Padova University, Padova, Italy

Department of Information Engineering

Ph.D. School in Information Engineering

Section: *Bioengineering* - Cycle: *XXVIII*

Deterministic and Stochastic Modeling of Human Papillomavirus Gene Regulatory Network

Headmaster of the school: Ch.mo Prof. Matteo Bertocco

Coordinator: Prof. Giovanni Sparacino

Supervisor: Ch.ma Prof.ssa Gianna Maria Toffolo

PhD candidate: Ing. Alberto Giaretta

TO MY PARENTS
TO MARY

Because if this travel was full of satisfactions and became music...
it was thanks to you...
I will always bring the melody of these years with me...
in my heart

Deterministic and Stochastic Modeling of Human Papillomavirus Gene Regulatory Network

Abstract

In this thesis a novel stochastic and deterministic mathematical model of Human papillomavirus (HPV) gene regulatory network was developed. The novelty of this project is both on methodological and biological /clinical site. The former is in line with the current challenge in recent years to have a holistic view of the basics regulatory mechanisms interconnected to form a complex machinery, where complex patterns can arise, only form the interconnection of basics modules. In fact, HPV offers a case of study of great interest in molecular systems biology. It involves a number of relevant regulatory mechanisms (e.g. transcription, translation, promoter modulation, polyadenylation regulation, splicing,) connected together to form a complex network, albeit its genome is relatively simple, thus suitable for an accurate deterministic and even stochastic modeling.

HPVs cause a series of diseases of the cutaneous and mucosal epithelium, ranging from minor lesions to precancerous cervical lesions and cervical cancer, which is considered one of the most common cancer in the women worldwide. Therefore, on the biological/clinical aspect the development of a mathematical model of HPV gene expression, is of great interest in order to dispose of an in silico simulator useful to achieve a better comprehension of the complex gene regulatory network, and capable to predict different scenarios from the first stages of viral infection

up to a cervical cancer condition. As far as we know, there is no model of HPV gene regulation available in literature.

A new synthesis of the HPV molecular biology with especial regard to gather/infer from literature the parameters useful for designing a dynamical model, and to shed light in what is still lacking in the biological literature, was preformed. The biological knowledge was translated into a stochastic model in terms of biochemical reactions. In particular, we modeled the HPV early and late promoters that account for the transcripts and proteins evolution during the entire viral life cycle. Even the post-transcriptional and post-translational modifications were modeled in order to properly capture the complex viral regulation known from literature. As far as we know, it is the first time a stochastic model accounts for the complex post-transcriptional control, modeling the splicing and polyadenylation sites regulation, and connect this latter to the transcriptional control layer, mediated by the promoters activities, in order to explore complex patterns that can arise only from the interconnection of different control layers.

The Master Equation (ME) of the system was considered in order to predict and investigate its stochastic behavior. Because of the complex system structure it wasn't possible to solve the whole ME analytically, hence numerical exact simulations were performed by means of the Gillespie's algorithm. A quasi-equilibrium approximation of the ME was developed in order to get a deterministic approximation of the model.

The model structure together with the fixed parameters we have gathered/inferred from literature was able to fit a dataset consistent of the

early promoter activity and to qualitatively reproduce the main dynamical behavior of two of the most important regulatory transcripts during viral late phase.

Different in silico experiments were designed to opportunely explore both the capability of the stochastic model to follows the deterministic predictions, when in fast fluctuations regimen, and to discover complex stochastic patterns, that can arise through the interconnection of the transcriptional and post-transcriptional control layers.

In general, both the stochastic and deterministic formulation of the model showed the capability to reproduce the HPV gene expression dynamics, during the entire viral life cycle, in good agreement with the current biological knowledge.

Modellizzazione Deterministica e Stocastica della rete di regolazione genica dello Human papillomavirus

SOMMARIO

In questa tesi é stato sviluppato un nuovo modello deterministico e stocastico della rete di regolazione genica dello Human papillomavirus (HPV).

Gli aspetti di novitá del modello ricadono sia sull'aspetto metodologico che clinico /biologico. Per quanto riguarda il primo aspetto il progetto é in linea con l'attuale sfida, presente in questi ultimi anni, di ottenere una visione olistica dei meccanismi regolatori di base interconnessi a formare un sistema complesso, all'interno del quale dinamiche complesse possono scaturire solo tramite un'interconnessione di moduli base. In linea con questo, l'HPV si pone come un caso di studio di grande interesse nella systems biology molecolare. Esso comprende una serie di importanti meccanismi regolatori (ad esempio trascrizione, traduzione, modulazione dei promotori, regolazione della poliadenilazione, splicing,...) connessi tra di loro al fine di formare una complessa rete, sebbene il genoma virale sia relativamete semplice e quindi adatto per sviluppare accuratamente una modellistica deterministica e persino stocastica.

L'HPV puó causare una serie di malattie della cute e della mucosa epiteliale, che spaziano da lesioni minori fino a lesioni pre-cancerogene e cancro al collo dell'utero, considerato uno dei principali tipi di cancro che affligge la popolazione femminile. Conseguentemente, da un punto di vista clinico/biologico lo sviluppo di un modello matematico

dell'espressione genica dell'HPV, é di grande interesse al fine di disporre di un simulatore in silico utile per raggiungere una migliore comprensione della complessa rete di regolazione genica e capace di predire scenari differenti a partire dalle prime fasi dell'infezione virale fino all'evoluzione del cancro. Da quello che sappiamo finora, non esiste alcun modello reperibile in letterature della regolazione genica dell'HPV.

E' stata effettuata una nuova sintesi della biologia molecolare dell'HPV con particolare riguardo al raccogliere/inferire da letteratura i parametri utili al fine di progettare un modello dinamico, e al fine di evidenziare le conoscenze biologiche mancanti in letteratura. La conoscenza biologica é stata tradotta in un modello stocastico in termini di un sistema di reazioni biochimiche. In particolare, sono stati modellati il primo ed il secondo promotore per tener conto della evoluzione dei trascritti e delle proteine durante l'intero ciclo virale. Sono state modellate anche le regolazioni post-trascrizionali e post-traduzionali al fine di catturare appropriatamente la complessa regolazione virale, nota in letteratura. Da quello che sappiamo finora, é la prima volta che un modello stocastico tiene conto del complesso controllo post-trascrizionale, modellando i siti di regolazione dello splicing e della poliadenilazione, e connettendo questi ultimi al modulo di controllo trascrizionale mediato dall'attivitá dei promotori, al fine di esplorare dinamiche complesse che possono scaturire solo dal modellare l'interconnessione dei singoli stadi di controllo.

E' stata considerata la Master Equation (ME) del sistema al fine di predire a investigare il suo comportamento stocastico. A causa della complessa struttura del sistema non é stato possibile risolvere analiticamente la ME, perció sono state effettuate simulazioni numeriche esatte facendo uso dell'algoritmo di Gillespie. Un'approssimazione al quasi-equilibrio della ME é stata sviluppata al fine di ottenere un modello deterministico.

Il modello sviluppato, fissando i parametri trovati/inferiti da letteratura, é stato in grado di fittare un dataset inerente l'attivitá dell'early promoter e riprodurre qualitativamente il principale comportamento dinamico di due dei piu' importanti trascritti sviluppati durante la fase terminale del ciclo virale.

Differenti esperimenti in silico sono stati progettati per esplorare opportunamente sia la capacitá del modello stocastico di inseguire le predizioni del modello deterministico, in regime di fluttuazioni veloci, che per scoprire dinamiche complesse che possono originarsi dall'interconnessione dei moduli di regolazione trascrizionale e post-trascrizionale.

In generale, sia la formulazione deterministica che stocastica del modello hanno mostrato la capacitá di riprodurre, in buon accordo con l'attuale conoscenza biologica, la dinamica dell'espressione genical durante l'intero ciclo virale.

Contents

1
Introduction

1.1 Complexity in cell biology

Cellular systems can be amazingly complex and composed by a huge variety of molecular actors, that interact each other in order to accomplish different and articulated tasks, orchestrated in multiple layers of control.

There is increasingly interest in a deep understanding of gene expression, representing that process by which the DNA information sequence is decoded into structure and function. As stated by the central dogma of molecular genetics, each gene, representing the minimal information element of DNA, is transcribed into complementary sequences, called mRNAs, that are usually translated into proteins, representing the main functional components of the cell as they preform work, they control metabolism and they are responsible for the regulation of DNA transcription. For non-protein coding genes the product is a functional RNA having regulatory tasks inside the cell, from transcriptional to translational control.

Gene expression is modulated among different processes consisting of transcriptional, translational, post-transcriptional and post-translational regulation. Transcriptional regulation accounts for the conversion of the genes information into mRNAs. This process is initiated by a particular DNA sequence called promoter which activity, controlling the mRNA production, can be modulated by transcription factors (usually proteins) that, in turn, can bind DNA upstream the promoter. In some cases, a promoter accounts for the regulation of more genes simultaneously, producing a pre-mRNA codifying the complementary information of the entire cluster of genes. This is actually very frequent in prokaryotes and viruses, but also in eukaryotes. The pre-mRNA undergoes post-transcriptional regulation, such as splicing and polyadenylation, in order to generate mature mRNAs for every single gene. To complete the picture, proteins can undergo to post-translational modification, during or after their biosynthesis, in order to adapt the response of the cell to external stimuli.

It is true goal of systems biology to investigate the functional properties of a molecular system state which is defined and regulated by its components as well as by the network of their associations. Within this context, understanding of the system structure, such as gene regulatory networks or signaling pathways, and of its dynamics, both in quantitative and qualitative terms, are key milestones to understand the system in its entireness. Each individual process (e.g. transcription, mRNA degradation, translation, protein degradation, modulation of simple promoter activity, post-transcriptional and post-translational modifications, cell membrane ions channels,...) of a complex molecular system has been extensively studied at individual level and mathematical models of different complexity were developed to reproduce their dynamic behavior. The current challenging in these last years is to understand the combined effects of individual regulatory processes on the functioning of complex molecular system. In fact, to quote from T. Szkely Jr.

"The recent paradigm of systems biology sets out to examine biological phe-

nomena at the systems level. This is in contrast to the widespread approach of reductionism, whereby researchers attempt to understand entire systems by studing them one small component at a time. The reductionist approach has given us valuable insights and a detailed understanding of the molecular components of biological processes. However, it is becoming clear that we need a complementary approach that takes a holistic view of these processes by looking at their systems-level dynamics: this is systems biology. Reductionism implicitly assumes that the entire system is just the sum of its parts, which is not necessarily true: complex patterns can arise from collections of simple components. The challenge of systems biology is to understand extremely complex systems without breaking them into easy-to-digest parts." (Szekely Jr. and Burrage (2014))

Even if on the one hand the goal of systems biology is to develop formal abstractions capable to capture the biological reality, on the other hand the main aims are to truly understand the mechanisms that stay beneath specific diseases and eventually the identification and/or design of an appropriate approach appropriately antagonize or even defeat the disease. This can be accomplished by the identification of drugs that can counterbalance the effects of the disease, or to appropriately design a gene therapy in order to defeat the principal sources of the disease (e.g. pathogens like viruses and bacteria or a DNA alteration), namely at DNA level. In particular gene therapies are very promising and usually can be implemented through the usage of reprogrammed viruses. In all this, the proper methodologies of systems biology can be of great help.

1.1.1 Heterogeneity in biological systems

Heterogeneity is a key property of biological systems at all scales: from molecular up to the population level. During the past century and half the contributions of "nature vs nurture" to heterogeneity evoked endless debate. Nowadays, thanks to new developmental technologies and experimental procedure that allow to investigate biological systems at all the scales up to the capability to observe microscopic regulatory mechanisms following the single cell dynamics, a clearer idea behind the heterogeneity sources came out. We can classify the heterogeneity as arising from three sources: genetic (nature), environmental (nurture) and stochastic (chance) (Szekely Jr. and Burrage (2014)).

The genetic heterogeneity has its roots in the Darwinism and actually consists in the well accepted idea that cells and animals with different genes should clearly be different.

However, even isogenic organisms or populations of cells can be very different and the causes have to be sought in the other two heterogeneity sources.

The environmental heterogeneity (usually referred as extrinsic especially in gene expression context as we'll see better in the next section) is defined, by exclusion, as the heterogeneity from neither genetic nor intrinsic sources. Actually this kind of heterogeneity depends on the context and it can be: the external environment of a cell population, accounting for the pH or nutrient level in which a cell population is propagated; the internal environment of the cells themselves accounting the numbers and positions within each cell of shared gene expression machinery such as ribosomes; cell states accounting the cellular cycle progression and the cellular aging (Szekely Jr. and Burrage (2014)). Hence, we could say that extrinsic noise usually account for the additional variability we see when following multiple cells (i.e., cell to cell variability).

The stochastic heterogeneity, also referred as intrinsic heterogeneity, arises from thermal fluctuations at the level of individual molecules. It affects in general the whole gene expression machinery, from the transcriptional and translational regulations to post-transcriptional and post-translational regulations (Szekely Jr. and Burrage (2014), Elowitz et al. (2002)). It ensures that the biochemical reactions occur randomly in time in the intra and extra cellular environment. Therefore, at a single gene level the intrinsic noise can be defined as the extent to which the activities of two identical copies of that gene, in the same intracellular environment, fail to correlate (Elowitz et al. (2002)). Hence they exhibit different behaviors even in the absence of the other heterogeneity sources.

1.1.2 Stochasticity in gene regulatory networks

Of great interest in unveiling the control logic among the numerous cellular mechanisms is the achieving of a profound comprehension of its gene regulatory networks, as it represents the core of cellular regulation machinery from which all the other control layers are orchestrated.

Both the topolgy and the dynamical structure of a complex gene regulatory network can be well described by a deterministic modeling approach able to predict its behavior in the context of a large system size (high copy numbers of expressed mRNAs and proteins, and either the case of large cell volumes or a large cell population size) and fast promoter and splicing kinetics. However, especially in the past fifteen years it has been extensively and deeply demonstrated the importance of the inherent stochastic nature of the gene expression in its basic mechanisms, as mentioned in the previous section, and by extension of a complex system as a gene regulatory network can be. Stochasticity in gene expression (also known as stochastic noise) arises from fluctuations in transcriptional, post-transcriptional, translational and post-translational processes. These phenomena can lead identical cells exposed to the same environmental conditions to show significant variation in molecular content. The comprehension of this varibility is of great interest since, on the one hand, can lead to detrimental effects on cellular function with potential implications for disease. On the other hand the stochasticity can be advantageous and provides the flexibility needed by cells to adapt to fluctuating environments or respond to sudden stresses, and a mechanism by which population heterogeneity can be established during cellular differentiation and development (Kaern et al. (2005)). As a result, a deterministic approach in modeling a molecular system cannot capture the potentially significant effects that cause stochasticity in gene expression, hence stochastic modeling and simulation methods are necessary in order to appropriately describe the true dimension of the molecular interactions. It is also of remarkable interest to develop experimental and data-analysis techniques to allow the study of stochasticity in complex regulatory systems and to deeply understand the augmented dynamical behavior driven by the stochastic noise.

In literature the stochasticity related to gene expression are usually thought in terms of finite-number effects. This accounts for a fundamental relationship between system size and noise: namely that noise tends to increase when the size of the system is decreased. This can happen when the cell volume changes but more interestingly and usual it can be interpreted as that, in general, when N denotes the chemical species molecular abundance, a decrease in abundance results in a characteristic $1/\sqrt{N}$ scaling of the noise (Kaern et al. (2005), Scott (2013)). Hence, lower the chemical species copy number, higher the stochastic noise. This is surely an important feature to have in mind about the noise contribution with respect the system size. However, not less important, in addition to a large system size, a second requirement for a strong effect of a molecular-level noise on gene expression is to have slow transitions between promoter states (e.g. in eukaryotic gene expression, for which the presence of nucleosomes and the packing of DNA-nucleosome complexes into chromatin makes promoter often inaccessible to the transcriptional machinery and the corresponding flctuations of the promoter states can be quite slow) (Kaern et al. (2005)).

1.1.3 Modeling of gene regulatory networks

The usual ”protocol” to follow in modeling a cellular system, in particular a gene regulatory network, is the following

- Develop a synthesis of the molecular biology with all the information about the network topology and possibly all the other crucial elements to develop a dynamical model (transcripts and proteins half lives, reactions rate constants, promoters activity,...)

- The first modeling step is to translate the biological knowledge into a set of biochemical equations about the promoter activities, splicing sites, transcription synthesis and degradations,... and so on.

Formally this is already a stochastic model since the chemical reactions are acutally markov chains.

- Translate the biochemical set of equations in terms of Chemical Master Equation

- Eventually approximate the system as a master equation (ME) diffusion limit, in terms of Fokker-Planck Equations (FPE) or stochastic differential equations (SDE). This is usually performed in that cases the exact numerical simulation of the master equation (performed by using the Gillespie's exact stochastic algorithm (Gillesple (1977), Gillespie (2002), Gillespie (2007), Gillespie (2013)) are particularly intensive, because the time scales of the various reactions involved can be very different (Gillespie (2013), Kepler and Elston (2001)).

- Develop a deterministic model by directly translating the biochemical equations by using the mass kinetics law or by a quasi-equilibrium approximation of the model master equation (Kepler and Elston (2001)).

1.1.4 Chemical Master Equation

Given its importance, in this thesis, we dedicate this little subsection to the Chemical Master Equation (CME).
Chemical Master Equation (CME) is currently the golden standard for modeling the stochastic behavior or chemical and biochemical systems.

When molecules of a well-stirred mixture of N molecular species $\{S_1, \ldots, S_N\}$ interact through M chemical reactions $\{R_1, \ldots, R_n\}$, the state vector of the system, $\mathbf{X}(t) \in \mathbb{R}^N$, accounting for the copy number of each molecular species S_i at time t, changes stochastically because of the inherent randomness of molecular collisions.

For each chemical species and reactions it is possible to define a state-change vector $\boldsymbol{\nu_j} \in \mathbb{R}^N$defined as follows

$$\begin{aligned}\nu_{ij} \triangleq & \textit{ the change in the number of } S_i \textit{ molecules produced by one } R_j \textit{ reaction},\\ & j = 1, \ldots, M;\, i = 1, \ldots, N\end{aligned} \tag{1.1}$$

If the molecules are confined to a fixed volume and kept at constant temperature it is possible to prove there exists a function, $h_j(\mathbf{x})$, called propensity function, such that

$$\begin{aligned}h_j(\mathbf{x})\,dt \triangleq & Pr[R_j \textit{ reaction will occur in the system in the next}\\ & \textit{infinitesimal time interval } [t, t+dt)], \quad j = 1, \ldots, M\end{aligned} \tag{1.2}$$

The propensity function h_j, together with the state-change vector, $\boldsymbol{\nu_j}$, completely characterizes reactions R_j, $j = 1, \ldots, M$.

Using only eq. 1.1 and 1.2 and the laws of probability theory, it is possible to prove that the probability $P(\mathbf{x}, t|\mathbf{x}_0, t_0)$, that $\mathbf{X}(t) = \mathrm{x}$ given the initial condition $\mathbf{X}(t_0) = \mathrm{x}_0$ for $t \geq t_0$, obeys the chemical master equation

$$\frac{\partial}{\partial t} P(\mathbf{x}, t|\mathbf{x}_0, t_0) = \sum_{j=1}^{M} [h_j(\mathbf{x} - \nu_j)\, P(\mathbf{x} - \nu_j,\, t|\mathbf{x}_0, t_0) - h_j(\mathbf{x})\, P(\mathbf{x}, t|\mathbf{x}_0, t_0)] \tag{1.3}$$

The previous equations imply that the system state $\mathbf{X}(t)$ performs a "random walk" on the integer lattice in the N dimensional species population space; in mathematical terms, $\mathbf{X}(t)$ is a jump Markov process (Gillesple (1977), Gillespie (2002), Gillespie (2007), Gillespie (2013)).

From now on we'll denote the chemical master equation with "ME" instead "CME".

1.2 Aim of the thesis

The first aim of the thesis is to explore deterministic and stochastic modeling methods in order to understand and investigate general problems related to the basic mechanisms of gene expression, such as the deterministic and stochastic design and response of cellular promoters, splicing sites, polyadenylation sites, ... Secondly the aim consists in investigating and understanding the modules responses after their interconnection to form a complex regulatory system to be in line with the current holistic paradigm of systems biology. We will accomplish this aim by studying the gene regulatory network of the human papillomavirus (HPV).

The second aim is to develop a novel model of the HPV gene expression under different biological conditions and to have insights in its regulatory mechanisms, such as transcriptional and post-transcriptional regulation, and their interconnections in order to explain the complex regulation the virus can exhibit during the infection.

