n, y)) =\\
&\qquad 1 - (\nu(n) + y\mu + y\kappa(n) + (1 - y)g(n))\Delta t + o(\Delta t),\\
&P(X(t + \Delta t) = \text{ anything else } \mid X(t) = (n, y)) = o(\Delta t). \qquad (7.1)
\end{aligned}$$

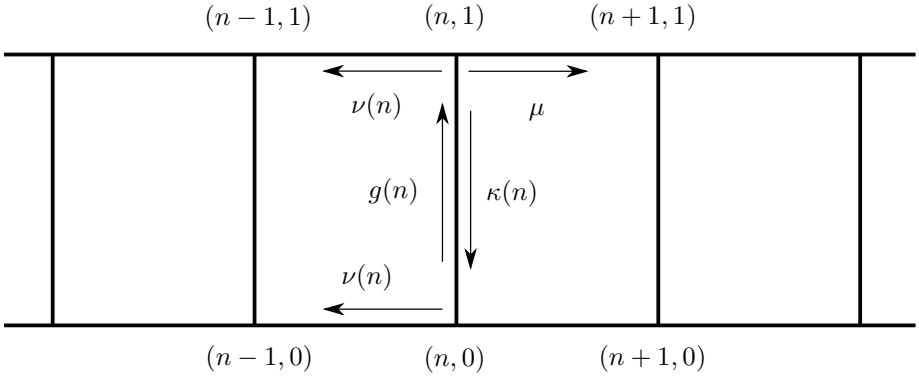

Figure 7.2: Visualization of the state space as a strip. The possibles transitions are represented by the arrows with corresponding rates.

To simplify the notations, let us write

$$p_{(n,y)}(t) := P(X(t) = (n, y)).$$

The corresponding chemical master equation is hence (see (4.2))

$$\begin{aligned}
\frac{d}{dt}\, p_{(n,0)}(t) =& \nu(n+1)\, p_{(n+1,0)}(t) + \kappa(n)\, p_{(n,1)}(t) \\
& - (\nu(n) + g(n))\, p_{(n,0)}(t), \\
\frac{d}{dt}\, p_{(n,1)}(t) =& \mu\, p_{(n-1,1)}(t) + \nu(n+1)\, p_{(n+1,1)}(t) + g(n)\, p_{(n,1)}(t) \\
& - (\mu + \nu(n) + \kappa(n))\, p_{(n,1)}(t)
\end{aligned}$$

or in compact form

$$\begin{aligned}
\frac{d}{dt}\, p_{(n,y)}(t) =& y\mu\, p_{(n-1,y)}(t) + \nu(n+1)\, p_{(n+1,y)}(t) \\
& + (1-y)\kappa(n)\, p_{(n,1-y)}(t) + yg(n)\, p_{(n,1-y)}(t) \\
& - (y\mu + \nu(n) + y\kappa(n) + (1-y)g(n))\, p_{(n,y)}(t)
\end{aligned} \tag{7.2}$$

If the state space is bounded, the boundary condition reads

$$\begin{aligned}
\frac{d}{dt}\, p_{(\Lambda,1)}(t) =& y\mu\, p_{(\Lambda-1,y)}(t) + g(n)\, p_{(\Lambda,0)}(t) \\
& - (\nu(\Lambda) + \kappa(\Lambda))\, p_{(\Lambda,1)}(t),
\end{aligned}$$

and $p_{(n,y)}(t) \equiv 0$ for $n > \Lambda$.

A convenient visualization of the system is given in the following Figure 7.2. This representation will be very useful for the method of transfer matrices to compute the invariant measure presented in Chapter 10.

8

TRANSIENT BEHAVIOUR OF THE SYSTEM

In this section, we give equations for the moments of $N(t)$ and the probability to be ON $P(Y(t) = 1)$. Since $Y(t)$ is a binary random variable, $P(Y(t) = 1)$ is also the moments of every order ≥ 1 of $Y(t)$, and hence we use the general theory discussed in Chapter 4.

8.1 BOUNDED AND UNBOUNDED STRIPS

Let us recall the difference between the bounded and unbounded strips in terms of propensity functions. In the bounded case $\{0, 1, \ldots, \Lambda\} \times \{0, 1\}$, we prevent the protein production at the upper boundary by setting the propensity function $a_1(n, 1)$ to zero if $n \geq \Lambda$, or in compact form $a_1(n, y) = y\mu\ 1_{\{n<\Lambda\}}$, where the indicator function $1_{\{n<\Lambda\}}$ is 1 if $n < \Lambda$ and 0 otherwise. The resulting equations for the bounded case are quite similar to those of the unbounded case but with additional boundary terms.

8.2 EVOLUTION OF THE MEAN AND PROBABILITY TO BE ON

The equation (4.4) yields in our case for the mean number of proteins

$$\frac{d}{dt}\mathbb{E}(N(t)) = \mathbb{E}(a_1(N(t), Y(t))) - \mathbb{E}(a_2(N(t), Y(t))),$$

hence in the unbounded case

$$\frac{d}{dt}\mathbb{E}(N(t)) = \mu\ \mathbb{E}(Y(t)) - \mathbb{E}(\nu(N(t))), \tag{8.1}$$

and in the bounded case, since $\mathbb{E}(a_1(N(t), Y(t))) = \mu\mathbb{E}(Y(t)\ 1_{\{N(t)<\Lambda\}})$,

$$\frac{d}{dt}\mathbb{E}(N(t)) = \mu\ \big(\mathbb{E}(Y(t)) - P\left((N(t), Y(t)) = (\Lambda, 1)\right)\big) - \mathbb{E}(\nu(N(t))).$$

For the state of the promoter, equation (4.4) become

$$\frac{d}{dt}\mathbb{E}(Y(t)) = \mathbb{E}\big(a_3(N(t), Y(t))\big) - \mathbb{E}\big(a_4(N(t), Y(t))\big).$$

It can be written in both cases as

$$\frac{d}{dt}P(Y(t) = 1) = \mathbb{E}(g(N(t)) - \mathbb{E}\Big(\big(\kappa(N(t)) + g(N(t))\big)Y(t)\Big), \tag{8.2}$$

the difference between bounded and unbounded strip being contained in the right side of the equation.

Higher moments can be computed similarily using Theorem 2 of Part II.

9

ASYMPTOTIC BEHAVIOUR OF THE SYSTEM IN SOME SIMPLE CASES

9.1 THE METHOD OF GENERATING FUNCTIONS

In some special cases, it is possible to give an exact solution using the method of generating functions. The method is explicited in Hornos *et al.* (2005), where the authors provide the invariant distribution in the case where $\nu(n) = \nu \cdot n$, $\kappa(n) = \kappa \cdot n$ and $g(n) = g$, i. e. the degradation is proportional to the number of proteins, there is a linear negative feedback and no positive feedback, and a formula for the case whithout feedback is given in Peccoud & Ycart (1995).

In the following, we always consider the unbounded strip $\mathbb{N} \times \{0, 1\}$ as the state space. We introduce the marginal generating functions defined on $[0, 1] \times \mathbb{R}_+$

$$\alpha(z,t) := \sum_{n=0}^{\infty} p_{n,1}(t) z^n \quad \text{and} \quad \beta(z,t) := \sum_{n=0}^{\infty} p_{n,0}(t) z^n,$$

in other terms $\alpha(z,t) = \mathbb{E}(z^{N(t)} Y(t))$ and $\beta(z,t) = \mathbb{E}(z^{N(t)}(1 - Y(t)))$. Since $Y(t)$ can only take the values 0 or 1, this formalism is here more appropriate than the usual generating function $G(s,t)$ discussed in Section 4.5, and

$$G((s_1, s_2), t) = \alpha(s_1, t) + s_2 \beta(s_1, t).$$

The analogous of equation (4.5) are given by

$$\begin{aligned} \frac{\partial \alpha(z,t)}{\partial t} =& \mu\ (z-1)\ \alpha(z,t) + (z^{-1} - 1)\ \mathbb{E}\big(z^{N(t)} \nu(N(t)) Y(t)\big) \\ &+ \mathbb{E}\big(z^{N(t)} g(N(t))(1 - Y(t))\big) - \mathbb{E}\big(z^{N(t)} \kappa(N(t)) Y(t)\big), \qquad (9.1) \\ \frac{\partial \beta(z,t)}{\partial t} =& (z^{-1} - 1)\ \mathbb{E}\big(z^{N(t)} \nu(N(t)) Y(t)\big) \\ &+ \mathbb{E}\big(z^{N(t)} \kappa(N(t)) Y(t)\big) - \mathbb{E}\big(z^{N(t)} g(N(t))(1 - Y(t))\big). \qquad (9.2) \end{aligned}$$

In the case where the propensity functions are polynomials, we use the relation

$$\begin{aligned} \mathbb{E}\big(z^{N(t)} N(t)^k Y(t)\big) &= \left(z \frac{\partial}{\partial z}\right)^k \alpha(z,t), \\ \mathbb{E}\big(z^{N(t)} N(t)^k (1 - Y(t))\big) &= \left(z \frac{\partial}{\partial z}\right)^k \beta(z,t). \qquad (9.3) \end{aligned}$$

This relation is a reformulation of (4.6).

In most cases, the transient behaviour of the system can not be computed and we focus on the equilibrium. We write

$$\alpha(z) = \lim_{t\to\infty} \alpha(z,t) \quad \text{and} \quad \beta(z) = \lim_{t\to\infty} \beta(z,t).$$

The equilibrium equations are given by (9.1) and (9.2) with the left hand side set to zero.

9.2 THE CASE WHITHOUT FEEDBACK

The most simple case is the case whithout feedback. It has been analyzed in depth in Peccoud & Ycart (1995) and used recently for example in Raj *et al.* (2006) or in Iyer-Biswas *et al.* (2007).

Let us consider the case without feedback and with linear degradation, i. e. the propensity functions are given by

Reaction R_j	Propensity function a_j	State-change vector v_j
$R_1 : \emptyset \xrightarrow{y\cdot\mu} \mathcal{P}$	$a_1(n,y) = y\mu$	$v_1 = (1,0)$
$R_2 : \mathcal{P} \xrightarrow{\nu(n)} \emptyset$	$a_2(n,y) = \nu n$	$v_2 = (-1,0)$
$R_3 : \mathcal{O}_0 \xrightarrow{g(n)} \mathcal{O}_1$	$a_3(n,y) = g\cdot(1-y)$	$v_3 = (0,1)$
$R_4 : \mathcal{O}_1 \xrightarrow{\kappa(n)} \mathcal{O}_0$	$a_4(n,y) = \kappa\cdot y$	$v_1 = (0,-1)$

At equilibrium and using the relations (9.3) and the equations (9.1) and (9.2), $\alpha(z)$ and $\beta(z)$ have to satisfy

$$0 = \mu\ (z-1)\ \alpha(z) + \nu\ (1-z)\frac{d}{dz}\alpha(z) + g\ \beta(z) - \kappa\ \alpha(z), \tag{9.4}$$

$$0 = \nu\ (1-z)\frac{d}{dz}\beta(z) + \kappa\ \alpha(z) - g\ \beta(z). \tag{9.5}$$

We sum these equations and isolate α in equation (9.5) to become

$$\begin{aligned} 0 &= \mu\ (z-1)\ \alpha(z) + \nu\ (1-z)\frac{d}{dz}(\alpha(z)+\beta(z)), \\ \alpha(z) &= \frac{g}{\kappa}\beta(z) - \frac{\nu}{\kappa}(1-z)\frac{d}{dz}\beta(z). \end{aligned} \tag{9.6}$$

Plugging α in the first equation yields after some algebra

$$\begin{aligned} 0 = &-\frac{\nu^2}{\kappa}(1-z)^2\ \frac{d^2}{dz^2}\beta(z) + (z-1)\left(\frac{\mu\ \nu}{\kappa}(z-1) - \frac{\nu\ (\kappa+g+\nu)}{\kappa}\right)\frac{d}{dz}\beta(z) \\ &+ (z-1)\frac{\mu g}{\kappa}\beta(z). \end{aligned}$$

For $0 < z < 1$, we divide this equation by $-\frac{\nu^2}{\kappa}(z-1)$,

$$0 = (z-1)\ \frac{d^2}{dz^2}\beta(z) + \left(\frac{\kappa+g+\nu}{\nu} - \frac{\mu}{\nu}(z-1)\right)\frac{d}{dz}\beta(z) - \frac{\mu g}{\nu^2}\beta(z).$$

With the change of variable $w = \frac{\mu}{\nu}(z-1)$ and $\gamma(w) := \beta\left(\frac{\nu}{\mu}w + 1\right)$, the equation simplify to

$$0 = \frac{\mu}{\nu}w\frac{d^2}{dw^2}\gamma(w) + \frac{\mu}{\nu}\left(\frac{\kappa+g+\nu}{\nu} - w\right)\frac{d}{dw}\gamma(w) - \frac{\mu g}{\nu^2}\gamma(w)$$

or equivalently

$$0 = w\frac{d^2}{dw^2}\gamma(w) + \left(\frac{\kappa+g+\nu}{\nu} - w\right)\frac{d}{dw}\gamma(w) - \frac{g}{\nu}\gamma(w),$$

which is the canonical form of Kummer's equation, see for example Abramowitz & Stegun (1964) p. 504. The two independent solutions of this equation are

$${}_1F_1\left(\frac{g}{\nu}, \frac{\kappa+g+\nu}{\nu}, w\right) \quad \text{and} \quad U\left(\frac{g}{\nu}, \frac{\kappa+g+\nu}{\nu}, w\right).$$

${}_1F_1$ is the confluent hypergeometric function

$${}_1F_1(a,b,w) = \sum_{n=0}^{\infty}\frac{(a)_n}{(b)_n}\frac{w^n}{n!},$$

where $(a)_n$ denotes the Pochammer symbol or rising factorial

$$(a)_n = a(a+1)\cdots(a+n-1) = \frac{\Gamma(a+n)}{\Gamma(a)},$$

and U is given by

$$U(a,b,w) = \frac{\pi}{\sin(\pi b)}\left(\frac{{}_1F_1(a,b,w)}{\Gamma(1+a-b)\Gamma(b)} - w^{1-b}\ \frac{{}_1F_1(1+a-b,2-b,w)}{\Gamma(a)\Gamma(2-b)}\right).$$

Hence,

$$\gamma(w) = c_1 \cdot {}_1F_1\left(\frac{g}{\nu}, \frac{\kappa+g+\nu}{\nu}, w\right) + c_2 \cdot U\left(\frac{g}{\nu}, \frac{\kappa+g+\nu}{\nu}, w\right),$$

and since $\beta(z) = \gamma\left(\frac{\mu}{\nu}(z-1)\right)$

$$\beta(z) = c_1 \cdot {}_1F_1\left(\frac{g}{\nu}, \frac{\kappa+g+\nu}{\nu}, \frac{\mu}{\nu}(z-1)\right) + c_2 \cdot U\left(\frac{g}{\nu}, \frac{\kappa+g+\nu}{\nu}, \frac{\mu}{\nu}(z-1)\right).$$

The function $U\left(\frac{g}{\nu}, \frac{\kappa+g+\nu}{\nu}, \frac{\mu}{\nu}(z-1)\right)$ is unbounded near $z = 1$ since $\frac{\kappa+g+\nu}{\nu} > 1$, but β has to satisfy the relation

$$\lim_{z\to 1}\beta(z) = \lim_{t\to\infty} P(Y(t) = 0) \leq 1,$$

thus c_2 has to be 0 and

$$\beta(z) = c_1 \cdot {}_1F_1\left(\frac{g}{\nu}, \frac{\kappa+g+\nu}{\nu}, \frac{\mu}{\nu}(z-1)\right). \tag{9.7}$$

Plugging $\beta(z)$ in the equation (9.6) yields

$$\begin{aligned}\alpha(z) =& c_1 \frac{g}{\kappa}\left({}_1F_1\left(\frac{g}{\nu}, \frac{g+\kappa}{\nu}+1, \frac{\mu(z-1)}{\nu}\right)\right.\\ &\left. + \frac{\mu(z-1)}{g+\kappa+\nu}\ {}_1F_1\left(\frac{g}{\nu}+1, \frac{g+\kappa}{\nu}+2, \frac{\mu(z-1)}{\nu}\right)\right)\end{aligned} \tag{9.8}$$

Since $\alpha(1)+\beta(1)=1$, the normalization factor c_1 is given by $c_1 = \dfrac{\kappa}{g+\kappa}$.

Theorem 3 *The generating function of the number of proteins at equilibrium*

$$G_N(z) := \lim_{t\to\infty} \mathbb{E}(z^{N(t)})$$

is the hypergeometric function

$$G_N(z) = {}_1F_1\left(\frac{g}{\nu}, \frac{g+\kappa}{\nu}, \frac{\mu}{\nu}(z-1)\right).$$

Proof: From (9.7) and (9.8) with $c_1 = \dfrac{\kappa}{g+\kappa}$, we get

$$\begin{aligned}G_N(z) =& \alpha(z)+\beta(z) = {}_1F_1\left(\frac{g}{\nu}, \frac{g+\kappa}{\nu}+1, \frac{\mu(z-1)}{\nu}\right)\\ &+ \frac{\mu\ g}{\kappa\ (g+\kappa+\nu)}(z-1)\ {}_1F_1\left(\frac{g}{\nu}+1, \frac{g+\kappa}{\nu}+2, \frac{\mu(z-1)}{\nu}\right).\end{aligned}$$

To simplify the notations, let us write

$$a = \frac{g}{\nu}, \quad b = \frac{g+\kappa}{\nu} \quad \text{and} \quad w = \frac{\mu}{\nu}(z-1),$$

and the last equation above become

$${}_1F_1(a, b+1, w) + \frac{a}{b(b+1)} w\ {}_1F_1(a, b+2, w).$$

The following relations between hypergeometric functions, see Abramowitz & Stegun (1964) p. 507–508,

$$\begin{aligned}\frac{a}{b(b+1)} w\ {}_1F_1(a+1, b+2, w) =& (\frac{w}{b}-1){}_1F_1(a, b+1, w) + {}_1F_1(a-1, b, w),\\ \frac{w}{b}{}_1F_1(a, b+1, w) =& {}_1F_1(a, b, w) - {}_1F_1(a-1, b, w),\end{aligned}$$

allow to compute

$$\begin{aligned}G_N(z) =& {}_1F_1(a, b+1, w) + \frac{a}{b(b+1)} w\ {}_1F_1(a, b+2, w)\\ =& \frac{w}{b}{}_1F_1(a, b+1, w) + {}_1F_1(a-1, b, w)\\ =& {}_1F_1(a, b, w) = {}_1F_1\left(\frac{g}{\nu}, \frac{g+\kappa}{\nu}, \frac{\mu}{\nu}(z-1)\right).\end{aligned}$$

□

From the generating functions $\beta(z)$ and $G_N(z)$, we can easily derive the equilibrium distribution of the process.

Theorem 4 *The equilibrium distributions*

$$\pi_{(n,\ \cdot\ )} = \lim_{t\to\infty} P(N(t)=n) \quad \textit{and} \quad \pi_{(n,y)} = \lim_{t\to\infty} P((N(t),Y(t))=(n,y))$$

are given by

$$\begin{aligned}
\pi_{(n,\ \cdot\ )} =& \frac{\left(\frac{g}{\nu}\right)_n}{\left(\frac{g+\kappa}{\nu}\right)_n \cdot n!} \left(\frac{\mu}{\nu}\right)^n {}_1F_1\left(\frac{g}{\nu}+n, \frac{\kappa+g}{\nu}+n, -\frac{\mu}{\nu}\right), \\
\pi_{(n,0)} =& \frac{\frac{\kappa}{\kappa+g}\cdot\left(\frac{g}{\nu}\right)_n}{\left(\frac{g+\kappa}{\nu}+1\right)_n \cdot n!} \left(\frac{\mu}{\nu}\right)^n {}_1F_1\left(\frac{g}{\nu}+n, \frac{\kappa+g}{\nu}+1+n, -\frac{\mu}{\nu}\right), \\
\pi_{(n,1)} =& \frac{g+\nu n}{\kappa}\pi_{(n,0)} - \frac{\nu\ (n+1)}{\kappa}\pi_{(n+1,0)} \\
=& \pi_{(n,\ \cdot\ )} - \pi_{(n,0)}.
\end{aligned}$$

The factorial moments

$$\begin{aligned}
e_n &:= \lim_{t\to\infty} \mathbb{E}\left(N(t)\ (N(t)-1)\cdots\ (N(t)-n+1)\right), \\
e_{n,1} &:= \lim_{t\to\infty} \mathbb{E}\left(N(t)\ (N(t)-1)\cdots\ (N(t)-n+1)\cdot Y(t)\right), \\
e_{n,0} &:= \lim_{t\to\infty} \mathbb{E}\left(N(t)\ (N(t)-1)\cdots\ (N(t)-n+1)\cdot (1-Y(t))\right)
\end{aligned}$$

can be derived from the generating functions.

Theorem 5 *The factorial moments are given by*

$$\begin{aligned}
e_n =& \frac{\left(\frac{g}{\nu}\right)_n}{\left(\frac{g+\kappa}{\nu}\right)_n} \left(\frac{\mu}{\nu}\right)^n \\
e_{n,1} =& \frac{\left(\frac{g}{\nu}\right)_{n+1}}{\left(\frac{g+\kappa}{\nu}\right)_{n+1}} \left(\frac{\mu}{\nu}\right)^n \\
e_{n,0} =& \frac{\kappa}{\kappa+g}\cdot\frac{\left(\frac{g}{\nu}\right)_n}{\left(\frac{g+\kappa}{\nu}+1\right)_n} \left(\frac{\mu}{\nu}\right)^n
\end{aligned}$$

Proof: The factorial moments e_n and $e_{n,0}$ are easily computed from the relations

$$e_n = \frac{d^n}{dz^n} G_N(z)\Big|_{z=1} \quad \text{and} \quad e_{n,0} = \frac{d^n}{dz^n}\beta(z)\Big|_{z=1}.$$

Furthermore,

$$\begin{aligned}
e_{n,1} = e_n - e_{n,0} =& \frac{\left(\frac{g+\kappa}{\nu}+n-\frac{g+\kappa}{\nu}\cdot\frac{\kappa}{\kappa+g}\right)}{\left(\frac{g+\kappa}{\nu}\right)_{n+1}} \left(\frac{g}{\nu}\right)_n \left(\frac{\mu}{\nu}\right)^n \\
=& \frac{\left(\frac{g}{\nu}\right)_{n+1}}{\left(\frac{g+\kappa}{\nu}\right)_{n+1}} \left(\frac{\mu}{\nu}\right)^n.
\end{aligned}$$

□

Let us write explicitely some quantities of interest that result from these relations. First the probability to be ON is asymptotically

$$\lim_{t\to\infty} P(Y(t)=1) = \lim_{t\to\infty} \mathbb{E}(Y(t)) = e_{0,1} = \frac{g}{g+\kappa},$$

and the mean number of proteins is

$$\lim_{t\to\infty} \mathbb{E}(N(t)) = e_1 = \frac{\mu}{\nu}\ \frac{g}{g+\kappa}.$$

The first quantity can also be deduced from equations (9.4) or (9.5) and its interpretation is the asymptotic probability to be ON on the simple two-state Markov process of switching from ON to OFF with rate κ and from OFF to ON with rate g.

When the promoter is ON, the system behaves like a simple birth and death process with birth rate μ and death rate $\nu\ n$. The asymptotical mean of such a process is given by $\frac{\mu}{\nu}$, see for example Feller (1968) p. 461. The quantity $\lim_{t\to\infty} \mathbb{E}(N(t))$ can thus be interpreted as the mean number of proteins if the system were always in state ON times the fraction of time the system is in state ON.
The covariance between the number of proteins and the state of the promoter is

$$\begin{aligned}\lim_{t\to\infty} \mathrm{Cov}(N(t),Y(t)) =& e_{11} - e_1\cdot e_{01} = \frac{\mu}{\nu}\left(\frac{g(g+\nu)}{(g+\kappa)(g+\kappa+\nu)} - \left(\frac{g}{g+\kappa}\right)^2\right)\\ =& \frac{\mu}{\nu}\ \frac{g(g+\nu)(g+\kappa) - g^2(g+\kappa+\nu)}{(g+\kappa)^2(g+\kappa+\nu)}\\ =& \frac{\mu g\kappa}{(g+\kappa)^2(g+\kappa+\nu)}.\end{aligned}$$

Similarily for the variance of the number of proteins

$$\begin{aligned}\lim_{t\to\infty} \mathrm{Var}(N(t)) =& e_2 - e_1^2 + e_1\\ =& \left(\frac{\mu}{\nu}\right)^2\ \frac{\frac{g}{\nu}}{\left(\frac{g+\kappa}{\nu}\right)}\ \left(\frac{\left(\frac{g}{\nu}+1\right)}{\left(\frac{g+\kappa}{\nu}+1\right)} - \frac{\left(\frac{g}{\nu}\right)}{\left(\frac{g+\kappa}{\nu}\right)}\right) + \frac{\mu}{\nu}\ \frac{\frac{g}{\nu}}{\left(\frac{g+\kappa}{\nu}\right)}\\ =& \frac{\kappa g}{(g+\kappa)^2}\ \frac{\mu^2}{\nu(g+\kappa+\nu)} + \frac{\mu g}{\nu\ (g+\kappa)}.\end{aligned}$$

The variance to mean ratio of the number of proteins is bigger than 1, hence the process is over-dispersed, more precisely

$$\lim_{t\to\infty} \frac{\mathrm{Var}(N(t))}{\mathbb{E}(N(t))} = 1 + \frac{\kappa}{g+\kappa}\ \frac{\mu}{g+\kappa+\mu}.$$

Notice that the above quantity happens to be

$$\lim_{t\to\infty} \frac{\mathrm{Var}(N(t))}{\mathbb{E}(N(t))} = 1 + \lim_{t\to\infty} P(Y(t)=1)\cdot \mathrm{Cov}(N(t),Y(t)),$$

but we do not see any intuitive explanation for this relation.

The partial variances $\lim_{t\to\infty} \mathrm{Var}(N(t)Y(t)) = e_{2,1}-e_{1,1}^2+e_{1,1}$ and $\lim_{t\to\infty} \mathrm{Var}(N(t)(1-Y(t))) = e_{2,0} - e_{1,0}^2 + e_{1,0}$ can also be obtained as complicated fractions of the parameters.