1.3 Human Papillomavirus: A novel model within the actual context of systems biology

HPVs cause a series of diseases of the cutaneous and mucosal epithelium, ranging from minor lesions (e.g. benign warts on the hand and feet) to precancerous cervical lesions and cervical cancer which is considered one of the most common cancer in the women worldwide. Infection with HR-HPV generally starts in the basal layer of the mucosal epithelium. In the normal epithelium and in low grade lesions, HPV is generally found in an episomal form and its entire genome is sequentially expressed and involved in the regulation processes during the progression of viral life cycle, as the infected basal epithelial cell differentiates and moves to upper keratinized epithelial layers. At variance, in high grade lesions and in cancer, multiple copies of the HPV genome may integrate in the cellular chromosomes in a manner that promotes and sustains cancer progression.The vast majority of HPV infections are cleared within 2 years, and long term persistence of the infection is rare. Continued stimulation of host cell growth during these rare cases of persistence can result in pre-malignant cervical lesions. These lesions might deteriorate further to cervical cancer if left untreated.

HPV offers a case of study of great interest in molecular systems biology and clinical studies. It involves a number of relevant regulatory mechanisms (e.g. transcription, translation, promoter modulation, poly adenylation regulation, splicing,) connected together to form a complex network, albeit its genome is relatively simple, thus suitable for an accurate deterministic and even stochastic modeling. It contains two main promoters, regulating the early and late phases of viral transcription. Their regulation and timing of promoters activation was investigated between the end of 90s and the beginning of 2000. However, a comprehensive understanding of their switching, of the stochastic effects in their host cell differentiation-dependent activity, and in more general terms of the dynamic evolution of viral transcripts and proteins involved, finely regulated by a complex post- transcriptional and post- translational control, is still lacking.

The development of a mathematical model of HPV genes and proteins regulation, is of great interest in order to achieve a better comprehension of the basic mechanisms and their interconnection regulating the first stages of viral infection evolution up to a cervical cancer condition, making the HPV a very interesting and actual system to study, perfectly in line with the current paradigm of systems biology, as argued before.

There is not so much literature about mathematical modeling of viruses gene regulatory networks. The only virus having a long modeling history is the HIV for which both deterministic and stochastic models where performed over the years (Singh and Weinberger (2009), Singh et al. (2010), Weinberger et al. (2008)). There exists also an interesting model on Epsetin-Barr virus (Werner et al. (2007a), Werner et al. (2007b)) and on HTLV-1 (Corradin et al. (2010), Corradin et al. (2011)).

As far as we know there is no model of HPV gene regulation available in literature, apart a first heuristic/deterministic model on the early promoter regulation we have developed (Giaretta et al. (2015)) and that we will present in chapter 3.

1.4 Overview

In this thesis we develop a novel model of HPV gene regulatory network with the purpose to condense the complex and wide biological knowledge on different clinical conditions including HPV lytic replication in a normal epithelium (both in the basal infection and in the whole viral cycle progression) and integrated HPV into host cell chromosomes leading to cancer development.

The first aim of the project is consistent with the development of a mathematical model able to predict, both stochastically and deterministically, the dynamical behavior of early and late promoters and of all the major viral transcripts and proteins during the entire life cycle. To do that, the condensed biological knowledge will be translated into a set of biochemical reactions stochastically modeling the molecular biology. The system will be based on the Chemical Master Equation (CME). This formalism allows to describe the system of biochemical reactions in terms of the evolution of the probability distribution of the system state variables, related to the chemical species copy numbers. Since the CME will be impossible to solve analytically for such a complex system, its whole dynamical behavior will be investigated with the aid of Gillespies stochastic algorithm (golden standard for simulating stochastic chemical systems), able to reproduce the exact numerical solution of the CME. Nevertheless a simplified assumption consistent with considering the stochasticity associated to the only promoters and splicing sites markov chains will permit to study their associated master equation in order to appropriately design these important ”controllers” in gene expression and to opportunely tune their parameters to achieve the whole dynamical behavior consistent with the biology. Finally, a Quasi-Equilibrium approximation will be performed in order to provide a deterministic revisitation of the stochastic system, under conditions of fast promoters and splicing sites fluctuations.

The second aim of the model is to achieve a better comprehension of the basic mechanisms and their interconnections regulating the early and late stages of viral infection evolution up to an HPV integrated condition. The stochastic predictions, by fixing/inferring parameters from literature, will permit to shed light into the stochasticity the HPV gene expression can exhibit, identifying the importance of a stochastic formulation for the system and find interesting behaviors a deterministic modeling dimension cannot achieve. Consequently, different in silico simulations will be carried out in order to predict and investigate the biological/clinical scenarios from the beginning of the infection up to integrated HPV, typically present in cancer stages.

The last aim of the project is to use the model to identify the major regulatory mechanisms under different biological conditions in order to properly design an experiment that will be used in order to validate the model, as will be argued in the final discussion.

In what follows, a brief description of the next chapters is reported.

In Chapter 2, a new synthesis on HR-HPV molecular biology condensing the transcriptional, post-transcriptional, translational and post- translational regulation underlying the early and the late viral promoters is presented, placing the emphasis in collecting/inferring the main available data from literature (mRNA and protein half-lifes, promoters time usages, protein and transcripts appearance post infection,) that are useful to design a dynamical model and placing emphasis on what is still lacking in literature that could be important in order to properly validate and optimize the structure of a dynamical model about HPV gene expression regulation.

In Chapter 3, a first heuristic/deterministic model we have developed (Giaretta et al. (2015)) in collaboration with the Department of Molecular

Medicine (DMMD) of the University of Padova, based on Ordinary Differential Equations (ODE), is presented with the purpose to describe the early promoter regulation in a context of either large cell population size and fast promoter fluctuations.

In Chapter 4, a second more formal and complete stochastic and deterministic model both regarding the early and the late promoter, developed in collaboration with the Elston Lab at the School of Medicine at Chapel Hill, University of North Carolina (UNC), is presented. This is done by first translating the molecular biology of two promoters gene expression regulation into a stochastic model presented in a set of biochemical reactions highligting the involved chemical species as chemical reactants and products and the correspondent rate constants.

In Chapter 5, we present the Chemical Master Equation of the model reported in Chapter 4.

In Chapter 6, we present the Quasi-Equilibrium Approximation of the developed stochastic formulation as a useful deterministic version of the developed model.

In Chapter 7, we report the dataset available from literature, the in silico simulations we plan to show and the correspondent fixed parameters we have gathered/inferred from literature.

In Chapter 8, we present a set of in silico experiments in different biological and clinical conditions to test the model capability in predicting the dynamical behaviors of all its of gene expression mechanisms (i.e. transcription, splicing, ...) that is in good agreement with the priori biological knowledge presented in literature and highlighting the complex patterns that can arise from the interconnection of different layers of control in gene expression. Some predictions is also reported to explore the augmented dynamical behavior associated with the intrinsic noise that only a stochastic formalism can predict.

In Chapter 9, a discussion inherent the achieved modeling results and their goodness with the present available biological knowledge is presented. Moreover immediate and long term future developments will be argued.

2
Human Papillomavirus (HPV)

2.1 Introduction: HPVs and diseases

HPVs (human papillomaviruses) are small double stranded DNA viruses. They infect epithelial cells and cause a variety of lesions ranging from common warts to cervical neoplasia and cancer. To date, more than 150 human papillomavirus types have been completely sequenced (Doorbar et al. (2012)).

The most well studied HPV types are the mucosal Alpha types that can cause cervical cancer. Cervical cancer is the second most common cancer among women worldwide. Globally there are around 530,000 new cases and 275,000 deaths due to cervical cancer annually (Raybould (2011)). For centuries, cervical carcinoma has been recognised to behave as a sexually transmitted disease and in the mid 1970s it was proposed that there was an aetiological link with HPV. Infection with HPV is now recognised as an essential factor for the development of cervical cancer. Importantly, these viruses are also associated with cancers at other sites, including the penis in men, the anal transformation zone, the tonsils, oropharynx and base of tongue (Doorbar (2006)).

2.1.1 High- and low-risk types

The Alpha HPVs are divided into cutaneous and mucosal types. These latter are further subdivided into high-risk and low-risk HPVs (LR-HPVs). A very studied low-risk type is HPV-11 and among the cutaneous Alpha types we can find HPV-2 which causes common warts and HPV-3 and -10 which cause flat warts. The low-risk mucosal types do not typically cause neoplasia.

Carcinomas associated with the high-risk HPV types (HR-HPVs) are, however, a far more significant burden. Twelve HPVs (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, and 59) are defined by the World Health Organisation (WHO) as being high-risk cancer causing types (Doorbar et al. (2012)). The most studied and aggressive version of HR-HPVs are HPV-16, -18 and -31. In particular, HPV-16 and -18 are responsible for approximately 70% of cervical cancer cases globally. Current vaccination programs serve to protect women against these latter two types infection. Nevertheless the vaccine is not a cure. It is important to develop new antiviral therapies to defeat the viral infection; for this reason a deep understanding of its molecular biology regulatory mechanisms could be vital in order to better understand its life cycle and gene expression regulation in order to develop a functioning antiviral therapy.

The model we will develop in the next chapters will account for HR-HPVs, with particular reference to HPV-16, -18 and -31.

2.1.2 Overview of the HPV Genome structure

All papillomaviruses have a common genetic structure and a nonenveloped icosahedral capsid. The viral particles consist of a circular double-stranded DNA molecule of about 8,000 bp that is bound to cellular histones and contained in a protein capsid composed of 72 pentameric capsomers (Cobo (2012)). The papillomavirus genome contains approximately eight open reading frames (ORFs) that are all transcribed from a single DNA strand. This latter contains two main polycistronic promoters and two polyadenylation sites. HPV genome is divided into three regions as shown in Fig.2.1.1 (Cobo (2012), Johansson and Schwartz (2013), Doorbar (2006), Doorbar et al. (2012)). The only elements shared by all members of papillomavirus genus are the presence of an upstream regulatory region (URR), the early proteins E1 and E2, and the late proteins L1 and L2. The first region is a noncoding upstream regulatory region of 4001,000 bp named as upper regulatory region or long control region (LCR). This region contains the early core promoter that regulates DNA replication by controlling the transcription of the ORFs (Cobo (2012)). The second region is the early region which contains the late core promoter ORFs that encode proteins necessary for viral DNA replication and oncogenesis transformation, named as ORFs E1, E2, E4, E5, E6 and E7. The three latter genes are the oncogenes, even if the last two (E6 and E7) are the most important and studied (Cobo (2012)). The third and last region consists of a late region which encodes viral structural proteins, including the major capsid proteins necessary for productive viral replication (L1 and L2) (Cobo (2012)). In particular the early promoter initiates and regulates a polycistronic primary transcript encoding for all the early proteins, while the late promoter initiates and regulates a polycistronic primary transcript that encodes for the late proteins (L1 and L2) and all the early genes but E6 and E7 (i.e. E1, E2, E4, E5).

2.1.3 Principal functions of the HPV genes/proteins

The early genes and hence proteins are regulatory in function, namely they are deputated for the viral DNA replication, transcriptional regulation, cell cycle control, cell signalling and apoptosis control, structural modifications, etc..., while the late genes are important for the assemble of the viral DNA into virions. In what follows we describe the main func-

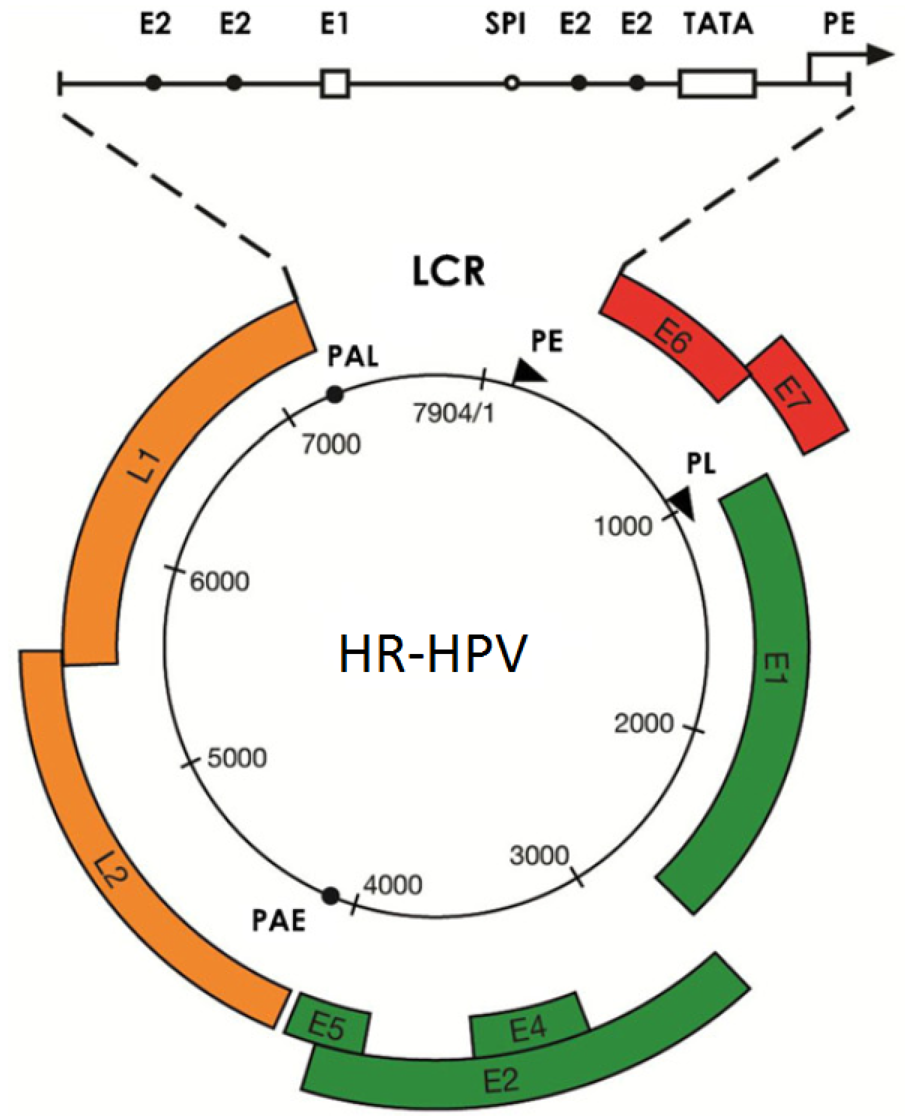

Figure 2.1.1: We show the HR-HPV genome structure. Its genome is shown as a black circle with the early (PE) and late (PL) promoters marked by arrows. The six early ORFs [E1, E2, E4 and E5 (in green) and E6 and E7 (in red)] are expressed from either PE or PL at different stages during epithelial cell differentiation. The late ORFs [L1 and L2 (in yellow)] are also expressed from PL, following a change in splicing patterns, and a shift in polyadenylation site usage [from early polyadenylation site (PAE) to late polyadenylation site (PAL)]. All the viral genes are encoded on one strand of the double-stranded circular DNA genome. The long control region (LCR) is enlarged to allow visualization of the E2-binding sites and the TATA element of the PE promoter. The location of the E1- and SP1-binding sites is also shown.

tions of HPV genes and proteins.

2.1.3.1 E1 PROTEIN

E1 is a viral protein of 73 kDa that is required for viral replication. This protein binds to a specific DNA sequence and assembles into hexameric complexes through E2 protein, in order to enhance the replication efficiency. The resultant complex has helicase activity necessary for oligomerization. E1 also interacts with replication protein A (RPA), which results in the rapid stabilization of single-stranded DNA generated by E1 helicase activity (Cobo (2012)).

2.1.3.2 E2 PROTEIN

E2 is a viral protein of 40 kDa to 45 kDa, depending on the HPV type. Expression of E2 protein in human cells results in the repression of transcription from the viral promoter. E2 also plays an important role in the production of new, replicated viral DNA with mitotic chromosomes for the distribution of this DNA in the divided cells. Moreover, E2 interacts with L2 and leads to the amplification of viral DNA to facilitate the production of new viral progeny (Cobo (2012)).

2.1.3.3 E4 PROTEIN

E4 gene is located in the E region, overlapping with E2, and is a heterogeneous protein. The functions of the E4 protein are oligomerization, phosphorylation and proteolytic cleavage. E4 also plays a role in supporting viral genome amplification, the regulation of late gene expression and the control of viral maturation (Cobo (2012)). It is produced by a transcript usually called $E1\hat{\ }E4$.

2.1.3.4 E5 PROTEIN

This protein is considered to be a transforming protein and enhances the potential of immortalization of E6 and E7 proteins. In addition, this protein enhances the activity of epidermal growth factor receptor (EGFR). With EGFR, E5 could interfere with several signal transduction pathways, including the mitogen-activated protein (MAP) kinase pathway (Cobo (2012)).

2.1.3.5 E6 protein

E6 is a viral protein with molecular weight around 18 kDa (Ayeda Ayed (2010)). The main function of E6 protein is its ability to bind and degrade the tumour suppressor protein p53 through the protein ligase E6 associated protein. This activity results in the inhibition of the transcriptional activity of p53 and the inhibition of apoptosis. E6 protein binds to several cellular proteins such as the proteins involved in cell polarity and motility, tumour suppressors and inducers of apoptosis, as well as the proteins for DNA replication and repair factors. Finally, E6 induces the expression and activity of telomerase and further cell immortalization (Cobo (2012)).

2.1.3.6 E7 protein

E7 is a viral protein that can be processed in three different isoforms with different molecular weights, probably because it undergoes different posttranslational modifications (Valdovinos-Torres et al. (2008)). The different isoforms are described as E7a1 (17.5 kDa), E7a (17 kDa) and E7b (16 kDa) (Valdovinos-Torres et al. (2008)). E7 binds with the tumour suppressor protein p105Rb that leads to the loss of p105Rb control over E2F transcription factors. E7 can also bind to p107 and p130. These interactions could produce the immortalization of cells and abrogate normal responses to DNA damage (Cobo (2012)).

2.1.3.7 L1 and L2 proteins

L1 is the major structural protein of papillomavirus. L1 are highly immunogenic, present conformational virus-neutralizing epitopes and could be used to detect HPV antibodies in the sera of patients with high specificity. However, L2 is the minor capsid protein of papillomavirus. L2 contributes to the interaction of the virion with the cell surface (Cobo (2012)).

2.2 Viral life cycle and viral infection

Infection is generally belevied (especially for HR-HPVs) to start in the basal layer of the mucosal epithelium to which HPV gains access through small wounds or abrasions in the mucosal epithelium. The virus initially gets to the basal lamina, and subsequently interacting with heparin sulphate proteoglycans and possibly also laminin (Doorbar (2006), Doorbar et al. (2012), Johansson and Schwartz (2013)). Structural changes in the virion capsid, facili- tate transfer to a secondary receptor on the basal keratinocyte, which is necessary for virus internalization and subsequent transfer of the viral genome to the nucleus (Doorbar (2006), Doorbar et al. (2012)). Once internalised, virions undergo endosomal transport, uncoating, and cellular sorting. The L2 protein-DNA complex ensures the correct nuclear entry of the viral genomes, while the L1 protein is retained in the endosome and ultimately subjected to lysosomal degradation (Doorbar (2006), Doorbar et al. (2012)).

2.2.1 Genome maintenance and cell proliferation

Infection is followed by an initial phase of genome amplification, and then by maintenance of the viral episome at low copy number (Doorbar (2006), Doorbar et al. (2012)). The copy number in the basal layer lesions is often proposed around 100/200 copies per cell (**?**).

The viral replication proteins E1 and E2 are thought to be essential for this initial amplification phase; in fact it was suggested they are expressed at the same time and they are the sole proteins we can find just 4 hours post infection (McKinney et al. (2015), Ozbun (2002)). E2 also regulates viral transcription, and has multiple binding sites in the viral LCR (long control region or upstream regulatory region [URR]), and (during viral DNA replication) can recruit the viral E1 helicase to a specific E1 binding motif in the viral origin of replication, making efficient the DNA replication. The level of replication must be tightly controlled during the maintenance phase of the viral life cycle. An interesting feature characterizing the DNA replication efficiency is the ratio E1/E2, which in the mantainance and proliferation phase is around 0.9-1.2 (McKinney et al. (2015), Ozbun and Meyers (1998)). The ratio is small since initially the genome replicates to low levels Johansson and Schwartz (2013).

It has been speculated that the use of a viral DNA helicase (i.e., E1), which is distinct from the cellular replication helicases (MCM proteins), allows viral DNA replication to be disconnected from cellular DNA replication during genome establishment and amplification ((Doorbar (2006), Doorbar et al. (2012), Blakaj et al. (2009)). Moreover, another proposed role for E2 is the regulation of accurate genome partitioning during basal cell division. In HPVs, other E2 binding proteins appear to be involved in the tethering of viral episomes to the cellular chromatin during cell division ((Doorbar (2006), Doorbar et al. (2012)).

The precise role of the HPV E6 and E7 proteins in infected basal cells is uncertain for the low-risk HPV types, but it is clear for high-risk HPV types, where they are important in driving cell proliferation in the basal and parabasal cell layers, especially at cervical sites where neoplasia can occur (Doorbar (2006), Doorbar et al. (2012)).

The remaining early genes (i.e. E4, E5) are not important at this stage and probably they are not transcribed or just in negligible quantities. The late genes (i.e. L1 and L2) are not transcribed at this stage but only in the final infection stages, thanks to the activation of the late promoter.

2.2.2 From genome maintenance to genome amplification

As the infected cells in the basal layer divide, the daughter cells migrate to the upper cell layer and start to differentiate (Johansson and Schwartz (2013)). The cells that commit differentiation, reaching the top stratum of the epithelium, are destined to die. This is counterproductive for the

HPVs, that have infected the cell, because they can't replicate themselves anymore, once the cell is dead. To overcome this apparent drawback, the HPV amplifies its genome replication before the infected cell death, hence guaranteeing a numerous progeny. It must be precised that the differentiation is just an apparent disadvantage for the virus, being the most efficient, if not the only one, way to spread out the infection to the neighborhood cells.

In embarking on the road of differentiation, the cell downregulates the expression of cellular factors that are required for replication of the viral genome. This downregulation is counteracted by E6 and E7 whose function is to stimulate cell cycle re-entry in the mid-epothelial layers in order to allow genome amplification (Doorbar (2006), Doorbar et al. (2012)).

In low-risk HPV infection, the basal cell proliferation is regulated by the presence of growth factors, as seen in uninfected cells. The primary role of E6 and E7 viral proteins is to drive cell cycle entry above the basal layer in order to facilitate HPV genome amplification. This is dependent on the fact E7 can bind to pRb family member p130 and to displace this latter and the associated E2Fs transcriptional repressors from target promoters required for S-phase gene expression. The transcriptional activators E2F1, 2 and 3 can the occupy the vacant sites in the latter promoters, hence stimulating expression of the host cell genes necessary for DNA replication and cell cycle progression.

In high-risk HPV infection the mechanism is the same presented in low-risk HPV infection with the main difference that E7 displaces p105 (Rb protein member) instead of p130. Another interesting dynamical feature of both low and high risk infection, consists in a compromising function of MDM in degrading p53, which subsequently leads to an increase in p53 abundance that tends to induce the cell cycle arrest. This latter is counteracted by E6 protein that mediates the degradation of p53 by means of the ubiquitin-proteasome pathway.

We have ascertained that E6 and E7 can promote the infected cell reenter in the mid-epithelial layers. However their expression in the upper epithelial layers allows to re-enter S-phase and to enhanc of viral genome copy number. To enhance viral genome copy number it is also required a strong amplification of the replication proteins E1 and E2. These latter are upregulated thanks to the late promoter which in turn is primarily activated by the cellular differentiation program. During this phase, to enhance the replication efficiency the E1/E2 ratio increased up to 2 fold for HR-HPV 31 (Ozbun and Meyers (1998)). HPV genome amplification persists as the differentiating cells move from S to G2 phase of the cell cycle, with viral genome amplification occuring primarily in G2 phase after cellular DNA replication has been completed (Banerjee et al. (2011), Wang et al. (2009)). Besides E1 and E2, also E4 and E5 are thought to have some marginal role indirectly contributing to genome amplification success by modifying the cellular environment. E5 regulation is not well understood, given its expression profile particularly elusive, mainly due to the absence of reliable E5-specific antibodies. However, it is thought to give an important contribution to genome amplification success through its ability to stabilize EGFR and to enhance EGF signalling and MAP kinase activity (Genther et al. (2003), Fehrmann et al. (2003), Straight et al. (1993)) and to modulate both ERK 1/2 and p38 indipendently of EGFR (Crusius et al. (2000)).

The MAP Kinases ERK 1/2 are also important regulators of E1 accumulation inside the nucleus of the cell. This could be important, as reported in recent works suggesting that an accumulation of E1 in the nucleus should increase viral DNA replication at the expense of cellular replication through induction of a DNA damage response (Doorbar (2006)).

About E4, given its primary function in virus release it has a secondary role in this phase, optimizing the amplification indirectly thanks to its growth arrest function.

2.2.3 PACKAGING AND RELEASING PHASE

The final step of the viral life cycle involves the production of L2, to exit the cell from its life cycle, and the expression of L1 to allow genome packaging in order to assembly the icosohedral viral capsid in the nucleus able to protect the virus in the extracellular environment. Virus maturation occurs in the most superficial, dying keratinocytes, which lose mitochondrial oxidative phosphorylation and convert from a reducing to an oxidizing environment just before virus release. Assembled particles contain 360 molecules of L1 arranged into 72 pentameric capsomeres, with a smaller and variable number of L2 molecules, which can occupy capsomeres at the 5-fold axis of symmetry (Buck et al. (2008)).

Besides L1 and L2 the third protein of pivotal importance is E4. In this phase and the previous one it has been accumulated in order to contribute to virion release and infectivity in the upper epithelial layers. In fact, it assembles into amyloid fibers (McIntosh et al. (2008)) that disrutp keratin structure and compromise the normal assembly of the cornified enevelope (Brown et al. (2006)) of the upper keratinized strata, allowing the newly formed viruses to spread out in the neighborhood cells.

In fig. 2.2.1 we report the classical evolution of the viral proteins during the entire viral life cycle.

2.2.4 PRECANCEROUS AND CANCEROUS LESIONS PROGRESSION

The HPV infection starts in the keratinocyte forming the basal layer of the epithelium and ends with that cells, embarking the road of differentiation, that reach the upper epithelial strata. In carcinogenic progression the next stage, after this ”classical” infection cycle, consists in the development of low and high grade precancerous lesions (Cervical Intraepithelial Neoplasia or CIN), and eventually Squamous Cell Carcinoma (SCC) or adenocarcinoma (ADC) conditions. In particular CIN condition is classified into three different grades: CIN-1, -2, -3.

CIN-1 is the first precancerous stage and is confined to the basal 1/3 of the epithelium (Kumar Vinay (2007)). It is the least risky stage and is usually reversible by the immune response or if opportunely treated. Moreover, CIN-1 is the only precancerous lesion that typically retains the ability to mantain a ”normal” HPV life cycle as in the classical infection (maybe with just a higher oncogenes production) and produce viral par-

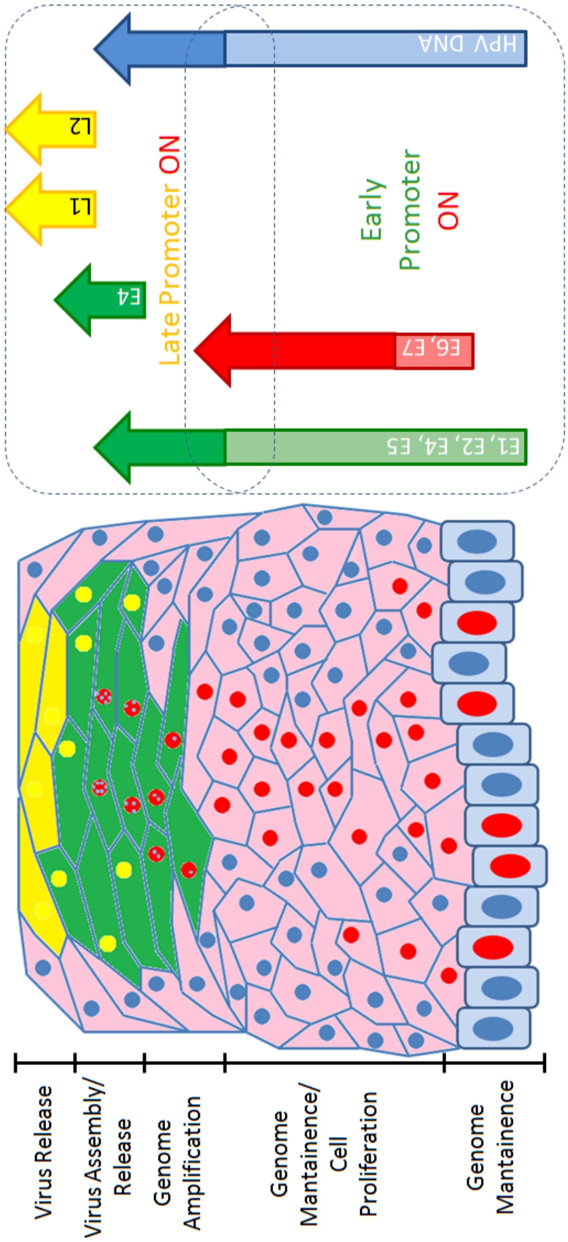

Figure 2.2.1: HPV starts the infection in the basal epithelium. During genome manteinence the first genes to be expressed are E_1 and E_2 to maintain a reservoir of viral DNA. When the cells start proliferating the oncogenes are produced in higher amount to restart the cell cycle. Upon differentiation and during the genome amplification phase the late promoter is activated and E_1 and E_2 are produced in higher amount in order to amplify the viral DNA. Even E_4 is produced given its importance in breaking down the cell membrane ath the end of the viral life cycle. During virus assembly and release fase the viral capsid genes are expressed in order to build up the viral capsid. At the end of the life cycle a new population of virus spread out, ready to infect the neighbourhood cells.

ticles in the upper epithelial strata (Johansson and Schwartz (2013),Kumar Vinay (2007)).

CIN-2 is a moderate dysplasia confined to the basal 2/3 of the epithelium and the most of the infected cells produce basically just the oncogenes and only in the very last epithelial strata there is a chance for the viral life cycle to be completed (Johansson and Schwartz (2013),Kumar Vinay (2007))

CIN-3 is a severe dysplasia extended to the whole epithelium without the normal viral life cycle and with, practically, the sole expression of the oncogenes, that are upregulated than in the normal infection. This stage is very difficult to be reversed and usually lead to invasive cancer. In general, persistent high-grade disease such as CIN-2, -3 are associated with an increasing risk of genome integration into the host cell chromosome and progression to cancer (Johansson and Schwartz (2013),Kumar Vinay (2007)).

The spatio-temporal progression of viral transcripts and proteins through the epithelium is completely deregulated in CIN-2,-3 and cervical cancer conditions. It is generally thought that levels of E6 and E7 expression increase from CIN-1 to CIN-3, and that these changes in gene expression directly underlie the neoplastic phenotype.

2.3 Early Promoter transcriptional regulation

The early promoter transcriptional regulation can be performed by either an endogenous way or an exogenous way. In the former the viral proteins are able to modulate the promoter activity, while in the second the early promoter is regulated by transcription factors produced by the infected cell. Both the ways are carried out in the LCR, which does not contain any open reading frames. LCR can be divided into three regions: 5′ LCR (bordering at the termination codon of the L1 gene), central LCR, and 3′ LCR (bordering at the transcription start site of the E6 gene), as shown in fig. 2.3.1 (Bernard (2002), Bernard (2013)). The 5′ LCR encodes the transcription termination signal of the late transcripts, the 3′ LCR con-

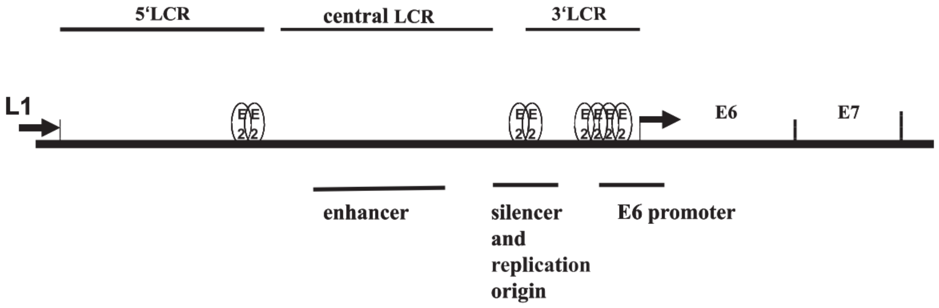

Figure 2.3.1: General organization of the long control region (LCR) of all alphapapillomaviruses.

tains the replication origin and the early promoter and the central LCR contains binding sites for the host cell enhancers and silencers transcription factors that exogenously regulate the early promoter (Bernard (2002), Bernard (2013)).

2.3.1 Endogenous regulation

Like in many promoters which makes use of the RNA polymerase II, the transcription start site is recognised by the pre-initiation complex composed of TFIID, which binds the TATA box (TATAAA, 31 to 26 bp upstream of the transcription start site), and five other general transcription factors (GTFs). 37-32 bp upstream of the TATA box there is a G rich hexamer, which is a binding site for SP1. This latter is a transcription factor necessary for the early promoter activation (Bernard (2002), Bernard (2013), Thierry (2009)), as shown in fig. 2.3.2.

The most important regulator of the early promoter endogenous modulation is the E2 viral protein, produced by the same promoter, that can bind in form of a dimer (DE2) upstream the promoter. More precisely, there are four specific E2 binding sites (E2BS) inside the LCR upstream the early promoter (Muller and Demeret (2012), Soeda et al. (2006), Bernard (2002), Demeret et al. (1997), Bernard (2013), Thierry (2009)), as shown in fig. 2.3.2. Two of these sites (E2BS #1 and #2) are about 50 bp upstream of the transcription start site, positioned exactly between the Sp1 binding site and the TATA box; a third site (E2BS #3) is about another 100 bp further upstream, and a fourth site (E2BS #4) about halfway between the end of L1 and E2BS #3. The E2BS #1, #2 and #3 control the repression of the early promoter, while E2BS #4 control its activation(Muller and Demeret (2012), Soeda et al. (2006), Demeret et al. (1997), Bernard (2013), Tan et al. (1994)). The four E2 binding sites have similar affinities (Demeret et al. (1997)), hence the E2 dimer (DE2) can bind with more or less the same probability all of its binding sites. However, the binding sites have different stabilities and precisely the half lives of the complexes DE2-DNA are ordered as follow: $t_h^{E2BS\#4} > t_h^{E2BS\#2} > t_h^{E2BS\#3} > t_h^{E2BS\#1}$ (Demeret et al. (1997)). These half lives contain the information of the DE2 dimer dissociation rate constant from its binding site in the DNA (Dukhovich (2002)). The dissociation rate constant is proportional to the dissociation probability of the DE2 dimer from its E2BS in the DNA. Consequently, from the previous information we can observe that it's more difficult to detach from E2BS #4, controlling the promoter activation, than respect to the other three binding sites, controlling the promoter repression.

The transcription starts when the pre-initiation complex (GTFs, TFIID, RNA polymerase II) and the Sp1 transcription factor are bound upstream the promoter. Transcripts that result from this event encode all the early genes, hence E2 gene, too. E2 protein translated from its transcript can bind, as a dimer DE2, to the promoter binding sites E2BS #i^{th}. E2 can displaces Sp1, binding E2BS #1, #2 and #3, thus leading to a down-modulation of the early promoter (Bernard (2002), Bernard (2013)). At a high concentration of E2 protein, E2 does not only displace Sp1 but also TFIID, as it binds to E2BS #1, leading to efficient repression of the promoter (Bernard (2002), Bernard (2013)). It is interesting to notice that, the necessity of a high E2 concentration to efficiently bind the E2BS #1 is consistent, with the previously reported experimental evidence, that this latter binding site has the lowest dissociation half life for the DE2 dimer (i.e. the DE2 dimer dissociation probability is very high, thus only when in high concentration we can be sure it will be bound most of the time). This is

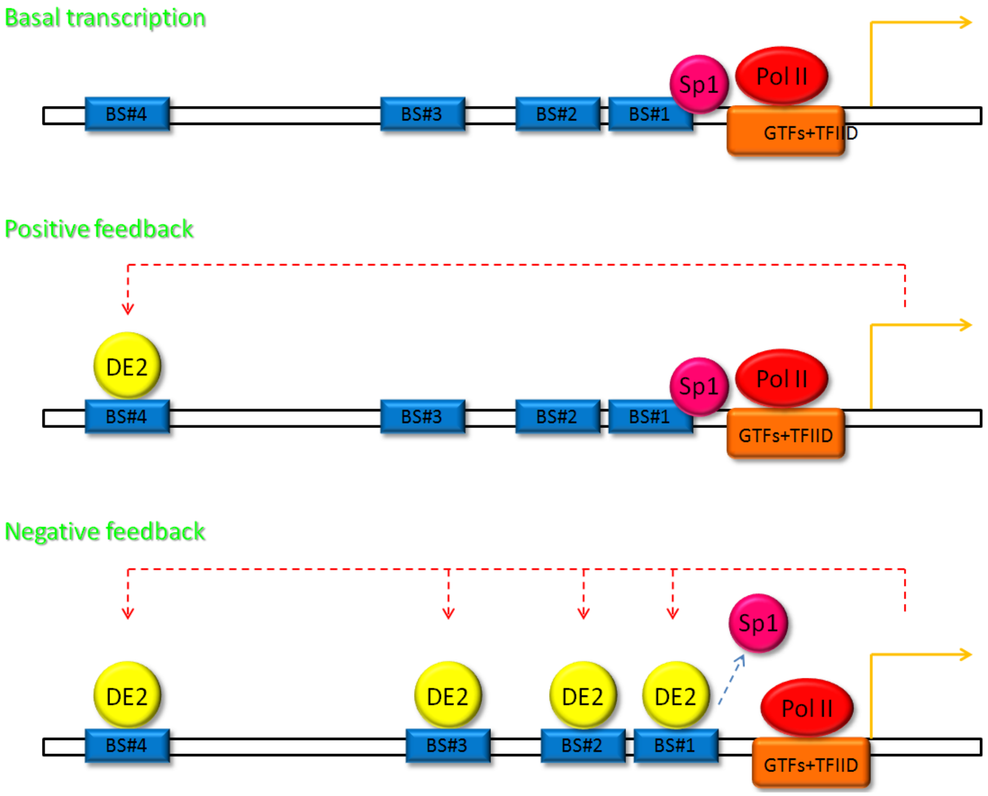

Figure 2.3.2: Endogenous minimal model of HPV transactivation. A basal transcription starts when $Sp1$ human transcription factor binds the early promoter. When E_2 is in low concentration it induces a positive feedback binding, in form of a dimer, to $BS4$. When E_2 is in high concentration it binds, in form of a dimer to all the binding sites inducing a negative feedback on the early promoter.

clearly a negative feedback loop that keeps the early promoter at a low steady-state activity, as increased promoter activity raises also the level of E2 transcripts and consequently of E2 proteins.

Besides the negative feedback regulation, it is widely accepted in literature that when E2 is in low concentration it can positively regulate the early promoter. In fact, as previously reported, E2BS #4 has the lowest dissociation probability and being the latter binding site associated to the positive regulation of the early promoter, this is an experimental proof that when E2 is in low concentration the positive feedback should win over the negative one. Actually, even if we know that E2 can positively regulate the early promoter, the mechanism behind still remains not well understood, yet. However, a quite recent work had partially shed light on this important question. It was proven that high levels of the co-activator CBP/p300 and the transcription factor CEBP/alpha can synergize with E2 in order to strongly activate the early promoter up to 12 fold increase (Krüppel et al. (2008)). This shows the possibility to have a strong positive feedback on the early promoter regulation. CEBP/alpha and CBP/p300 were shown to increase during the keratinocyte differentiation (Krüppel et al. (2008)) where the late promoter is activated and negatively regulates the early promoter. Nevertheless, this latter regulation performed by the late promoter does not mean the early promoter must be completely shut down, but just strongly controlled in order to contain the oncogenes expression. Probably this strong positive regulation of the early promoter can happen in the first phase of the differentiation. This could be consistent with the sustained oncogenes expression in CIN-1,-2 with respect the normal viral life cycle. In fact, p300 expression is high in the suprabasal layers of HPV16 CIN-1,-2 pre-cancerous conditions (Krüppel et al. (2008)).

The consequence of these various mechanisms is a feedback that balances between strong repression, which would lead to termination of the viral life cycle, and strong expression with a more fulminant course of the HPV infection.

In Cahpter 7 we will report an interesting data set about the early promoter activity of HPV-16 and -18 in function of E2 protein concentration (Hou et al. (2002)). This data set takes into account the sole repression function of E2. Unfortunately we haven't any available data about E2 positive regulation. However, this data set is an interesting and important state of the art of the biological knowledge about the early promoter functioning and it will be used in to accurately tune the model of the promoter activity we have designed and show it is able to fit the experimental data.

Another interesting and important regulatory feature of the early promoter is the E2 regulation mediated by E1. In particular, E1 is able to stabilize E2 by a direct protein- protein interaction (King et al. (2011)). In general, in HPV-16 and -18, E2 has an half-life ranging from 45 min to 3 h (Bellanger et al. (2001), McBride (2013), Taylor et al. (2003)). However, it was shown the E2 half-life can increase, in the presence of E1, from 2.6 h to 3.5 h in 293 T cells line, and from 3.6 h to 5.7 h in C33A cervical carcinoma cell line.