Remark 1 *The case without feedback is of particular importance since it can be used to model a system where the promoter is not self-regulated but regulated by another control substance, present in much higher quantities and thus approximated deterministically, see Chapter 11.*

9.3 LINEAR POSITIVE FEEDBACK

In a similar way, one can find the generating functions when the positive feedback is linear, i. e.

$$a_3(n, y) = (g_0 + g_1\ n)\ (1 - y).$$

Here we will see the limitations of this method. The partial generating function β and the generating function G_N can again be written in terms of hypergeometric functions, but here things become more complicated.

At equilibrium, equations (9.1) and (9.2) are

$$\begin{aligned} 0 =& \mu\ (z-1)\alpha(z) + (\nu - \nu\ z)\frac{d}{dz}\alpha(z) + g_0\ \beta(z) + z g_1 \frac{d}{dz}\beta(z) - \kappa\ \alpha(z), \\ 0 =& (\nu - \nu\ z - g_1\ z)\frac{d}{dz}\beta(z) + \kappa\alpha(z) - g_0\beta(z). \end{aligned}$$

Again, summing these equations yields

$$0 = \mu\ (z-1)\alpha(z) + \nu\ (1-z)\frac{d}{dz}(\alpha(z) + \beta(z)),$$

but now the equation for α is

$$\alpha(z) = \frac{(\nu + g_1)z - \nu}{\kappa}\frac{d}{dz}\beta(z) + \frac{g_0}{\kappa}\beta(z), \tag{9.9}$$

and combining these equations and dividing by $-\frac{\nu(z-1)}{\kappa\ (g_1+\nu)}$ yields after some algebra

$$\begin{aligned} 0 =& \left(z - \frac{\nu}{g_1+\nu}\right)\frac{d^2}{dz^2}\beta(z) + \left(\frac{g_0+\kappa}{g_1+\nu} + 1 - \frac{\mu}{\nu}\left(z - \frac{\nu}{g_1+\nu}\right)\right)\frac{d}{dz}\beta(z) \\ & - \frac{\mu\ g_0}{\nu\ (g_1+\nu)}\beta(z). \end{aligned}$$

With the change of variable $w = \frac{\mu}{\nu}\left(z - \frac{\nu}{g_1+\nu}\right)$ and $\gamma(w) = \beta\left(\frac{\nu}{\mu}w + \frac{\nu}{g_1+\nu}\right)$, the above equation can be written after dividing by $\frac{\mu}{\nu}$ in the canonical form of Kummer's equation

$$0 = w\frac{d^2}{dw^2}\gamma(w) + \left(\frac{g_0+\kappa}{g_1+\nu} + 1 - w\right)\frac{d}{dw}\gamma(w) - \frac{g_0}{g_1+\nu}\gamma(w),$$

like in the case without feedback. Hence $\beta(z)$ is a linear combination of

$${}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu} + 1, \frac{\mu}{\nu}\left(z - \frac{\nu}{g_1+\nu}\right)\right)$$

and

$$U\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}+1, \frac{\mu}{\nu}\left(z-\frac{\nu}{g_1+\nu}\right)\right).$$

Since U has a pole at $z = \frac{\nu}{g_1+\nu} < 1$, the coefficient multiplying U in the combination has to be 0 and

$$\beta(z) = c_1\ {}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}+1, \frac{\mu}{\nu}\left(z-\frac{\nu}{g_1+\nu}\right)\right), \tag{9.10}$$

with c_1 a constant to be determined from the condition $\alpha(1)+\beta(1)=1$. Equation (9.9) gives

$$\begin{aligned}\alpha(z) =& c_1\frac{\mu((\nu+g_1)z-\nu)g_0}{\nu\ \kappa\ (g_0+g_1+\kappa+\nu)}\ {}_1F_1\left(\frac{g_0}{g_1+\nu}+1, \frac{g_0+\kappa}{g_1+\nu}+2, \frac{\mu}{\nu}\left(z-\frac{\nu}{g_1+\nu}\right)\right)\\ &+ c_1\frac{g_0}{\kappa}\ {}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}+1, \frac{\mu}{\nu}\left(z-\frac{\nu}{g_1+\nu}\right)\right).\end{aligned} \tag{9.11}$$

Since the generating functions are now series in $\left(z-\frac{\nu}{g_1+\nu}\right)$ and no more in $(z-1)$, the normalizing constant and the expressions for the factorial moments are more complicated, yielding power series in $\left(1-\frac{\nu}{g_1+\nu}\right)$. However, we can give a closed form for the generating function of the number of proteins $G_N(z) = \alpha(z)+\beta(z)$ quite similar to the case without feedback.

Theorem 6 *The generating function of the number of proteins at equilibrium*

$$G_N(z) := \lim_{t\to\infty} \mathbb{E}(z^{N(t)})$$

and the partial generating functions $\alpha(z)$ and $\beta(z)$ can be written as the hypergeometric functions

$$\begin{aligned}G_N(z) =& \frac{{}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}, \frac{\mu}{\nu}\left(z-\frac{\nu}{g_1+\nu}\right)\right)}{{}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}, \frac{\mu}{\nu}\left(1-\frac{\nu}{g_1+\nu}\right)\right)},\\ \alpha(z) =& \frac{{}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}, \frac{\mu}{\nu}\left(z-\frac{\nu}{g_1+\nu}\right)\right) - \frac{\kappa}{g_0+\kappa}\ {}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}+1, \frac{\mu}{\nu}\left(z-\frac{\nu}{g_1+\nu}\right)\right)}{{}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}, \frac{\mu}{\nu}\left(1-\frac{\nu}{g_1+\nu}\right)\right)},\\ \beta(z) =& \frac{\kappa}{g_0+\kappa}\ \frac{{}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}+1, \frac{\mu}{\nu}\left(z-\frac{\nu}{g_1+\nu}\right)\right)}{{}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}, \frac{\mu}{\nu}\left(1-\frac{\nu}{g_1+\nu}\right)\right)}.\end{aligned}$$

Proof: From equations (9.10) and (9.11), we get

$$\begin{aligned}\frac{G_N(z)}{c_1} =& \frac{g_0+\kappa}{\kappa}{}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}+1, \frac{\mu}{\nu}\left(z-\frac{\nu}{g_1+\nu}\right)\right)\\ &+\frac{\mu\ ((\nu+g_1)z-\nu)\ g_0}{\nu\ \kappa\ (g_0+\kappa+g_1+\nu)}{}_1F_1\left(\frac{g_0}{g_1+\nu}+1, \frac{g_0+\kappa}{g_1+\nu}+2, \frac{\mu}{\nu}\left(z-\frac{\nu}{g_1+\nu}\right)\right).\end{aligned}$$

To simplify the notations, let us write

$$a = \frac{g_0}{g_1+\nu}, \quad b = \frac{g_0+\kappa}{g_1+\nu} \quad \text{and} \quad w = \frac{\mu}{\nu}\left(z-\frac{\nu}{g_1+\nu}\right),$$

and the last equation above become

$$\frac{G_N(z)}{c_1} = \frac{b}{b-a}\ {}_1F_1(a, b+1, w) + \frac{w\ a}{(b-a)\ (b+1)}\ {}_1F_1(a, b+2, w).$$

The following relations between hypergeometric functions hold, see Abramowitz & Stegun (1964) p. 507–508,

$$\begin{aligned}\frac{a\ w}{b+1}\ {}_1F_1(a+1, b+2, w) =& (w-b)\ {}_1F_1(a, b+1, w) + b\ {}_1F_1(a-1, b, w),\\ w\ {}_1F_1(a, b+1, w) =& b\ {}_1F_1(a, b, w) - b\ {}_1F_1(a-1, b, w),\end{aligned}$$

and thus

$$\begin{aligned}\frac{G_N(z)}{c_1}\ (b-a) &= b\ {}_1F_1(a, b+1, w) + \frac{w\ a}{b+1}\ {}_1F_1(a, b+2, w)\\ &= b\ {}_1F_1(a, b+1, w) + (w-b)\ {}_1F_1(a, b+1, w) + b\ {}_1F_1(a-1, b, w)\\ &= w\ {}_1F_1(a, b+1, w) + b\ {}_1F_1(a-1, b, w)\\ &= b\ {}_1F_1(a, b, w).\end{aligned}$$

With $G_N(1) = 1$,

$$c_1 = \frac{b-a}{b\ {}_1F_1\left(a, b, \frac{\mu}{\nu}\left(1 - \frac{\nu}{g_1+\nu}\right)\right)},$$

and from equation (9.10) we get

$$\beta(z) = \frac{(b-a)}{b}\ \frac{{}_1F_1\left(a, b, w\right)}{{}_1F_1\left(a, b, \frac{\mu}{\nu}\left(1 - \frac{\nu}{g_1+\nu}\right)\right)}.$$

Concerning α, we simply use the fact that $\alpha(z) = G_N(z) - \beta(z)$. □

The asymptotic distributions of the quantities of interest

$$\pi_{(n,\ \cdot\ )} = \lim_{t\to\infty} P(N(t) = n) \quad \text{and} \quad \pi_{(n,y)} = \lim_{t\to\infty} P((N(t), Y(t)) = (n, y))$$

can easily be computed by taking the derivatives at $z = 0$ of the corresponding generating functions and we have

$$\begin{aligned}\pi_{(n,\ \cdot\ )} =& \frac{\left(\frac{g_0}{g_1+\nu}\right)_n}{\left(\frac{g_0+\kappa}{g_1+\nu}\right)_n}\ \left(\frac{\mu}{\nu}\right)^n\ \frac{{}_1F_1\left(\frac{g_0}{g_1+\nu}+n, \frac{g_0+\kappa}{g_1+\nu}+n, -\frac{\mu}{g_1+\nu}\right)}{{}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}, \frac{\mu}{\nu}\left(1 - \frac{\nu}{g_1+\nu}\right)\right)},\\ \pi_{(n,0)} =& \frac{\kappa}{\kappa+g_0}\ \frac{\left(\frac{g_0}{g_1+\nu}\right)_n}{\left(\frac{g_0+\kappa}{g_1+\nu}+1\right)_n}\ \left(\frac{\mu}{\nu}\right)^n\ \frac{{}_1F_1\left(\frac{g_0}{g_1+\nu}+n, \frac{g_0+\kappa}{g_1+\nu}+n+1, -\frac{\mu}{g_1+\nu}\right)}{{}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}, \frac{\mu}{\nu}\left(1 - \frac{\nu}{g_1+\nu}\right)\right)},\\ \pi_{(n,1)} =& \pi_{(n,\ \cdot\ )} - \pi_{(n,0)}.\end{aligned}$$

In a similar way, but now taking derivatives at $z = 1$, the factorial moments e_n, $e_{n,0}$ and $e_{n,1}$ defined in the preceding section 9.2 are given by

$$e_n = \frac{\left(\frac{g_0}{g_1+\nu}\right)_n}{\left(\frac{g_0+\kappa}{g_1+\nu}\right)_n} \left(\frac{\mu}{\nu}\right)^n \frac{{}_1F_1\left(\frac{g_0}{g_1+\nu}+n, \frac{g_0+\kappa}{g_1+\nu}+n, \frac{\mu}{\nu}\left(1-\frac{\nu}{g_1+\nu}\right)\right)}{{}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}, \frac{\mu}{\nu}\left(1-\frac{\nu}{g_1+\nu}\right)\right)},$$

$$e_{n,0} = \frac{\kappa}{\kappa+g_0} \frac{\left(\frac{g_0}{g_1+\nu}\right)_n}{\left(\frac{g_0+\kappa}{g_1+\nu}+1\right)_n} \left(\frac{\mu}{\nu}\right)^n \frac{{}_1F_1\left(\frac{g_0}{g_1+\nu}+n, \frac{g_0+\kappa}{g_1+\nu}+n+1, \frac{\mu}{\nu}\left(1-\frac{\nu}{g_1+\nu}\right)\right)}{{}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}, \frac{\mu}{\nu}\left(1-\frac{\nu}{g_1+\nu}\right)\right)},$$

$$e_{n,1} = e_n - e_{n,0}.$$

Now the probability to be OFF is

$$\pi_{(\,\cdot,\,0)} = \lim_{t\to\infty} P(Y(t)=0) = \frac{\kappa}{\kappa+g_0} \frac{{}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}+1, \frac{\mu}{\nu}\left(1-\frac{\nu}{g_1+\nu}\right)\right)}{{}_1F_1\left(\frac{g_0}{g_1+\nu}, \frac{g_0+\kappa}{g_1+\nu}, \frac{\mu}{\nu}\left(1-\frac{\nu}{g_1+\nu}\right)\right)},$$

and expressions for the variances and covariances can also be obtained in the same manner but they are tedious and we do not write them explicitely.

9.4 LIMITATIONS OF THE METHOD OF GENERATING FUNCTIONS

Although it provides a powerful tool for analytic description, the method of generating functions is very particular in the sense that a little change in the form of one of the feedback propensity function can induce major changes in the generating function, and for each particular propensity function one has to derive the whole set of equations anew.

Furthermore, an explicit form for the generating function can only be found when the feedback propensity functions are simple, either constant or linear in the protein numbers. In practice, the feedback propensity functions are related to the number of sites in the promoter on which the proteins bind, either directly in monomer form or in more complicated bound forms like dimers or higher order polymers.

In the case of monomers, the propensity functions are polynomial in the number of proteins with order equal to the number of sites of the promoter. The equations for the generating functions contain derivatives of the same order, hence in the case of order two and higher, the generating functions, given the equations can be solved, are no longer simple hypergeometric functions.

When the feedback is given by bound forms of the protein, the intermediate reaction has to be taken into account. Usually the typical time scale of such reactions are much more rapid than the other reactions in the system and we can apply the technique also used in the slow-scale stochastic simulation, see Cao *et al.* (2005b) and Cao *et al.* (2005a). The detailed validation of this accelerated technique is discussed in Zhang & Yin (1997), Yin & Zhang (2002) and Yin & Zhang (2005). The feedback propensity functions are in this case given by mean-field approximations of a polynom of the bound form of the protein for a given protein number, usually a function of the protein number present in the system that can not be written in a closed form, and hence the equations

for the generating functions can not be solved. We will discuss in details the case of dimers in Chapter 17.

In Chapter 10 we provide a flexible method to compute the invariant measure for arbitrary feedback propensity functions.

10

INVARIANT MEASURE AND THE METHOD OF TRANSFER MATRICES

10.1 INTRODUCTION

As discussed in Section 9.4, the method of generating functions is only adapted for very particular academic examples. In practice, the feedback propensity functions are usually much more complicated. In this section we develop a flexible method that allows the use of biologically more revelant feedback loops.
For this method, we will use the limitation that the state space is finite, i. e. there is a maximal number of proteins Λ that can not be surpassed. This restriction leads to significant mathematical simplifications, and is biologically meaningful taking into account the finite volume of a cell. Moreover, this simplication can not be avoided for practical computation since we will have to find a left eigenvector at the boundary, and it can be showed that the solutions of the system restricted to the finite space on $\{0, 1, \ldots, \Lambda\} \times \{0, 1\}$ converges to the invariant distribution of the system on the space $\mathbb{N} \times \{0, 1\}$ under suitable asumptions, see Pasquier (2008) for details.

We recall that in this case the propensity function a_1 at the boundary is $a_1(\Lambda, y) = 0$ and the CME (7.2) for $p_{(\Lambda,1)}$ reads

$$\begin{aligned}\frac{d}{dt}\, p_{(\Lambda,1)}(t) =& y\mu\; p_{(\Lambda-1,y)}(t) + g(n)\; p_{(\Lambda,0)}(t) \\ &- (\nu(\Lambda) + \kappa(\Lambda))\; p_{(\Lambda,1)}(t),\end{aligned}$$

and $p_{(n,y)}(t) \equiv 0$ for $n > \Lambda$.
Since every state is recurrent, the finiteness of the state space ensures that every state is positive recurrent and we know from the usual theory of continuous-time Markov chains that the Kolmogorv forward and backward systems are equivalent and that there exist an invariant distribution, see for example Brémaud (1999), p. 338, and Norris (1997), p. 123.

Let us write the row vector with $2(\Lambda + 1)$ components

$$p(t) := (p_{(0,0)}(t), p_{(0,1)}(t), p_{(1,0)}(t)p_{(1,1)}(t), p_{(2,0)}(t), \ldots, p_{(\Lambda,0)}(t), p_{(\Lambda,1)}(t)).$$

The CME (7.2) can be written in compact form as

$$\frac{d}{dt}p(t) = p(t)Q, \tag{10.1}$$

where Q is the generator matrix

$$Q = \begin{bmatrix} R_0 & U & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ Q_1 & R_1 & U & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 0 & Q_n & R_n & U & 0 & \cdots & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & Q_{\Lambda-1} & R_{\Lambda-1} & U \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & Q_\Lambda & R_\Lambda \end{bmatrix},$$

with the 2×2-matrices $U = \begin{bmatrix} 0 & 0 \\ 0 & \mu \end{bmatrix}$, $D_n = \begin{bmatrix} \nu(n) & 0 \\ 0 & \nu(n) \end{bmatrix}$ and

$$R_n = \begin{bmatrix} -(g(n) + \nu(n)) & g(n) \\ \kappa(n) & -(\kappa(n) + \nu(n) + \mu) \end{bmatrix}$$

for $1 \leq n \leq \Lambda - 1$, and the boundaries $R_0 = \begin{bmatrix} -g(0) & g(0) \\ \kappa(0) & -(\kappa(0) + \mu) \end{bmatrix}$,

$$D_\Lambda = \begin{bmatrix} \nu(\Lambda) & 0 \\ 0 & \nu(\Lambda) \end{bmatrix} \text{ and } R_\Lambda = \begin{bmatrix} -(g(\Lambda) + \nu(\Lambda)) & g(\Lambda) \\ \kappa(\Lambda) & -(\kappa(\Lambda) + \nu(\Lambda)) \end{bmatrix}.$$

The transient behaviour of the system can be found to be

$$p(t) = p(0)e^{Qt},$$

by integrating equation (10.1), with $p(0)$ the initial distribution that we do not specify here. Due to the size and shape of the matrix Q, this representation does not help much for practical computation.

Both for theoretical and computational purposes, it would be very helpful to find an invariant distribution of the time-comtinuous Markov process. The above exponential representation is not satisfactory, since the eigenvalues of Q are not obvious. In the following, we present a method based on transfer matrices, quite similar to the method presented in Fournier *et al.* (2007) and based on an idea developed in Bolthausen & Goldsheid (2000).

Let us rethink our state space as a succession of layers of the form $\{(n, 0), (n, 1)\}$, $0 \leq n \leq \Lambda$, define $p_n(t) = (p_{(n,0)}(t), p_{(n,1)}(t))$ the probability vector of a layer and rewrite equation (10.1) layerwise as

$$\begin{aligned} \frac{d}{dt}p_0(t) =& p_0(t)R_0 + p_1(t)D_1, \\ \frac{d}{dt}p_n(t) =& p_{n-1}(t)U + p_n(t)R_n + p_{n+1}(t)D_{n+1}, \quad 0 < n < \Lambda, \\ \frac{d}{dt}p_\Lambda(t) =& p_{\Lambda-1}(t)U + p_\Lambda(t)R_\Lambda. \end{aligned}$$

To solve this equation seems impossible, or at least the form of the solution should not be simpler than the exponential form given above. However, an explicit solution can be found for the invariant distribution.

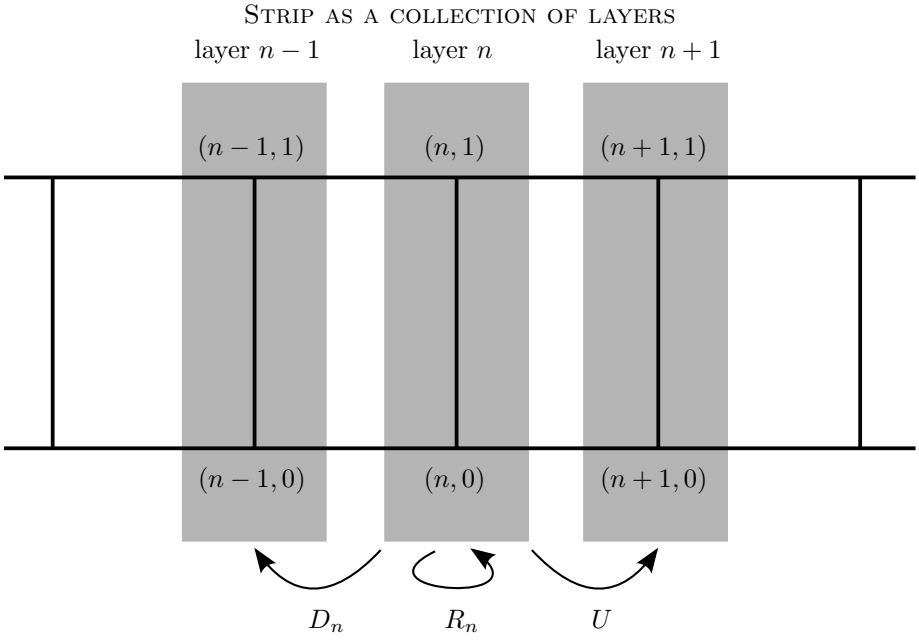

Figure 10.1: The layerwise vizualisation of the strip. The matrices D_n, R_n and U are defined in the main text.

10.2 TRANSFER MATRICES AND INVARIANT MEASURE

As said before, the invariant distribution exists and is unique since every state is recurrent and the state space is finite.

Let $\pi = (\pi_{(0,0)}, \pi_{(0,1)}, \pi_{(1,0)}, \pi_{(1,1)}, \pi_{(2,0)}, \dots, \pi_{(\Lambda,0)}, \pi_{(\Lambda,0)})$ be this invariant distribution, and let us write it layerwise as $\pi_n = (\pi_{(n,0)}, \pi_{(n,1)})$, as in Figure 10.1. The distribution π solves the above equation with the left hand side set to 0, i. e.

$$\begin{aligned} 0 =& \pi_0 R_0 + \pi_1 D_1, && (10.2)\\ 0 =& \pi_{n-1} U + \pi_n R_n + \pi_{n+1} D_{n+1}, \quad 0 < n < \Lambda, && (10.3)\\ 0 =& \pi_{\Lambda-1} U + \pi_\Lambda R_\Lambda. && (10.4) \end{aligned}$$

The idea is to look for transfer matrices α_n such that $\pi_n = \pi_{n+1}\alpha_n$, and to find a suitable first vector π_Λ at the boundary to compute the π_n recursively. Plugging the relation $\pi_n = \pi_{n+1}\alpha_n$ into equations (10.2) and (10.3),

$$\begin{aligned} 0 =& \pi_0 R_0 + \pi_1 D_1,\\ 0 =& \pi_{n-1} U + \pi_n R_n + \pi_{n+1} D_{n+1} = \pi_n(\alpha_{n-1} U + R_n) + \pi_{n+1} D_{n+1}, \quad 0 < n < \Lambda, \end{aligned}$$

the transfer matrices matrices α_n have to satisfy

$$\begin{aligned} \alpha_0 =& - D_1 R_0^{-1},\\ \alpha_n =& - D_{n+1}(\alpha_{n-1} U + R_n)^{-1}, \quad 0 < n < \Lambda, && (10.5) \end{aligned}$$

provided that the matrices R_0 and $\alpha_{n-1}U+R_n$ are invertible, $0<n<\Lambda$. In the following theorem, we show that this is the case, solve the matrix valued continuous fraction of equations (10.5) and provide a closed form for the invariant distribution π of the time-continuous Markov process.

Theorem 7 *The invariant distribution of the time-continuous Markov process* $\{X(t)=(N(t),Y(t))\}_{t>0}$ *restricted on the strip* $\{0,1,\ldots,\Lambda\}\times\{0,1\}$ *is*

$$\pi_n=\frac{w_\Lambda\alpha_{\Lambda-1}\alpha_{\Lambda-2}\cdots\alpha_n}{Z_\Lambda}\ \textit{for}\ 0\le n<\Lambda,\ \textit{and}\ \pi_\Lambda=\frac{w_\Lambda}{Z_\Lambda},\tag{10.6}$$

with $w_\Lambda=(\kappa(\Lambda),g(\Lambda)+\nu(\Lambda))$, *the transfer matrices* α_n *given by*

$$\alpha_n=\frac{\nu(n+1)}{\mu}\begin{bmatrix}\frac{\kappa(n)+\mu}{g(n)+\nu(n)} & 1\\ \frac{\kappa(n)}{g(n)+\nu(n)} & 1\end{bmatrix},\ 0<n<\Lambda,\quad \alpha_0=\frac{\nu(1)}{\mu}\begin{bmatrix}\frac{\kappa(0)+\mu}{g(0)} & 1\\ \frac{\kappa(0)}{g(0)} & 1\end{bmatrix},$$

and the normalization constant

$$Z_\Lambda=w_\Lambda\cdot(1,1)^T+\sum_{j=0}^{\Lambda-1}w_\Lambda\alpha_{\Lambda-1}\alpha_{\Lambda-2}\cdots\alpha_j\cdot(1,1)^T.$$

Proof: Let us define recursively $w_n=w_{n+1}\alpha_n$, starting from $n=\Lambda-1$, and show that the equations (10.2), (10.3) and (10.4) are satisfied for the vectors w_n, $0\le n\le\Lambda$. For $0<n<\Lambda$, the matrices $\alpha_{n-1}U+R_n$ are invertible since

$$\begin{aligned}\alpha_{n-1}U+R_n&=\begin{bmatrix}0 & \nu(n)\\ 0 & \nu(n)\end{bmatrix}+\begin{bmatrix}-(g(n)+\nu(n)) & g(n)\\ \kappa(n) & -(\kappa(n)+\nu(n)+\mu)\end{bmatrix}\\&=\begin{bmatrix}-(g(n)+\nu(n)) & g(n)+\nu(n)\\ \kappa(n) & -(\kappa(n)+\mu)\end{bmatrix}\end{aligned}$$

with determinant $\mu\ (g(n)+\nu(n))>0$, and

$$(\alpha_{n-1}U+R_n)^{-1}=\frac{-1}{\mu\ (g(n)+\nu(n))}\begin{bmatrix}\kappa(n)+\mu & g(n)+\nu(n)\\ \kappa(n) & g(n)+\nu(n)\end{bmatrix},$$

thus they satisfy

$$\alpha_n=-D_{n+1}\left(\alpha_{n-1}U+R_n\right)^{-1}.$$

At the boundary $n=0$, α_0 is simply $-D_1R_0^{-1}$. Hence the transfer matrices α_n, $0\le n<\Lambda$, solve the matrix valued continuous fraction (10.5).