HPV16 E2 is turned over via the proteasome following ubiquitinylation pathway (Taylor et al. (2003)) and therefore E1 presumably stabilises E2 by preventing this turnover. Recent studies have proven that the Brd4 C-terminal domain can inhibit the interaction between E2 and the proteosome pathway, thus increasing E2 stability. It is possible that E1 acts in a similar manner preventing E2 degradation (King et al. (2011)).

Since E2 also interacts with some components of the SUMO proteins (King et al. (2011)), it is alternatively possible that the Sumoylation of the E1 and E2 proteins is, in turn, involved in E2 protein turnover regulation (King et al. (2011)). Probably, even the E1/E2 tetramer complex formation stabilizes E2 due its importance in recruiting E1 at the origin of replication and increasing the E1 binding specificity. Another important function of E1 co-regulation is that it increases the E2 transcription up to 2-4 fold (King et al. (2011)). This is probably accomplished at a post-transcriptional level since the early promoter produces a pre-mRNA encoding all the early genes. Unfortunately, the exact mechanism hasn't been found yet, nevertheless we'll show in chapter 4 how we'll handle the modeling of this important dynamical feature.

2.3.1.1 Regulation of viral DNA replication

The double-stranded circular DNA replicates as a multi-copy extrachromosomal plasmid in the nucleus of infected cells. The replication of papillomavirus DNA is initiated from the origin of replication consisting of binding sites for E1 and E2 proteins (Bernard (2013), Kurg (2009)).

The papillomavirus origin of replication consists of three E2 binding sites (E2BS), from which only one is absolutely required for replication, and an AT rich region containing a cluster of E1 protein binding sites (E1BS) (Kurg (2009), Robert L. Garcea (2007)). In particular, E1 is the main initiator protein of HPV DNA replication. E1 is responsible for recognition of the replication origin, as well as for subsequent unwinding of the double helix (Kurg (2009), Robert L. Garcea (2007)). The viral E1 protein is an ATP-dependent helicase which binds as a dimer to a pair of its binding sites, but E1 by itself binds to the origin with low specificity. However, in the presence of E2 the specificity is increased. In fact, the E1 and E2 proteins form a tetramer complex, called E1/E2 (a dimer composed by one DE1 and one DE2 dimers), through multiple protein-protein interactions and bind cooperatively with high specificity to adjacent binding sites in the origin of replication.

In the next step, additional E1 molecules are added by displacing E2 from the DNA-bound complex in an ATP-dependent manner. Subsequently, two additional E1 molecules are recruited to the origin, which results in the formation of two E1 trimers on the ori, followed by formation of two hexamers in the presence of ATP. E1 hexameric complex has the DNA helicase activity which is able to unwind the DNA and start the viral DNA replication (Kurg (2009), Robert L. Garcea (2007)). It is not well understood what kind of regulation the early promoter can be subjected when the E1/E2 complex is bound to the ori. It was proposed a model for HPV-11 that suggests there is a minimal replication at the same time of the transcription in the basal cells, where only the early promoter is activated, while in differentiating keratinocyte the transcription is stopped during the replication (Hartley and Alexander (2002)). However, despite this latter model, since the E1/E2 complex binds to the origin of replication, made of E1 binding sites and E2BS just related to the promoter repression, it is reasonable to hypothesize a transcriptional repression when E1/E2 is bound. The dissociation constant for E1/E2 from its DNA binding site was found to be $k_d = 2.2[nM]$ for HPV-11 (Chao et al. (1999)). However, it can be inferred a similar, if not equal, dissociation constant for high-risk HPV-16, -18, too. Unfortunately, the dissociation constant of the E1/E2 complex into its DE1 and DE2 dimers seems to be unavailable in literature and the same for its half life. However, since E2 is more stable in the presence of E1, as reported in the previous paragraph, it is reasonable to assume E1/E2 half life higher than E2 monomer half life and probably higher than DE2 dimer.

2.3.2 Exogenous regulation by the infected cell

The central LCR is divided into two parts: *i)* the silencer part (bordering the 3'LCR to the right) containing binding sites for exogenous transcription factors (TFs) coming from the infected cell that have the capability to strongly repress the early promoter activity; *ii)* the enhancer part (bordering the 5' LCR to the left and the silencer LCR part to the right) containing binding sites for exogenous TFs from the host cell that can strongly activate the early promoter.

In fig. 2.3.3 is shown the complete LCR TFs. There are around 15 TFs among enhancers and silencers, however half of them are not well understood yet. In what follows we are reporting just the principal enhancers and silencers.

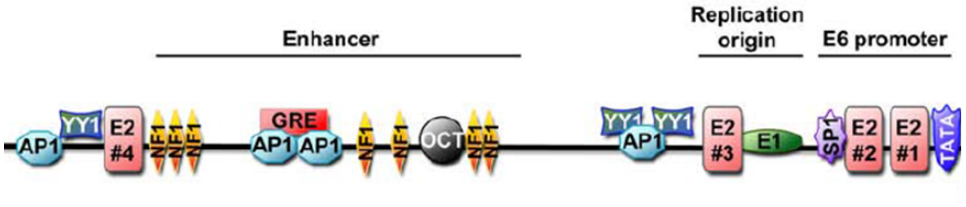

Figure 2.3.3: Regulation of the early promoter, mediated by exogenous transcription factors that can enhance or silence the transcription.

2.3.2.1 AP-1 enhancer

AP-1 is a heterodimer composed of subunits of the fos and jun gene family and its typical AP-1 binding DNA element is the sequence TGANTCA (Bernard (2002)).

AP-1 is considered the principal activator of the HPV enhancers and it could be important in modulating the transcriptional activity even during the epithelial differentiation, hence during the whole viral life cycle (Bernard (2013)).

2.3.2.2 NFI enhancer

NFI constitutes a family of factors derived from four genes, NFI-A, B, C and X. NFI factors occur in dimeric form, with the two subunits potentially derived from the four different NFI genes and also from different spliced products, even if in the epithelial cells NFI is mostly composed by NFI-C subunits, while in non-epithelial cell it is mostly composed by NFI-X subunits (Bernard (2013), Bernard (2002)). Its binding sites represent the most conspicuous binding elements of the central LCR and is a cluster of half-palindromic TTGGCT/A sequences.

NFI is a transcriptional weak activator, that appears to play an important role even in the modulation of HPV16 carcinogenesis; moreover, its binding sites activate HPV enhancers in synergy with other factors (i.e. there is a cooperation between the TFs) (Bernard (2013)).

2.3.2.3 Oct1 enhancer

Oct1 is a member of POU factors and its recognition binding site is represented by the DNA sequence ATGCAAAT. Oct1 motifs exist in the HR-HPV LCR and it is usually very close (2 bp) from a hal-palindromic NFI binding site. Oct1 is not really an enehancer, but it seems to tether NFI to the enhancer LCR zone rather than being a transcriptional activatore by itself (Bernard (2013)).

2.3.2.4 TEF1 enhancer

TEF1 is an enhancer not well understood yet. It is known that it binds multiple sites in HPV-16 LCR enhancer zone, especially the motifs ACATACCG and ACATATTT. TEF1 requires a cofactor (not identified yet), that does not directly bind the DNA and helps TEF1 in contributing to the epithelial specific function (Bernard (2013), Bernard (2002)).

2.3.2.5 YY1 silencer

YY1 is a transcription factor that can exhibit positive and negative controls and usually binds the DNA sequence CCGCCATNTT. It can activate or repress transcription directly, by interacting with histone-deacetylases or histone-acetyltransferase, or indirectly by interacting with other transcription factors such as $C/EBP\beta$ (Bernard (2013), Bernard (2002)). However, there are numerous binding sites inside the LCR where YY1 negatively regulates the promoter. An opportune modulation of YY1 can counteract the transcriptional stimulation by enhancers such as AP-1, NFI, Oct1 and TEF-1 (Bernard (2013)).

2.3.2.6 CBP and CBP/p300

CBP (CREB binding protein) and p300 (also called CBP/p300) are cellular co-activators of transcription, usually involved in control of HPV gene expression. They are co-activators for the exogenous cellular transcription factors, such as AP1, C/EBP but also for the endogenous E2 viral protein (Krüppel et al. (2008)).

CBP/p300 is active in several biological processes, such as differentiation control, apoptosis and cell cycle regulation (Krüppel et al. (2008)). It can also synergize with E2 as briefly explained in the following paragraph.

2.3.2.7 C/EBPs regulators

$C/EBP\alpha$, β and δ are three transcription factors of the C/EBPs family, that can usually bind the sequence ATTGCGCAAT and have different effects on HPVs. $C/EBP\delta$ and $C/EBP\alpha$ are repressors for both HPV16 and HPV11 (Bernard (2013)). Besides, $C/EBP\alpha$ bound at the position 480, downstream the early promoter, can sinergize with the E2 viral protein and CBP/p300 cofactor in order to strongly activate the early promoter (Krüppel et al. (2008)).

A C/EBP site in the LCR silencer zone of HPV18 is capable to sinergize with YY1 and repress the promoter in the HeLa cells line (Bernard (2013)).

C/EBPbeta, together with its isoforms, is also proposed to be one of the most important regulator in the activation of the Late promoter, as better discussed in later paragraphs.

2.3.2.8 CDP silencer

CDP (that stands for CCAAT-displacement protein) is a transcription factor able to repress the transcription and at the same time negatively regulate the DNA replication (**?**). This latter function is possible thanks to its binding sites (sequences TATAATAAT and TACAATAAT) that overlap the E1 binding site. Hence CDP can interfere with the normal E1 function in replicating the viral DNA (Bernard (2013)).

CDP is an interesting silencer even during the differentiation, since it is usually decreased in differentiated epithelial cells (Bernard (2013)).

2.4 Post-transcriptional regulation

In this section we report the biological state of the art knowledge about the post-transcriptional regulation of HPV.

Both the early and the late promoters produce a polycistronic pre-mRNA. The arely pre-mRNA encodes for all the early genes while the late pre-mRNA encodes for the late genes and for all the early genes but the oncogenes. Thus, HPV splicing regulation is really very important for understanding how the different mRNAs are generated from the respective pre-mRNA and modulated during the entire viral life cycle.

We will refer to HPV-16 splicing regulation, if not differently specified, since it is the most studied and understood in terms of post-transcriptional and post-translational regulations. Nevertheless, HPV-16 splicing regulation seems to be preatty similar to the other HPV-18 and -31.

2.4.1 The major 3′ splice site

SA3358 is the major 3′ splie site used in HPV-16. It is maybe, as far as we know, the most important and commonly used splicing site, having a central role in the production of the early mRNAs encoding E6, E7, E4 and E5 but preventing the production of E1 and E2 mRNAs (Johansson and Schwartz (2013), Somberg and Schwartz (2010), Schwartz (2013)). It is also used for the production of the late genes (Johansson and Schwartz (2013), Schwartz (2013)), as we will see later. The region between SA3358 and SD3632 contains the binding sites for different SR proteins that act as splicing enhancers and silencers, as shown in fig. 2.4.1, that are SR. Splice site SA3358 is enhanced by binding of SRSF1 splicing factor (formerly known as ASF1/SF2) to the downstream enhancer site and seems to be inhibited by the binding of SRSF9 (formerly known as SRp30C) and SRSF3 (formerly known as SRp20) splicing factors to the downstream enhancer and silencer splicing sites, respectively.

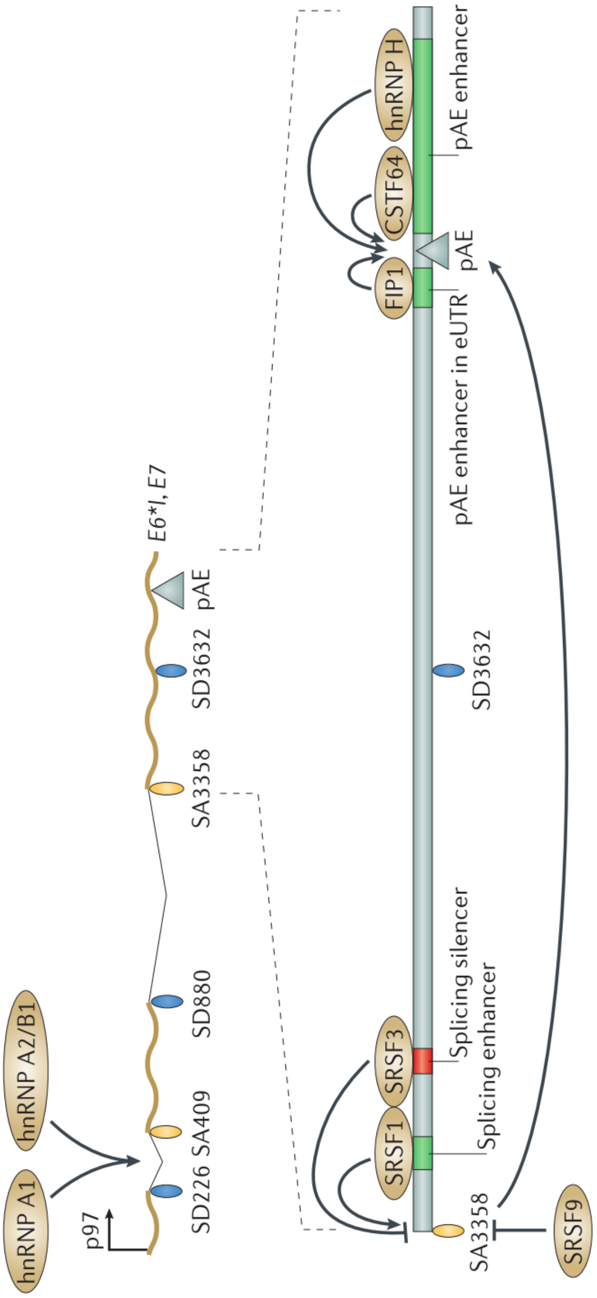

Figure 2.4.1: Splicing regulation of the major 3' splicing site; splicing regulation of the oncogenes between splicing sites SD226 and SA409. Positive regulation of early polyadenylation signal pAE.

2.4.2 E1, E2 and E4 splicing regulation

As far as we know, the most important splicing mechanism that regulates E2, E4 and E1 is regarding the mutual exclusivity between the major splicing site SA3358 and SA2709 site. Splicing site SA3358 generates E4 but not E2, since SA3358 competes with SA2709 E2 splicing site (Johansson and Schwartz (2013)). Conversly, when the splicing site SA2709 is activated E2 mRNA is generated but not E4 mRNA. Both E1 and E2 mRNAs are derived from the same pre-mRNA (Chow et al. (2010)) and when SA3358 is activated the E1 mRNA production is prevented, as well.

A better comprehension of E1, E2 and E4 splicing regulation is required in order to achieve a better understanding on how their transcripts are regulated during both the early and late stage of the infection.

We could observe that since the 3′ end of the E1 ORF overlaps the 5′ end of the E2 ORF (Chow et al. (2010)) there is a possibility they are alternatively transcribed (i.e. they couldn't be generated at the same time). Nevertheless, we know that E1 and E2 are present at the same time (King et al. (2011)), probably due to their important co-regulations previously argued, and actually they appear to be synchronous to each other (Ozbun and Meyers (1998)).

2.4.3 Oncogenes splicing regulation

Probably the most important knowledge about the oncogenes splicing modulation is that high levels of SRSF1, in the basal layers of the epithelium, drives E6 and E7 mRNAs expression eventually reducing E2 expression. This is because high levels of SRSF1 stimulate splicing at SA3358 at the expense of splicing to the E2 mRNAs site S2709, because of their mutual exclusivity (Johansson and Schwartz (2013)).

E6 and E7 are not produced at the same time due to another splicing mechanism involving a cluster of splicing sites upstream SA3358 site. In fact, splicing between SD226 and SA409, by retention of the first intron, generates mRNAs that are translated to E6*I (an E6 protein variant) and E7 proteins (by leaky scanning (Stacey et al., 1995), by a ribosomal shunting mechanism (Remm et al., 1999) or by a translation-reinitiation mechanism (Tang et al., 2006)) whereas mRNAs that remain unspliced between SD226 and SA409 produce full-length E6 (Schwartz (2013)). Splicing between SD226 and SA409 are stimulated by heterogeneous nuclear ribonucleoprotein hnRNP A1 and hnRNP A2 (as shown in fig. 2.4.1), whose expression is downregulated in differentiated keratinocytes.

The post-transcriptional regulation of the oncogenes is regulated even by the epidermal growth factor (EGF). High levels of EGF lead to the activation of ERK1-ERK2 pathway, resulting in the inhibition of splicing between SD226 and SA409, thus leadng to E6 mRNAs. Conversly, a reduction in EGF favours E6*I and E7 mRNAs production.

2.4.4 Early Polyadenilation regulation

The main role of the early polyadenylation signal, pAE, is to efficiently regulate the expression of the late genes L1 and L2. This is crucial in order to prevent a too early production of the late proteins that could jeopardise the ability of the virus to persist long enough to be spread out in the neighborhood cells or to be transmitted to another individual. The activity of the pAe is enhanced by the binding of human factor FIP1 to the early 3′ UTR (eUTR) located upstream the pAE, but it is also enhanced by downstream stimulatory elements such as hnRNP H and the polyadenylation factor cleavage stimulation factor 64 kDa subunit (CSTF4) (Johansson and Schwartz (2013), Schwartz (2013)). The schematic regulation of the early polyadenylation by these latter factors is reported in (2.4.1).

The last, and maybe most important, interesting mechanism in regulating the pAE is mediated by the viral protein E2. This can be produce by both the early and late promoters. This latter produces E2 in higher concentrations with respect the early promoter can do (as better explained in the paragraph about the late promoter regulation). When E2 reaches enough high concentrations, can repress the pAE, allowing readthrough into the late region of the HPV genome, paving the way for the production of the late mRNAs L1 and L2 (Johansson and Schwartz

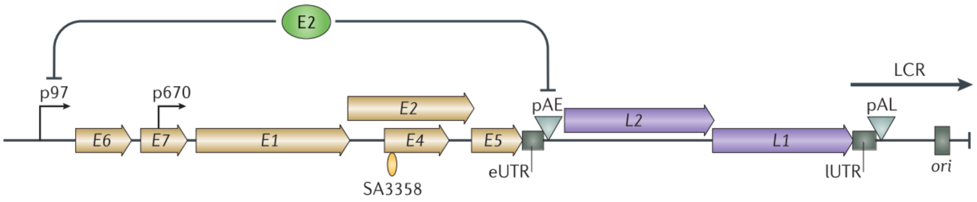

Figure 2.4.2: E_2 negatively regulate the early promoter and inhibits the early polyadenylation signal pAE, paving the way for L_1 and L_2 transcription.

(2013)). Nevertheless, we have to specify that E2 does not bind directly to some pAE upstream or downstream binding site. It targets the general factor CPSF30 that is required for polyadenylation of the majority of the mRNAs in the cell (Johansson et al. (2012)). Therefore E2 acts as a co-factor in inhibiting the early polyadenylation pAE. The main mechanisms mediated by E2 are reported in fig. 2.4.2.

We recall the packaging of viral genomes and virus release phases to report a last interesting fact linked to the polyadenylation regulation. Actually, the completion of the HPV life cycle requires to exit the cell from the cell cycle, the expression of L2 and of L1 to allow genome packaging of the viral genetic makeup. This definitely requires a post-transcriptional regulation leading to transcripts that are polyadenylated at the late polyadenilation site (pAL) rather than the early site (pAE). This results in a switch from the late promoter production of $E1\hat{}E4$ and E5 mRNAs (besides E1 and E2) to a production of $E1\hat{}E4$, L1 and L2 mRNAs (besides E1 and E2) (Doorbar (2005), Doorbar et al. (2012)).

2.4.5 Late genes splicing regulation

The most important mechanism in regulating the late genes transcription is probably the regulation of the early polyadenylation site pAE. Nevertheless, we report the other poorly understand additional regulations.

The expression of the late genes seems to be linked to the downregulation of different SR proteins during keratinocyte terminal differentiation. L1 mRNA is exclusively spliced to the splice sites SD3632 and SA5639 (Johansson and Schwartz (2013), Schwartz (2013)). Conversly with the idea that late genes are expressed when SR proteins are downregulated, it was found out that SRSF9 and SRSF1 can enhance splicing to the 3′ splice site SA5639, thus promoting L1 mRNA production. In a similar manner hNRNP A2/B1 cand induce HPV late gene expression, while hNRNP A1 inhibts late gene expression (Johansson and Schwartz (2013), Schwartz (2013)).

SD3632 is suppressed by some splicing factors both upstream and downstream the splice site but they haven't been identified yet. What is for sure is that SD3632 is totally inhibited in in cervical cancer cells due to their lack of differentiation (Johansson and Schwartz (2013), Schwartz (2013)).