Plugging w_Λ and $w_{\Lambda-1}=w_\Lambda\alpha_{\Lambda-1}$ in the right hand side of (10.4), w_Λ has to be a left eigenvector of the matrix $\alpha_{\Lambda-1}U+R_\Lambda$ for the eigenvalue 0. This is the case since $w_\Lambda=(\kappa(\Lambda),g(\Lambda)+\nu(\Lambda))$ and

$$\alpha_{\Lambda-1}U+R_\Lambda=\begin{bmatrix}-(g(\Lambda)+\nu(\Lambda)) & g(\Lambda)+\nu(\Lambda)\\ \kappa(\Lambda) & -\kappa(\Lambda)\end{bmatrix}.$$

The vectors

$$w_\Lambda,\ w_{\Lambda-1}=w_\Lambda\cdot\alpha_{\Lambda-1},\ \ldots,\ w_n=w_\Lambda\cdot\alpha_{\Lambda-1}\cdots\alpha_n,\ \ldots,\ w_0=w_\Lambda\cdot\alpha_{\Lambda-1}\cdots\alpha_0$$

thus solve the system (10.2), (10.3) and (10.4). Moreover, each vector has only strictly positive components and the vector

$$w=(w_{(0,0)},w_{(0,1)},w_{(1,0)},w_{(1,1)},w_{(2,0)},\ldots,w_{(\Lambda,0)},w_{(\Lambda,0)})$$

where $(w_{(n,0)}, w_{(n,1)}) = w_n$ is an invariant measure of the process, and thus proportional to the unique invariant distribution π, i. e. $\pi = \frac{w}{Z_\Lambda}$ with the normalization factor $Z_\Lambda = \sum_{n=0}^{\Lambda} w_n \cdot (1,1)^T$. □

Formula (10.6) provides a closed form of the invariant measure that is not adapted to numerical computation since for large Λ both the numerator and denominator rapidly diverge.

10.3 NORMALIZATION ALGORITHM

To allow numerical computation we will use the following normalization algorithm. Let $S = \{(0,1) + t(1,-1), t \in [0,1]\}$ denote the line segment between $(0,1)$ and $(1,0)$ and $\|\cdot\|$ the 1-norm $\|w\| := w \cdot (1,1)^T$. The idea is to introduce without computational costs the vectors $\tilde{v}_n$ on S that will permit to easily recover the invariant measure. The algorithm consists in three simple steps:

STEP 1: Define $\tilde{v}_n$ for $n = \Lambda - 1$ to 0 as

$$\tilde{v}_\Lambda := \frac{w_\Lambda}{\|w_\Lambda\|}, \quad \text{and} \quad \tilde{v}_n := \frac{\tilde{v}_{n+1}\alpha_n}{\|\tilde{v}_{n+1}\alpha_n\|}.$$

STEP 2: Given the $\tilde{v}_n$, define $v_0 = \tilde{v}_0$ and, for $n = 1$ to Λ, set

$$v_n := \frac{\tilde{v}_n}{\|\tilde{v}_n\alpha_{n-1}\| \cdot \|\tilde{v}_{n-1}\alpha_{n-2}\| \cdots \|\tilde{v}_1\alpha_0\|}.$$

STEP 3: Compute the steady state distribution as

$$\pi_n = \frac{v_n}{V_\Lambda}, \qquad \text{where} \quad V_\Lambda := \sum_{i=0}^{\Lambda} v_i \cdot (1,1)^T.$$

It immediately results from their definition that the vectors $\tilde{v}_n$ and v_n have the properties:

- For all $0 \le n \le \Lambda$, $\|\tilde{v}_n\| = 1$, i. e. each $\tilde{v}_n$ lies on the line segment $S \subset \mathbb{R}^2$ between the points $(0,1)$ and $(1,0)$.
- $\tilde{v}_{n+2}\alpha_{n+1}\alpha_n = \tilde{v}_{n+1}\alpha_n \cdot \|\tilde{v}_{n+2}\alpha_{n+1}\|$ yields

  $$v_n = \frac{\tilde{v}_\Lambda \alpha_{\Lambda-1}\alpha_{\Lambda-2}\cdots\alpha_n}{\|\tilde{v}_\Lambda\alpha_{\Lambda-1}\| \cdot \|\tilde{v}_{\Lambda-1}\alpha_{\Lambda-2}\| \cdots \|\tilde{v}_{n+1}\alpha_n\| \cdot \|\tilde{v}_n\alpha_{n-1}\| \cdots \|\tilde{v}_1\alpha_0\|}.$$

  The denominator of the above expression is independent of n and $\tilde{v}_\Lambda$ is proportional to w_Λ. Hence v_n is proportional to the invariant measure π_n,

  $$\pi_n = \frac{v_n}{V_\Lambda}, \qquad \text{where} \quad V_\Lambda := \sum_{i=0}^{\Lambda} v_i \cdot (1,1)^T.$$

- The 1-norm of v_n is $\|v_n\| = \dfrac{\|\tilde{v}_n\|}{\|\tilde{v}_n\alpha_{n-1}\| \cdots \|\tilde{v}_1\alpha_0\|} = \dfrac{1}{\|\tilde{v}_n\alpha_{n-1}\| \cdots \|\tilde{v}_1\alpha_0\|}$.

Theorem 8 below provides conditions under which the normalization constant V_Λ remains bounded as Λ is large. The function $\nu(n)$ gives the monomer degradation rates for n proteins, and is assumed to be increasing with $\nu(0) = 0$, and strictly positive for $n \geq 1$. Usually, $\nu(n)$ is taken to be a constant times n, here we assume the less restrictive condition that $\inf_{n\geq 1} \nu(n)/n$ is stricly positive to allow situations where for example proteins that are present in bound forms (dimer, trimer,...) can not be degradated, or situations where $\nu(n)/n \to \infty$ as $n \to \infty$.

To prove this Theorem, we use the following Lemma:

Lemma 2 *If $\inf_{n\geq 1} \nu(n)/n$ is stricly positive, there exists $k > 0$ depending only on μ, ν, κ, g (and not on Λ) such that for all $n \geq 1$, $\|\tilde{v}_n \alpha_{n-1}\| \geq nk$.*

Proof: Each $\tilde{v}_j$ lies in S and depends on Λ. To break this dependence, we prove the results for an arbitrary vector $v = (t, 1-t) \in S$, $t \in [0, 1]$.

$$v\alpha_n = \frac{\nu(n)}{\mu}(\star, t + (1-t)) = \frac{\nu(n)}{\mu}(\star, 1),$$

with $\star > 0$ for all n. Hence, uniformly in S,

$$\frac{\|v\alpha_n\|}{n} \geq \frac{\nu(n)}{n\mu} \geq \inf_{n\geq 1} \frac{\nu(n)}{n\mu} =: k > 0.$$

□

Hence we can give bounds uniformly in Λ:

Theorem 8 *If $\inf_{n\geq 1} \nu(n)/n$ is stricly positive, there exists $M > 0$, depending only on μ, ν, κ, g (and not on Λ), such that*

$$1 \leq V_\Lambda = \sum_{n=0}^{\Lambda} \|v_n\| \leq M.$$

Proof: Notice that $\|v_0\| = \|\tilde{v}_0\| = 1$. Moreover Lemma 2 yields

$$V_\Lambda = 1 + \sum_{n=1}^{\Lambda} \|v_n\| \leq 1 + \sum_{n=1}^{\infty} \frac{k^{-n}}{n!} = e^{1/k} =: M.$$

□

This Theorem shows that our algorithm is well suited for computational purposes, and permits to show with a tightness argument that the unbounded chain defined on $\mathbb{N} \times \{0, 1\}$ is ergodic with the invariant distribution given by the limit as $\Lambda \to \infty$ of the above distribution, for details see Pasquier (2008).

A computer implementation of the algorithm in Matlab is provided in Section C.1 of Appendix C.

11

THE PROPENSITY FUNCTIONS

11.1 PRODUCTION AND DEGRADATION RATES

The protein production propensity function, $a_1(n, y) = \mu \cdot y$ in the unbounded case or $a_1(n, y) = \mu \cdot y \cdot 1_{\{n<\Lambda\}}$ in the bounded case, is here in the ON state a constant combining the transcription and translation phases, see Appendix A Sections A.2 and A.3 for more details.

In the bounded case, it is interesting to note that the method of transfer matrices also works, with a slight modification, when the production propensity function depends on n. Suppose that $a_1(n, y) = \mu(n) \cdot y$, equation (10.5) has to be changed in

$$\begin{aligned} \alpha_0 &= -D_1 R_0^{-1}, \\ \alpha_n &= -D_{n+1}(\alpha_{n-1}U_{n-1} + R_n)^{-1}, \quad 0 < n < \Lambda, \end{aligned} \tag{11.1}$$

with the matrice U being changed to

$$U_{n-1} = \begin{bmatrix} 0 & 0 \\ 0 & \mu(n-1) \end{bmatrix}.$$

It is easy to see, mimicking the proof of Theorem 7, that the sequence of matrices

$$\alpha_n = \frac{\nu(n+1)}{\mu(n)} \begin{bmatrix} \frac{\kappa(n)+\mu(n)}{g(n)+\nu(n)} & 1 \\ \frac{\kappa(n)}{g(n)+\nu(n)} & 1 \end{bmatrix}, \ 0 < n < \Lambda, \quad \alpha_0 = \frac{\nu(1)}{\mu(0)} \begin{bmatrix} \frac{\kappa(0)+\mu(0)}{g(0)} & 1 \\ \frac{\kappa(0)}{g(0)} & 1 \end{bmatrix},$$

solves the recurrence equation (11.1), and the invariant measure can be computed in the same way as in Chapter 10, with the condition of Theorem 8 adapted to $\inf_{n\geq 1} \dfrac{\nu(n)}{n \cdot \mu(n)}$ to be strictly positive.
The case of a production propensity being a function of n could be used in the case where the transcription or translation phase is affected by the protein number, for example by a saturation effect.

The protein degradation propensity function $\nu(n)$ is usually modeled as proportional to the protein number, $\nu(n) = \nu \cdot n$ for a degradation rate ν. The algorithm presented in Chapter 10 can however be used for a large variety of functions, the only requirement being that $\inf_{n\geq 1} \dfrac{\nu(n)}{n}$ is strictly positive.
Usually the degradation propensity function is supposed to be linear, $\nu(n) = \nu\, n$, but when the produced proteins bound in a further reaction to multimers, the degradation can also be modeled as a degradation rate times the number of monomers instead of the number of proteins, as in Chapter 12.

11.2 FEEDBACK REGULATED BY EXTERNAL FACTORS

When the protein that binds to the promoter is not the protein synthetized by the genetic module but an external factor considered at a constant concentration, the promoter dynamics can be modeled as governed by a constant feedback. The computational and analytic methods presented in Section 9.2 can be used, but it is often more convenient for practical computation to use the algorithm developed in Chapter 10 even in this simple case.

11.3 SELF REGULATED GENE

The negative feedback case has been studied extensively in the literature with simple propensity functions, see for example Hornos *et al.* (2005). Here we suppose it to be constant in all the examples studied.

We are mainly interested in the autocatalytic case and try to show what can possibly happen in the presence of positive feedback, for example bimodal invariant distributions.

12

POSITIVE FEEDBACK THROUGH DIMERS

12.1 DESCRIPTION

In this Chapter, we discuss the special case where positive feedback is provided through the intermediate reaction of dimerization. Biologically, dimers bind highly cooperatively to activate the promoter. Here the protein is a GAL4 type protein and the promoter has 5 binding sites, similar to the engineered module studied experimentally in Imhof *et al.* (2000).

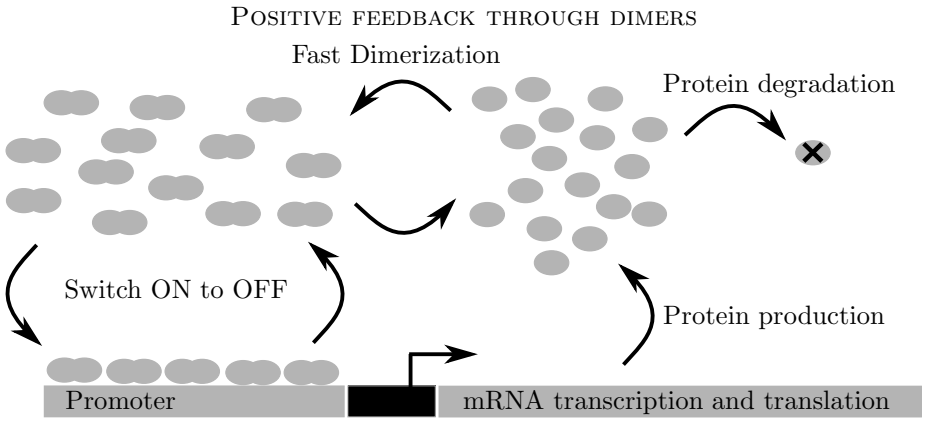

Figure 12.1: Schematical description of the case where the module is auto-regulated by dimers. Dimers bind and unbind cooperatively to the promoter to launch transcription, the produced protein can form dimers or can be degradated. The reactions involved in the dimerization process are supposed to be much faster than the other.

Mathematically, the whole system does not correspond yet to the genetic module presented in this Part. The state space is much bigger since it has not only to take into account the number of proteins and state of the promoter, but also the number of dimers. The dimerization is often thought to be a much faster reaction than the other reactions in the system, thus we will call the corresponding time-continuous Markov chain $\{X^\varepsilon(t)\}_{t\geq 0}$ where $X^\varepsilon(t) = (N^\varepsilon(t), Y^\varepsilon(t), D^\varepsilon(t))$ and ε emphasizes the fact that the propensity function governing the changes in $D^\varepsilon(t)$ are of order $1/\varepsilon$ where ε is near 0.

More precisely, the evolution of $N^\varepsilon(t)$ and $Y^\varepsilon(t)$ is the same as before, except that the positive feedback propensity function is now governed by the number of dimers $D^\varepsilon(t)$. The feedback function contains two distinct parts. The first one is the basal activity $g_0 > 0$, reflecting the fact that even in the absence of protein, the promoter can possibly become active and start a transcriptional burst. The second one accounts for the likelihood that dimers are bound to the promoter. Since dimers are thought to bind and unbind highly cooperatively, we use a stochastic Hill type model and the contribution to the feedback propensity function is then proportional to the number of dimers to the power 5.

We restrict to the bounded state space

$$(N^\varepsilon(t), Y^\varepsilon(t), D^\varepsilon(t)) \in \{0, 1, \dots, \Lambda\} \times \{0, 1\} \times \{0, 1, \dots, \lfloor \Lambda/2 \rfloor\},$$

where $\lfloor \Lambda/2 \rfloor$ is the integer part of $\Lambda/2$. The reactions and propensity functions can be summarized as follows, with $\mathcal{M}$ denoting a protein in monomer form and $\mathcal{D}$ in dimer form:

Summary of the process $\{X^\varepsilon(t)\}_{t\geq 0}$		
Reaction R_j	Propensity function a_j	Change vector v_j
$R_1 : \emptyset \to \mathcal{P}$	$a_1(n, y, d) = y\mu \cdot 1_{\{n<\Lambda\}}$	$v_1 = (1, 0, 0)$
$R_2 : \mathcal{P} \to \emptyset$	$a_2(n, y, d) = \nu(n - 2d)$	$v_2 = (-1, 0, 0)$
$R_3 : \mathcal{O}_0 \to \mathcal{O}_1$	$a_3(n, y, d) = (g_1\ d^5 + g_0)(1 - y)$	$v_3 = (0, 1, 0)$
$R_4 : \mathcal{O}_1 \to \mathcal{O}_0$	$a_4(n, y, d) = \kappa\ y$	$v_4 = (0, -1, 0)$
$R_5 : 2\mathcal{M} \to \mathcal{D}$	$a_5^\varepsilon(n, y, d) = \frac{c_+}{\varepsilon}(n - 2d)(n - 2d - 1)$	$v_5 = (0, 0, 1)$
$R_6 : \mathcal{D} \to 2\mathcal{M}$	$a_6^\varepsilon(n, y, d) = \frac{c_-}{\varepsilon} d$	$v_6 = (0, 0, -1)$

Some authors use a factor $\frac{1}{2}$ in a_5^ε for combinatorial reasons, here we include it in c_+ for notational simplicity. For more details on the dimerization process we refer to Part V.

12.2 ELIMINATION OF THE FAST VARIABLE

When some propensity functions are much bigger than the other, the corresponding reactions occur much more frequently than the reactions corresponding to slower propensity functions. It is often the case that the slow variables are responsible for the most important features of the system, while the faster one are secondary. In stochastic simulation, consequences can be very unpleasant, since the trajectory almost always jump between states of the fast variables, while

time does not move forward and the simulation is stuck and do not reach the prescripted ending time in a reasonable computational time. Cao *et al.* (2005b) developed a theory for the slow-scale stochastic simulation, mainly based on heuristic arguments, while Zhang & Yin (1997) provides the rigorous probabilistic framework to handle the issue.

We first translate the heuristic argument of Cao *et al.* (2005b) in our case. The propensity functions in the original system involving the fast variable are $a_2(n, y, d) = \nu(n - 2d)$ and $a_3(n, y, d) = (g_1\ d^5 + g_0)(1 - y)$. Let Δt be a time small compared to the typical reaction time of a slow variable but huge compared to the reaction time of the fast variable. If the slow variables are in state $N^\varepsilon(t) = n$ and $Y^\varepsilon(t) = y$, the reactions to occur between t and $t + \Delta t$ are very likely to be all reactions of the fast variable, and a sensible choice for an aggregated propensity function for the slow process taking into account the fluctuations of the fast variable would be to take the average value of $a_2(n, y, D^\varepsilon(s))$ and $a_3(n, y, D^\varepsilon(s))$ between t and $t + \Delta t$. If we consider Δt to be large enough compared to the typical reaction time of the fast variable, it can be well approximated by $\mathbb{E}_n(a_2(n, y, D^\varepsilon(\infty)))$ and $\mathbb{E}_n(a_3(n, y, D^\varepsilon(\infty)))$, where $D^\varepsilon(\infty)$ is a random variable distributed with the equilibrium distribution of the dimer number with n proteins and $\mathbb{E}_n$ denotes the mean under this distribution. In our case $\mathbb{E}_n(a_2(n, y, D^\varepsilon(\infty))) = \nu(n-2\mathbb{E}_n)$ and $\mathbb{E}_n(a_3(n, y, D^\varepsilon(\infty))) = (g_1\ \mathbb{E}_n^5+g_0)(1-y)$, where $\mathbb{E}_n^5$ is the fifth moment of the dimer number at equilibrium for a fixed amount n of proteins.

Let us now write things formally to show that for small ε, the above approximation can be used. Since we are only interested in the variables $N^\varepsilon(t)$ and $Y^\varepsilon(t)$, let us define $\{\bar{X}^\varepsilon(t)\}_{t\geq 0}$ the marginal process $\bar{X}^\varepsilon(t) = (N^\varepsilon(t), Y^\varepsilon(t))$ only keeping track of the slow variables and $\{\bar{X}(t)\}_{t\geq 0}$ the time continuous Markov process $\bar{X}(t) = (\bar{N}(t), \bar{Y}(t))$ governed by the propensity functions summarized in the table below. Notice that $\{\bar{X}^\varepsilon(t)\}_{t\geq 0}$ is not Markovian.

Summary of the process $\{\bar{X}(t)\}_{t\geq 0}$		
Reaction R_j	Propensity function $\bar{a}_j$	Change vector v_j
$R_1 : \emptyset \ \rightarrow \mathcal{P}$	$\bar{a}_1(\bar{n}, \bar{y}) = \bar{y}\mu \cdot 1_{\{\bar{n}<\Lambda\}}$	$v_1 = (1, 0)$
$R_2 : \mathcal{P} \ \rightarrow \emptyset$	$\bar{a}_2(\bar{n}, \bar{y}) = \nu(\bar{n} - 2\mathbb{E}_{\bar{n}})$	$v_2 = (-1, 0)$
$R_3 : \mathcal{O}_0 \ \rightarrow \mathcal{O}_1$	$\bar{a}_3(\bar{n}, \bar{y}) = (g_1\ \mathbb{E}_{\bar{n}}^5 + g_0)(1 - \bar{y})$	$v_3 = (0, 1)$
$R_4 : \mathcal{O}_1 \ \rightarrow \mathcal{O}_0$	$\bar{a}_4(\bar{n}, \bar{y}) = \kappa\ \bar{y}$	$v_4 = (0, -1)$

In the above table, $\mathbb{E}_{\bar{n}}$ and $\mathbb{E}_{\bar{n}}^5$ denote respectivelly the mean and the fifth moment of the dimer number for a fixed amount of $\bar{n}$ proteins at equilibrium. The propensity functions $\bar{a}_2$ and $\bar{a}_3$ can be computed using the methods of Sections

18.3 and 18.4.

Theorem 2. 3 in Zhang & Yin (1997) states that the process $\{\bar{X}^{\varepsilon}(t)\}_{t\geq 0}$ converges in distribution to the Markov process $\{\bar{X}(t)\}_{t\geq 0}$ as $\varepsilon \to 0$. The process $\{\bar{X}(t)\}_{t\geq 0}$ is a special case of the genetic module studied throughout this Part and hence its exact invariant distribution can easily be computed using the methods of Chapter 10. Moreover, since the dimerization is supposed to be much faster than the other reactions, it provides a good approximation to the quantity of interest, namely the equilibrium distribution of the non Markovian process $\{\bar{X}^{\varepsilon}(t)\}_{t\geq 0}$.

12.3 COMPUTATIONAL RESULTS

In Figure 12.2, we show a special parameter regime where the invariant distribution exhibits bimodality, both for the probability to be ON and for the marginal probability of the protein number.

The somehow disturbing behaviour for low protein number is due to the

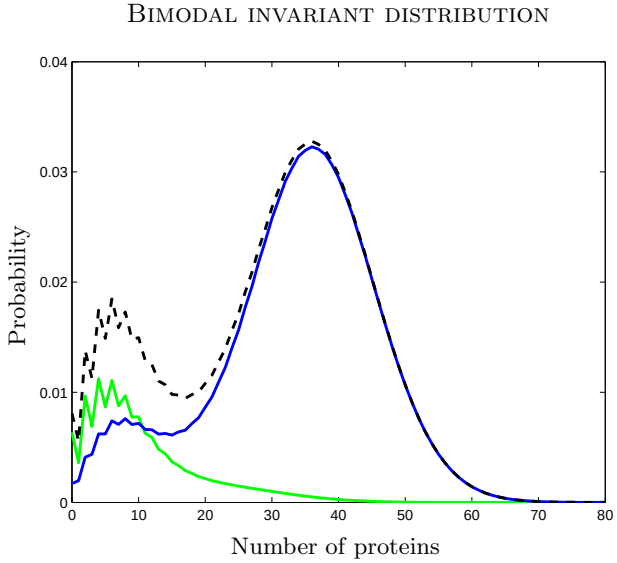

Figure 12.2: Invariant measure of the process $\{\bar{X}(t)\}_{t\geq 0}$ with the following parameters: $\mu = 550$, $\nu = 170$, $g_1 = 0.01$, $g_0 = 150$, $\kappa = 200$, $c_+/c_- = 5/3$. The green curve represents the invariant measure in the OFF state $\pi_n(0)$, the blue curve in the ON state $\pi_n(1)$ and the black curve is the sum of the two other curves or in other words the marginal probability of the number of proteins.

fact that in this model we only take into account the degradation of protein monomers and suppose that the dimers can not be degradated, hence for $\bar{n} \leq 20$ the function $\nu(\bar{n} - 2\mathbb{E}_{\bar{n}})$ is not increasing as would be expected from a degradation propensity function. The positive feedback propensity function

$$\bar{a}_3(\bar{n}, \bar{y}) = g_1 \mathbb{E}_{\bar{n}}^5 + g_0$$

increases very rapidly and we plot it in semilogarithmic scale.

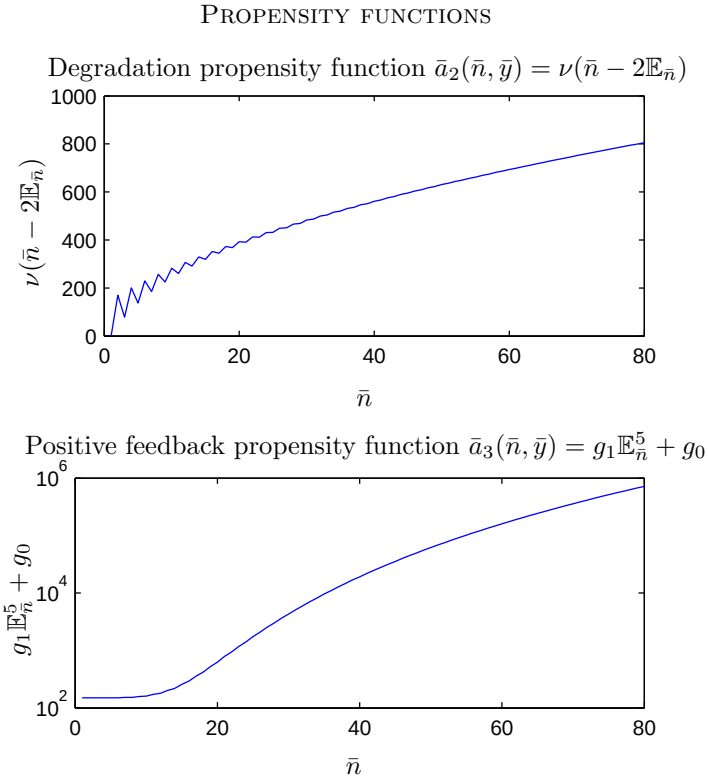

Figure 12.3: The propensity functions $\bar{a}_2(\bar{n},\bar{y})$ and $\bar{a}_3(\bar{n},\bar{y})$, parameters are those of Figure 12.2.