L2 mRNA splicing regulation is understood even less than L1. It is spliced from splice donor SD880 to splice acceptor SA3358. Its translation is repressed when the molecular complex composed by the heterogeneous nuclear ribonucleoprotein K (hnRNP K), poly(C)-binding protein 1 (PCBP1) and PCBP2 is bound to L2 coding region (Johansson and Schwartz (2013), Schwartz (2013)).

We will not report the regulation of the late polyadenylation site (pAL) since its biological knowledge is poor and its regulatory importance is not well understood yet.

2.4.6 Splicing factor SRSF1

Given the importance of SRSF1 in regulating the major 3′ HPV splicing site we dedicate a little section to its molecular biology.

SRSF1 is an important protein, of the Serine/Arginine-rich (SR) proteins family, for the regulation of constitutive and alternative splicing of cellular pre-mRNAs. It has an half life of about 1.5 h in T cells (Moulton et al. (2014)). Overexpression of SRSF1 has been reported in various tumors types and this has consequences for the alternative splicing profile expressed in tumor cells.

We have seen it is involved in the regulation of the HPV major 3′ SA3358 splice site and it is vital to effciently produce the regulatory elements (E1 and E2) and the oncogenes (E6 and E4). However it is important to drive E4 expression during the cellular differentiation, at expense of E2 and E1, and to enhance the late genes (L1 and L2) transcription.

It was found that SRSF1 can autoregulate its own expression and this involves multiple layers of post-transcriptional and translational control.

SRSF1 has a negative auto feedback and seems that this auto-regulatory dynamics is quite slow, around 3 days (Sun et al. (2010)). Its autoregulation is mediated by its alternative splicing regulation, translational control and potential contribution of some miRNAs not identified yet. In particular, the 3′ UTR is necessary and sufficient for autoregulation (Sun et al. (2010)). Probably nuclear SFRSF1 affetcs the mRNA composition of its own transcript, which in turn affects how efficiently it is translated in the cytoplasm (Sun et al. (2010)).

It can enhance splicing at SA5639 splicing site (Somberg and Schwartz (2010)), controlling the expression of L1 mRNA in the late phases of the viral life cycle. In particular, SRSF1 has a molecular weight of around 30 kDa and to have some effects on L1 regulation its molar concentration should vary between 0.2 and 2 [nM] (Somberg and Schwartz (2010)).

Interestingly, it can also increase up to 2-3 fold upon epithelial differentiation (Mole et al. (2009)), hence during the late promoter usage. In this way it can up-regulate the expression of E4 and L1 during the late stages of the infection.

2.4.7 Late Promoter regulation

The discovery of the late promoter existence is pretty recent. The paper reviews about the molecular biology of Human papillomavirus began to support the possibility of the existence of a such promoter around the beginning of 2000 (Bernard (2002)). The late promoter, located inside the E7 ORF, controls the expression of the early genes, with the exception of the oncogenes, and the late genes in host cells that have started terminal differentiation (Kukimoto et al. (2006)). The tight linkage of viral late-gene expression and epithelial differentiation suggets that differentiation-specific cellular factors are involved in the regulation of the late promoter. Unfortunately, little information is available in literature about the identity of these factors and hence about the regulation of the late promoter itself. In this paragraph we gather the most important knowledge about this important promoter in order to have some minimal insights in its transcriptional regulation in order to be able to develop at least a first minimal model as we'll see in chapter 4.

Most of the transcriptional regulators of the late promoter are not even discovered yet. Nevertheless, it is pretty sure that one of the most important regulators is the $C/EBP\beta$ transcription factor (as previously reported). It was found that $C/EBP\beta$ and its isoforms have different binding sites in the LCR where can help in repressing the eraly promoter (Gunasekharan et al. (2012)) and two binding sites upstream the late promoter region (Kukimoto et al. (2006)). When $C/EBP\beta$ binds to the late promoter binding sites the late promoter transcription starts and can be enhanced (Kukimoto et al. (2006), Gunasekharan et al. (2012)).

It has been shown that a rearrangement of chromatin occurs around the late promoter region on epithelial differentiation (Longworth and Laimins (2004)). $C/EBP\beta$ recruits SWI/SNF chromatin complex to the upstream region of the HPV late promoter. This mechanism induces rearrangement of the nucleosome structure around the promoter, resulting in the appearance of a nucleosome-free region where transcription initiates (Kukimoto et al. (2006), Kowenz-Leutz and Leutz (1999)). This is an interesting molecular evidence since the chromating remodeling can take time, hence the activation of the transcription mediated by $C/EBP\beta$ could insert a delay in the activation of the late promoter.

The gene that encodes for $C/EBP\beta$ is intronless, hence it is transcribed into a single mRNA that in turn, through the usage of alternative start codons, gives rise to three isoforms: two full length $C/EBP\beta$ proteins LAP and LAP* referred to as liver-enriched transcriptional activator proteins and the LIP proteins, a short repressive isoforms. LIP lacks the transactivating domain found in the N terminus of LAP and LAP*, and is unable to recruit histone acetyltransferases to activate transcription. Usually, LIP dimerizes with LAP in order to form the LIP-LAP heterodimer acting as a strong repressor of the late promoter activity (Gunasekharan et al. (2012)).

The most recent biological idea of a model about the transcriptional regulation of the late promoter in HR-HPV 31 is the following one. In undifferentiated cells, LIP-LAP heterodimers predominate over LAP-LAP dimers, thus repressing the late promoter activity. Upon differentiation LAP-LAP homodimers predominate, resulting in a strong activation of the late promoter (Gunasekharan et al. (2012)). In fact, it was found that $C/EBP\beta$ (in its LAP forms) is expressed in high levels in the middle and upper stratum spinosum of the epithelium, where the cells are differentiating and the HPV late genes are actively transcribed (Kukimoto et al. (2006)).

We think we could extend the latter biological idea of the late promoter regulation to the other HR-HPVs in particular to HPV-18 and -16 whose structure is similar to HPV-31.

The late promoter produces the early genes with the exception of the oncogenes in higher amount compared to what the early promoter does. Moreover, there are experimental evidences indicating the slow increase in time of E1 and E2 mRNAs produced by the late promoter. They could take more than half of the total differentiation time to reach their maximum level (Ozbun and Meyers (1998)), where the differentiation time is among 6 and 16 days for in vitro systems. This is consistent with the regulatory mechanism of the late genes (L1 and L2) mediated by E2, namely when E2 reaches enough concentration the early polyadenilation site pAE is repressed, paving the way for the late genes production as explained in section (2.4.4). The late genes are procued in the terminal phase of the viral life cycle therefore there is the necessity of a delay in their production and this is probably guaranteed by a slow rate increase of E2 protein accumulation. This could be done by a stabilization of E2 protein mediated for example by the ubiquitin proteasome pathway as speculated in literature (King et al. (2011)). Another possible mechanism to slowly reach high E2 protein levels consists in a slow increase to higher level of its transcript, as well. As far as we know from the literature, the latter mechanism seems to be the dominant one.

To slowly increase the transcripts conversion from the primary transcript we could speculate the possibility of a delay in the post- transcriptional regulation. This is surely possible because of the SA3358 major 3′ splicing site is controlled by SRSF1, in turn activated by E2. Higher the E2 level is, higher SRSF1 should be, resulting in an increased probability of SA3358 to be activated. This imply a higher conversion of E4 mRNA at the expense of E1 and E2 mRNAs. However, it is very likely that the primary transcript in vivo is processed, converted and degraded (the non converted part) with a very fast dynamics, from around 0.4 minutes (Audibert et al. (2002)) up to 1.2 hours (Hicks et al. (2005)). This is the kinetics for eukaryotic cells splicing where the transcription is mediated by RNA pol II. Probably the virus will follows the same kinetics in splicing regulation being a DNA virus, therefore it makes use of the DNA replication machinery of the host cell and of the RNA pol II for its transcriptional regulation. It is most likely the slow dynamical increase in mRNAs produced by the late promoter is due to a slow modulation increase in its activity. This is consistent with the dynamical behavior of its major activator, the transcription factor $C/EBP\beta$. It was found out that $C/EBP\beta$ has an half life between 2 and 6 h (Sears and Sealy (1994). Nevertheless, in an interesting experiment where was induced differentiation for 6 days, mediated by calcium, was shown an increase in $C/EBP\beta$ up to 4 days upon the beginning of differentiation (Maytin and Habener (1998)). This evidence, compared with the transcription factor half-life, suggests a presence of some kind of regulatory mechanism in delaying the growth of $C/EBP\beta$. In fact, recently was experimentally proven that $C/EBP\beta$-LAP*/LAP expression has a marked increase during differentiation of monocytes. This increase can be delayed up to 3 days and is essentially due to MEK-RSK-dependent signalling cascade (Huber et al. (2015)). Interestingly, this regulation doesn't occur at a transcriptional level but at post-translational regulation. This latter can strongly stabilize LAP*/LAP proteins which levels in the absence of differentiation-inducing stimuli are almost completely degraded in 12 h, while upon differentiation the complex half-life was shown to be $\geq 24h$ (Huber et al. (2015)). In some cases it takes up to 3 days to reach the steady state condition. All the previous evidences where in the context of experiments with a total duration of a week, we can observe that $C/EBP\beta$ takes most of the differentiation time to reach its steady state level.

In a recent study it was found the importance of p63 protein (of the oncosuppressor protein p53 family) in the activation of the late promoter. p63 is one of the major regulator of the keratynocyte differentiation acting in a crosstalk with Notch1, codifying for membrane proteins, another important gene for achieving a good differentiation. These two important proteins in the control of differentiation act in order to negatively repress each other (Nguyen et al. (2006), Dotto (2009)). During normal keratynocyte differentiation, p63 is decrease to barely detectable levels and Notch is in high concentration. Although p63 levels also decrease upon differentiation in HPV-positive cells, the p63 levels are retained at higher levels than in normal keratinocytes. It is necessary a minimum level of p63 to activate the late promoter (Mighty and Laimins (2011). It was speculated the presence of a binding site for p63 to the late promoter but it was not found so far. Nevertheless, this show the complexity of late promoter regulation and even if, as far as we know from the current biology, the major regulator seems to be $C/EBP\beta$, the regulation is probably orchestrated by a lot of other co-factors that in turn could accelerate or most likely delay the promoter activation.

2.5 Integrated HPV

Persistent infections with high-risk human papillomaviruses (HPVs) induce dysplastic lesions of the lower genital tract. Some of these lesions eventually progress to invasive cancers, particularly of the uterine cervix.In many advanced preneoplastic cervical lesions and most derived carcinomas, HPV genomes are found to be integrated into the host cell chromosomes.Although HPV integration seems to play an important role in the progression of cervical dysplasia, the underlying mechanisms are still unclear. DNA integration occurs in the presence of Double Strand Breaks (DSBs). DSBs occur in regions of DNA in which the DNA repair process has failed. Regions that harbour DNA instability, known as Chromosome Fragile Sites (CFSs), are distributed throughout the genome. Studies have reported an increased frequency of HR-HPV integration in regions of DNA that contains CFSs (Raybould (2011), Schmitz et al. (2012), Shin et al. (2014)).

Different studies have demonstrated that upon viral integration, different parts of the HPV genome are disrupted; fragments containing E_2, E_4 and E_5 ORFs are missing whereas the entire E_1, E_6 and E_7 ORFs are integrated and retained. These phenomena bring to an over expression of the oncogenes since E_2 is not produced anymore, resulting in a decreasing of the negative feedback strength (Raybould (2011), Schmitz et al. (2012)).

2.6 Antiviral therapies proposals

In the past years different antiviral therapy proposals were investigated.

It was proposed to target E_1 and E_2 fro the development of antiviral agents. This is because of the replication of the viral genome is accomplished by the viral E_1 and E_2 proteins in concert with the cellular DNA replication machinery (Bernard (2002)).

L_1 and L_2 are other possible targets since they are responsible for the viral capsid formation. If we could stop their synthesis the viral DNA wouldn't have a protecting envelope in the extracellular environment, resulting in the impossibility for the virus to spread out the infection (Seth (2008)).

When the virus is integrated E_2 ORFs are disrupted, resulting in an overexpression of the E_6 and E_7 oncogenes. In these cases, targeting these latter or the cellular proteins mediating their oncogenicity is clearly the only antiviral possible approach. E_7 from high-risk HPV types stimulates cells to undergo DNA synthesis by binding to Rb family members and promoting their degradation whereas E_6 prevents growth arrest and apoptosis by promoting degradation of the p53 tumour suppressor. It has been shown that downregulation of E_6 and E_7 expression, by transfection of a functional E_2 or by small interfering RNAs, results in restoration of the p53 and Rb pathways and the subsequent induction of cellular senescence (Seth (2008)).

Another proposed therapy aims preventing assembly of the initial E_1E_2-ori complex, in order to treat HPV lesions, in which the viral genome is maintained in episomal form, such as condylomas. Assembly of the E_1E_2ori complex depends on the interaction of E_1 and E_2 with DNA and on a critical proteinprotein interaction between the TAD of E_2 and the C-terminal helicase domain of E_1. Actually, a small molecule inhibitor of the E_1E_2 protein interaction has been identified (**?**).

Finally it was proposed to target $Sp1$ human transcription factor, too, due to its importance in activating the early promoter (Seth (2008)).

Even if different antiviral therapies were proposed during the past years, there is no antiviral drug currently available for the treatment of HPV associated diseases.

3

A first deterministic model for HPV early promoter regulation

In this chapter we present a first deterministic model of the early promoter regulation, we have developed (Giaretta et al. (2015)). In what follows we will report the model as in the paper with the same nomenclature.

To develop this model we have followed an heuristic approach without really translating the biology in a set of biochemical reactions as reported in chapter 1. We have mathematically translate the biology in terms of a compartmental-logic approach (Cobelli Claudio (2008)). In this model we consider the only early promoter regulation mediated by E_1 and E_2. We account for the early promoter repression by the late promoter, modeling mE_1 as a forcing function.

3.1 Model topology

The main mechanisms of HPV early promoter model designed here are summarized and depicted in Fig. 3.1.1:

- The early promoter controls the primary polycistronic transcript x. In the early stages of the viral life cycle, transcription of x generates, by alternative splicing, mRNAs encoding all the early genes.
- The spliced mRNA E_2 (mE_2) encodes for E_2 protein.
- The spliced mRNA E_1 (mE_1) encodes for E_1 protein.
- E_2 protein is the main regulator of the early promoter. It generates a slight positive feedback effect when present in low concentration, a negative feedback effect when present in high concentration, eventually inhibiting the early promoter activity.
- E_1 acts with a positive regulation enhancing the mE_2 transcript and with a negative regulation on E_2 degradation, hence increasing its stability.

3.2 Model Equations

Model equations, based on mass action, are:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = S_x(E_2) - k_s\, x \tag{3.1}$$

$$\frac{\mathrm{d}mE_1}{\mathrm{d}t} = k_{1s}(t)\, ks\, x - \delta_{1m}\, mE_1 \tag{3.2}$$

$$\frac{\mathrm{d}mE_2}{\mathrm{d}t} = k_{2s}(E_1)\, ks\, x - \delta_{2m}\, mE_2 \tag{3.3}$$

$$\frac{\mathrm{d}E_1}{\mathrm{d}t} = \beta_1\, mE_1 - \delta_{1p}\, E_1 \tag{3.4}$$

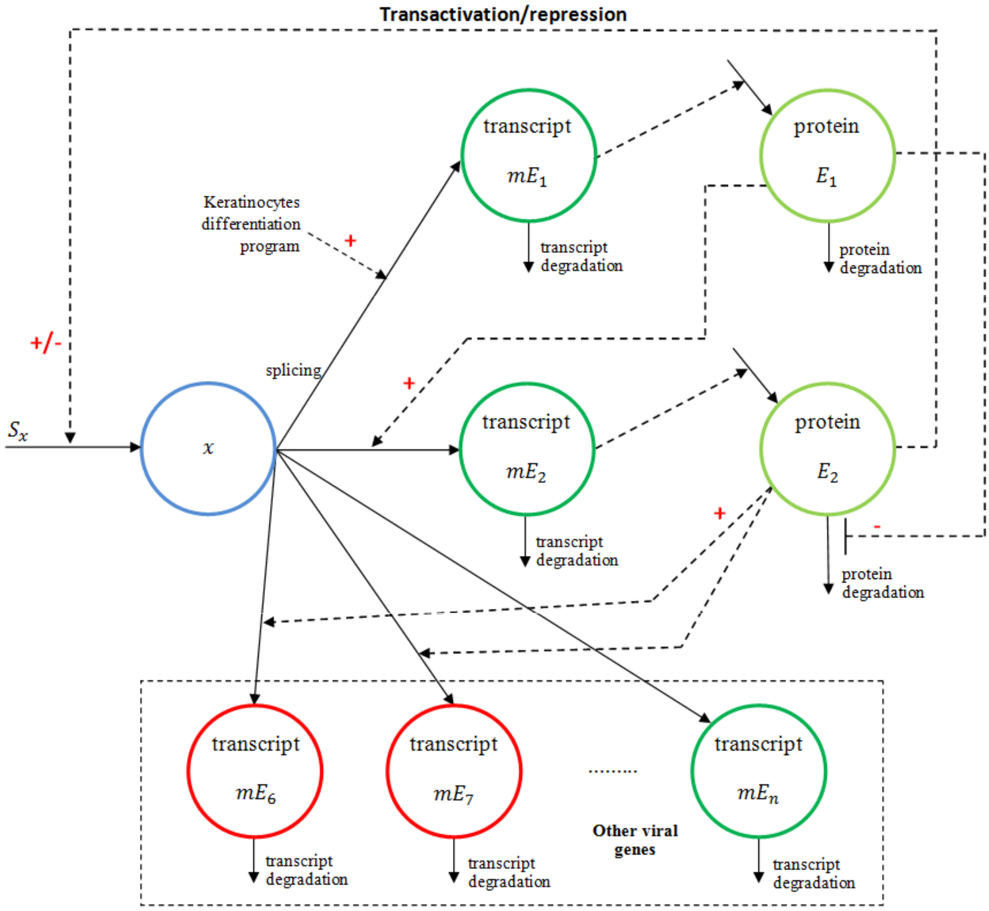

Figure 3.1.1: Background knowledge on the HPV-16 gene circuit. Solid arrows represent fluxes, dashed arrows controls. In blue the primary transcript, in green the transcripts (and their respective proteins) of early non-oncogenes and in red the early oncogenes.

$$\frac{\mathrm{d}E_2}{\mathrm{d}t} = \beta_2\, mE_2 - \delta_{2p}(E_1)\, E_2 \tag{3.5}$$

$$\frac{\mathrm{d}mE_i}{\mathrm{d}t} = k_{is}(E_1)\, x - \delta_{im}\, mE_i, \quad i = 4,5,6,7 \tag{3.6}$$

where the state variables are the concentrations [nM] of: the primary transcript x, mRNAs mE_i, i = 1,2,4,5,6,7, proteins E_1 and E_2.

S_x refers for the transcription of x enhanced by low values of E_2 concentration and repressed by elevated E_2 concentration (Fig. 8.2.1A) (as explained in chapter 2) according to 3.7.

$$S_x(E_2) = \begin{cases} S_b + \frac{a_1\, E_2^{q1}}{\lambda_1^{q1} + E_2^{q1}} & E_2 < E_2^{th} \\ \\ \frac{a_2\, \lambda_2^{q2}}{\lambda_2^{q2} + (E_2 - E_2^{th})^{q2}} & E_2 > E_2^{th} \end{cases} \tag{3.7}$$

$k1s$ represents the splicing flux for the mE_1 transcript (Fig. **??**B) and is modeled as a time variant forcing function (3.8), tightly bounded to the host cell differentiation program as reported in. In particular, the variable t in is the time evolution associated to the keratinocytes differentiation program.

$$k_{1s}(t) = \begin{cases} k_{1s}^{min} + \frac{a_3\, t^{q3}}{\lambda_3^{q3} + t^{q3}} & t < t_h \\ \\ \frac{a_4\, \lambda_4^{q4}}{\lambda_4^{q4} + (t - t_h)^{q4}} & t > t_h \end{cases} \tag{3.8}$$

k_{2s} accounts for the splicing flux for the mE_2 transcript (3.9) and it is positively regulated by E_1.

$$k_{2s}(E_1) = \frac{(f_1 - 1)1, k_{2s}^{min}}{1 + exp\left(\frac{\lambda_5 - E_1}{\sigma_1}\right)} + k_{2s}^{min} \tag{3.9}$$

δ_{2p} is the degradation of E_2 (3.10) which is negatively regulated by E_1.

$$\delta_{2p}(E_1) = (f_2 - 1)\, \delta_2^{min} \left(1 - \frac{1}{1 + exp\left(\frac{\lambda_6 - E_1}{\sigma_2}\right)}\right) \tag{3.10}$$

β_1 and β_2 are the rate constants for E_1 and E_2 protein translation assumed to be linearly related to the cognate mRNAs, see (3.4, 3.5). δ_{im} and δ_{ip} are degradations of transcripts and proteins (see (3.2-3.6)), assumed to be first order processes with the only exception of δ_{2p}, see (3.10).