Figure 12.4 shows the simulated asymptotic empirical distribution of the non Markovian process $\{\bar{X}^\varepsilon(t)\}_{t\geq 0}$, in other words the distribution of

$$\lim_{t\to\infty}(N^\varepsilon(t), Y^\varepsilon(t)),$$

for various values of ε, the parameter that controls the speed of the fast dimerization. As explained above, the simulation of the process is computationally very intensive since most of the simulated reactions only affect the fast variable $D^\varepsilon(t)$ and we are mainly interested in the slow variables $N^\varepsilon(t)$ and $Y^\varepsilon(t)$. The values of ε are 10^{-2} and 10^{-4}. For each value, simulation was performed over 100 trajectories and about 10^6 stochastic events, from wich we stored the last 10^4 with a change in N^ε or Y^ε in at least one of the trajectories to compute the empirical marginal distribution of $\{\bar{X}^\varepsilon(t)\}_{t\geq 0}$. Simulation for $\varepsilon = 10^{-2}$ took a few minutes while for $\varepsilon = 10^{-4}$ it took about ten hours, and with a result that is not quite satisfactory. The respective empirical distributions are compared with the exact distribution of the fast approximation, whose computation took less than one second. Parameters are the same as in Figure 12.2.

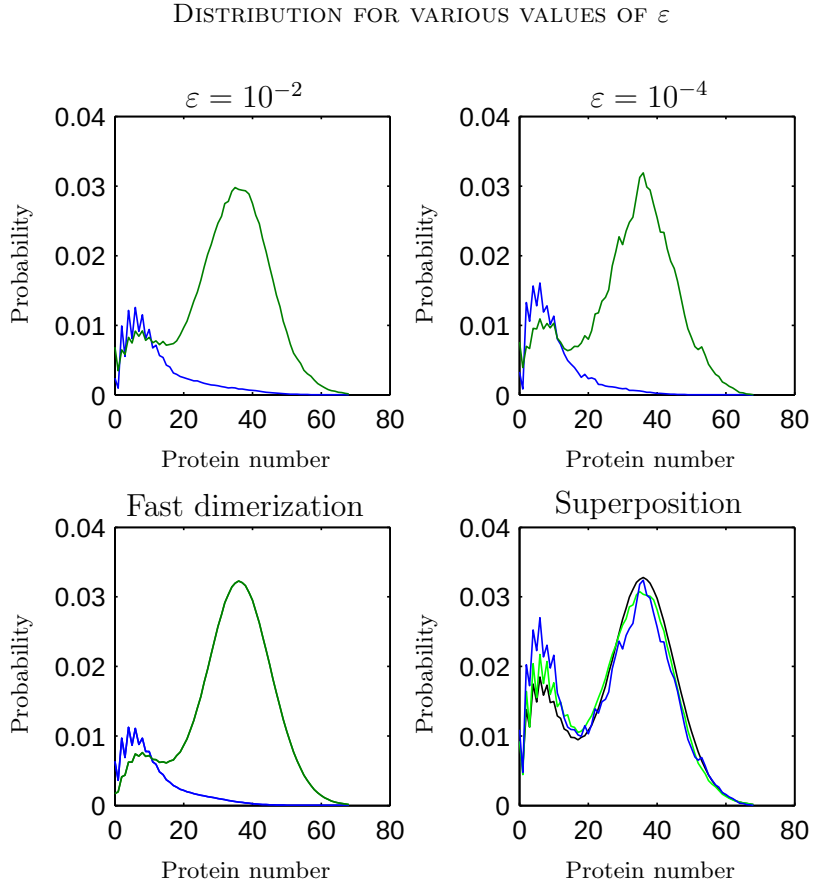

Figure 12.4: Comparison of the distribution of $(N(t), Y(t))$, with parameters other as ε being those of Figure 12.2. The upper line shows the empirical distribution of the simulation of the full process for $\varepsilon = 10^{-2}$ (left) and $\varepsilon = 10^{-4}$ (right). The figure at the bottom left represent the computed invariant measure with the quasi-equilibrium fast approximation. In these three plots, the green curve is the ON regime and the blue one the OFF regime. When $\varepsilon = 10^{-2}$, the rates of the dimerization are commensurate with the other rates and there is no reason to do a fast dimerization approximation. However, the simulation shows a quite similar behaviour as in the fast approximation. When $\varepsilon = 10^{-2}$, simulation took about ten hours and the result seems very poor, indeed even the simulation with $\varepsilon = 10^{-2}$ shows better agreement with the fast approximation. The computation of the steady-state distribution for the fast approximation required less than a second. The last figure on the bottom right displays the marginal probability of the protein number, for $\varepsilon = 10^{-2}$ in green, $\varepsilon = 10^{-4}$ in blue and in the fast approximation in black.

13

SEMI-STOCHASTIC COUPLING AND TIME DELAY

Throughout this Chapter, we consider the unbounded state space $E = \mathbb{N} \times \{0, 1\}$.

13.1 SEMI-STOCHASTICITY

In their works on the epidemic of schistosomiasis, Nasell & Hirsch (1972) and Nasell & Hirsch (1973) propose a mathematical model with semi-stochastic coupling. The authors consider a Markov chain quite similar to the one of our model except that they replace in the infinitesimal evolution equations every external random variables by their mean value.

If we adapt it to our genetic module, the variable $N(t)$ is replaced by its mean value in the equations concerning the change in the promoter state while the variable $Y(t)$ in the equations concernig a change in the protein number. The evolution equations (7.1) become

$$
\begin{aligned}
&P(X(t+\Delta t) = (n+1, y) \mid X(t) = (n, y)) = \mathbb{E}(Y(t))\mu\Delta t + o(\Delta t),\\
&P(X(t+\Delta t) = (n-1, y) \mid X(t) = (n, y)) = \nu(n)\Delta t + o(\Delta t),\\
&P(X(t+\Delta t) = (n, 1-y) \mid X(t) = (n, y)) =\\
&\qquad\qquad (y\kappa\left(\mathbb{E}(N(t))\right) + (1-y)g\left(\mathbb{E}(N(t))\right)\Delta t + o(\Delta t),\\
&P(X(t+\Delta t) = (n, y) \mid X(t) = (n, y)) =\\
&\qquad\qquad 1 - (\nu(n) + y\mu + y\kappa(\mathbb{E}(N(t)) + (1-y)g(\mathbb{E}(N(t)))\Delta t + o(\Delta t),\\
&P(X(t+\Delta t) = \text{anything else} \mid X(t) = (n, y)) = o(\Delta t).
\end{aligned}
$$

The resulting process is now a time-nonhomogeneous Markov process. It can be shown that the pair $(\mathbb{E}(N(t)), \mathbb{E}(Y(t)))$ converges to a limit

$$(\mathbb{E}(N(\infty)), \mathbb{E}(Y(\infty))) \in \mathbb{R}_+ \times (0, 1),$$

and that in the case where the degradation function is linear in n, $\nu(n) = \nu\, n$, the stationary distribution is Poisson with parameter $\dfrac{\mathbb{E}(Y(\infty))\ \mu}{\nu}$, hence the mean and variance are asymptotically $\dfrac{\mathbb{E}(Y(\infty))\ \mu}{\nu}$. It has been shown experimentally and theoretically that when the mean gene expression is increasing as a function of some inducer, the variance can exhibit a peak, see for example Pedraza & van Oudenaarden (2005). This kind of behaviour can not be predicted qualitatively when the process is Poisson since in this case mean and variance are the same.

As discussed in Section 5.1, a condition for the system with $Y(t)$ replaced by

$\mathbb{E}(Y(t))$ to be similar to the original system is fast promoter kinetics. The role of the promoter is very particular since a single stochastic event like a transition from OFF to ON can fire a very important global event, here a transcriptional burst. For these reasons a semi-stochastic approach in the sense of Nasell & Hirsch (1972) is not appropriate for our model and we will hence conserve the external variable $Y(t)$ as stochastic and replace only $N(t)$ by its average when involved as an external variable. With a slight misuse of notation, the word *semi-stochastic* will refer in the following to this second approach.

13.2 TIME DELAY

In a recent paper, Goutsias & Kim (2006) considered a semi-stochastic model in the gene regulation setting, but did not involve the promoter in their analysis, and introduced biologically meaningful delays in feedback interactions. In other words, they replaced $\kappa(N(t))$ and $g(N(t))$ by expressions involving their expected values with some delay θ, namely $\kappa(\mathbb{E}(N(t-\theta)))$ and $g(\mathbb{E}(N(t-\theta)))$. The biological significance of the time delay can be thought as follows: proteins move around at random and need some time to reach the neighborhood of the promoter, and θ might represent the average time needed. In our setting, when the promoter dynamics is not taken into account, the limiting behaviour of $N(t)$ is again Poisson and the model do not predict accurately the propagation of noise in gene expression levels.

We will therefore modify the evolution equations (7.1) in the semi-stochastic setting with time delay for interactions with $N(t)$ as an external variable. Our new Markov process $\{X^{\text{sstd}}(t)\}_{t\geq 0}$, with $X^{\text{sstd}}(t) = (N(t), Y(t))$, evolves according to

$$
\begin{aligned}
&P(X^{\text{sstd}}(t+\Delta t) = (n+1,y) \mid X^{\text{sstd}}(t) = (n,y)) = Y(t)\mu\Delta t + o(\Delta t),\\
&P(X^{\text{sstd}}(t+\Delta t) = (n-1,y) \mid X^{\text{sstd}}(t) = (n,y)) = \nu(n)\Delta t + o(\Delta t),\\
&P(X^{\text{sstd}}(t+\Delta t) = (n,1-y) \mid X^{\text{sstd}}(t) = (n,y)) =\\
&\qquad\qquad (y\kappa\left(\mathbb{E}(N(t-\theta))\right) + (1-y)g\left(\mathbb{E}(N(t-\theta))\right))\,\Delta t + o(\Delta t),\\
&P(X^{\text{sstd}}(t+\Delta t) = (n,y) \mid X^{\text{sstd}}(t) = (n,y)) =\\
&\quad 1 - (\nu(n) + y\mu + y\kappa\left(\mathbb{E}(N(t-\theta))\right) + (1-y)g\left(\mathbb{E}(N(t-\theta))\right))\Delta t + o(\Delta t),\\
&P(X^{\text{sstd}}(t+\Delta t) = \text{anything else} \mid X^{\text{sstd}}(t) = (n,y)) = o(\Delta t).
\end{aligned}
$$

The propensity functions are now time nonhomogeneous, and additional difficulties arise concerning the asymptotical behaviour of the process.

13.3 TIME-NONHOMOGENEOUS MARKOV PROCESSES

In the semi stochastic approach with time delay, the propensity functions concerning the promoter $a_3(n,y)$ and $a_4(n,y)$ are replaced by functions depending on the time and the state of the promoter, for simplicity we will call them $a_3(t,y) = g(\mathbb{E}(N(t-\theta)))\,(1-y)$ and $a_4(t,y) = \kappa(\mathbb{E}(N(t-\theta)))\,y$. If $g(\mathbb{E}(N(t-\theta)))$ and $\kappa(\mathbb{E}(N(t-\theta)))$ converge to a limit g_∞ respectivelly κ_∞, we define the process $\{X^0(t)\}_{t\geq 0}$ to be the time homogeneous Markov process generated by the

same propensity functions as $\{X^{\text{sstd}}(t)\}_{t\geq 0}$ except that $a_3(t,y)$ and $a_4(t,y)$ are replaced by their limit $a_3(\infty,y) = g_\infty\ (1-y)$ and $a_4(\infty,y) = \kappa_\infty\ y$. Suppose that $\{X^0(t)\}_{t\geq 0}$ possesses the invariant distribution π. One can naively think that in this case the invariant distribution of $\{X^{\text{sstd}}(t)\}_{t\geq 0}$ is automatically π. Indeed, the behaviour of a time-nonhomogeneous Markov process can be more surprising. If the convergence to g_∞ or κ_∞ is not fast enough, it is still possible that π is not the invariant distribution of $\{X^{\text{sstd}}(t)\}_{t\geq 0}$.

In Theorem 15 of Appendix B, we cite a criterion from Abramov & Liptser (2004) ensuring convergence. This criterion entails a condition that holds automatically in our case and a further condition on the speed of convergence of $g(\mathbb{E}(N(t-\theta)))$ and $\kappa(\mathbb{E}(N(t-\theta)))$, namely that the integrals

$$\int_0^\infty (\sqrt{g(\mathbb{E}(N(t-\theta)))} - \sqrt{g_\infty})^2 dt \quad \text{and} \quad \int_0^\infty (\sqrt{\kappa(\mathbb{E}(N(t-\theta)))} - \sqrt{\kappa_\infty})^2 dt$$

are finite. In the following, we will discuss a particuliar case of interest where the criterion holds, allowing to compute the steady state distribution of $\{X^{\text{sstd}}(t)\}_{t\geq 0}$ by only considering the much simpler time homogeneous process $\{X^0(t)\}_{t\geq 0}$.

13.4 THE CASE OF LINEAR POSITIVE FEEDBACK

In this section, we focus on the case where the autocatalytic propensity function $a_3(t,y) = g(\mathbb{E}(N(t-\theta)))\ (1-y)$ is given by an affine function g and the negative feedback function is constant. We will prove the result with the slightly less restrictive condition that the function g is continuously differentiable, increasing over $\mathbb{R}_+$ and such that $g(0) > 0$ and $g(x)/x$ is decreasing, allowing possibly for other type of positive feedback. We suppose the degradation propensity function to be linear in n, $a_2(n,y) = \nu\ n$. The time-nonhomogeneous Markov process $\{X^{\text{sstd}}(t)\}_{t\geq 0}$ describing the process is summarized in the following table.

Summary of the process $\{X^{\text{sstd}}(t)\}_{t\geq 0}$		
Reaction R_j	Propensity function a_j	Change vector v_j
$R_1 : \emptyset \ \to \mathcal{P}$	$a_1(n,y) = y\mu$	$v_1 = (1,0)$
$R_2 : \mathcal{P} \ \to \emptyset$	$a_2(n,y) = \nu n$	$v_2 = (-1,0)$
$R_3 : \mathcal{O}_0 \ \to \mathcal{O}_1$	$a_3(t,y) = g(\mathbb{E}(N(t-\theta)))\ (1-y)$	$v_3 = (0,1)$
$R_4 : \mathcal{O}_1 \ \to \mathcal{O}_0$	$a_4(t,y) = \kappa\ y$	$v_4 = (0,-1)$

As discussed in Section 13.3, $g(\mathbb{E}(t-\theta))$ has to converge to g_∞ fast enough to ensure that the equilibrium distribution of the process equals π, the equilibrium distribution of the simpler process $\{X^0(t)\}_{t\geq 0}$, more precisely that the condition

$$\int_0^\infty (\sqrt{g(\mathbb{E}(N(t-\theta)))} - \sqrt{g_\infty})^2 dt < \infty \tag{13.1}$$

is satisfied. To show that this is the case, we first discuss step by step the behaviour of $\mathbb{E}(N(t-\theta))$, $t \geq 0$. To simplify the notation, we will write

$$E(t) := \mathbb{E}(N(t)) \quad \text{and} \quad G(t) := \mathbb{E}(Y(t)),$$

whose evolution equations (8.1) for the mean and (8.2) for the probability to be ON become

$$\begin{aligned} \frac{dE(t)}{dt} &= \mu G(t) - \nu E(t), \\ \frac{dG(t)}{dt} &= g(E(t-\theta))(1-G(t)) - \kappa G(t). \end{aligned} \tag{13.2}$$

In Lemma 3, we prove that the evolution equations defining the system are well defined, providing a unique solution, then we show in Lemma 4 that the system converges to the unique biologically meaningful critical point of the system and in Lemma 5 that the speed of convergence is exponential. The methods used are adapted from Gabriel *et al.* (1981). Finally, using our hypothesis on the function g, it is easy to conclude that the condition (13.1) holds, and the main result is stated in Theorem 9.

The theory of delayed differential equations is very different from the usual theory of differential equations, here the initial condition is no more a point in the finite dimensional space $\mathbb{R}_+ \times [0,1]$ but a continuous nonegative function $E(t)$ over the interval $[-\theta, 0]$ and a value $G(0) \in [0,1]$. To solve the system (13.2), we have to first integrate the second equation over the interval $[0,\theta]$, then plug the solution in the first equation and integrate using the variation of constant over the interval $[0,\theta]$ and begin the whole procedure anew over the interval $[\theta, 2\theta]$ with initial condition given by $E(t)$ over the interval $[0,\theta]$, and so on.

Lemma 3 *Existence and unicity*
For any initial condition $E(t)$ non-negative and continuous over $[-\theta, 0]$ and $0 \leq G(0) \leq 1$, there exists a unique solution of the system (13.2) defined over $[0, +\infty)$. Furthermore,

$$0 < E(t) \leq \max\{E(0), \mu/\nu\},\ t \geq 0,$$

and

$$0 < G(t) < 1,\ t > 0.$$

Proof: For any initial condition $0 \leq G(0) \leq 1$ and $E(t)$ non-negative and continuous over $[-\theta, 0]$, (13.2) admits obviously a unique solution over $[0,\theta]$. If $G(0) > 0$, then by continuity G remains strictly positive over some open intervall to the right of 0. If $G(0) = 0$, then according to the second equation of (13.2), $\dot{G}(0) > 0$ and the same conclusion holds. The same reasoning shows that $G < 1$ over some open intervall to the right of 0. Clearly, if they exist, $t_0 = \inf\{t \in (0,\theta],\ G(t) = 0\}$ and $t_1 = \inf\{t \in (0,\theta],\ G(t) = 1\}$ are both strictly positive. By definition, $\dot{G}(t_0) \leq 0$ and by continuity, $G(t_0) = 0$. The second equation of (13.2) entails $\dot{G}(t_0) > 0$. We have a similar contradiction for t_1, thus $0 < G(t) < 1$ over $(0,\theta]$. The variation of constant formula entails

$$0 < E(t) \leq \max\{E(0), \mu/\nu\} \text{ over } (0,\theta].$$

Iterating the procedure provides existence and unicity of a solution defined over $[0, +\infty)$ and the preceding inequalities are preserved. □

We are interested in the possible equilibria of (13.2) in $\mathbb{R}_+ \times [0,1]$, i.e. the solutions (E_0, G_0) in $\mathbb{R}_+ \times [0,1]$ of

$$0 = \mu G_0 - \nu E_0, \quad 0 = g(E_0)(1 - G_0) - \kappa G_0.$$

Clearly $G_0 = 0$ and $G_0 = 1$ lead to contradictions. We thus have $0 < G_0 < 1$ and consequently $E_0 > 0$. Plugging $G_0 = \frac{\nu}{\mu} E_0$ in the second equation yields

$$\frac{g(E_0)}{E_0}(\frac{\mu}{\nu} - E_0) = \kappa.$$

If $g(0) > 0$ and $\frac{g(E)}{E}$ is decreasing over $(0, +\infty)$, then $\frac{g(E)}{E}(\frac{\mu}{\nu} - E)$ is strictly decreasing. Since it starts at $+\infty$ and becomes ultimately negative, we conclude to the existence of a unique solution $(E_0, G_0) \in \mathbb{R}^2 \times [0,1]$.

We will use the fluctuation Lemma 9 given in Appendix B to prove the convergence to the critical point (E_0, G_0).

Lemma 4 *Convergence*
For any initial condition $E(t)$ non-negative and continuous over $[-\theta, 0]$ and $0 \leq G(0) \leq 1$, the unique solution $(E(t), G(t))$ converges to (E_0, G_0) as $t \to \infty$.

Proof: The fluctuation Lemma 9 and the monotonicity of g imply that

$$\begin{aligned} 0 &\geq \mu \underline{G} - \nu \underline{E}, \qquad 0 \geq g(\underline{E})(1 - \underline{G}) - \kappa \underline{G} \\ 0 &\leq \mu \overline{G} - \nu \overline{E}, \qquad 0 \leq g(\overline{E})(1 - \overline{G}) - \kappa \overline{G}. \end{aligned} \tag{13.3}$$

We prove the last inequality to exemplify the method. We choose $t_n \uparrow +\infty$ so that $G(t_n) \to \overline{G}$ and $\dot{G}(t_n) \to 0$ as $n \to +\infty$. Since the sequence $\{E(t_n)\}_{n \geq 1}$ is bounded, there exists a subsequence $(t_{n_k})_{k \geq 1}$ so that $E(t_{n_k})$ converges as $k \to \infty$ to a certain value that we call E_∞. Evaluating the equation for G over the subsequence $(t_{n_k})_{k \geq 1}$ and letting $k \to +\infty$, we get

$$0 = g(E_\infty)(1 - \overline{G}) - \kappa \overline{G} \leq g(\overline{E})(1 - \overline{G}) - \kappa \overline{G}$$

since g is increasing and $\overline{G} \leq 1$. The proof of the other inequalities in (13.3) is similar.
We already know that $0 < \underline{G}, \underline{E} > 0$ and $\overline{G} < 1$, and (13.3) entails $\underline{G} \leq \frac{\nu}{\mu} \underline{E}$ and $\overline{G} \geq \frac{\nu}{\mu} \overline{E}$, and in particular $\overline{E} \leq \frac{\mu}{\nu} \overline{G} < \frac{\mu}{\nu}$. Consequently

$$\frac{\kappa \nu}{\mu} \overline{E} \leq \kappa \overline{G} \leq g(\overline{E})(1 - \overline{G}) \leq g(\overline{E})(1 - \frac{\nu}{\mu} \overline{E}),$$

hence

$$\kappa \leq \frac{g(\overline{E})}{\overline{E}}(\frac{\mu}{\nu} - \overline{E}).$$

Repeating the same argument for $\underline{E}$, one gets

$$\kappa \geq \frac{g(\underline{E})}{\underline{E}}(\frac{\mu}{\nu} - \underline{E}).$$

By assumption, $\frac{g(E)}{E}(\frac{\mu}{\nu} - E)$ is decreasing, so that the two last equations then give that $\overline{E} \leq E_0$ and $\underline{E} \geq E_0$. Clearly we have $\underline{E} = \overline{E} = E_0$, so that $E(t)$ converges as $t \to +\infty$. According to Lemma 3.1 in Coppel (1965), $E(t)$ and its first two derivatives being bounded on $[\theta, +\infty)$, we have $\lim_{t\to+\infty} \dot{E}(t) = 0$ and the relation $\dot{E} = \mu G - \nu E$ entail the convergence of $G(t)$ as $t \to +\infty$. □

From this Lemma, we deduce that the propensity function

$$a_3(t, y) = g(\mathbb{E}(N(t-\theta)))\ y$$

converges to $g_\infty\ y = g(E_0)\ y$ as $t \to \infty$. To show that the convergence speed is exponential, we use Theorem 16 cited in Appendix B.

Lemma 5 *Exponential convergence*
The convergence of $(E(t), G(t))$ *to* (E_0, G_0) *is exponential.*

Proof: Near a critical point, the asymptotic behaviour of the system is determined by the asymptotic behaviour of the linearized system

$$\begin{bmatrix} \dot{E}(t) \\ \dot{G}(t) \end{bmatrix} = A \cdot \begin{bmatrix} E(t) \\ G(t) \end{bmatrix} + B \cdot \begin{bmatrix} E(t-\theta) \\ G(t-\theta) \end{bmatrix},$$

where A and B are the matrices

$$A = \begin{bmatrix} -\nu & \mu \\ 0 & -(g(E_0)+\kappa) \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 \\ g'(E_0)(1-G_0) & 0 \end{bmatrix}.$$

We show that all roots λ of the characteristic equation $\det(A + e^{-\lambda\theta}B - \lambda I) = 0$ have negative real parts. The characteristic equation is here

$$\lambda^2 + \lambda(\nu + g(E_0) + \kappa) + \nu(g(E_0)+\kappa) - \mu g'(E_0)(1-G_0)e^{-\lambda\theta} = 0,$$

and all roots λ of this equation have negative real part if and only if all roots z of

$$H(z) := (z^2 + pz + q)e^z + r = 0$$

have negative real parts, with

$$p := (\nu + g(E_0) + \kappa)\theta, \quad q := \nu(g(E_0)+\kappa)\theta^2, \quad r := -\mu g'(E_0)(1-G_0)\theta^2,$$

and the change of variable $z := \lambda\theta$. According to Theorem 16 given in the Appendix B, since $r < 0$ and

$$p^2 = 2q + \nu^2\theta^2 + (g(E_0)+\kappa)^2\theta^2 \geq 2q,$$

we have to check that $-q < r < 0$ and $r\sin(a_2)/(pa_2) \leq 1$, where a_2 is the unique root of the equation $\cot(a) = (a^2 - p)/q$ which lies in the interval $(2\pi, 3\pi)$. The second inequality is clear since $r/p < 0$ and $\sin(x)/x \geq 0$ on $(2\pi, 3\pi)$. For the inequality $-r < q$, notice that $g'(x) \leq g(x)/x$ since $g(x)/x$ is decreasing, and using the equilibrium equation $g(E_0)(1-G_0) = \kappa G_0 = \frac{\nu\kappa}{\mu}E_0$, we have

$$-r = \mu g'(E_0)(1-G_0)\theta^2 \leq \mu\frac{g(E_0)}{E_0}(1-G_0)\theta^2 = \nu\kappa\theta^2 < \nu(\kappa + g(E_0))\theta^2 = q.$$

Hence all roots of the characteristic equation have negative real parts and the system is asymptotically stable. Since our system is autonomous, asymptotic stability implies uniform asymptotic stability. According to theorem 4.6 in Halanay (1966), the convergence is exponential. □

Using the exponential convergence of $\mathbb{E}(N(t-\theta))$ and the hypothesis on g, it is now easy to show that condition (13.1) is satisfied.

Theorem 9 *The time-nonhomogeneous process* $\{X^{sstd}(t)\}_{t\geq 0}$ *has the same invariant distribution* π *as the simple time homogeneous process* $\{X^0(t)\}_{t\geq 0}$.