We can observe that in this first attempt we have condensed, for simplicity, the late promoter higher production of E_1 mRNA as a forcing function (dependent on time) accounting for the E_1 mRNA synthesis, namely $k_{1s}(t)$.

In what follows the rate constants of the system. In the results chapter we will show an in silico experiment in order to investigate the behavior of this first HPV model.

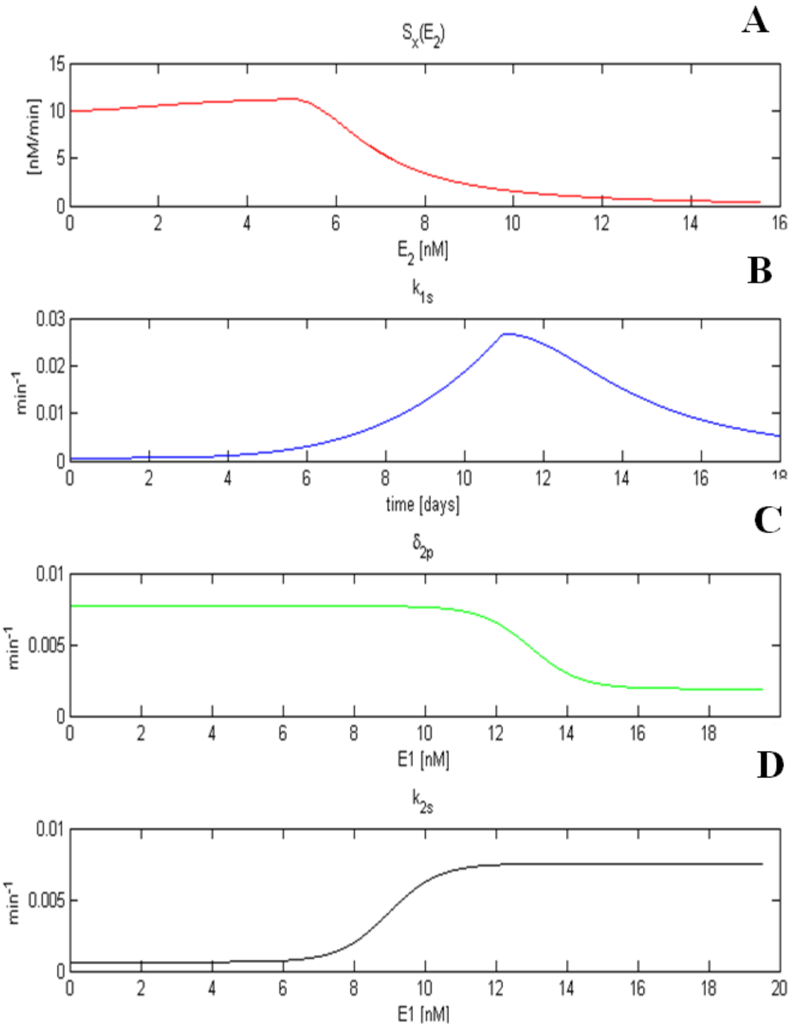

Figure 3.2.1: A. Synthesis of the primary transcript regulated by E_2 (7). B. Rate constant for mE_1 splicing regulated by cellular differentiation (8). C. E_2 degradation mediates by E_1 concentration (9). D. Rate constant for mE_2 splicing positively regulated by E_1 concentration (10).

What we have reported here was just a first heuristic attempt in modeling the basics viral regulatory network. In the next chapters we will formally develop a novel and much more complete model in terms of biochemical reactions and a stochastic and deterministic formulation.

4
Stochastic Model Development

4.1 Introduction to the model structure

The model structure of the HPV gene regulatory network is made of two main subsystems in cascade: early and late promoter. The early promoter part is composed of a regulatory core managed by E_1, E_2 and the splicing factors, and a secondary subsystem about the oncogenes control. These latter don't have any direct regulatory function on the regulatory core. The late promoter is not controlled by the early promoter but is activated by the differentiation program of the cell. Nevertheless the late promoter is able to have a negative regulatory function on the early promoter.

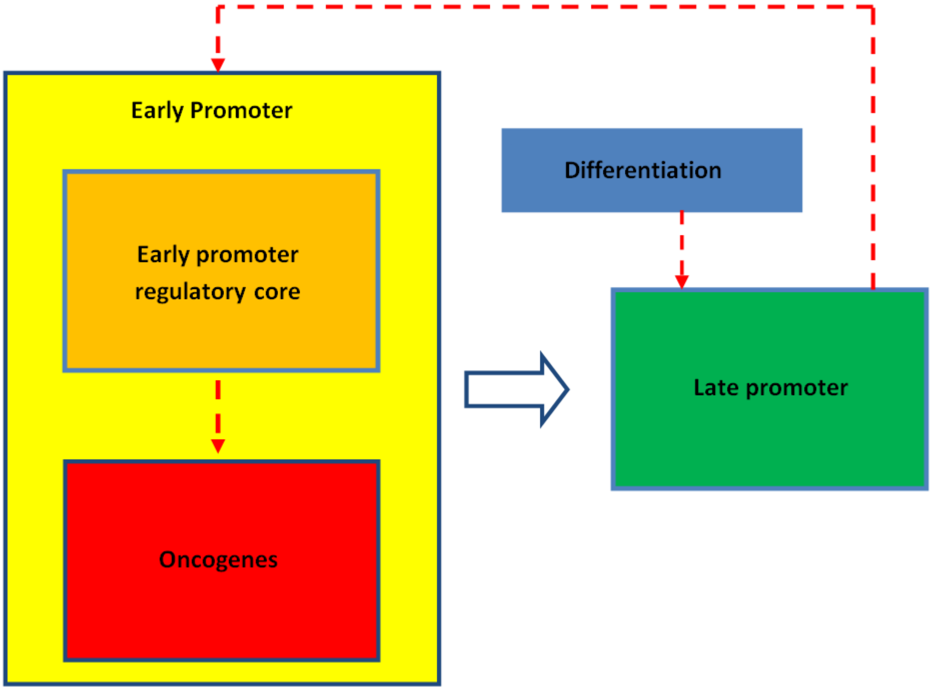

Figure 4.1.1: Schematic representation of the HPV regulatory network main structure.The virus has a regulatory core mediated by E_1 and E_2 inducing the transactivation. The oncogenes are controlled by the regulatory core. The late promoter negatively regulates the erly promoter but it is not regulated by this latter.

In this chapter, as in chapter 5 and 6, we denote, where not differently specified, with "E" and "L" apexes the chemical species derived by the early and late promoter, respesctively. These species are in common between the two promoters and are E_1, E_2, DE_1, DE_2, E_1E_2. Let's call this notation, the "EL" notation. This notation won't be applied neither to their degradations nor to the association/dissociations rate constants for the multimeric complexes. This is because E_1 and E_2 have the same kinetics properties regardless the promoter.

4.2 Modeling the eraly promoter

In this section we do not use the "EL" notation to keep the treatment as general as possible, since the early promoter can be controlled by the late promoter, too.

In modeling the early promoter regulation we do not consider the LCR regulation from the infected cell transcription factors (TFs) that can modulate the early promoter (EP) activity. We consider these latter in higher concentration with respect to the viral transcripts and proteins, in order to neglect their dynamical contributions. We assume two binding sites, in place of four, for the E_2 dimer (DE_2) binding to EP. We also consider the binding of the tetramer E_1/E_2 to the DE_2 binding sites regulating in turn the EP activity in the same manner DE_2 does. With two binding sites and two molecules , which can bind them, we consider five possible promoter states denoted by P_i^E,$i = 0, 1, 2, 3, 4$. In particular, when the dimer DE_2 and the tetramer E_1E_2 are in low concentration it is more likely there is only one binding site occupied most of the time, conversely when they are in high concentration both the binding sites will be occupied most of the time.

In particular we will have a basal transcription when no molecule is bound, a positive feedback when only a DE_2 dimer molecule is bound (i.e. low DE_2 concentration) while a negative feedback when a E_1E_2 tetramer is bound to the replication origin, as argued in chapter 2, or when two molecules are bound (two DE_2 dimers or one DE_2 dimer and one E_1E_2 tetramer). For simplicity, we do not distinguish which binding site is occupied when just one molecule is bound (otherwise, we should consider other two additional states) and which one is occupied by the dimer or the tetramer when two mixed binding occurs. Additionally, in our model DE_2 or E_1/E_2 binding occurs non-cooperatively in this system, meaning that the binding of an individual DE_2 dimer or E_1/E_2 tetramer occurs entirely independently of the others.

The biochemical reactions for the EP states can be written as the markov chain in Fig. 4.5.1. The state P_0^{EP} accounts for basal transcription, state P_1^{EP} accounts for positive feedback (only one DE_2 bound to the promoter), and states P_2^{EP}, P_3^{EP} and $_4^{EP}$ account for negative feedback. In particular, P_2^{EP} takes account for two DE_2 molecules bound to the promoter, P_3^{EP} the condition with E_1E_2 bound to the origin of replication and able to negatively affect the promoter, and finally P_4^{EP} accounts for the mixed effects (i.e., two molecules bound, one DE_2 dimer and one E_1E_2 tetramer).

In the markov chain in Fig. 4.5.1 k_1 and k_5 are the forward rates to transit from the basal transcription state, P_0^{EP}, to the states P_1^{EP} and P_3^{EP} respectively, while k_2 and k_6 are the reverse rates for coming back to the basal transcription state; k_3 and k_9 are the forward rates to transit from the positive feedback state P_1^{EP} to the negative feedback states P_2^{EP} and P_4^{EP}, respectively, while k_4 and k_{10} are the reverse rates for coming back to P_1^{EP} state. Finally, k_7 is the forward rate to transit from the state P_3^{EP} to the P_4^{EP} state and k_8 is the correspondent reverse constant. In particular, the forward rates are multiplied by the concentration of the DE_2 dimer or the E_1/E_2 tetramer to take into account for the positive and negative feedback behaviors on the promoter.

The EP markov chain shows a closed cycle between the EP states ($P_0^{EP} \rightarrow P_1^{EP} \rightarrow P_4^{EP} \rightarrow P_3^{EP} \rightarrow P_0^{EP}$) to maintain a physical equilibrium consis-

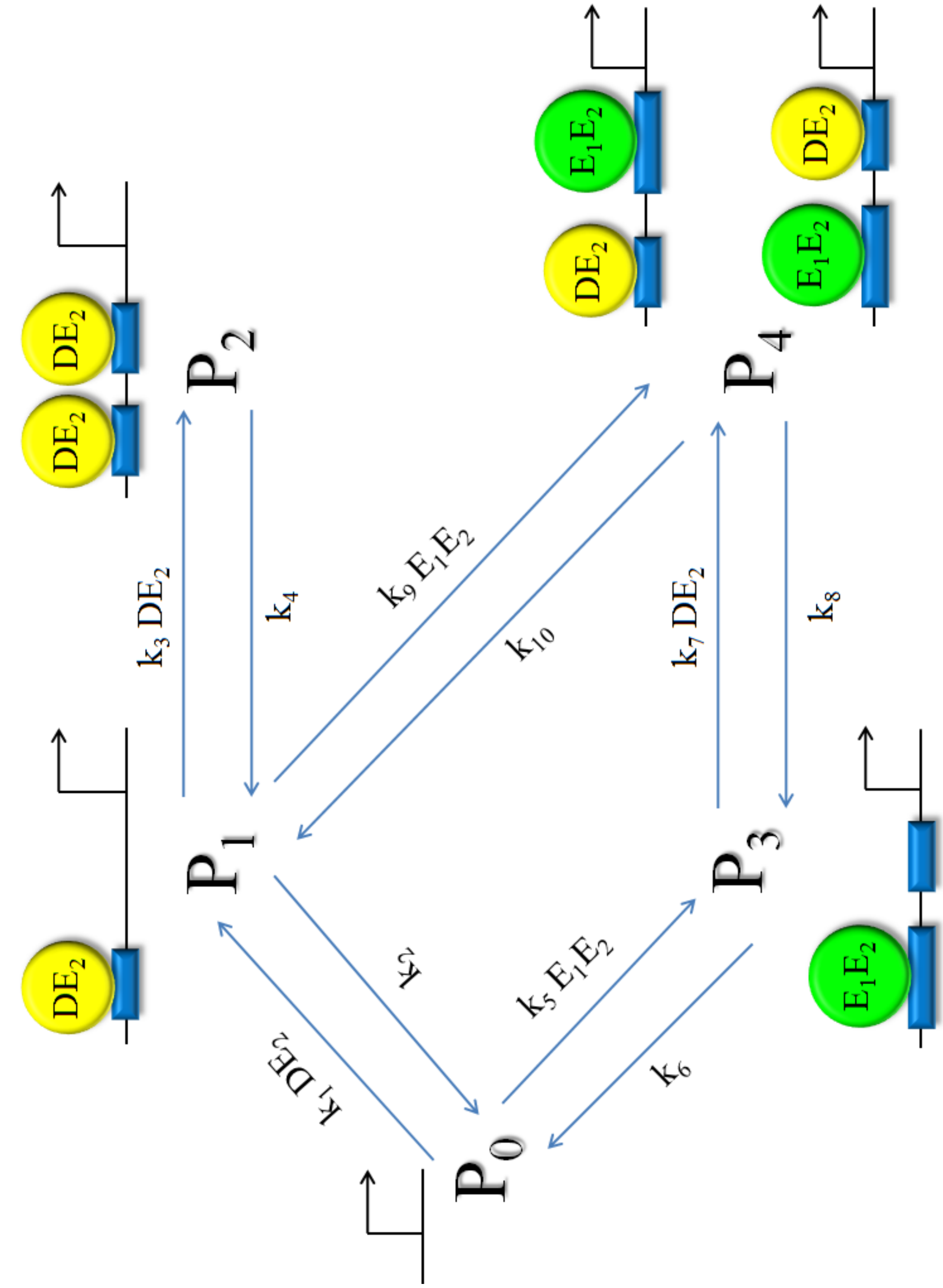

Figure 4.2.1: Early promoter (EP) Markov chain modeling the EP chemical states modulated by the binding of DE_2 and E_1E_2. In particular state P_0 accounts for basal transcription, state P_1 accounts for the positive feedback and states P_2, P_3 and P_4 account for the negative feedback.

tency we have imposed the detailed balance condition

$$k_9\, E_1/E_2\, p_1^{EP} = k_{10}\, p_4^{EP} \tag{4.1}$$

$$k_7\, DE_2\, p_3^{EP} = k_8\, p_4^{EP} \tag{4.2}$$

$$k_5\, E_1 E_2\, p_0^{EP} = k_6\, p_3^{EP} \tag{4.3}$$

$$k_1\, DE_2\, p_0^{EP} = k_2\, p_1^{EP} \tag{4.4}$$

where p_i^{EP} are the probabilities to be in the $i-th$ state of the early promoter markov chain in Fig.4.5.1.
Solving the previous system we can obtain the following constraint among the EP markov chain parameters

$$k_1\, k_6\, k_8\, k_9 \;=\; k_2\, k_5\, k_7\, k_{10} \tag{4.5}$$

4.3 Markov Chain models of splicing site and promoter splicing factor

In this section we will not use the "EL" notation since the splicing modeling we present here is in common to both the promoters.

In chapter 2 we have seen the importance of HPV splicing regulation to properly modulate the viral proteins during the entire life cycle, strictly linked to the differentiation program of the infected cell. In this section we model the splicing regulation related to the early promoter regulatory "core". We develop a minimal model with a unique splicing site (SS_1) condensing the $SA3358$ and $SA2709$ splicing sites given their mutual exclusivity. We do not consider the regulation, at SA3358, of the remaining splicing factors (e.g. SRSF9m SRSF3), because they are not important as SRSF1; besides their temporal evolution is not well understood. Hence, in modeling this splicing mechanism we consider two states for SS_1 denoted by SS_1^i, $i = 0, 1$. When SS_1 is occupied (state SS_1^1) we assume that $SRSF1$ splicing factor is bound to $SA3358$ with $SA2709$ inactive, while the opposite when SS_1 is unoccupied (state SS_1^0), as shown in Fig. **??**. The biochemical reactions for the SS_1 states can be written as the following two state markov chain

$$SS_1^0 + SRSF1 \underset{k_{off}^{SS_1}(E_1)}{\overset{k_{on}^{SS_1}}{\rightleftharpoons}} SS_1^1 \tag{4.6}$$

where $k_{on}^{SS_1}$ and $k_{off}^{SS_1}$ (E_1) are the rate constants for SRSF1 binding and dissociation, respectively. $k_{off}^{SS_1}$ (E_1) depends on the E_1 concentration. Because it is clear that E_1 enhances the E_2 transcription but it is not clear what mechanism by which this control is performed, we have assumed a Hill functional response such that

$$k_{off}^{SS_1}(E_1) = \tilde{k}_{off}^{SS_1} + \frac{\hat{k}_{off}^{SS_1} E_1^{n_{SS_1}}}{\lambda_{SS_1}^{n_{SS_1}} + E_1^{n_{SS_1}}} \tag{4.7}$$

where λ_{SS_1} is the Michaelis constant and denotes the E_1 concentration at which $k_{off}^{SS_1}$ (E_1) is half its maximum value $\hat{k}_{off}^{SS_1}$. The Hill coefficient, n_{SS_1}, determines the sharpness of the transition about λ_{SS_1}. Thus, higher the concentration of E_1, higher the probability to get SS_1^1, hence having a higher E_2 transcription. $\tilde{k}_{off}^{SS_1}$ represents a minimal basal detachment rate constant for SS_1 splicing site. From literature we know that when the splicing occurs at SA3358 splicing site, actually even E_1 is not transcribed. Therefore we consider E_1 controlled by SS_1 splicing site in the same manner E_2 was. In this case we have to observe that E_1 can positively regulates itself and even this thing was suggested in literature (see chapter 2).

The HPV splicing regulation is very complex and not completely understood yet. The choice to make a minimal model of the splicing regulation considering only the splicing sites $SA3358$ and $SA2709$ is due to their importance in modulating E_2, representing the main regulator of the early promoter. We consider the sole $SRSF1$ splicing factor as, for the current biological knowledge, it represents one of the most significant splicing regulators thanks to its capability in driving the oncogenes expression (we'll model them in the next section) eventually diminishing E_2 transcription. Moreover, $SRSF1$ is in turn positively regulated by E_2, as shown in Fig. 3. All these features make $SRSF1$ an interesting molecular actor in the context of the HPV post-transcriptional regulation, capable to directly connect this latter to the transcriptional regulation. To model the $SRSF1$ activation we consider a two state promoter (P^{SRSF1}), denoted by the states P_i^{SRSF1}, $i = 0, 1$, with a positive regulation induced by the E_2 binding. The biochemical reactions for the P^{SRSF1} states can be written as the following two state markov chain

$$P_0^{SRSF1} \underset{k_{off}^{SRSF1}}{\overset{k_{on}^{SRSF1}(E_2)}{\rightleftharpoons}} P_1^{SRSF1} \tag{4.8}$$

where $k_{on}^{SRSF1}(E_2)$ and k_{off}^{SRSF1} are the rate constants controlling the two states of the SRSF1 promoter. $k_{on}^{SRSF1}(E_2)$ is modeled as a Hill functional response such that

$$k_{on}^{SRSF1}(E_2) = \frac{\hat{k}_{on}^{SRSF1}\ E_2^{n_{SRSF1}}}{\lambda_{SRSF1}^{n_{SRSF1}} + E_2^{n_{SRSF1}}} \tag{4.9}$$

where λ_{SRSF1} is the Michaelis constant and denotes the E_2 concentration at which k_{on}^{SRSF1} (E_2) is half its maximum value $\hat{k}_{on}^{SRFS1}$. The Hill coefficient, n_{SRSF1}, determines the sharpness of the transition about λ_{SRSF1}. Thus, higher the concentration of E_2, higher the probability to get P_1^{SRSF1}, hence having a higher $SRSF1$ transcription. We have modeled $k_{on}^{SRSF1}(E_2)$ as a function of E_2 because this latter doesn't directly bind SRSF1 promoter. It binds as a co-regulator to another transcription factors which in turn binds the promoter (chapter 2, SRSF1 section).

The biochemical reactions relating the $SRSF1$ transcription are the following

$$\emptyset \underset{\delta^{mSRSF1}}{\overset{S_{mSSRSF1}(P_i^{SRSF1})}{\rightleftharpoons}} mSRSF1 \tag{4.10}$$

where $mSRSF1$ denotes the $SRSF1$ mRNA, δ^{mSRSF1} is the $mSRSF1$ degradation rate, and $S_{mSRSF1}(P_i^{SRSF1})$ is the $mSRSF1$ synthesis dependent on the two promoter states

$$S_{mSRSF1}(P_i^{SRSF1}) = \begin{cases} s_0^{SRSF1} & if \quad P_i^{SRSF1} = P_0^{SRSF1} \\ s_1^{SRSF1} & if \quad P_i^{SRSF1} = P_1^{SRSF1} \end{cases} \tag{4.11}$$

when E_2 is bound to the promoter (P_1^{SRSF1} state) $mSRSF1$ is produced with s_1^{SRSF1} rate, otherwise (P_0^{SRSF1} state) $mSRSF1$ is produced with s_0^{SRSF1}, with the constraint $s_1^{SRSF1} > s_0^{SRSF1}$ to account for the positive regulation induced by E_2. The transcript is then translated into protein

$$\emptyset \underset{\delta^{SRSF1}}{\overset{k_p^{SRSF1}\, mSRSF1}{\rightleftharpoons}} SRSF1 \tag{4.12}$$

where production rate of the $SRSF1$ protein is assumed to be proportional to $mSRSF1$ through the rate constant k_p^{SRSF1}, and δ^{SRSF1} is the $SRSF1$ degradation rate. It is known from literature that the $SRSF1$ auto negative feedback is mostly implemented through both translational efficiency and post-transcriptional regulation. Since our purpose is to keep a minimal model of the HPV post-transcriptional regulation, considering that SRSF1 splicing regulation is complex and not well characterized in the HPV context, we choose to consider the sole translational efficiency as the principal mechanism accountable for the auto negative regulation on $SRSF1$. To model this latter, we assume the formation of a heterodimer between $SRSF1$ transcript and its own protein (as reported in chapter 2) with the constraint of a higher degradation than $SRSF1$ transcript

$$\emptyset \underset{k_{rms}}{\overset{k_{fms}}{\rightleftharpoons}} mSRSF1/SRSF1 \xrightarrow{\delta_{mp}^{SRSF1}} \emptyset \tag{4.13}$$

where k_{fms} and k_{rms} are the association and dissociation rate constants for the heterodimer $mSRSF1/SRSF1$, respectively; while δ_{mp}^{SRSF1} represents its degradation rate, with the constraint $\delta_{mp}^{SRSF1} > \delta^{mSRSF1}$ to model the auto negative feedback on $SRSF1$ in terms of translational efficiency.

In fig. 4.3.1 we show the schematic representation of the splicing regulation we have modeled so far.

4.4 ADDITIONAL REACTIONS

Once the control part of the transcriptional and post-transcriptional regulation of the early promoter "core" (i.e. promoters regulation and the major splicing site regulations) has been modeled, what remains to model are the biochemical reactions pertaining to the synthesis, degradation of the viral transcripts, proteins and the dimers and tetramers formation, dissociation and degradation. The early promoter regulates the production of the early primary transcript pM^E

$$\emptyset \xrightarrow{S_{pM^E}(P_i^E)} pM^E \tag{4.14}$$

where S_{pM^E} (P_i^E) is the early pM^E synthesis dependent on the EP states

$$S_{pM^E}(P_i^E) = \begin{cases} s_0^{pM^E} & if \quad P_i^E = P_0^{EP} \\ s_1^{pM^E} & if \quad P_i^E = P_1^{EP} \\ s_2^{pM^E} & if \quad P_i^E \in \{P_2^{EP}, P_4^{EP}\} \\ s_3^{pM^E} & if \quad P_i^E = P_3^{EP} \end{cases} \tag{4.15}$$

with the constraint $s_1^{pM^E} > s_0^{pM^E} > s_3^{pM^E} \geq s_2^{pM^E}$ to account for the positive feedback ($s_1^{pM^E}$), the negative feedback ($s_2^{pM^E}$ and $s_3^{pM^E}$ account for weak and strong negative feedback, respectively), and the basal transcription ($s_0^{pM^E}$), to be consistent with the previous treatment about the early promoter markov chain.

The primary transcript is then converted into its transcripts

$$pM^E \xrightarrow{k_m^{E_1^E}(SS_1^i)} mE_1^E \tag{4.16}$$

$$pM^E \xrightarrow{k_m^{E_2^E}(SS_1^i)} mE_2^E \tag{4.17}$$

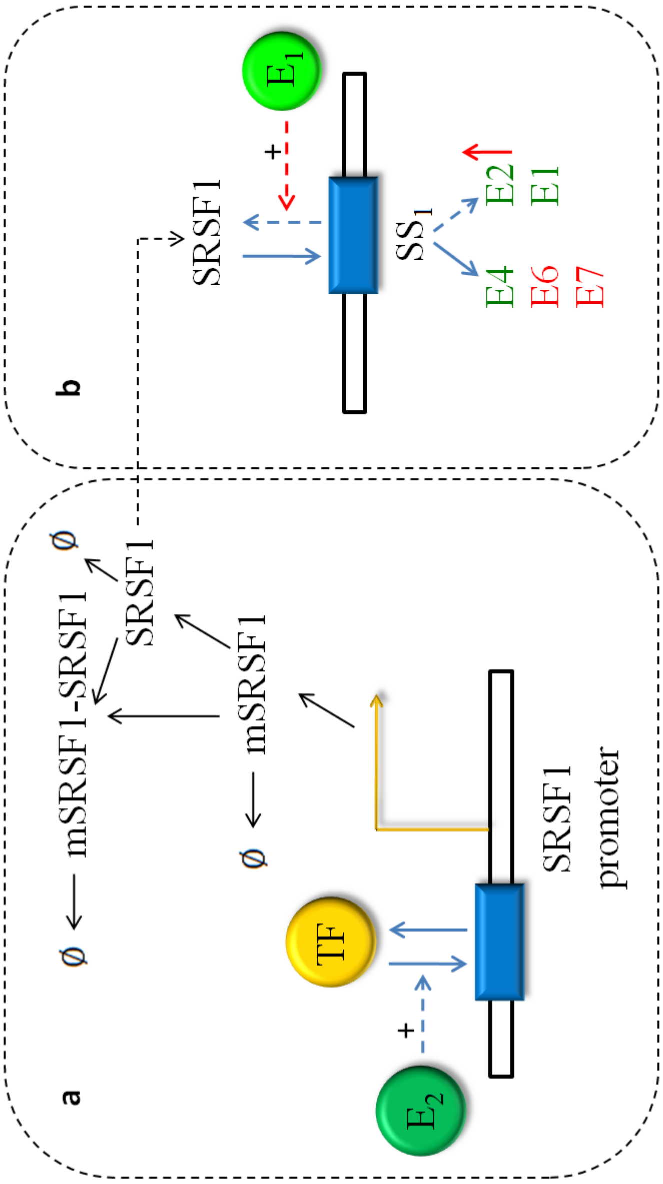

Figure 4.3.1: Minimal model of HPV post-transcriptional regulation. a. transcriptional and translational regulation of the splicing factor $SRSF1$. $E2$ can co-activate (by binding to a transcription factor TF) the $SRSF1$ promoter. The formation of the tetramer $mSRSF1/SRSF1$ with a higher degradation with respect to $mSRSF1$ is assumed to model the negative auto feedback on $SRSF1$ in terms of translational efficiency mechanism. b. Modeling of the splicing site SS_1 regulation relating to the E_2 and E_1 post-transcriptional regulation through the $SRSF1$ binding. The off rate is dependent on the E_1 concentration to account for the E_2 enhancement in the presence of E_1.

$$mE_1^E \xrightarrow{\delta_m^{E_1^E}} \emptyset \tag{4.18}$$

$$mE_2^E \xrightarrow{\delta_m^{E_2^E}} \emptyset \tag{4.19}$$

$$pM^E \xrightarrow{k_s^E} \emptyset \tag{4.20}$$

where mE_1^E and mE_2^E are the transcripts of E_1^E and E_2^E early genes transcribed by the EP, respectively. $k_m^{E_1^E}(SS_1)$ and $k_m^{E_2^E}(SS_1^i)$ are the rate constants relating to the primary transcript (pM^E) conversion into mE_1^E and mE_2^E, respectively. In particular, $k_m^{E_1^E}(SS_1^i)$ and $k_m^{E_2^E}(SS_1^i)$ are dependent on the splicing site (SS_1^i) state through the relationship

$$k_m^{E_1^E}(SS_1^i) = \begin{cases} \hat{k}_m^{E_1^E} & if \quad SS_1^i = SS_1^0 \\ 0 & if \quad SS_1^i = SS_1^1 \end{cases} \tag{4.21}$$

$$k_m^{E_2^E}(SS_1^i) = \begin{cases} \hat{k}_m^{E_2^E} & if \quad SS_1^i = SS_1^0 \\ 0 & if \quad SS_1^i = SS_1^1 \end{cases} \tag{4.22}$$

where $\hat{k}_m^{E_1^E}$ and $\hat{k}_m^{E_2^E}$ are the mE_2 and mE_1 conversion rates, respectively, from the primary transcript when the splicing site is unoccupied (therefore splicing occurs at SA2709 splicing site). When SS_1 is occupied then $k_m^{E_1^E}(SS_1) = 0$ and $k_m^{E_2^E}(SS_1) = 0$. this is because both E_1 and E_2 are not transcribed when splicing occurs at SA3358 splicing site. The rate k_s^E takes into account for the conversion of the primary transcript into the early transcripts we are not interested to model: E_4^E, E_5^E, E_6^E , E_7^E. However, we will model the oncogenes in the next section. The transcripts are then translated into proteins

$$\emptyset \xrightarrow{k_p^{E_1}\, mE_1^E} E_1^E \tag{4.23}$$

$$E_1^E \xrightarrow{\delta_p^{E_1}} \emptyset \tag{4.24}$$

$$\emptyset \xrightarrow{k_p^{E_2^E}\ mE_2^E} E_2^E \tag{4.25}$$

$$E_2^E \xrightarrow{\delta_p^{E_2}} \emptyset \tag{4.26}$$

where E_1^E and E_2^E are the proteins translated from the transcripts mE_1^E and mE_2^E, respectively. The translation rates for these latter are proportional to their transcripts through the constants $k_p^{E_1^E}$ and $k_p^{E_2^E}$, respectively. $\delta_p^{E_1}$ and $\delta_p^{E_2}$ are the degradation rates for E_1^E and E_2^E, respectively. E_1^E and E_2^E can form dimers DE_1^E and DE_2^E, respectively, as well as these latter can form a tetramer $E_1E_2^E$

$$E_1^E + E_1^E \underset{k_r^{DE_1}}{\overset{k_f^{DE_1}}{\rightleftharpoons}} DE_1^E \xrightarrow{\delta_{DE_1}} \emptyset \tag{4.27}$$

$$E_2^E + E_2^E \underset{k_r^{DE_2}}{\overset{k_f^{DE_2}}{\rightleftharpoons}} DE_2^E \xrightarrow{\delta_{DE_2}} \emptyset \tag{4.28}$$

$$DE_1^E + DE_2^E \underset{k_r^{E_1E_2}}{\overset{k_f^{E_1E_2}}{\rightleftharpoons}} E_1E_2^E \xrightarrow{\delta_{E_1E_2}} \emptyset \tag{4.29}$$

where $k_f^{DE_1}$, $k_r^{DE_1}$ are the forward and reverse rate constants regarding the formation of the dimer DE_2^E, respectively and $k_f^{DE_2}$, $k_r^{DE_2}$ are the forward and reverse rate constants regarding the formation of the dimer DE_1^E, respectively. $k_f^{E_1E_2}$ and $k_r^{E_1E_2}$ are the forward and reverse rate constants regarding the formation of the tetramer $E_1E_2^E$, respectively; $\delta_{E_1E_2}$, δ_{DE_2} and δ_{DE_1} are the degradation rates for the tetramer and the dimers, respectively. To take into account of the stabilization effect mediated by E_1^E on E_2^E, we have assumed the constraint $\delta_{E_1E_2} < \delta_{E_2}$.

4.4.1 Modeling the oncogenes

So far, we have modeled the only regulatory "core". It accounts for only the two genes E1 and E2, since, as far as we know from literature, they seem to be the only elements, originating from the virus, capable to regulate the early promoter activity. About the splicing regulation we have made the choice to design a minimal model of just the major 3′ splicing site SA3358 (in turn controlled by E2) and of the splicing site SA2709. These latter were modeled as a unique splicing site thanks to their mutual exclusivity.

In this section we present a model extension, to the early promoter ″core″, in order to account for the oncogenes regulation. We observe that the oncogenes, as far as we know from the current biology, do not have any regulatory roles on the early promoter. This means we don't have to change the early promoter markov chain but simply update the model for the post-transcriptional regulation.

We won't model neither E4 nor E5 genes. About E5, it is still a very elusive gene and, even if its function is nowadays clearer (especially about its role in the apoptosis pathway) than in the past, its transcriptional regulation is not understood at all. About E4 the reason is that its main role is during the late phase of the infection when produced by the late promoter and its contribution to the early phase is negligible.

To model the splicing of the oncogenes we consider two splicing sites. The first is actually SS_1. In fact, when SS_1 is unoccupied we have the already modeled the production of E1 and E2, and now when it is occupied we will consider the production of E6 and E7.

Nevertheless, E6 and E7 transcripts are also spliced between the splicing donor site SD226 and the splicing acceptor site SA409 (chapter 2). E6 and E7 mRNAs are not spliced at the same time: E7 is transcribed by retention of the first intron between SD226 and SA409, while E6 is the mRNA which remains unspliced between SD226 and SA409.

To model the alternative production of the oncogenes we need a second splicing site, SS_2, designed as a two state model. The new post-transcriptional regulation model is shown in Fig. 4.4.1.

Thanks to the alternative production of the oncogenes, we assume that if SS_2 site is occupied, then we have the retention of the first intron, between SD226 and SA409, with the consequent production of E7 mRNA.

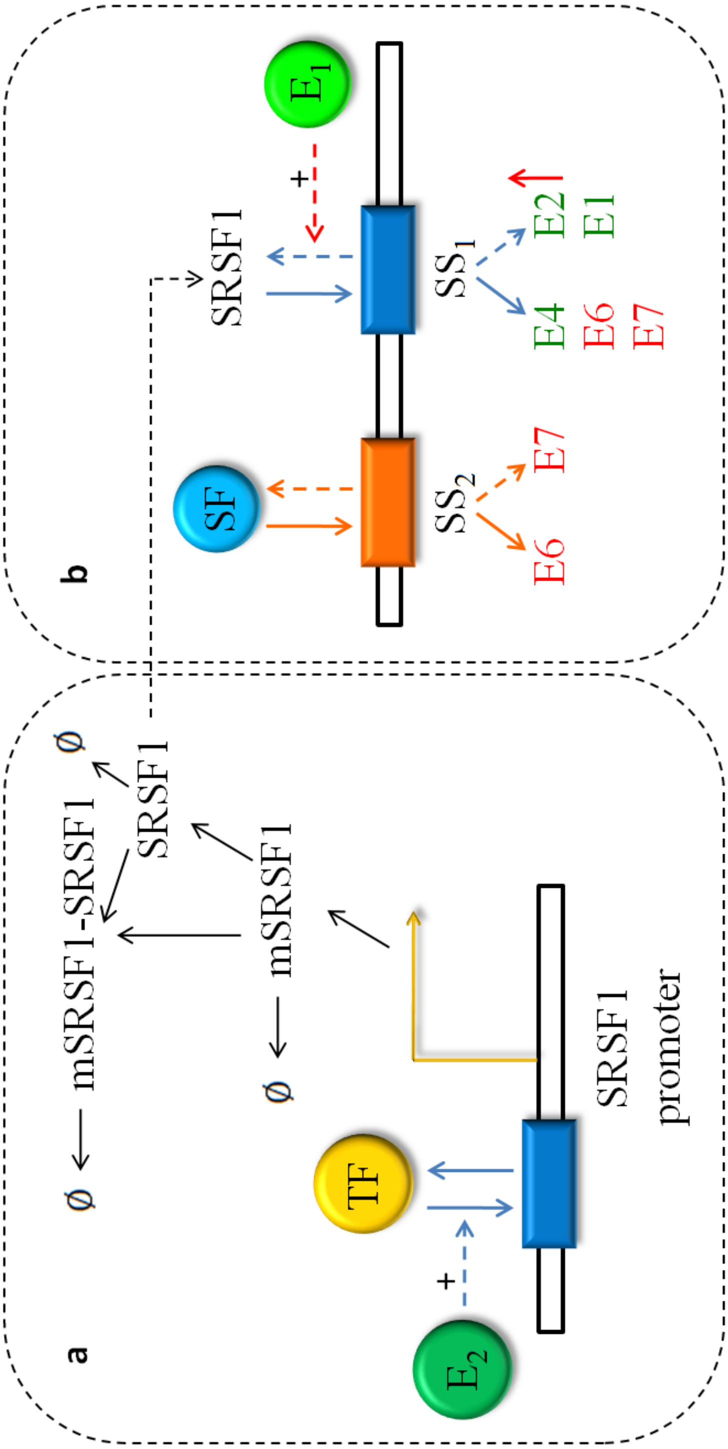

Figure 4.4.1: Same model for the splicing regulation presented before, with the addition of SS_2 splicing site to model their alternative production.

Conversly, when the splicing site is not occupied, E6 mRNA is produced. In this model we don't consider the modulation effect of the heterogeneous nuclear ribonucleoprotein hnRNP A1 and A2 that can stimulate the splicing between SD226 and SA409 (chapter 2) since their modulation during time is not understood yet. Probably they account for a finer modulation of SS_2. Anyway, we can get a good qualitative behavior even without considering their possible co-regulations.

The biochemical reactions for the new splicing site SS_2 can be written as the following two state markov chain

$$SS_2^0 \underset{k_{off}^{SS_2}}{\overset{k_{on}^{SS_2}}{\rightleftharpoons}} SS_2^1 \tag{4.30}$$

where SS_2^0 and SS_2^1 are two states (free and occupied, respectively) of the splicing site SS_2, and where $k_{on}^{SS_2}$ and $k_{off}SS_2$ are the rate constants for switching from the unoccupied state to the occupied one and vice versa, respectively.

The conversion rates of the primary transcript into E6 and E7 mRNAs are now function of both SS_1 and SS_2 splicing sites and are written as follows

$$k_m^{E6}(SS_1^i, SS_2^i) = \begin{cases} 0 & if \quad SS_1^0 = SS_1^0 \\ \hat{k}_m^{E6} & if \quad SS_1^i = SS_1^1 \wedge SS_2^i = SS_2^0 \end{cases} \tag{4.31}$$

$$k_m^{E7}(SS_1^i, SS_2^i) = \begin{cases} 0 & if \quad SS_1^i = SS_1^0 \\ \hat{k}_m^{E7} & if \quad SS_1^i = SS_1^1 \wedge SS_2^i = SS_2^1 \end{cases} \tag{4.32}$$

in the previous definitions a necessary condition for the oncogenes is the activation of the splicing site SA3358, hence the site SS_1 has to be occupied by the SRSF1 splicing factor in order to drive the oncogene production, as reported in (chapter 2).

4.5 Modeling Late Promoter

The late promoter (LP) is quite intriguingly since it is controlled by the differentiation program of the cell, hence by forcing functions exogenous to the viral system.

In modeling the late promoter we consider the two $C/EBP\beta$ dimer binding sites in order to regulate its activity (chapter 2). We do not consider the binding of the different $C/EBP\beta$ isoforms (LIP, LAP, LAP* ,(section late promoter chapter 2)) to the promoter, since we won't model their translation (as specified below). We do not account for other regulators or co-regulators of the late promoter such as p63. This is because all the other possible regulatory factors are not known or their action mechanism is not currently understood. We do not consider any auto-feedback effect on the promoter since there is no evidence in the literature.

The late promoter produces its transcripts very slowly (Ozbun and Meyers (1998)) following the differentiation program of the infected cell. We also know that $C/EBP\beta$ recruits the SWI/SNF chromatin complex to the upstream region of the late promoter. This mechanism induces rearrangement of the nucleosome structure around the promoter, resulting in the appearance of a nucleosome-free region where transcription initiates. This recruitment, mediated by $C/EBP\beta$, could insert also a quite strong delay (Kaern et al. (2005)) in the late promoter activation as well as other possible regulators or co-regulators could induce. To account for these possible delays, in the upstream promoter regulatory region, we will consider an intrinsic delay in the promoter markov chain as we'll see later.

With two binding sites and only the $C/EBP\beta$ dimer that can bind to them, we consider a three state promoter. The states are denoted by $P_i^L,\ i = 0,\,1,\,2$, as shown in the following markov chain

$$P_0^{LP} + DC/EBP\beta \underset{k_2^{LP}}{\overset{k_1^{LP}}{\rightleftharpoons}} p_1^{LP} + DC/EBP\beta \underset{k_4^{LP}}{\overset{k_3^{LP}}{\rightleftharpoons}} P_2^{LP} \tag{4.33}$$

where $DC/EBP\beta$ is the $C/EBP\beta$ dimer that binds to the late promoter.