Proof: According to Theorem 15 in Appendix B, we only have to show that condition (13.1) holds. Using the positiveness and boundedness of $g(E(t-\theta))$, $0 < g(0) \leq g(E(t-\theta)) \leq g(\mu/\nu)$, and the expansion

$$\sqrt{g(E(t-\theta))} - \sqrt{g(E_0)} = \frac{g(E(t-\theta)) - g(E_0)}{\sqrt{g(E(t-\theta))} + \sqrt{g(E_0)}},$$

we have

$$\frac{\mid g(E(t-\theta)) - g(E_0) \mid}{2\sqrt{g(\mu/\nu)}} \leq \mid \sqrt{g(E(t-\theta))} - \sqrt{g(E_0)} \mid \leq \frac{\mid g(E(t-\theta)) - g(E_0) \mid}{2\sqrt{g(0)}}$$

and condition (13.1) is in our case equivalent to

$$\int_0^\infty \big(g(E(t-\theta)) - g(E_0)\big)^2 dt < \infty.$$

Let ε be positive, $\varepsilon < E_0$ and T_ε be such that $\mid E(t-\theta) - E_0 \mid < \varepsilon$ for all $t \geq T_\varepsilon$. Using the mean value theorem, for all $t \geq T_\varepsilon$, there exists a ξ_t in the interval delimited by $E(t-\theta)$ and E_0 such that

$$\mid g(E(t-\theta)) - g(E_0) \mid = g'(\xi_t) \mid E(t-\theta) - E_0 \mid .$$

Since for all $t \geq T_\varepsilon$, $E_0 - \varepsilon < \min(E(t-\theta), E_0)$, and furthermore $0 \leq g'(x) \leq g(x)/x$ and $g(x)/x$ is decreasing,

$$\begin{aligned} \mid g(E(t-\theta)) - g(E_0) \mid = g'(\xi_t) \mid E(t-\theta) - E_0 \mid &\leq \frac{g(\xi_t)}{\xi_t} \mid E(t-\theta) - E_0 \mid \\ &\leq \frac{g(E_0-\varepsilon)}{E_0-\varepsilon} \mid E(t-\theta) - E_0 \mid =: L_\varepsilon \mid E(t-\theta) - E_0 \mid, \end{aligned}$$

and finally with the exponential convergence of $E(t-\theta)$ to E_0, condition (13.1) holds

$$\begin{aligned} \int_0^\infty \big(g(E(t-\theta)) - g(E_0)\big)^2 dt \leq &\int_0^{T_\varepsilon} \big(g(E(t-\theta)) - g(E_0)\big)^2 dt \\ &+ L_\varepsilon \int_{T_\varepsilon}^\infty \big(E(t-\theta) - E_0\big)^2 dt < \infty. \end{aligned}$$

□

Remark 2 *When the positive feedback rate* $g(\mathbb{E}(N(t-\theta)))$ *is such that* $\frac{g(x)}{x}$ *is increasing, for example when* g *is a polynomial of degree* ≥ 2*, there can possibly exist several biologically meaningful equilibrium points and it can not be excluded that for some initial conditions the solutions of equation (13.2) oscillate endlessly.*

When g *is constant but the negative feedback* $\kappa(\mathbb{E}(N(t-\theta)))$ *is an increasing function of* $\mathbb{E}(N(t-\theta))$*, the biologically meaningful equilibrium point is unique but similar application of the fluctuation lemma as in the proof of Lemma 4 yields the trivial observation that* $\underline{E} \leq E_0 \leq \overline{E}$*, and oscillating solutions can not be excluded in this case either.*

Part IV

A SELF REGULATED GENE NETWORK

14

INTRODUCTION

In this Part, we describe a more elaborated self-regulated gene network consisting of three genes. Designed to efficiently control transgene expression, the network has been engineered and experimented in Imhof *et al.* (2000).

14.1 NEED FOR REGULATORY NETWORKS IN GENE THERAPY

One of the first gene therapy treatment with humans was carried out in 2000 on children suffering on the fatal X-SCID (Severe Combined Immunodeficiency), also known as "bubble boy" disease, caused by a fault copy of a gene responsible for the production of an immune protein. A transgene was introduced in a patient's cell to produce the missing protein. The first results were so encouraging that the therapy received large mediatic echo and was often refered to as the "miracle of gene therapy", but after two years the trial had to be stopped since some of the patients developed leukemia. To ensure high production, the transgene's unregulated promoter was engineered to be very likely in the ON state. The insertion of the transgene in DNA occuring at random, the promoter was in some cases so near to another gene that it acted as a promoter for this gene too, boosting its protein production and leading to an overproduction that caused leukemia. More details can be found in Kaiser (2003).

This tragic example emphasizes the need for regulatory networks, and the network described here might give a tight control of the promoter through an external inducer, doxycycline, a safe drug with a long history of use in humans.

14.2 DESCRIPTION OF THE NETWORK

The network consists of three genes, the repressor, the activator and the transgene. The first one encodes for a bacterial tetracycline repressor (TetR) that binds to and inhibits the promoter of the two other genes, see Figure 14.1. The activity of the repressor is inhibited by doxycycline, a safe drug that can be used without problem on a daily basis. This small antibiotic molecule binds to the repressor, preventing it to bind to the promoters of the two other genes. The number of repressor protein is supposed to be stable since produced by an unregulated promoter, and large enough to almost completely inhibit transgene production in the absence of doxycycline.

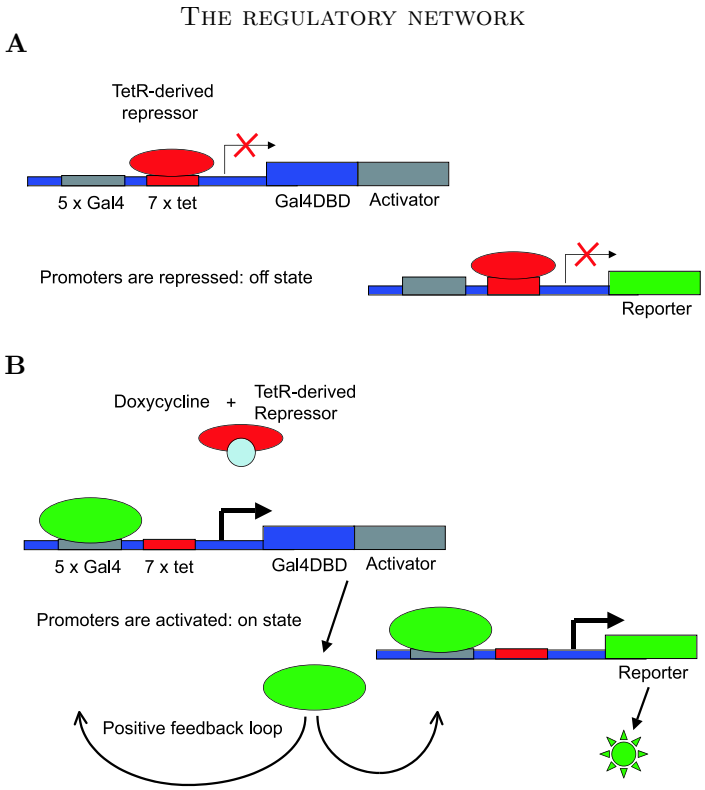

Figure 14.1: Schematic representation of the regulation system. **A**. In the absence of doxycycline, TetR repressors are bound to the promoters of the activator and the transgene. Hence both genes are almost silent due to strong repression. **B**. The adjunction of doxycycline prevent the TetR repressor to bind to the Tet operator site, allowing for production of the Gal4 activator protein. Boosted by the positive feedback loop, Gal4 production increases and leads to a maximal activation of the transgene and hence of the GFP production. Withdrawal of doxycyline re-establishes quickly the state **A** by resuming the TetR repression.

The activator and the transgene are two genetic modules similar to the one studied in Part III. The activator produces a Gal4 protein, responsible for the regulation of both its own promoter and the transgene's one. More precisely the two promoters are engineered to be alike, with each 5 Gal4 binding sites, the activator being autocatalytic while the transgene is regulated by the activator's product. The transgene produce a therapeutic protein, GFP.

In the absence of doxycycline, the network is almost completely silenced, a feature of obvious interest to overcome the problems discussed in Section 14.1.

15

MATHEMATICAL MODEL OF THE NETWORK

We are mostly interested in the behaviour of the transactivator and the transgene, hence we will model the network as a time-continuous Markov process $\{X(t)\}_{t\geq 0}$,

$$X(t) = (N(t), Y(t), M(t), Z(t)),$$

where $N(t)$ is the number of activator protein (Gal4) at time t, $Y(t)$ the state of the activator's promoter, $M(t)$ the number of transgene protein (GFP) at time t and $Z(t)$ the state of the transgene's promoter. The repressor is modeled deterministically using Hill kinetics and act on the transition rates of $\{X(t)\}_{t\geq 0}$. Let us describe the three components one by one.

15.1 REPRESSOR

As said before, we model the repressor effect as deterministic, which is equivalent to suppose the extrinsic noise, here the random fluctuation of the number of repressor and doxycycline molecules, attains a chemical equilibrium. Since the repressor product is supposed to be at a stable level, we are mainly interested in the effects of the adjunction of doxycycline to the probability that a repressor binds to the Tet operator of the two other genes.

Cooperative ligand binding

We first consider the reaction of a ligand $\mathcal{L}$ binding to multiple binding sites on a polymer $\mathcal{P}$, following Bromberg & Dill (2002), Chapter 28, where cooperative ligand binding is discussed. If the polymer has k binding sites, for $1 \leq i \leq k$, the reaction

$$\mathcal{P} + i\mathcal{L} \rightleftharpoons \mathcal{P}\mathcal{L}_i,$$

represent a bound of i ligand molecule to i sites on the polymer $\mathcal{P}$, and $\mathcal{P}\mathcal{L}_i$ means that the i sites are occupied. This reaction has equilibrium constant

$$K_i = \frac{[\mathcal{P}\mathcal{L}_i]}{[\mathcal{P}][\mathcal{L}]^i},$$

where $[\,\cdot\,]$ stands for the concentration. To simplify notations, we call $X = [\mathcal{L}]$. Since the total concentration of polymer, either bound or not, is $\sum_{j=0}^{k} [\mathcal{P}\mathcal{L}_j]$, the fraction of $\mathcal{P}$ molecules that are i-liganded is hence

$$\frac{[\mathcal{P}\mathcal{L}_i]}{\sum_{j=0}^{k} [\mathcal{P}\mathcal{L}_j]} = \frac{[\mathcal{P}\mathcal{L}_i]}{[\mathcal{P}](1 + K_1\ X + K_2\ X^2 + \cdots + K_k\ X^k)} =: \frac{[\mathcal{P}\mathcal{L}_i]}{[\mathcal{P}]\ Q(X)}.$$

The polynomial $Q(X) = 1 + \sum_{i=1}^{k} K_i \ X^i$ appearing in the denominator is called *binding polynomial*. The mean number of occupied sites $M(X)$ can be written as the logarithmic derivative of Q with respect to the logarithm of X, since

$$M(X) = \frac{\sum_{i=0}^{k} i \ [\mathcal{PL}_i]}{\sum_{i=0}^{k} [\mathcal{PL}_i]} = \frac{X \ Q'(X)}{Q(X)} = \frac{d \ln(Q(X))}{d \ln(X)}.$$

The two extreme cases of ligand binging are independent binding and fully cooperative binding. In the first case, every site behaves independently and hence the concentration of i-liganded site is given by

$$[\mathcal{PL}_i] = \binom{k}{i} [\mathcal{PL}_1]^i,$$

and the binding polynomial is $Q(X) = (1 + K \ X)^k$, with $K = K_1$, and the average number of ligands bound per $\mathcal{P}$ molecule is $M(X) = \dfrac{kKX}{1 + KX}$.

In the full cooperative case or Hill model, either none or all of the sites are occupied, the only possible reaction is

$$\mathcal{P} + k\mathcal{L} \rightleftharpoons \mathcal{PL}_k$$

with equilibrium constant K and the binding polynomial is given by $Q(X) = 1 + KX^k$, while the average number of ligands bound per $\mathcal{P}$ is

$$M(X) = \frac{kKX^k}{1 + KX^k}. \tag{15.1}$$

All kind of intermediate models are possible, but since binding to an operator is believed to be very cooperative, we will use the simple Hill model. It remains to determine the concentration of ligand that is here the concentration of repressor that are not bound to a doxycycline molecule.

Reaction of TetR repressor and doxycycline

The reaction between doxycyline $\mathcal{D}ox$ and the TetR repressor $\mathcal{R}$ is

$$\mathcal{R} + \mathcal{D}ox \rightleftharpoons \mathcal{RD},$$

where the $\mathcal{RD}$ complex consits of a doxycycline molecule bound to a repressor. The equilibrium constant is

$$K_{\mathcal{RD}} = \frac{[\mathcal{RD}]}{[\mathcal{R}][\mathcal{D}ox]}.$$

We are here interested in the proportion of free $\mathcal{R}$ molecules that can bind to the k_{op} binding sites of the operators of the two other genes and play the role of the ligand in equation (15.1). The average number of bound sites is therefore

$$M_r([\mathcal{R}]) = \frac{k_{\text{op}} K_{\text{op}} [\mathcal{R}]^{k_{\text{op}}}}{1 + K_{\text{op}} [\mathcal{R}]^{k_{\text{op}}}},$$

with K_{op} the equilibrium constant of the reaction between TetR and the operator, and the fraction of operator sites free of repressor is

$$\frac{k_r - M_r([\mathcal{R}])}{k_r} = \frac{1}{1 + K_{\text{op}}[\mathcal{R}]^{k_{\text{op}}}}.$$

To write this fraction as a function of the doxycycline concentration, we call $[\mathcal{R}_{\text{tot}}]$ the concentration of repressor present in the system in any form, either free, bound to doxycycline or bound to an operator, so that

$$[\mathcal{R}_{\text{tot}}] = [\mathcal{R}] + K_{\mathcal{RD}}[\mathcal{R}]\ [\mathcal{D}ox] + k_{\text{op}}K_{\text{op}}[\mathcal{R}]\ [\mathcal{O}p],$$

where $[\mathcal{O}p]$ denotes the concentration of operators and is negligible compared to the other quantities, and hence

$$[\mathcal{R}_{\text{tot}}] \approx [\mathcal{R}]\ (1 + K_{\mathcal{RD}}[\mathcal{D}ox]).$$

The average fraction of sites free of repressor can be written as a function of the doxycycline concentration,

$$F_{\mathcal{R}_{\text{tot}}}([\mathcal{D}ox]) = \frac{k_r - M_r([\mathcal{R}])}{k_r} = \frac{(1 + K_{\mathcal{RD}}[\mathcal{D}ox])^{k_{\text{op}}}}{(1 + K_{\mathcal{RD}}[\mathcal{D}ox])^{k_{\text{op}}} + K_{\text{op}}[\mathcal{R}_{\text{tot}}]^{k_{\text{op}}}}. \quad (15.2)$$

As an illustration of the shape of a Hill function, Figure 15.1 shows the plot $F_{\mathcal{R}_{\text{tot}}}([\mathcal{D}ox])$ as a function of the doxycycline concentration.

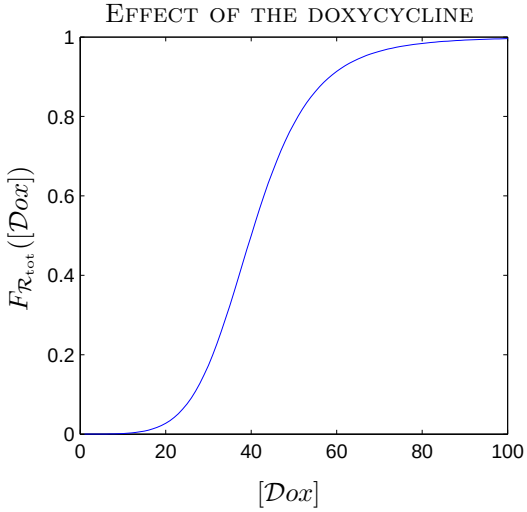

Figure 15.1: Plot of the average fraction of operator sites free of repressor as a function of doxycycline. The function is given by equation (15.2), with the parameters $k_{\text{op}} = 7$, $K_{\mathcal{RD}} = 0.1$, $K_{\text{op}} = 1$ and $[\mathcal{R}_{\text{tot}}] = 5$.

15.2 ACTIVATOR AND TRANSGENE

The transactivator is a genetic module similar to the one studied in Part III, with the propensity functions described in the table below. The strength of the

positive feedback propensity function depends on the doxycycline level through the average fraction of sites free of repressor $F_{\mathcal{R}_{\text{tot}}}([\mathcal{D}ox])$ of equation (15.1), more precisely

$$g_{\mathcal{D}ox}(n) = F_{\mathcal{R}_{\text{tot}}}([\mathcal{D}ox]) \cdot g(n), \tag{15.3}$$

and there is no negative feedback, κ is constant.

The transgene behaves in a similar way as the activator except that it is not autocatalytic, but its propensity function to switch from OFF to ON is a function similar of the activator's product. Its propensity function to go from OFF to ON is

$$g^T_{\mathcal{D}ox}(n) = F_{\mathcal{R}_{\text{tot}}}([\mathcal{D}ox]) \cdot g^T(n), \tag{15.4}$$

where n is the product of the activator.

15.3 PROCESS DESCRIBING THE NETWORK

As discussed at the begining of this Chapter, we model the network as a time-continuous Markov process $\{X(t)\}_{t\geq 0}$,

$$X(t) = (N(t), Y(t), M(t), Z(t)).$$

We recall that $N(t)$ is the number of activator protein (Gal4) at time t, $Y(t)$ the state of the activator's promoter, $M(t)$ the number of transactivator protein (GFP) at time t and $Z(t)$ the state of the transgene's promoter. The state space of this process is hence $\mathbb{N} \times \{0,1\} \times \mathbb{N} \times \{0,1\}$ in the unbounded case, and in the bounded case we replace $\mathbb{N}$ by the corresponding finite spaces. Let $\mathcal{P}$ denote protein and $\mathcal{O}$ promoter for the activator, and $\mathcal{P}^T$ respectivelly $\mathcal{O}^T$ the same for the transgene. The reactions and propensity functions defining the Markov process $\{X(t)\}_{t\geq 0}$ are summarized in the table below.

Reaction R_j	Propensity function a_j	Change vector v_j
R_1: $\emptyset \rightarrow \mathcal{P}$	$a_1(n,y,m,z) = y\mu \cdot 1_{\{(n+1,y,m,z)\in E\}}$	$v_1 = (1,0,0,0)$
R_2: $\mathcal{P} \rightarrow \emptyset$	$a_2(n,y,m,z) = \nu(n)$	$v_2 = (-1,0,0,0)$
R_3: $\mathcal{O}_0 \rightarrow \mathcal{O}_1$	$a_3(n,y,m,z) = g_{\mathcal{D}ox}(n)(1-y)$	$v_3 = (0,1,0,0)$
R_4: $\mathcal{O}_1 \rightarrow \mathcal{O}_0$	$a_4(n,y,m,z) = \kappa\ y$	$v_4 = (0,-1,0,0)$
R_5: $\emptyset \rightarrow \mathcal{P}^T$	$a_5(n,y,m,z) = z\mu^T \cdot 1_{\{(n,y,m+1,z)\in E\}}$	$v_5 = (0,0,1,0)$
R_6: $\mathcal{P}^T \rightarrow \emptyset$	$a_6(n,y,m,z) = \nu^T(m)$	$v_6 = (0,0,-1,0)$
R_7: $\mathcal{O}^T_0 \rightarrow \mathcal{O}^T_1$	$a_7(n,y,m,z) = g^T_{\mathcal{D}ox}(n)(1-z)$	$v_7 = (0,0,0,1)$
R_8: $\mathcal{O}^T_1 \rightarrow \mathcal{O}^T_0$	$a_8(n,y,m,z) = \kappa^T\ z$	$v_8 = (0,0,0,-1)$

The functions $g_{\mathcal{D}ox}(n)$ and $g^T_{\mathcal{D}ox}(n)$ are given in equations (15.3) and (15.4).

16

A MEAN FIELD EXAMPLE

16.1 DESCRIPTION

In this chapter, we consider in more details a mean field model based on two examples discussed in Part III. At the activator level, we consider the autocatalytic model where positive feedback occurs through the intermediate reaction of fast dimerization, as in Chapter 12. Since experimentally the transgene and the transactivator are inserted at random at different loci, we choose to model the effect of the transactivator on the transgene as semi-stochastic with a time delay θ accounting for the average time the activator's Gal4 proteins take to reach the neighborhood of the transgene. The difference between this model and the example discussed in Chapter 13 is that here we do not replace the external variable by its mean value but we take a mean field approximation of the feedback propensity function, see the description below for more details.

We are mainly interested in the marginal bivariate distributions of the protein number and state of the promoter, once for the activator and once the transgene, and not on the full distribution of the process $\{X(t)\}_{t>0}$. More precisely, we split the time-continuous Markov process $\{X(t)\}_{t\geq 0}$,

$$X(t) = (N(t), Y(t), M(t), Z(t)),$$

into the two marginal processes

$$\{X^A(t)\}_{t\geq 0}, \qquad X^A(t) = (N(t), Y(t))$$

and

$$\{X^T(t)\}_{t\geq 0}, \qquad X^T(t) = (M(t), Z(t)),$$

and look for the invariant distributions of $\{X^A(t)\}_{t\geq 0}$ and $\{X^T(t)\}_{t\geq 0}$.

Since now the reactions concerning $\{X^A(t)\}_{t\geq 0}$ and $\{X^T(t)\}_{t\geq 0}$ are decoupled, the two marginal processes behaves as the simple genetic module described in Part III and the method of transfer matrices of Chapter 10 can be used to compute their respective equilibrium distribution. Indeed, from the chemical master equation (CME) of the full process, one can derive the CME of one of the marginal processes by summing over all possible values of the other marginal process. Since they are decoupled, each term involving the other marginal process vanishes, and we can thus handle the two processes separatly, using the evolution patterns described in the following tables.

Summary of the marginal process $\{X^A(t)\}_{t\geq 0}$		
Reaction R_j	Propensity function a_j	Change vector v_j
R_1: $\emptyset \rightarrow \mathcal{P}$	$a_1(n,y) = y\mu \cdot 1_{\{(n+1,y)\in E\}}$	$v_1 = (1,0)$
R_2: $\mathcal{P} \rightarrow \emptyset$	$a_2(n,y) = \nu \cdot n$	$v_2 = (-1,0)$
R_3: $\mathcal{O}_0 \rightarrow \mathcal{O}_1$	$a_3(n,y) = g_{\mathcal{D}ox}(n)(1-y)$	$v_3 = (0,1)$
R_4: $\mathcal{O}_1 \rightarrow \mathcal{O}_0$	$a_4(n,y) = \kappa\ y$	$v_4 = (0,-1)$

The degradation rate is here a constant ν multiplying the number of proteins, taking into account for protein degradation in monomer or dimer form. The positive feedback rate function $g_{\mathcal{D}ox}(n)$ is given by

$$g_{\mathcal{D}ox}(n) = F_{\mathcal{R}_{\text{tot}}}([\mathcal{D}ox]) \cdot (g_1 \cdot \mathbb{E}_n^5 + g_0),$$

with $F_{\mathcal{R}_{\text{tot}}}$ the Hill type function defined in equation (15.2) and $\mathbb{E}_n^5$ is the fifth moment at equilibrium of the number of dimers when a total of n proteins are present in the system, see Chapter 12 and Part V for more details on fast dimerization. The fifth moment correspond to the 5 sites where Gal4 can bind to the promoter.

Summary of the marginal process $\{X^T(t)\}_{t\geq 0}$		
Reaction R_j	Propensity function a_j	Change vector v_j
R_5: $\emptyset \rightarrow \mathcal{P}^T$	$a_5(m,z) = z\mu^T \cdot 1_{\{(m+1,z)\in E\}}$	$v_5 = (1,0)$
R_6: $\mathcal{P}^T \rightarrow \emptyset$	$a_6(m,z) = \nu^T \cdot m$	$v_6 = (-1,0)$
R_7: $\mathcal{O}_0^T \rightarrow \mathcal{O}_1^T$	$a_7(t,z) = \mathbb{E}(g_{\mathcal{D}ox}(N(t-\theta)))(1-z)$	$v_7 = (0,1)$
R_8: $\mathcal{O}_1^T \rightarrow \mathcal{O}_0^T$	$a_8(m,z) = \kappa^T\ z$	$v_8 = (0,-1)$

The degradation rate is here proportional to the number of transgene proteins. In this artificial network, the promoters of the activator and the transgene are engineered alike, hence the propensity functions describing their behaviour should be similar. The transactivator and the transgene are located at different loci, hence we choose a mean field approximation with time delay to describe the

propensity function to activate the transgene's promoter and take into account the fact that the transgene is not regulated by its own product, and we set

$$g^T_{\mathcal{D}ox}(t) = \mathbb{E}(g_{\mathcal{D}ox}(N(t-\theta))).$$

For concrete computation, we restrict to the equilibrium regime in the finite state space setting $E = \{0, 1, \ldots, \Lambda\} \times \{0, 1\}$ for both marginal processes. We treat the doxycycline concentration level as a fixed parameter, and compute the quantities of interest for various concentration levels.

The process $\{X^T(t)\}_{t>0}$ is time-nonhomogeneous due to the function $g^T_{\mathcal{D}ox}(t) = \mathbb{E}(g_{\mathcal{D}ox}(N(t-\theta)))$, and we have to check that the conditions of Theorem 15 in Appendix B are fulfilled, in our case that

$$\int_0^\infty \left(\sqrt{g^T_{\mathcal{D}ox}(t)} - \sqrt{g^T_{\mathcal{D}ox}(\infty)}\right)^2 dt < \infty.$$

With the special shape of the function $g^T_{\mathcal{D}ox}(t)$, the same analysis as in Chapter 13 can not be carried out. Nevertheless, we can prove convergence using the fact that an ergodic Markov process on a finite state space converges geometrically to its invariant distribution. More precisely, let $\pi^A(n, y)$ denote the marginal invariant distribution of the activator process,

$$\pi^A(n, y) = \lim_{t\to\infty} P(X^A(t) = (n, y)).$$

This quantity can be computed using the method of Chapter 10.

Theorem 10 *The time-nonhomogeneous process $\{X^T(t)\}_{t>0}$ has the same invariant distribution $\pi^T(m, z)$, $0 \le m \le \Lambda$, $z = 0, 1$, as the simple time homogeneous process having the same propensity functions but $g^T_{\mathcal{D}ox}(t)$ replaced by*

$$g^T_{\mathcal{D}ox}(\infty) = \sum_{y=0,1}\sum_{n=0}^{\Lambda} g_{\mathcal{D}ox}(n)\pi^A(n, y).$$

Proof: Using Doeblin condition, see for example Rosenthal (1995) Section 6.2, there exist positive constants C and $\alpha \in (0, 1)$ such that

$$|P(X^A(t) = (n, y)) - \pi^A(n, y)| < C\ \alpha^t.$$

Equivalently, if f is a bounded nonnegative function over $\{0, 1, \ldots, \Lambda\}$,

$$\Big|\sum_{y=0,1}\sum_{n=0}^{\Lambda} f(n)P(X^A(t) = (n, y)) - \sum_{y=0,1}\sum_{n=0}^{\Lambda} f(n)\pi^A(n, y)\Big| < 2\ C\ \alpha^t \sum_{n=0}^{\Lambda} f(n).$$

Applying this inequality to the function

$$g^T_{\mathcal{D}ox}(t) = \sum_{y=0,1}\sum_{n=0}^{\Lambda} g_{\mathcal{D}ox}(n)P(X^A(t-\theta) = (n, y))$$

yields

$$|g^T_{\mathcal{D}ox}(t) - g^T_{\mathcal{D}ox}(\infty)| < K\ \alpha^t, \qquad 0 < K < \infty.$$

Since $g^T_{\mathcal{D}ox}(t)$ is strictly positive and bounded, using the same arguments as in Theorem 9 Chapter 13, one can show the equivalence

$$\int_0^\infty \left(\sqrt{g^T_{\mathcal{D}ox}(t)} - \sqrt{g^T_{\mathcal{D}ox}(\infty)}\right)^2 dt < \infty \iff \int_0^\infty \left(g^T_{\mathcal{D}ox}(t) - g^T_{\mathcal{D}ox}(\infty)\right)^2 dt < \infty,$$

and the right side holds since

$$\int_0^\infty \left(g^T_{\mathcal{D}ox}(t) - g^T_{\mathcal{D}ox}(\infty)\right)^2 dt < K \int_0^\infty e^{2\ln(\alpha)t} dt < \infty.$$

The theorem is thus proved since the other conditions in Theorem 15 in Appendix B are automatically fulfilled in our setting. □

Hence we can also compute the marginal equilibrium distribution π^T in a simple way using the method of Chapter 10.

16.2 COMPUTATIONAL RESULTS

We show here what can typically happen for this network, using the Matlab script given in Section C.2 of Part C. The function $F_{\mathcal{R}_{\text{tot}}}([\mathcal{D}ox])$ is plotted in Figure 15.1 with the same parameters as we use here. The other parameters are chosen to exhibit some interesting features of the model. Moreover, we take the same parameter for the activator and the transgene to emphasize the role of positive feedback for the activator and of the mean field regulation for the promoter. The parameters are given in the following table.

Activator	Transgene
$\mu =$ 800	$\mu^T =$ 800
$\nu =$ 9	$\nu^T =$ 9
$g_1 = 10^{-4}$	not applicable
$g_0 =$ 150	not applicable
$\kappa =$ 500	$\kappa^T =$ 500

Figure 16.1 displays the equilibrium mean and variance of $N(t)$, $M(t)$, $Y(t)$ and $Z(t)$. The mean protein production is higher for the transgene due to the mean field propensity function that ensure a high probability to be ON for the transgene even when there are few activator's protein. The most interesting feature is that the response to doxycycline exhibits a peak in the variance of the number of protein for the activator while for the transgene the response is graded and quite commensurable with the mean. More details are given in the discussion below the figure.

For the activator, the coefficient of variation and the variance to mean ratio also exhibit a peak for the activator, while for the transgene the response decreases

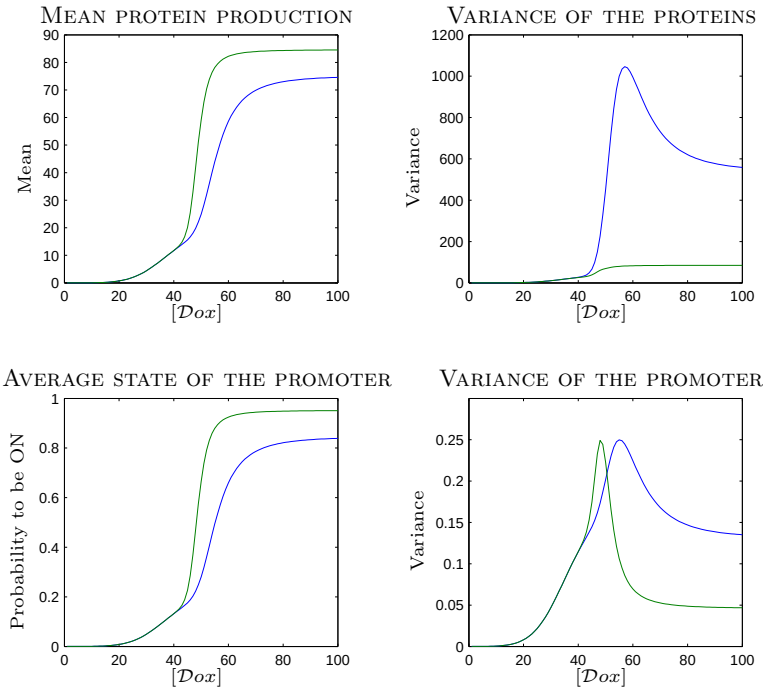

Figure 16.1: In each plot, the curves related to the transactivator are in blue and those related to the transgene in green. On the top left, the mean protein production is plotted against the doxycycline concentration level. It appears that for low levels of doxycycline, here until about 45, the two curves can not be distinguished. This feature occurs in every plots of this Figure and of Figure 16.2. For high levels of doxycyline, the transgene produces more proteins than the activator. On the top right, the variance is plotted against the doxycycline level. In the autocatalytic case, one can clearly see a peak in the variance, while in the case of the transgene, variance is commensurable with the mean for levels higher than 50, see Figure 16.2. The difference of behaviour is obvious when looking at Figures 16.3 and 16.4. On the bottom left, the probability to be ON is plotted against the doxycycline level. The transgene reaches maximal variance before the activator, corresponding to the level where the probability to be ON and the probability to be OFF are equal. At high doxycycline concentration, the transgene shows less variance, which can be explained when looking at the Figures 16.3 and 16.4, where we see that the transgene is more likely to be ON at high levels.

with adjunction of doxycycline. For doxycycline concentration levels higher than 50, the variance to mean ratio for the transgene is almost 1, which possibly indicate a poissonian behaviour, and the covariance goes to 0. These two features are not surprising when compared to the closed formulae of Section 9.2 where it is showed that the variance to mean ratio goes to 1 and the covariance goes to 0 as $g^T_{\mathcal{D}ox}(\infty)$ goes to ∞.

For the activator, the variance to mean ratio is substantially larger than 1, indicating for clusters, which is verified in the bimodal distributions of Figure 16.3

that shows that the protein number is likely to take either a value between 10 and 30 or a value between 70 and 110.

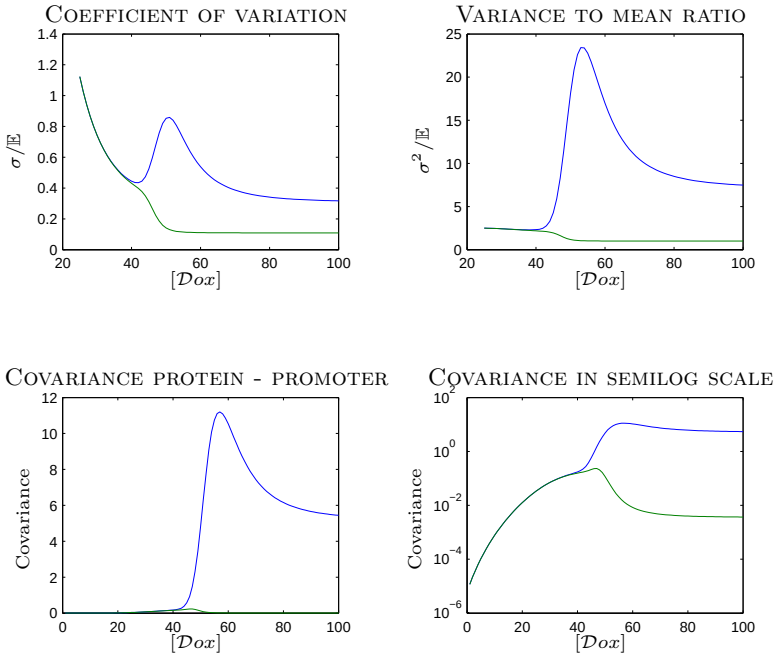

Figure 16.2: In each plot, the curves related to the transactivator are in blue and those related to the transgene in green. The top left plot shows the evolution of the coefficient of variation of the protein number, defined as standard deviation over mean, and the top right plot the variance to mean ratio. For the activator, the curves show a peak, while for the transgene they decrease gradually with the doxycycline concentration level. For doxycycline concentrations over 50, the variance to mean ratio is very close to 1, indicating a poissonian type behaviour. The bottom line displays the covariance of the number of proteins and the state of the promoter, left in the normal scale and right in semilogarithmic scale. The two variables are positively correlated, strongly for the transgene and weakly for the transgene.

Figures 16.3 and 16.4 show the qualitative behaviour of the invariant distribution of $X^A(t) = (N(t), Y(t))$ respectivelly $X^T(t) = (M(t), Z(t))$. The autocatalytic activator displays bimodal distribution, and the transgene unimodal distributions with a mode moving forward as a response to the adjunction of doxycycline.

The transition from almost no production at a low doxycycline concentration to full production is very abrupt, indicating that the control of the network through the safe doxycycline drug is very tight, providing an effective genetic

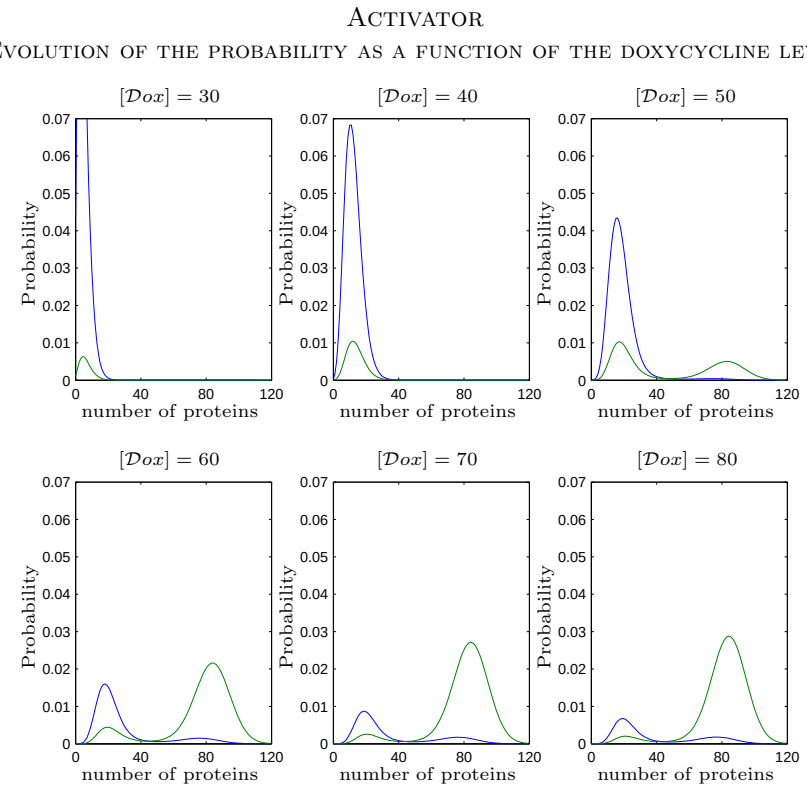

Figure 16.3: Invariant distribution of the number of proteins produced by the activator for some levels of doxycycline. The green curve represent the ON state of the promoter and the blue one the OFF state, and the doxycycline concentration is indicated above the plots. Most of the distributions are bimodal, which explains the peaks in Figures 16.1 and 16.2.

toggle switch. The parameters used here were chosen to emphasize the main features of the system. In a slightly different setting and with other parameters chosen to fit at best, we showed in Fournier *et al.* (2007) that experimental results could be reproduced qualitatively by our model.

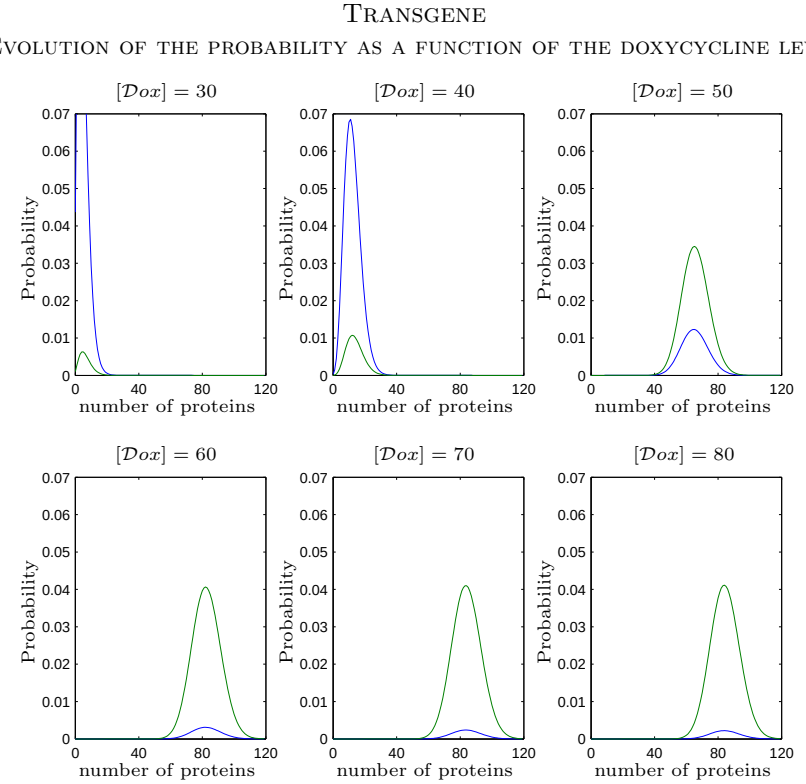

Figure 16.4: Invariant distribution of the number of proteins produced by the transgene for some levels of doxycycline. The green curve represent the ON state of the promoter and the blue one the OFF state, and the doxycycline concentration is indicated above the plots. The distributions are unimodal and move gradually to the right, which explains the graded response to doxycycline in Figures 16.1 and 16.2.

Part V

THE DIMERIZATION PROCESS

17

MATHEMATICAL MODEL OF THE DIMERIZATION

17.1 INTRODUCTION

Dimerization is the chemical union of two identical molecules. Here we consider a fixed amount of a protein $\mathcal{P}$ present either in form of a monomer or a dimer and the chemical reactions

$$\mathcal{M} + \mathcal{M} \underset{c_-}{\overset{c_+}{\rightleftharpoons}} \mathcal{D}$$

where $\mathcal{M}$ and $\mathcal{D}$ stand for monomer and dimer. This process is a simple birth and death process with reflecting barriers and one can easily give a formula for its invariant measure.

However, this formula involves special functions and is not very convenient for computation. In the following we give a recurrence formula for the moments of the invariant measure. The use of recurrence is particularly meaningful if one wishes to compute repeatedly the moments for 1 up to N proteins, as in Chapter 12, where the dimerization process is supposed to be much faster than the other reactions in the system, validating approximatively the hypothesis of a fixed amount of proteins.

17.2 INVARIANT MEASURE AND GENERATING FUNCTION

Let N denote the constant amount of protein, $N_2 := \lfloor N/2 \rfloor$ the integer part of $N/2$, representing the largest possible number of dimers, and $D(t)$ the number of dimers at time t. The number of monomers at time t is simply $N - 2D(t)$. The chemical master equation in this particular setting reads (see Theorem 1)

$$\begin{aligned}\frac{dP(D(t)=i)}{dt} =& c_+(N-2i+2)(N-2i+1)P(D(t)=i-1)\\ &+ c_-(i+1)P(D(t)=i+1)\\ &- (c_+(N-2i)(N-2i-1)+c_-i)P(D(t)=i).\end{aligned}$$

By the classical theory of Markov processes with finite state space, there exists a unique invariant measure, let us denote it π, and it has to satisfy

$$\begin{aligned}0 =& c_+(N-2i+2)(N-2i+1)\pi(i-1) + c_-(i+1)\pi(i+1)\\ &- (c_+(N-2i)(N-2i-1)+c_-i)\pi(i).\end{aligned} \tag{17.1}$$

Lemma 6 *For $0 \leq i < \lfloor N_2 \rfloor$, π satisfy the difference equation*

$$\pi(i) = q_i \pi(i+1)$$

with $q_i = \dfrac{c_-(i+1)}{c_+(N-2i)(N-2i-1)}$.

Proof: The boundary conditions in (17.1) are for $i = 0$

$$0 = c_- \pi(1) - c_+ N(N-1)\pi(0)$$

and for $i = N_2 - 1$

$$0 = c_+(N - 2N_2 + 2)(N - 2N_2 + 1)\pi(N_2 - 1) + c_- N_2 \pi(N_2),$$

hence the difference equation holds at the boundary.
If the relation holds for some $i-1$, $1 \leq i \leq N_2 - 2$, then it holds for i since plugging it into (17.1) yields

$$\begin{aligned} 0 =& c_+(N-2i+2)(N-2i+1)q_{i-1}\pi(i) + c_-(i+1)\pi(i+1) \\ & - (c_+(N-2i)(N-2i-1) + c_- i)\pi(i) \\ =& c_-(i+1)\pi(i+1) - c_+(N-2i)(N-2i-1)\pi(i), \end{aligned}$$

hence $\pi(i) = q_i \pi(i+1)$. □

Up to a proportionality factor, the invariant measure can therefore be written as

$$\pi(i) \propto (q_0 q_1 \cdots q_{i-1})^{-1} \propto \left(\frac{c_+}{c_-}\right)^i \frac{N!}{(N-2i)!\, i!}.$$

The proportionality factor is the normalization constant

$$Z_N(c_+, c_-) := \sum_{i=0}^{N_2} \left(\frac{c_+}{c_-}\right)^i \frac{N!}{(N-2i)!\, i!}.$$

The generating function

$$G(s) := \sum_{i=0}^{N_2} \pi(i)\ s^i$$

can be written as a quotient of confluent hypergeometric functions, that are in fact polynomials.

Theorem 11 *The generating function of the dimers number at equilibrium is given by*

$$G(s) = \begin{cases} s^{N_2} \dfrac{{}_1F_1\left(-N_2, \frac{1}{2}, -\frac{c_-}{4c_+ s}\right)}{{}_1F_1\left(-N_2, \frac{1}{2}, -\frac{c_-}{4c_+}\right)} & \text{if } N \text{ is even,} \\ \\ s^{N_2} \dfrac{{}_1F_1\left(-N_2, \frac{3}{2}, -\frac{c_-}{4c_+ s}\right)}{{}_1F_1\left(-N_2, \frac{3}{2}, -\frac{c_-}{4c_+}\right)} & \text{if } N \text{ is odd.} \end{cases}$$

Proof: For simplicity, we will write $\alpha = \frac{c_+}{c_-}$. From the above relations, summing the other way round and multiplying by $\frac{N_2!}{N_2!}$,

$$G(s) \propto \sum_{i=0}^{N_2} \frac{N!\ (\alpha\ s)^i}{(N-2i)!\ i!} = \frac{N!\ (\alpha\ s)^{N_2}}{N_2!} \sum_{j=0}^{N_2} \frac{N_2!}{(N-2(N_2-j))!\ (N_2-j)!} (\alpha\ s)^{-j}.$$

To transform this formula, we will use the following relations:

$$\begin{aligned}
(2j)! =& (2j)(2j-2)(2j-4)\cdots 2\cdot(2j-1)(2j-3)\cdots 1 \\
=& 2^j\ j!\cdot 2^j\ \frac{1}{2}\left(\frac{1}{2}+1\right)\cdots\left(\frac{1}{2}+j-1\right) = 4^j\ j!\ \left(\frac{1}{2}\right)_j \\
(2j+1)! =& 2\ \left(\frac{1}{2}+j\right)\ (2j)! = 4^{j+1}\ j!\ \left(\frac{1}{2}\right)_{j+1} = 4^j\ j!\ \left(\frac{3}{2}\right)_j \\
\frac{N_2!}{(N_2-j)!} =& (-1)^j\ (-N_2)(-N_2+1)\cdots(-N_2+j-1) \\
=& (-1)^j\ (-N_2)_j,
\end{aligned}$$

$(\cdot)_j$ denoting the Pochhammer symbol or rising factorial $(\cdot)(\cdot+1)\cdots(\cdot+j-1)$. In the case N is even, $N-2(N_2-j) = 2j$, and the generating function is proportional to

$$G(s) \propto \frac{N!\ (\alpha\ s)^{N_2}}{N_2!} \sum_{j=0}^{N_2} \frac{(-N_2)_j}{\left(\frac{1}{2}\right)_j} \cdot \frac{(-4\alpha s)^{-j}}{j!} = \frac{N!\ (\alpha\ s)^{N_2}}{N_2!}\ {}_1F_1\left(-N_2, \frac{1}{2}, -\frac{1}{4\alpha s}\right),$$

while if N is odd $N-2(N_2-j) = 2j+1$ and

$$G(s) \propto \frac{N!\ (\alpha\ s)^{N_2}}{N_2!} \sum_{j=0}^{N_2} \frac{(-N_2)_j}{\left(\frac{3}{2}\right)_j} \cdot \frac{(-4\alpha s)^{-j}}{j!} = \frac{N!\ (\alpha\ s)^{N_2}}{N_2!}\ {}_1F_1\left(-N_2, \frac{3}{2}, -\frac{1}{4\alpha s}\right).$$

The fact that $G(1) = 1$ completes the proof. □

Remark 3 *The generating function formalism is not very appropriated to compute π using the relation*

$$\pi(i) = \frac{d^i}{ds^i}\Big|_{s=0} \frac{G(s)}{i!}.$$

A numerical algorithm to compute efficiently the invariant measure can be found in Pasquier (2008).

18

RELATIONS BETWEEN PARTITION FUNCTIONS AND MOMENTS

The usual representation of the invariant measure presented in Section 17.2 is not adapted for our practical example of Chapter 12 where we need to compute moments repeatedly for all possible N between 0 and some large integer Λ. In this Chapter, we develop a recursive method for computing the moments for N based only on the successive means for $n < N$. We first discuss the theoretical relations between the partition function and the subgroup of involutions of N elements in Section 18.1, translate them in terms of mean and variance in Section 18.2, discuss the relations between higher moments and successive means in Section 18.3, and finally provide a simple numerical algorithm to compute recursively the means in Section 18.4.

18.1 PARTITION FUNCTION AND INVOLUTIONS

The partition function is closely related to the subgroup of involutions I_N of the permutation group of N elements. Involutions are permutations σ such that $\sigma \circ \sigma = id$, or equivalently a permutation that consists only in transpositions (exchanges of two elements) and fixed points. As illustrated in the example of Figure 18.1, we adopt the notation

$$\sigma = (j_1, k_1) \dots (j_i, k_i)$$

for the involution that exchanges j_m and k_m and keep the other points fix, $1 \leq m \leq i \leq N_2$, and $\text{fix}(\sigma) = N - 2i$ for its number of fixed points.

EXAMPLE OF AN INVOLUTION

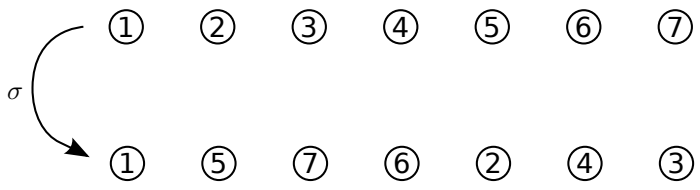

Figure 18.1: Involution exchanging 2 and 5, 3 and 7, 4 and 6, or in the notation described above $\sigma = (2, 5)(3, 7)(4, 6)$. This involution has 1 as unique fixed point.

The number of involutions with $N - 2i$ fixed points, or equivalently the number

of ways of choosing i transpositions in a set of N elements, is given by

$$\sum_{\sigma\in I_N} 1_{\{\mathrm{fix}(\sigma)=N-2i\}}(\sigma) = \binom{N}{2}\cdot\binom{N-2}{2}\cdots\binom{N-2(i-1)}{2}\cdot\frac{1}{i!}$$
$$= \frac{N!}{(N-2i)!\,i!\cdot 2^i}.$$

Hence we can rewrite the partition function as

$$Z_N(c_+,c_-) := \sum_{i=0}^{N_2}\left(\frac{2c_+}{c_-}\right)^i \frac{N!}{(N-2i)!\,i!\,2^i}$$
$$= \sum_{\sigma\in I_N}\left(\frac{2c_+}{c_-}\right)^{\frac{N-\mathrm{fix}(\sigma)}{2}} = \left(\frac{2c_+}{c_-}\right)^{N/2}\sum_{\sigma\in I_N}\left(\sqrt{\frac{c_-}{2c_+}}\right)^{\mathrm{fix}(\sigma)}.$$

According to Randrianarivony (1997), Z_N can be identified as the Taylor serie of a Stieltjes type continuous fraction. Here we consider a different approach and give simple recurrence relations for $Q_N(x) := \sum_{\sigma\in I_N} x^{\mathrm{fix}(\sigma)}$, where $x := \sqrt{\frac{c_-}{2c_+}}$. The partitions function expressed in this way is simply $Z_N(x) = \frac{Q_N(x)}{x^N}$.

Lemma 7 *$Q_N(x)$ satisfies the recurrence relation $Q_{N+1}(x) = x\,Q_N(x)+Q'_N(x)$, where prime denotes differentiation with respect to x.*

Proof: Writing $\sigma = (j_1,k_1)\dots(j_i,k_i)$ for the involution that exchanges i_m and j_m and keep the other points fixed, one see that each element of I_N is an element of I_{N+1} and furthermore if n is fixed by $\sigma = (j_1,k_1)\dots(j_i,k_i)\in I_N$, then there is an involution

$$\tilde\sigma := \sigma\circ(n,N+1) = (j_1,k_1)\dots(j_i,k_i)(n,N+1)\in I_{N+1},$$

or more precisely there is a natural bijection between the involutions of I_N that fixes the point n and the involutions of I_{N+1} that exchange n and $N+1$, hence we can partition I_{N+1} depending on whether the involution fixes $N+1$ or not

$$I_{N+1} = \{\sigma\in I_{N+1};\sigma \text{ fixes } N+1\}\cup\{\sigma\in I_{N+1};\sigma \text{ does not fix } N+1\}$$
$$= I_N \cup \bigcup_{\sigma\in I_N}\bigcup_{\substack{1\le n\le N,\\ \sigma \text{ fixes } n}}\{\sigma\circ(n,N+1)\}.$$

An involution in I_N seen as elements of I_{N+1} have one fixed point more than the same involution seen as an element of I_N, and the terms on the right-hand side of the $\cup$ have one fixed point less than the involution of I_N they are combined with. Therefore we can write Q_{N+1} as stated above

$$Q_{N+1}(x) = \sum_{\sigma\in I_{N+1}} x^{\mathrm{fix}(\sigma)} = \sum_{\sigma\in I_N} x^{\mathrm{fix}(\sigma)+1} + \sum_{\substack{\sigma\in I_N,\\ \mathrm{fix}(\sigma)\ge 1}}\sum_{1\le j\le \mathrm{fix}(\sigma)} x^{\mathrm{fix}(\sigma)-1}$$
$$= x\sum_{\sigma\in I_N} x^{\mathrm{fix}(\sigma)} + \sum_{\substack{\sigma\in I_N,\\ \mathrm{fix}(\sigma)\ge 1}}\mathrm{fix}(\sigma)x^{\mathrm{fix}(\sigma)-1}$$
$$= x\,Q_N(x) + Q'_N(x).$$

□

Remark 4 *The partition above also give the following relation between the number of elements in I_{N+1} and I_N : since $\sum_{\sigma\in I_{N+1}} 1 = \sum_{\sigma\in I_N}(1+fix(\sigma))$, we have*

$$\sum_{\sigma\in I_N} fix(\sigma) = \mid I_{N+1} \mid - \mid I_N \mid,$$

where $|\cdot|$ denotes the number of elements in the set.

Theorem 12 *The derivative of Q_N is given by $Q'_N(x) = N\cdot Q_{N-1}(x)$, and hence*

$$Q_{N+1}(x) = x\,Q_N(x) + N\cdot Q_{N-1}(x). \tag{18.1}$$

Proof: One can easily compute $Q_1(x) = x$ and $Q_2(x) = x^2+1$. If $Q'_N(x) = NQ_{N-1}(x)$ for some N, using Lemma 7 for the first and last equality and by the induction hypothesis for the second one, we have

$$\begin{aligned} Q'_{N+1}(x) =& Q_N(x) + xQ'_N(x) + Q''_N(x) = Q_N(x) + xNQ_{N-1}(x) + NQ'_{N-1}(x)\\ =&(N+1)Q_N(x), \end{aligned}$$

and hence, using Lemma 7 again, (18.1) holds. □

Remark 5 *A generating function for the polynomials $Q_N(x)$.*

Since the coefficient of Q_N rapidly increase, we try to define a kind of moment generating function for the Q_N as follow:

$$h_x(t) := \sum_{N=0}^{\infty} Q_N(x)\frac{t^N}{N!},$$

where we treat x as a parameter. Multiplying both sides of (18.1) by $\frac{t^N}{N!}$ and summing over all possible N leads to the equation

$$\underbrace{\sum_{N=1}^{\infty} Q_{N+1}(x)\frac{t^N}{N!}}_{=h'_x(t)-Q_1(x)} = \underbrace{x\sum_{N=1}^{\infty} Q_N(x)\frac{t^N}{N!}}_{=x(h_x(t)-1)} + \underbrace{\sum_{N=1}^{\infty} N\,Q_{N-1}(x)\frac{t^N}{N!}}_{=th_x(t)}.$$

With $Q_1(x) = x$, this implies that $h_x(t)$ satisfies the differential equation

$$h'_x(t) = (t+x)h_x(t)$$

with initial condition $h_x(0) = Q_0(x) = 1$, which has the unique solution

$$h_x(t) = e^{\frac{t^2}{2}+xt}$$

This function is exactly the moment generating function of a normal random variable $X\sim\mathcal{N}(x,1)$ with mean x and variance 1, and hence $Q_N(x)$ can also be computed as

$$Q_N(x) = \frac{d^N}{dt^N}\Big|_{t=0} e^{\frac{t^2}{2}+xt}$$

18.2 MEAN AND VARIANCE.

In this section, we give a recursive formula to easily compute the mean and variance of D under the invariant measure π. Let us call

$$y = \frac{c_+}{c_-} \qquad \text{and} \qquad x = \sqrt{\frac{c_-}{2c_+}}$$

so that we allow ourselves to freely switch from x to $y = x^{-2}/2$, and $Z(x)$ and $Z(y)$ both mean $Z(c_+, c_-)$ with c_+ and c_- expressed by x or y. The partition function defined in Section 17.2 in the y-notation is a polynomial of degree $N_2 = \lfloor N/2 \rfloor$, given by

$$Z_N(y) = \sum_{i=0}^{N_2} \frac{y^i}{(N-2i)!i!}$$

One can write the following two polynomials as combination of partition functions

$$\begin{aligned}
\sum_{i=1}^{N_2} \frac{iy^i}{(N-2i)!i!} =& y \sum_{i=1}^{N_2} \frac{iy^{i-1}}{(N-2-2(i-1))!(i-1)! \cdot i} \\
=& y \sum_{i=0}^{N_2-1} \frac{y^i}{(N-2-2i)!i!} = yZ_{N-2}(y) \\
\sum_{i=1}^{N_2} \frac{i^2y^i}{(N-2i)!i!} =& \sum_{i=2}^{N_2} \frac{i(i-1)y^i}{(N-4-2(i-2))!(i-2)! \cdot i \cdot (i-1)} \\
&+ \sum_{i=1}^{N_2} \frac{iy^i}{(N-2-2(i-1))!(i-1)! \cdot i} \\
=& y^2 Z_{N-4}(y) + yZ_{N-2}(y)
\end{aligned}$$

so that the mean and the second moment can be expressed in term of y as

$$\mathbb{E}_N(y) = \frac{1}{Z_N(y)} \sum_{i=1}^{N_2} \frac{iy^i}{(N-2i)!i!} = y\frac{Z_{N-2}(y)}{Z_N(y)} \tag{18.2}$$

$$\begin{aligned}
\mathbb{E}_N^2(y) =& \frac{1}{Z_N(y)} \sum_{i=1}^{N_2} \frac{i^2y^i}{(N-2i)!i!} = y^2\frac{Z_{N-4}(y)}{Z_N(y)} + y\frac{Z_{N-2}(y)}{Z_N(y)} \\
=& y\frac{Z_{N-2}(y)}{Z_N(y)}\Big(y\frac{Z_{N-4}(y)}{Z_{N-2}(y)} + 1\Big) = \mathbb{E}_N(y)\big(1 + \mathbb{E}_{N-2}(y)\big)
\end{aligned}$$

which becomes in the x-notation with

$$\mathbb{E}_N(x) = \frac{x^{-2}}{2}\frac{x^N}{x^{N-2}}\frac{N!}{(N-2)!}\frac{Q_{N-2}(x)}{Q_N(x)} = \frac{N(N-1)}{2}\frac{Q_{N-2}(x)}{Q_N(x)} \tag{18.3}$$

$$\mathbb{E}_N^2(x) = \mathbb{E}_N(x)\big(1 + \mathbb{E}_{N-2}(x)\big) \tag{18.4}$$

Combining (18.3) and (18.4), the variance is

$$\sigma_N^2(x) = \mathbb{E}_N^2(x) - \big(\mathbb{E}_N(x)\big)^2 = \mathbb{E}_N(x)\big(1 + \mathbb{E}_{N-2}(x) - \mathbb{E}_N(x)\big) \tag{18.5}$$

Remark 6 *This last equation confirms the intuitive fact that*

$$\mathbb{E}_N(x) - \mathbb{E}_{N-2}(x) \leq 1.$$

18.3 HIGHER MOMENTS

Generalization of 18.4 for higher moments requires some efforts... Let $P_{j+1}(i)$ denote the polynomial

$$P_{j+1}(i) := i \cdot (i-1) \cdots (i-2) \cdots (i-j) =: i^{j+1} - \sum_{l=1}^{j} a_{l,j} i^l.$$

With the convention that $\mathbb{E}_i = 0$ for $i < 0$, the higher moments can be computed as combinations of the means for lower total number of proteins.

Lemma 8 $\mathbb{E}_N(P_{j+1}(D)) = \mathbb{E}_{N-2j} \cdot \mathbb{E}_{N-2(j-1)} \cdots \mathbb{E}_{N-2} \cdot \mathbb{E}_N$.

Proof: We show that both terms are equal to $y^{j+1} \ \dfrac{Z_{N-2(j+1)}}{Z_N}$.

$$\begin{aligned}
Z_N \cdot \mathbb{E}_N(P_{j+1}(D)) =& \sum_{i=1}^{N_2} \frac{P_{j+1}(i) y^i}{(N-2i)!\ i!} = \sum_{i=1}^{N_2} \frac{i \cdot (i-1) \cdots (i-2) \cdots (i-j)\ y^i}{(N-2i)!\ i!} \\
=& \sum_{i=1}^{N_2} \frac{y^i}{(N-2i)!(i-j-1)!} \cdot 1_{\{i>j\}} \\
=& \sum_{i=j+1}^{N_2} \frac{y^i}{(N-2(i-j-1)-2(j+1))!\ (i-j-1)!} \\
=& \sum_{i=0}^{N_2-(j+1)} \frac{y^{i+j+1}}{(N-2(j+1)-2i)!\ i!} \\
=& y^{j+1} Z_{N-2(j+1)},
\end{aligned}$$

and with (18.2), we have

$$\begin{aligned}
\mathbb{E}_{N-2j} \cdot \mathbb{E}_{N-2(j-1)} \cdots \mathbb{E}_{N-2} \cdot \mathbb{E}_N =& y\ \frac{Z_{N-2(j+1)}}{Z_{N-2j}} \cdot y\ \frac{Z_{N-2j}}{Z_{N-2(j-1)}} \cdots y\ \frac{Z_{N-2}}{Z_N} \\
=& y^{j+1}\ \frac{Z_{N-2(j+1)}}{Z_N}.
\end{aligned}$$

□

From the preceding Lemma, we can give a formula for arbitrary moments:

Theorem 13 *The $j+1$-th moment of D is given by*

$$\mathbb{E}_N^{j+1} = \mathbb{E}_{N-2j} \cdot \mathbb{E}_{N-2(j-1)} \cdots \mathbb{E}_{N-2} \cdot \mathbb{E}_N + \sum_{l=1}^{j} a_{l,j} \mathbb{E}_N^l.$$

Proof: From the definition of the coefficients $a_{l,j}$,

$$\mathbb{E}_N(P_{j+1}(D)) = \mathbb{E}_N^{j+1} - \sum_{l=1}^{j} a_{l,j} \mathbb{E}_N^l,$$

hence with Lemma 8 the statement holds. □

We are mainly interested in the moments of low order and can write them down as functions of the means:

$$\begin{aligned}
\mathbb{E}_N^2 =& \mathbb{E}_N + \mathbb{E}_{N-2}\mathbb{E}_N \\
\mathbb{E}_N^3 =& \mathbb{E}_{N-4}\mathbb{E}_{N-2}\mathbb{E}_N + (1+2)\mathbb{E}_N^2 - (1\cdot 2)\mathbb{E}_N \\
=& \mathbb{E}_N + 3\cdot\mathbb{E}_N\mathbb{E}_{N-2} + \mathbb{E}_{N-4}\mathbb{E}_{N-2}\mathbb{E}_N \\
\mathbb{E}_N^4 =& \mathbb{E}_{N-6}\cdots\mathbb{E}_{N-2}\mathbb{E}_N + \frac{3\cdot 4}{2}\mathbb{E}_N^3 - (2+3+6)\mathbb{E}_N^2 + 3!\cdot\mathbb{E}_N \\
=& \mathbb{E}_N + 7\cdot\mathbb{E}_{N-2}\mathbb{E}_N + 6\cdot\mathbb{E}_{N-4}\mathbb{E}_{N-2}\mathbb{E}_N + \mathbb{E}_{N-6}\cdots\mathbb{E}_{N-2}\mathbb{E}_N \\
\mathbb{E}_N^5 =& \mathbb{E}_{N-8}\cdots\mathbb{E}_N + \frac{4\cdot 5}{2}\mathbb{E}_N^4 - (2+3+4+6+12)\mathbb{E}_N^3 \\
&+ (6+12+24)\mathbb{E}_N^2 - 4!\cdot\mathbb{E}_N \\
=& \mathbb{E}_{N-8}\cdots\mathbb{E}_{N-2}\mathbb{E}_N + 10\cdot\mathbb{E}_N^4 - 27\cdot\mathbb{E}_N^3 + 42\cdot\mathbb{E}_N^2 - 24\cdot\cdot\mathbb{E}_N \\
=& \mathbb{E}_N + 31\mathbb{E}_{N-2}\mathbb{E}_N + 33\cdot\mathbb{E}_{N-4}\mathbb{E}_{N-2}\mathbb{E}_N \\
&+ 10\cdot\mathbb{E}_{N-6}\mathbb{E}_{N-4}\mathbb{E}_{N-2}\mathbb{E}_N + \mathbb{E}_{N-8}\cdots\mathbb{E}_N
\end{aligned}$$

We can summarize some relation between the moments, with the notation

$$e_j := \mathbb{E}_{N-2j}\cdots\mathbb{E}_{N-4}\mathbb{E}_{N-2}\mathbb{E}_N,$$

	$\mathbb{E}_N$	$\mathbb{E}_N^2$	$\mathbb{E}_N^3$	$\mathbb{E}_N^4$
$\mathbb{E}_N = e_0$				
$\mathbb{E}_N^2 = e_1$	$+1$			
$\mathbb{E}_N^3 = e_2$	-2	$+3$		
$\mathbb{E}_N^4 = e_3$	$+6$	-11	$+6$	
$\mathbb{E}_N^5 = e_4$	-24	$+42$	-27	$+10$

Representation of the moments of low order as in Theorem 13.

	e_0	e_1	e_2	e_3	e_4
$\mathbb{E}_N$	1				
$\mathbb{E}_N^2$	1	1			
$\mathbb{E}_N^3$	1	3	1		
$\mathbb{E}_N^4$	1	7	6	1	
$\mathbb{E}_N^5$	1	31	33	10	1

Representation of the moments of low order in terms of e_j.

18.4 NUMERICAL COMPUTATION AND ASYMPTOTIC OF THE MEAN

Due to the fast increase of his coefficients, Q_N cannot be efficiently computed for large N. However, the computation of the mean only requires to compute the quotient Q_{N-2}/Q_N.

Defining $c_N(x) := \dfrac{Q_{N-1}(x)}{Q_N(x)}$, the recurrence (12) becomes

$$\frac{1}{c_{N+1}(x)} = x + N c_N(x), \tag{18.6}$$

and the ratio

$$\begin{aligned}
\frac{Q_{N-2}(x)}{Q_N(x)} =& c_{N-1}(x)\cdot c_N(x) = \frac{1}{N-1}\Big(\frac{1}{c_N(x)} - x\Big)c_N(x) \\
=& \frac{1}{N-1}\big(1 - x\, c_N(x)\big),
\end{aligned}$$

allowing to write the mean (18.3) and the second moment (18.4) as

$$\begin{aligned}\mathbb{E}_N(x) =& \frac{N}{2}\big(1 - x\, c_N(x)\big)\\ \mathbb{E}_N^2(x) =& \frac{N}{2}\big(1 - x\, c_N(x)\big)\Big(1 + \frac{N-2}{2}\big(1 - x\, c_{N-2}(x)\big)\Big)\\ =& \frac{N}{2}\big(1 - x\, c_N(x)\big)\Big(\frac{N}{2}\big(1 - x\, c_{N-2}(x) - x c_{N-2}(x)\big)\Big)\end{aligned}$$

With (18.5), since $\mathbb{E}_{N-2}(x) - \mathbb{E}_N(x) = -1 - \dfrac{N-2}{2} x c_{N-2} + \dfrac{N}{2} x c_N$, we have the inequality

$$N c_N(x) \geq (N-2) c_{N-2}(x).$$

Theorem 14 *$c_N(x) \to 0$ as $N \to \infty$.*

Proof: Suppose that the lim sup of the sequence of non-negative numbers $\{c_N(x)\}_{N\geq 1}$ is strictly positive,

$$\limsup_{N\to\infty} c_N(x) = a > 0.$$

Using relation (18.6), one gets

$$a = \limsup_{N\to\infty} c_N(x) = \frac{1}{x + \liminf\limits_{N\to\infty} N c_N(x)} =: \frac{1}{x+b},$$

where $b := \liminf\limits_{N\to\infty} N c_N(x)$ has to be finite. Isolating $c_N(x)$ in (18.6) yields

$$c_N(x) = \frac{1}{N c_{N+1}(x)} - \frac{x}{N}$$

and we get

$$a = \limsup_{N\to\infty} c_N(x) = \limsup_{N\to\infty}\left(\frac{1}{N c_{N+1}(x)} - \frac{x}{N}\right) = \frac{1}{\liminf\limits_{N\to\infty} N c_{N+1}(x)\cdot \frac{N+1}{N+1}} = \frac{1}{b}.$$

Since $x > 0$ and $b < \infty$, this leads to the contradiction $\dfrac{1}{x+b} = \dfrac{1}{b}$. □

The above Theorem leave to the somehow counterintuitive conclusion that the fraction of dimers is about $\frac{1}{2}$ for N large, more precisely

$$\lim_{N\to\infty} \frac{\mathbb{E}_N}{N} = \frac{1}{2},$$

for every set of positive parameters $c_+,\ c_-$.

Part VI

APPENDIX

A

SOME NOTIONS OF MOLECULAR BIOLOGY

In this appendix we briefly outline some important biological notions. This is intended as a practical guide for a mathematician without any background in biology and thus we do not claim neither completeness nor sharp precision. Useful supplementary informations can be found in Kimmel & Axelrod (2002), Waterman (2000) or in Ptashne (2004).

A.1 DEOXYRIBONUCLEIC ACID OR DNA

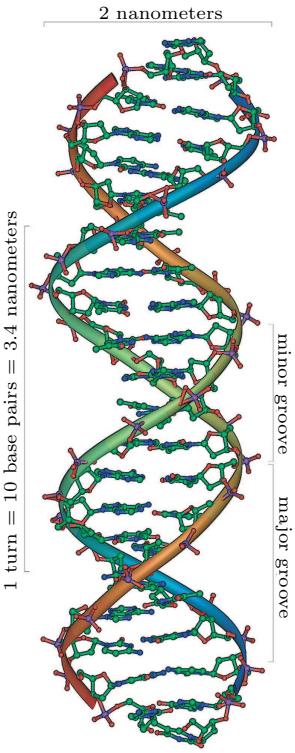

Figure A.1: Illustration of the structure of a part of the DNA double helix. (copylefted image, source: Wikipedia)

Deoxyribonucleic acid (DNA) is a polymer of simple units called nucleotides, consisting of a base, a sugar and a phosphate group. The backbone of the DNA strand is formed by alternating phosphate and sugar, and one of the four basis adenosine (A), cytosine (C), guanine (G) or thymine (T) is attached at each sugar. The bases can bind to each other according to the so-called pairing rule: an A base can only pair with a C base, while a G base can only pair with a T base. The pairing is due to weak hydrogen bonds, two bonds for AT and three bonds for GC, hence the pair AT can be broken with less energy than the pair GC.

As illustrated in Figure A.1, DNA appears in living organisms as two complementary strands, a bases sequence in one strand is bound to its negative sequence in the other strand according to the base pairing rule, for example the sequence $ACCATCGA$ is bound to $TGGTAGCT$. Geometrically, the two strands take the shape of a double helix. Moreover, each strand has a polarity or directionality, the direction goes from the end called 5' and to the end 3', and the two complementary strands run in opposite direction (the names 5' and 3' come from the way each sugar is attached with one phosphate, for more details see Ptashne (2004)).

One turn of the helix corresponds to 10 bases. The whole quantity of DNA present in a living cell consists of about 3×10^9 bases in a human cell and about 5×10^6 bases in the extensively studied intestinal bacterium *Escherichia coli*.

The fact that the weak hydrogen bonds can be broken or rejoined easily and the base pairing rule make perfect duplication possible. During replication the hydrogen bonds between the two strands are broken and each strand is used as a template to build two new double strands, as schematized in Figure A.2.

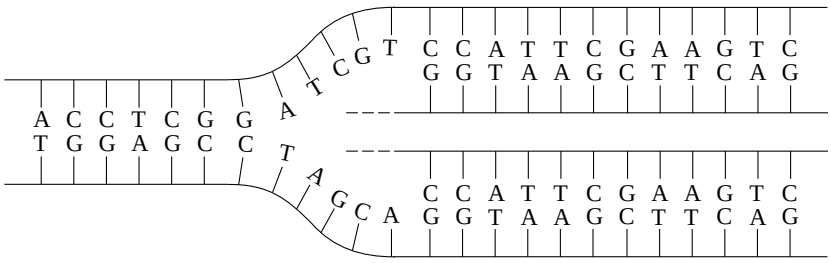

Figure A.2: Illustration of the DNA replication process.

A.2 GENES AND RNA

A gene is a subsequence of a DNA double strand that codes for a protein, typically about 1000 pairs of bases. The first step of gene expression is the transcription of one strand of gene into ribonucleic acid or RNA, a linear molecule. An enzyme called RNA polymerase binds near the beginning of a gene and initiate the transcription process. This process is very similar to the DNA replication process, except that only one strand of a subsequence of the DNA is copied and that the thymine base (T) is replaced by its unmethylated form called uracil (U), see Figure A.3.

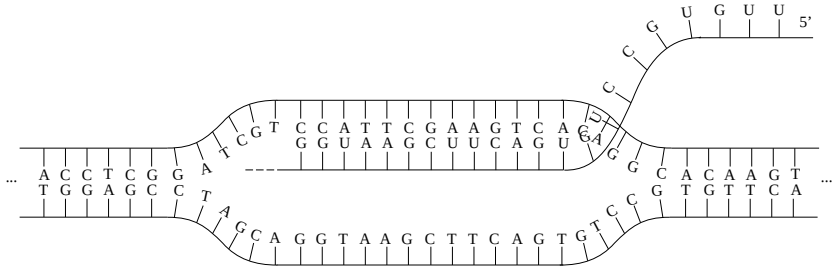

Figure A.3: Illustration of the RNA transcription process.

There are two types of RNA molecules, some function in the cell as end products while other called messenger RNA or short mRNA are devoted to synthesize proteins, see Section A.3.

A.3 GENETIC CODE, AMINO ACIDS AND PROTEINS

A protein is an organic compound made of amino acids and is synthesized during mRNA translation. An ordered triplet of the bases A, U, G, C is called a

codon. Roughly speaking, the genetic code is a mapping from the set of all 64 codons to the set containing 20 amino acids and a stop signal, and the mRNA translation can be thought of as the translation of the successive codons in the mRNA (beginning at the end 5') into amino acids using the genetic code. A special codon is AUG, that codes both for an amino acid and a start signal. The genetic code is not one-to-one, in accordance with the central dogma of molecular biology that states that no information can be transfered back from protein to either protein or nucleic acid. Some proteins have structural functions, other (enzymes) encourage chemical reactions.

A.4 GENE REGULATION, OPERATOR AND PROMOTER

Although the genetic material contained in each cell is the same, the morphology or function of different cells in an organism may be very distinct, for example a skin cell and a cardiac muscle cell. One explanation is that different genes are expressed. The control of the amount and timing of the gene product is referred to as gene regulation, for simplicity we only consider here gene regulation at the level of mRNA transcription. A gene can be ON or OFF, depending on the state of its operator, a gene ON means that the gene is expressed through successive mRNA transcription and translation and a gene OFF that mRNA transcription is prevented. We will indistinctly refer to a gene ON or OFF, a promoter ON or OFF or an operator ON or OFF.
The promoter is a region of DNA situated before the 5' end of a gene that is recognized by RNA polymerase to amorce transcription. The operator is a DNA region situated beside the promoter and that can interact with an activator or a repressor, both DNA binding proteins. The activator facilitate the interaction between the promoter and the RNA polymerase and thus enhance mRNA transcription, while the repressor binds to non coding DNA region very close to the promoter and block or interfere with the RNA polymerase progression on the strand. An interesting special case is self regulation, occuring when the gene product is itself his own activator or repressor.

B

SOME MATHEMATICS

In this appendix we cite some theorems that we use in the thesis.

B.1 CONVERGENCE OF TIME-NONHOMOGENEOUS MARKOV CHAINS

We consider a nonhomogeneous Markov chain $X(t)$ taking values in $\mathbb{N}$, of instantaneous transition matrix $Q_t = (q_t(i,j))_{i,j\in\mathbb{N}}$. The following Theorem is proved in Abramov & Liptser (2004).

Theorem 15 *Assume that we can find nonnegative constants $q(i,j)$ such that*

$$\sum_{j\neq i} q(i,j) < +\infty \text{ for fixed } i,$$

and for $i \neq j$,

$$\int_0^\infty (\sqrt{q_t(i,j)} - \sqrt{q(i,j)})^2 \mathrm{d}t < +\infty,$$

and

$$\int_{0\leq s\leq t,\ q(i,j)>0} q_s(i,j)\mathrm{d}s = \int_0^t q_s(i,j)\mathrm{d}s.$$

Let $Q_0 = (q(i,j))_{i,j\in\mathbb{N}}$, and let $X^0(t)$ be the related $\mathbb{N}$-valued Markov chain. Suppose that Q_0 is ergodic, that is that there is a unique probability measure π such that $\pi Q_0 = 0$ and

$$\lim_{t\to\infty} P(X^0(t) = j | X^0(s) = i) = \pi_j, \quad \forall s,\ i,\ j.$$

Then

$$\lim_{t\to\infty} P(X(t) = j | X(s) = i) = \pi_j, \quad \forall s,\ i,\ j.$$

B.2 FLUCTUATION LEMMA

The following result is a slight modification of Lemma 4.2 in Hirsch *et al.* (1985):

Lemma 9 *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be bounded and differentiable, $\dot{f}$ denoting its derivative. There exist increasing sequences $t_n \uparrow +\infty$ and $s_n \uparrow +\infty$, such that*

$$f(t_n) \to \overline{f},\ \dot{f}(t_n) \to 0, \text{ and } f(s_n) \to \underline{f},\ \dot{f}(s_n) \to 0$$

as $n \to +\infty$, where for a function f we denote

$$\overline{f} := \limsup_{t\to+\infty} f(t),\ \underline{f} := \liminf_{t\to+\infty} f(t).$$

B.3 ZEROS OF AN EXPONENTIAL POLYNOMIAL

We consider the exponential polynomial $H(z) = (z^2 + pz + q)e^z + r$, where p is real and positive, q is real and nonnegative, and r is real. The following Theorem is proved in Bellman & Cooke (1963), p. 449.

Theorem 16 *Denote by* a_k $(k \geq 0)$ *the sole root of the equation* $\cot(a) = (a^2 - q)/p$ *which lies on the interval* $(k\pi, k\pi + \pi)$. *We define the number* w *as follows:*

1. *if* $r \geq 0$ *and* $p^2 \geq 2q$, $w = 1$;
2. *if* $r \geq 0$ *and* $p^2 < 2q$, w *is the odd* k *for which* a_k *lies closest to* $\sqrt{q - p^2/2}$;
3. *if* $r < 0$ *and* $p^2 \geq 2q$, $w = 2$;
4. *if* $r < 0$ *and* $p^2 < 2q$, w *is the even* k *for which* a_k *lies closest to* $\sqrt{q - p^2/2}$.

Then, a necessary and sufficient condition that all roots of $H(z) = 0$ *lie to the left of the imaginary axis is that*

1. $r \geq 0$ *and* $r \sin(a_w)/(pa_w) < 1$ *or*
2. $-q < r < 0$ *and* $r \sin(a_w)/(pa_w) < 1$.

C

MATLAB CODES

In this Appendix, we provide the Matlab codes for some of the computational results presented graphically in the main text. To save some paper, we focus on the examples that can not be found elsewhere in the literature, skipping the simulation part that is a classic and some endless boring codes that I had the pleasure to write during the PhD.

C.1 ALGORITHM FOR THE INVARIANT MEASURE

The following code is a computer implementation of the algorithm described in Section 10.3.

```
function [piLambda,pLambda,EY,EN,VY,VN,Cov]=mes_inv(mu,nu,kappa,g,Lambda)

% algorithm to compute the invariant measure using the algorithm of
% Chapter 10
%
% entries: mu = scalar
%          nu = scalar or vector of length Lambda+1, if scalar nu(n)=nu*n
%          g = scalar or vector of length Lambda+1
%          kappa = scalar or vector of length Lambda+1
%          Lambda = maximal number of protein (boundary of the strip)
%
%
% outcome:  piLambda = invariant measure, piLambda(n,1:2)=[pi_n(0),pi_n(1)] (columnwise)
%            pLambda = marginal of the protein number
%                    =piLambda(:,1)+piLambda(:,2)
%            EY = probability to be ON
%            EN = mean protein production
%            VY = variance on the promoter
%            VN = variance of the protein production
%            Cov= covariance of N and Y
%
%
% CAUTION: a vector can not be indexed by 0, hence eg g(n+1) as a vector
%          means g(n) as a function


% conversion from scalar to vectors
if length(nu)==1
    nu=nu*(0:Lambda);
end
if length(g)==1
    g(1:Lambda+1)=g;
end
if length(kappa)==1
    kappa(1:Lambda+1)=kappa;
end

c=kappa+mu; % c and d are usefull quantities
d=nu+g;

% normalized left eigenvector at the boundary
w(Lambda+1,:)=[kappa(Lambda+1),d(Lambda+1)]/(kappa(Lambda+1)+d(Lambda+1));
```

```
norma=ones(1,Lambda); % initialize normalization to 1
for i=Lambda-1:-1:0
    alpha=nu(i+2)/mu*[c(i+1)/d(i+1),1;kappa(i+1)/d(i+1),1];
    w2=w(i+2,:)*alpha;
    norma(i+1)=sum(w2);
    w(i+1,:)=w2/norma(i+1); % tilde v in the algorithm
end
w(2:Lambda+1,:)=w(2:Lambda+1,:)./[cumprod(norma,2);cumprod(norma,2)]';
% v in the algorithm
piLambda=w/sum(sum(w)); % properly renormalized

pLambda=sum(piLambda,2)'; % other values of interest
EY=sum(piLambda(:,2));
EN=sum((0:Lambda).*pLambda);
VY=EY.*(1-EY);
VN=sum((0:Lambda).^2.*pLambda)-EN.^2;
Cov=sum((0:Lambda)*piLambda(:,2))-EN*EY;
```

C.2 MEAN FIELD MODEL FOR THE NETWORK

The following script has been used to compute the values appearing in the graphics of Section 16.2.

```
% code to compute the steady states for several level of doxycyline,
% model of Chapter 16

clear all

% maximal number of proteins
Lambda=200;

% Hill function of the doxycyline
Dox=1:100;
kop=7;
Krd=0.1;
Kop=1;
Rtot=5;
FDox=(1+Krd*Dox).^kop./((1+Krd*Dox).^kop+Kop*Rtot^kop);

% parameters of the activator
mu=800;
nu=9;
g1=10^(-4);
g0=150;
kappa=500;
% dimerization
cp=5;
cm=3;


% fifth moment of the fast dimerization, see Chapter 18
x=sqrt(cm/(2*cp));
c=zeros(Lambda+1,1);
c(2)=1/x;
E(1:2)=0;
for i=3:Lambda+1
    c(i)=1/(x+(i-2)*c(i-1));
    E(i)=(i-1)/2 *(1-x*c(i));
end
nufast=nu*((0:Lambda)-2*E);
e2=E.*[0 0 E(1:Lambda-1)];
e3=e2.*[0 0 0 0 E(1:Lambda-3)];
e4=e3.*[0 0 0 0 0 0 E(1:Lambda-5)];
e5=e4.*[0 0 0 0 0 0 0 0 E(1:Lambda-7)];
E5=E+31*e2+33*e3+10*e4+e5;

% propensity function g(n)=g1*E^5_n +g0
gn=g1*E5+g0;

% parameters of the transgene
muT=800;
nuT=9;
```

```
kappaT=500;


% preallocation for the activator
piAct=zeros(Lambda+1,2*length(Dox)); % matrix to store the invariant distribution
pAct=zeros(Lambda+1,length(Dox));% matrix to store the invariant distribution of proteins
EYAct=zeros(size(Dox)); % vector to store the successive values of proba to be on
ENAct=EYAct; % same for the mean protein number
VYAct=EYAct; % same for the variance of the promoter
VNAct=EYAct; % same for the the variance of the protein number
CovAct=EYAct; % same for the covariance N and Y for the promoter

% preallocation for the transgene
piTrans=zeros(Lambda+1,2*length(Dox)); % matrix to store the invariant distribution
pTrans=zeros(Lambda+1,length(Dox));% matrix to store the invariant distribution of proteins
EYTrans=zeros(size(Dox)); % vector to store the successive values of proba to be on
ENTrans=EYTrans; % same for the mean protein number
VYTrans=EYTrans; % same for the variance of the promoter
VNTrans=EYTrans; % same for the the variance of the protein number
CovTrans=EYTrans; % same for the covariance N and Y for the promoter


% computation for the successive levels of Dox

for i=1:length(Dox)
    % activator
    gDox=FDox(i)*gn;
    [piA,pA,EY,EN,VY,VN,Cov]=mes_inv(mu,nu,kappa,gDox,Lambda);
    piAct(:,2*i-1:2*i)=piA;
    pAct(:,i)=pA;
    EYAct(i)=EY;
    ENAct(i)=EN;
    VYAct(i)=VY;
    VNAct(i)=VN;
    CovAct(i)=Cov;
    % transgene
    gDoxT=gDox*pA';
    [piT,pT,EY,EN,VY,VN,Cov]=mes_inv(mu,nu,kappa,gDoxT,Lambda);
    piTrans(:,2*i-1:2*i)=piT;
    pTrans(:,i)=pT;
    EYTrans(i)=EY;
    ENTrans(i)=EN;
    VYTrans(i)=VY;
    VNTrans(i)=VN;
    CovTrans(i)=Cov;
end
```

BIBLIOGRAPHY

Abramov, V., & Liptser, R. 2004. On existence of limiting distribution for time-nonhomogeneous countable Markov process. *Queueing syst.*, **46**(3-4), 353–361.

Abramowitz, Milton, & Stegun, Irene A. 1964. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* ninth dover printing, tenth gpo printing edn. New York: Dover.

Acar, M., Mettetal, J.T., & van Oudenaarden, A. 2008. Stochastic switching as a survival strategy in fluctuating environments. *Nat. genet.*, **40**(Apr), 471–475.

Becskei, A., & Serrano, L. 2000. Engineering stability in gene networks by autoregulation. *Nature*, **405**(Jun), 590–593.

Becskei, A., Séraphin, B., & Serrano, L. 2001. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *Embo j*, **20**(10), 2528–2535.

Bellman, Richard, & Cooke, Kenneth L. 1963. *Differential-difference equations.* Mathematics in Science and Engineering. 6. New York-London: Academic Press, XVI, 462 p. .

Bishop, Amy L., Rab, Faiza A., Sumner, Edward R., & Avery, Simon V. 2007. Phenotypic heterogeneity can enhance rare-cell survival in stress-sensitive yeast populations. *Molecular microbiology*, **63**, 507–520(14).

Bolthausen, Erwin, & Goldsheid, Ilya. 2000. Recurrence and transience of random walks in random environments on a strip. *Commun. math. phys.*, **214**(2), 429–447.

Brémaud, Pierre. 1999. *Markov chains. Gibbs fields, Monte Carlo simulation, and queues.* Texts in Applied Mathematics. New York, NY: Springer. xviii, 444 p.

Bromberg, Sarina, & Dill, Ken A. 2002. *Molecular driving forces: Statistical thermodynamics in chemistry and biology.* Garland Publishing Inc,US.

Cao, Yang, Gillespie, Dan, & Petzold, Linda. 2005a. Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems. *J. comput. phys.*, **206**(2), 395–411.

Cao, Yang, Gillespie, Daniel T, & Petzold, Linda R. 2005b. The slow-scale stochastic simulation algorithm. *J chem phys*, **122**(1), 14116.

Coppel, W.A. 1965. *Stability and asymptotic behavior of differential equations.* Heath Mathematical Monographs. Boston: D. C. Heath and Company VIII, 166 p. .

Darvey, I. G., Ninham, B. W., & Staff, P. J. 1966. Stochastic models for second-order chemical reaction kinetics. the equilibrium state. *The journal of chemical physics*, **45**(6), 2145–2155.

Delbruck, Max. 1940. Statistical fluctuations in autocatalytic reactions. *The journal of chemical physics*, **8**(1), 120–124.

El-Samad, H., Kurata, H., Doyle, J. C., Gross, C. A., & Khammash, M. 2005. Surviving heat shock: control strategies for robustness and performance. *Proc natl acad sci u s a*, **102**(8), 2736–2741.

El Samad, Hana, Khammash, Mustafa, Petzold, Linda, & Gillespie, Dan. 2005. Stochastic modelling of gene regulatory networks. *Int. j. robust nonlinear control*, **15**(15), 691–711.

Erban, Radek, Kevrekidis, Ioannis G, Adalsteinsson, David, & Elston, Timothy C. 2006. Gene regulatory networks: a coarse-grained, equation-free approach to multiscale computation. *J chem phys*, **124**(8), 084106.

Fedoroff, Nina, & Fontana, Walter. 2002. Small numbers of big molecules. *Science*, **297**, 1129–1131.

Feller, W. 1968. *An introduction to probability theory and its applications. Vol I. 3nd ed.* Wiley Series in Probability and Mathematical Statistics. New York-London-Sydney: John Wiley and Sons, Inc. XVIII, 509 p. .

Fournier, T., Gabriel, J. P., Mazza, C., Pasquier, J., Galbete, J. L., & Mermod, N. 2007. Steady-state expression of self-regulated genes. *Bioinformatics*, **23**(December), 3185–3192.

Gabriel, Jean-Pierre, Hanisch, Herman, & Hirsch, Warren M. 1981. Dynamic equilibria of helminthic infections? *In: Quantitative population dynamics, Stat. Ecol. Ser. 13, 83-104 (1981).*

Gardiner, Crispin W. 1983. *Handbook of stochastic methods for physics, chemistry and the natural sciences.* Springer Series in Synergetics. 13. Berlin: Springer.

Gillespie, Daniel T. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, **22**(4), 403–434.

Gillespie, Daniel T. 1992. A rigorous derivation of the chemical master equation. *Physica a: Statistical and theoretical physics*, **188**(1-3), 404–425.

Gillespie, Daniel T. 2001. Approximate accelerated stochastic simulation of chemically reacting systems. *The journal of chemical physics*, **115**(4), 1716–1733.

Goutsias, John. 2007. Classical versus stochastic kinetics modeling of biochemical reaction systems. *Biophys. j.*, **92**(7), 2350–2365.

Goutsias, John, & Kim, Seungchan. 2006. Stochastic transcriptional regulatory systems with time delays: a mean-field approximation. *J comput biol*, **13**(5), 1049–1076.

Guido, Nicholas J., Wang, Xiao, Adalsteinsson, David, Mcmillen, David, Hasty, Jeff, Cantor, Charles R., Elston, Timothy C., & Collins, J. J. 2006. A bottom-up approach to gene regulation. *Nature*, **439**(7078), 856–860.

Halanay, Aristide. 1966. *Differential equations: Stability, oscillations, time lags.* Mathematics in Science and Engineering. 23. New York-London: Academic Press. XII, 528 p. .

Hirsch, Warren M., Hanisch, Herman, & Gabriel, Jean-Pierre. 1985. Differential equation models of some parasitic infections: Methods for the study of asymptotic behavior. *Commun. pure appl. math.*, **38**, 733–753.

Hornos, J., Schultz, D., Innocentini, G., Wang, J., Walczak, A., Onuchic, J., & Wolynes, P. 2005. Self-regulating gene: an exact solution. *Phys. rev. e*, **72**(5 Pt 1), 051907.

Imhof, M O, Chatellard, P, & Mermod, N. 2000. A regulatory network for the efficient control of transgene expression. *J gene med*, **2**(2), 107–116.

Isaacs, F. J., Hasty, J., Cantor, C. R., & Collins, J. J. 2003. Prediction and measurement of an autoregulatory genetic module. *Proc natl acad sci u s a*, **100**(13), 7714–7719.

Iyer-Biswas, Srividya, Hayot, F., & Jayaprakash, C. 2007. Transcriptional pulsing and consequent stochasticity in gene expression. *arxiv.org*.

Kærn, Mads, Elston, Timothy C., Blake, William J., & Collins, James J. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat. rev. genet.*, **6**, 451–464.

Kaiser, J. 2003. Gene therapy. Seeking the cause of induced leukemias in X-SCID trial. *Science*, **299**(Jan), 495.

Kepler, T. B., & Elston, T. C. 2001. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophys j*, **81**(6), 3116–3136.

Kimmel, Marek, & Axelrod, David E. 2002. *Branching processes in biology.* Interdisciplinary Applied Mathematics. 19. New York, NY: Springer.

Nasell, Ingemar, & Hirsch, Warren M. 1972. A mathematical model of some helminthic infections. *Commun. pure appl. math.*, **25**, 459–477.

Nasell, Ingemar, & Hirsch, Warren M. 1973. The transmission dynamics of schistosomiasis. *Commun. pure appl. math.*, **26**, 395–453.

Norris, J.R. 1997. *Markov chains.* Cambridge Series in Statistical and Probabilistic Mathematics. 2. Cambridge: Cambridge University Press. xvi, 237 p. .

Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D., & van Oudenaarden, A. 2002. Regulation of noise in the expression of a single gene. *Nat genet*, **31**(1), 69–73.

Pasquier, Jérôme. 2008. *Stochastic models for gene regulation and other models.* Ph.D. thesis, Université de Fribourg.

Peccoud, Jean, & Ycart, Bernard. 1995. Markovian modelling of gene product synthesis. *Theor. popul. biol.*, **48**(2), 222–234.

Pedraza, Juan M., & van Oudenaarden, Alexander. 2005. Noise propagation in gene networks. *Science*, **307**(5717), 1965–1969.

Ptashne, Mark. 2004. *A genetic switch, third edition, λ-phage revisited.* 3rd ed. edn. Cold Spring Harbor Laboratory Press.

Raj, Arjun, Peskin, Charles S., Tranchina, Daniel, Vargas, Diana Y., & Tyagi, Sanjay. 2006. Stochastic mrna synthesis in mammalian cells. *Plos biology*, **4**(10), 1707–1719.

Randrianarivony, Arthur. 1997. Fractions continues, q-nombres de Catalan et q-polynômes de Genocchi. *Eur. j. comb.*, **18**(1), 75–92.

Rosenthal, Jeffrey S. 1995. Convergence rates for markov chains. *Siam rev.*, **37**(3), 387–405.

Thattai, M., & van Oudenaarden, A. 2004. Stochastic gene expression in fluctuating environments. *Genetics*, **167**(1), 523–530.

Vilar, José M. G., Kueh, Hao Y., Barkai, Naama, & Leibler, Stanislas. 2002. Mechanisms of noise-resistance in genetic oscillators. *Proc natl acad sci u s a*, **99**(9), 5988–5992.

Waterman, Michael S. 2000. *Introduction to computational biology. Maps, sequences and genomes. Reprint.* Chapman and Hall/CRC.

Wilkinson, Darren James. 2006. *Stochastic modelling for systems biology.* Chapman & Hall/CRC Mathematical Biology and Medicine Series. Boca Raton, FL, 254 p.

Yin, G., & Zhang, Hanqin. 2002. Countable-state-space Markov chains with two time scales and applications to queueing systems. *Adv. appl. probab.*, **34**(3), 662–688.

Yin, G., & Zhang, Hanqin. 2005. Two-time-scale Markov chains and applications to quasi-birth-death queues. *Siam j. appl. math.*, **65**(2), 567–586.

Zhang, Q., & Yin, G. 1997. Structural properties of Markov chains with weak and strong interactions. *Stochastic processes appl.*, **70**(2), 181–197.

Zhou, Z.H., & Davies, S.W. 2004. Consequences of deterministic and stochastic modeling of a promoter. *Engineering in medicine and biology society, 2004. iembs '04. 26th annual international conference of the ieee*, **2**(Sept.), 2873–2876 Vol.4.

REMERCIEMENTS

Je tiens à remercier tout particulièrement mon directeur de thèse le Prof. Christian Mazza pour son ouverture d'esprit, sa simplicité de contact et la confiance qu'il m'accorde. Véritable ouragan scientifique, il m'a fallu parfois des semaines pour digérer les explications denses qu'il me donnait brièvement au café ou des pages de calcul pour vérifier une formule qui chez lui tenait en deux lignes.

Mes remerciements vont également à mon ami et collègue Jérôme Pasquier. Travaillant dans le même domaine, partageant le même bureau et bien souvent le même bar, nous avons tous deux profité d'une saine émulation et pu progresser bien plus rapidement.

C'est le Prof. Jean-Pierre Gabriel, par ses cours bien pensés, distillés à la manière théatrale dont il a le secret, qui m'a donné l'envie de poursuivre mes études par un doctorat. Ses qualités humaines et son expertise dans quantité de domaines, particulièrement l'étude des équations différentielles avec délai, m'ont été d'une très grande aide tout au long de ma thèse.

Le dialogue entre biologistes, pour qui souvent les mathématiques ne sont qu'un lointain souvenir désagréable de leurs années d'études, et les mathématiciens, dont la compréhension des phénomènes biologiques est très limitée, n'est pas toujours aisé. J'ai durant mon travail eu la chance de rencontrer le Prof. Nicolas Mermod de l'Université de Lausanne, qui par son ouverture d'esprit a permis un dialogue très constructif et m'a aidé à mieux appréhender certains phénomènes biologiques. Je tiens à m'excuser auprès de lui et de tous les biologistes pour les imprécisions dans les parties de ce travail concernant leur domaine.

La liste des gens qui m'ont marqué durant ces années est longue, mais j'aimerais encore citer le Prof. Dieter Rüegg, Dr. Richard Baltensperger, Dr. Marc-Adrien Schnetzer, Dr. Patrick Freund et Thomas Clerc.

Un merci spécial pour l'expert externe, le Prof. Stephan Morgenthaler de l'EPFL.

Sur un plan plus personnel, je tiens à remercier ma femme Anaïs et ma famille pour leur soutien constant, ainsi que mes amis, plus particulièrement ceux avec qui j'ai eu la joie de passer tant de moments inoubliables en montagne et qui m'ont permis de reprendre un grand bol d'oxygène qui souvent manque dans les locaux glauques de l'université.

thomasfournier

Born on November 23, 1979, in Nendaz, VS, CH.
Married to Anaïs Pitteloud Fournier, no children.

Education

2009 **Post Doc**, *Department of Biology, Unit of Ecology & Evolution*, University of Fribourg.

2003–2008 **Doctoral Degree**, *Mathematics Department*, University of Fribourg.

1999–2003 **Academic Degree**, *Mathematics Department*, University of Fribourg, *Diploma*.
Equivalent to a Master Degree.

1995–1999 **High School**, *Creusets & Spiritus Sanctus*, Sion & Brig, *Maturité Fédérale*.
Type C (Scientific), in german.

Experience

2008–2009 **Lecturer**, *Swiss college of agriculture*, Zollikofen.
half time

2008 **Lecturer**, *College of engineering and architecture*, Fribourg.
4 hours a week

2003–2008 **Teaching assistant**, *Mathematics Department*, University of Fribourg.
Teaching assistant for several courses
Statistical consultation (academic and industrial)
Responsible assistant for the department library

2006 **Substitute teacher**, *College de Gambach*, Fribourg.
Mathematics, 3rd year students, 4 hours a week, 3 months

2004 **Substitute teacher**, *College of Engineering and Architecture*, Fribourg.
Analysis and linear algebra, 7 hours a week, one month

2001–2003 **Tutor**, *Mathematics Department*, University of Fribourg.

Languages

French	**Mother tongue**	
German	**Fluent**	*Studied 3 years in a german speaking high school, including final exams*
English	**Fluent**	*Speak, read and write fluently*
Spanish	**Basic**	*Able to hold a conversation, made several trips to Latin America*

Computer skills

Math & Stat	Matlab, R, Octave, Mathematica, SPSS, Maple
Office	LaTeX, OpenOffice, Microsoft Office, FileMaker
Miscellanous	LabVIEW, Gimp, Adobe Illustrator & Photoshop, Inkscape
Platforms	Ubuntu (Linux), MacOSX, Microsoft Windows XP & Vista

Rue d'Or 17 • CH-1700 Fribourg
+41793302930 • ☎ +41265343042 • ✉ thomas.fournier@unifr.ch

Interests

Mountaineering	Freeride snowboard and ski, hike and bike
Music	Collection of about one thousand Jazz, Free Jazz, Funk and Electro vinyls
Travelling	Several trips to Latin America (Peru, Bolivia and Cuba)

Publications

T. Fournier, J. P. Gabriel, C. Mazza, J. Pasquier, J. L. Galbete, and N. Mermod. Steady-state expression of self-regulated genes. *Bioinformatics*, 23:3185–3192, December 2007.

T. Fournier, J. P. Gabriel, C. Mazza, J. Pasquier, J. L. Galbete, and N. Mermod. Stochastic models and numerical algorithms for a class of regulatory gene networks, 2008, Submitted for publication.

J.-L. Galbete, M. Buceta, T. Fournier, J-P. Gabriel, C. Mazza, J. Pasquier, and N. Mermod. Matrix attachment regions mediate stable expression by reversing gene silencing. Manuscript in preparation.

C. Garcion, M.-A. Schnetzer, R. Baltensperger, T. Fournier, J. Pasquier, J.-P. Gabriel, and J.-P. Métraux. Fire and microarrays: a fast answer to burning questions. *Trends in Plant Science*, 11:320–322, 2006.

PhD thesis

title	*A self-regulated gene network*
supervisors	Prof. Christian Mazza
description	In this thesis we discuss mathematically a class of simple self-regulated genes which are the building blocks for many regulatory gene networks. Instead of performing simulation using the Gillespie algorithm, we describe a closed formula for the steady state distribution of the corresponding Markov process and provide an efficient numerical algorithm for concrete computation. Several special cases of this simple self-regulated gene are discussed in detail, including positive feedback loops involving fast dimerization or time delay. We use these results to model a self-regulated network that works as a potent genetic switch, in agreement with experimental observations.

Diploma thesis

title	*Un problème de temps d'attente (A waiting time problem)*
supervisors	Prof. Jean-Pierre Gabriel
description	This diploma thesis discusses the famous "waiting time paradox" from the viewpoint of renewal processes. We give a detailled derivation of the renewal theorem to compute the exact asymptotic distribution of the waiting time between a fixed time t and the next renewal to occur and compare it with simulation for several nonnegative laws.

Referees

Prof. Christian Mazza
Département de Mathématiques
Chemin du Musée 23
1700 Fribourg
☎ +41263009181
christian.mazza@unifr.ch

Prof. Jean-Pierre Gabriel
Département de Mathématiques
Chemin du Musée 23
1700 Fribourg
☎ +41263009189
jean-pierre.gabriel@unifr.ch

Prof. Nicolas Mermod
Institute of Biotechnology
Université de Lausanne
1015 Lausanne
☎ +41216937616
nicolas.mermod@unil.ch