$DC/EBP\beta$ condenses the possible LAP-LAP homodimers or LAP-LAP* heterodimer that can bind the promoter. k_1^{LP} and k_2^{LP} are the forward rates to transit from the state P_0^{LP} to P_1^{LP} and from P_1^{LP} to P_2^{LP}, respectively. While, k_3^{LP} and k_4^{LP} are the backward rates to transit from P_1^{LP} to P_0^{LP} and from P_2^{LP} to P_1^{LP}, respectively. We assume a basal or absent transcription (P_0^{LP} state) when no $DC/EBP\beta$ molecule is bound, an intermediate transcription when just one site is occupied (P_1^{LP} state) and a high transcription level when both the $DC/EBP\beta$ binding sites are occupied (P_2^{LP} state). A scheme of the later promoter model is shown in fig. 4.5.1.

The parameters of the late promoter markov chain will be tuned in order to orchestrate the probabilities of each state to satisfy an intrinsic during $DC/EBP\beta$ slow transient growth, as we'll see in the results chapter. In this way, the parameters will not account for only the $DC/EBP\beta$ binding but they will condense other possible delays that could occur in the promoter activation as previously discussed (e.g. $C/EBP\beta$ mediated rearrangement of the nucleosome structure).

4.6 Splicing sites, splicing factors and polyadenylation sites

The late promoter can produce all the early genes with the exception of the oncogenes. Morover during the differentiation phase, E4 is produced at higher levels due to its importance in breaking down the cell membrane. Since E4 is driven by SA3358 in a mutually exclusive manner than respect to E_2 and E_1, then the splicing markov chain regulating the late promoter early genes is the same of the early promoter with the only SS_1 splicing site (late promoter does not produce the oncogenes). However, E_1 co-regulation, on the splicing site, is now dependent on the total E_1 produced by both the promoters. Hence we can rewrite the equation 4.6

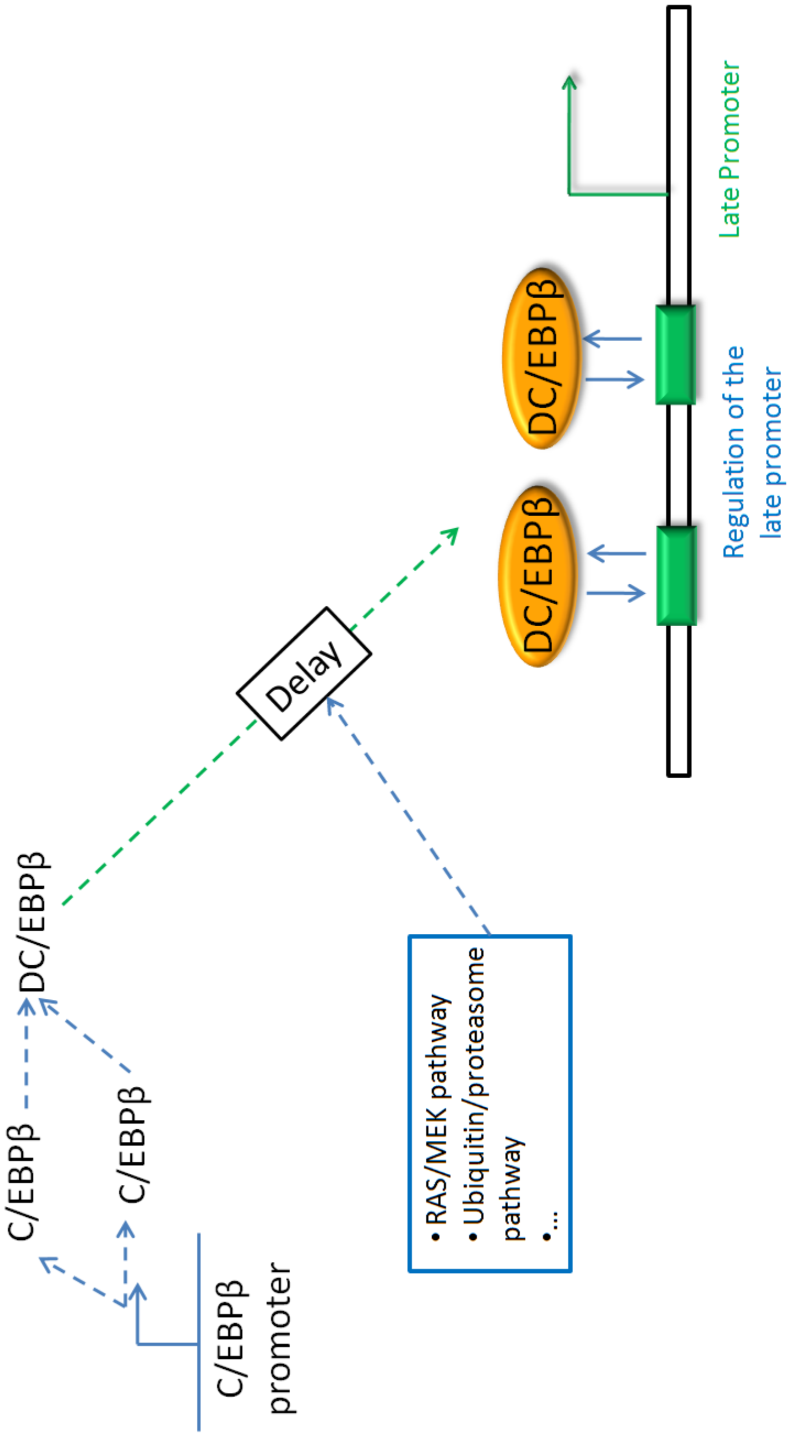

Figure 4.5.1: Late promoter model. $C/EBP\beta$ forms a dimer that is delayedto account for its strong stabilization mediated by different pathways. Its dimer can activate the late promoter by binding to it.

as follows

$$SS_1^0 + SRSF1 \underset{k_{off}^{SS_1}(E1^{TOT})}{\overset{k_{on}^{SS_1}}{\rightleftharpoons}} SS_1^1 \tag{4.34}$$

where all the other parameters and variables are the same as we have previously reported. $k_{off}^{SS_1}(E1^{TOT})$, modeled as a hill function, depends on $E1^{TOT} = E1^E + E1^L$, where $E1^E$ and $E1^L$ are the $E1$ proteins produced by the early and late promoter, respectively. We consider the contribution of the early promoter because, even if it is down regulated by the late one, it probably does not completely turn off at all. It is just more repressed.

SRSF1 promoter markov chain, reported in the equation **?**, is rewritten as follows

$$P_0^{SRSF1} \underset{k_{off}^{SRSF1}}{\overset{k_{on}^{SRSF1}(E_2^{TOT})}{\rightleftharpoons}} P_1^{SRSF1} \tag{4.35}$$

we see that the markov chain is the same as reported for the early promoter with the exception that we consider the total E_2, $E2^{TOT} \triangleq E2^E + E2^L$.

All the other reactions about the $mSRSF1$, $SRSF1$, $mSRSF1/SRSF1$ synthesis and degradation are the same we have reported in the previous sections.

About the late genes we know that when E2 is in enough high concentration the early polyadenylation site is inhibited, paving the way for L1 and L2 production. In a similar manner hNRNP A2/B1 cand induce HPV late gene expression, while hNRNP A1 inhibts late gene expression.

Moreover, splicing at SA5639 is enhanced by SRSF1 and SRSF9, resulting in the production of L1 mRNA. While L2 translation is repressed when the molecular complex composed by the heterogeneous nuclear ribonucleoprotein K (hnRNP K), poly(C)-binding protein 1 (PCBP1) and PCBP2 is bound to L2 coding region.

In modeling L_1 and L_2 regulation we won't consider all the previous reported mechanisms. This is because they are mediated by factors exogenous to the viral system and because their time evolution and co-regulation is not still clear. We will adopt the same approach we have done in modeling the early promoter. Therefore we assume all these factors in high concentration with respect to the viral endogenous regulation, in order to neglect their dynamics.

Hence, we will model the early polyadenylation (pAE) switch mediated by E2 in order to start producing the late transcripts. We know from literature that E2 does not bind directly to pAE. It binds to CPSF30 that, in turn, regulates pAE. We won't model CPSF30 dynamics and its binding to pAE, but just the E2 mediated control given its crucial function. Besides, we won't model FIP1, hnRNP H and CSTF4 co-regulation at pAE for the same reason we argued before.

Moreover we won't explicitly consider the SRSF1 positive regulation in splicing at SA5639. This because the E2 regulation at pAE should be enough, since E2 is inside a feedback loop with SRSF1, as well.

We have done several simplifying assumption but this is in line to keep the model structure as minimal as possible, modeling only the major endogenous regulatory mechanisms.

The polyadenylation two state markov chain is modeled as follows

$$pAE_0 \underset{k_{off}^{pAE}}{\overset{k_{on}^{pAE}(E2^{TOT})}{\rightleftharpoons}} pAE_1 \tag{4.36}$$

where pAE_0 corresponds to early polyadenylation signal activation, thus without the L1 and L2 mRNAs production, while pAE_1 regards the inhibition of pAE, thus permitting the late genes transcription. k_{off}^{pAE} and $k_{on}^{pAE}(E2^{TOT})$ are the pAE activation and inhibition rate constants, respectively. To model $k_{on}^{pAE}(E2^{TOT})$ we have assumed a Hill functional response such that

$$k_{on}^{pAE}(E2^{TOT}) = \frac{\hat{k}_{on}^{pAE}\,(E2^{TOT})^{n_{pAE}}}{\lambda_{pAE}^{n_{pAE}} + (E2^{TOT})^{n_{pAE}}} \tag{4.37}$$

where λ_{pAE} is the Michaelis constant and denotes the $E2^{TOT}$ concentration at which $k_{on}^{pAE}(E2^{TOT})$ is half its maximum value $\hat{k}_{on}^{pAE}$. The Hill coefficient, n_{pAE}, determines the sharpness of the transition about λ_{pAE}. Thus, higher the concentration of $E2^{TOT}$ is, higher the probability to get pAE_1 is, paving the way for the late mRNAs production.

In fig. 4.6.1 we report the whole model accounting the post-transcriptional regulation (splicing and polyadenylation).

4.6.1 Additional reactions

As we have done for the early promoter, after we have modeled the control part of the transcriptional and post-transcriptional regulation, what remains to model are the biochemical reactions pertaining to the synthesis, degrdation of the transcripts, proteins and the dimers formation, dissociation and degradation.

These reactions will be nearly same of the early promoter with some differences. Let's report the remaining biochemical reactions.

The late promoter regulates the production of the late primary transcript pM^L

$$\emptyset \xrightarrow{S_{pM^L}(P_i^L)} pM^L \tag{4.38}$$

where S_{pM^L} (P_i^L) is the pM^L synthesis dependent on the LP states

$$S_{pM^L}(P_i^L) = \begin{cases} s_0^{pM^L} & if \quad P_i^L = P_0^L \\ s_1^{pM^L} & if \quad P_i^L = P_1^L \\ s_2^{pM^L} & if \quad P_i^L = P_2^L \end{cases} \tag{4.39}$$

with the constraint $s_0^{pM^L} < s_1^{pM^L} < s_2^{pM^L}$ to account for the gradual activation of the promoter by the slow increase of its main regulator $C/EBP\beta$.

The late primary transcript is then converted into its transcripts

$$pM^L \xrightarrow{k_m^{E1^L}(SS_1^i)} mE1^L \tag{4.40}$$

$$pM^L \xrightarrow{k_m^{E2^L}(SS_1^i)} mE2 \tag{4.41}$$

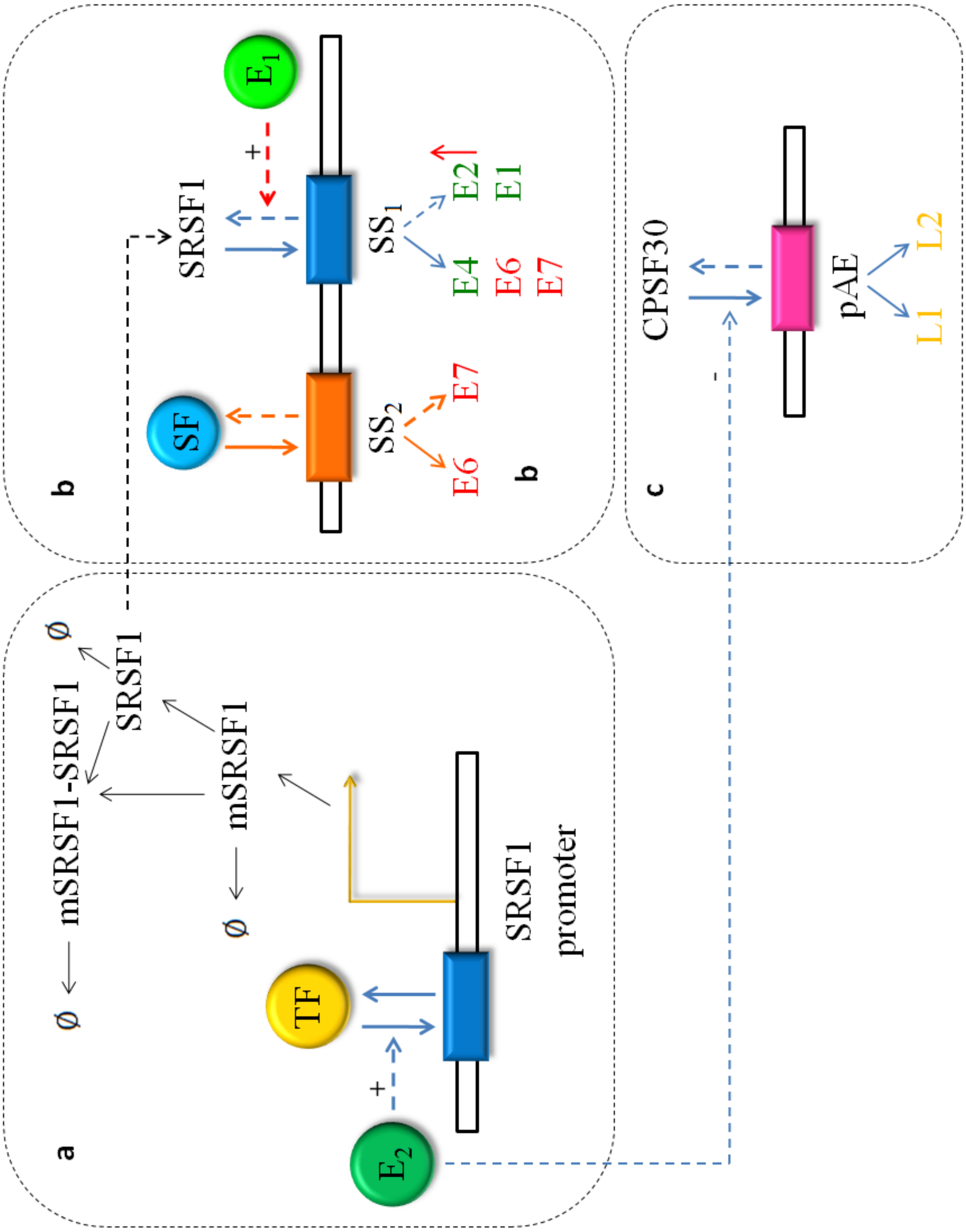

Figure 4.6.1: Same model of the splicing control we have reported so far, with the addition of the early polyadenylation site, pAE. When E_2 is in high concentration the polyadenylation site is inhibited, paving the way for L_1 and L_2 mRNAs.

$$pM^L \xrightarrow{k_m^{E1\hat{}E4^L}(SS_1^i)} mE1\hat{}E4 \tag{4.42}$$

$$pM^L \xrightarrow{k_m^{L1}(pAE_i,SS_3^i)} mL1 \tag{4.43}$$

$$pM^L \xrightarrow{k_m^{L2}(pAE_i,SS_3^i)} mL2 \tag{4.44}$$

$$mE1^L \xrightarrow{\delta_m^{E1^L}} \emptyset \tag{4.45}$$

$$mE2^L \xrightarrow{\delta_m^{E2^L}} \emptyset \tag{4.46}$$

$$mE4^L \xrightarrow{\delta_m^{E1\hat{}E4^L}} \emptyset \tag{4.47}$$

$$mL1^L \xrightarrow{\delta_m^{L1}} \emptyset \tag{4.48}$$

$$mL2^L \xrightarrow{\delta_m^{L2}} \emptyset \tag{4.49}$$

$$pM^L \xrightarrow{k_s^L} \emptyset \tag{4.50}$$

where $mE1^L$, $mE2^L$, $mE4^L$, $mL1$ and $mL2$ are the transcripts of $E1^L$, $E2^L$ and $E1\hat{}E4^L$ early genes produced by the late promoter, respectively, while $mL1$ and $mL2$ are the transcripts of $L1$ and $L2$ late genes produced by the late promoter, respectively. $k_m^{E1}(SS_1^i)$, $k_m^{E2}(SS_1^i)$, $k_m^{mE1\hat{}E4}(SS_1^i)$, $k_m^{L1}(pAE_i,\ SS_3^i)$ and $k_m^{L2}(pAE_i,\ SS_3^i)$ are the rate constants relating to the late primary transcript (pM^L) conversion into $mE1^L$, $mE2^L$, $mE1\hat{}E4^L$, $mL1$ and $mL2$, respectively. The last four reactions account for the transcripts degradation and $\delta_m^{E1^L}$, $\delta_m^{E2^L}$, $\delta_m^{E1\hat{}E4^L}$, δ_m^{L1} and δ_m^{L2} are the degradation constant rates for $mE1^L$, $mE2^L$, $mE4^L$, $mL1$ and $mL2$, respectively.

The conversion rates of the early genes, transcribed from the late pre-mRNA, are dependent on the splicing site (SS_1^i) state through the relationships

$$k_m^{E1^L}(SS_1^i) = \begin{cases} \hat{k}_m^{E1^L} & if \quad SS_1^i = SS_1^0 \\ 0 & if \quad SS_1^i = SS_1^1 \end{cases} \tag{4.51}$$

$$k_m^{E2^L}(SS_1^i) = \begin{cases} \hat{k}_m^{E2^L} & if \quad SS_1^i = SS_1^0 \\ 0 & if \quad SS_1^i = SS_1^1 \end{cases} \tag{4.52}$$

$$k_m^{E4^L}(SS_1^i) = \begin{cases} 0 & if \quad SS_1^i = SS_1^0 \\ \hat{k}_m^{E4^L} & if \quad SS_1^i = SS_1^1 \end{cases} \tag{4.53}$$

where $k_m^{E1^L}(SS_1^i)$, $k_m^{E2^L}(SS_1^i)$ and $k_m^{E2^L}(SS_1^i)$ are the $mE1^L$, $mE2^L$ and $mE1\hat{}4^L$ conversion rates, from the late primary transcript, respectively. From the previous formulae we can observe that E_1^L and E_2^L mRNAs are produced when the splicing site is free (splicing at SA2709 splicing site), while $E_1\hat{}E_4$ mRNA is produced when the splicing site is occupied (splicing at SA3358 splicing site).

Let's observe that the oncogenes are not considered since they are not transcribed from the late promoter.

The conversion rates of the late genes are dependent on the early polyadenylation site, pAE, as reported below

$$k_m^{L1}(pAE_i) = \begin{cases} 0 & if \quad pAE_i = pAE_0 \\ \hat{k}_m^{L1} & if \quad pAE_i = pAE_1 \end{cases} \tag{4.54}$$

$$k_m^{L2}(pAE_i) = \begin{cases} 0 & if \quad pAE_i = pAE_0 \\ \hat{k}_m^{L2} & if \quad pAE_i = pAE_1 \end{cases} \tag{4.55}$$

where L_1 and L_2 are transcribed with conversion rates $\hat{k}_m^{L1}$ and $\hat{k}_m^{L2}$ when the early polyadeylation signal is inhibited (pAE_1).

The rate k_s^L , in formula 4.50, takes into account for the conversion of the late primary transcript into the early transcripts we are not interested to model: E_5^L. From the literature, when the production of the late genes starts there is also a switch from the transcription of $E1\hat{}E4^L$ and E_5^L (besides E_1^L and E_2^L) to the transcription of $E1\hat{\ }E4^L$, L_1 and L_2. Therefore to account for this latter mechanism with the following definition.

$$k_s^L(pAE_i) = \begin{cases} \hat{k}_s^L & if \quad pAE_i = pAE_0 \\ 0 & if \quad pAE_i = pAE_1 \end{cases} \tag{4.56}$$

where $\hat{k}_s$ is the value for the conversion of the primary transcript into E5 when the polyadenylation site is not inhibited, otherwise $k_s^L(pAE) = 0$ and E_5^L is not produced anymore.

The transcripts are then translated into proteins

$$\emptyset \xrightarrow{k_p^{E_1^L}\, mE_1^L} E_1^L \tag{4.57}$$

$$E_1^L \xrightarrow{\delta_p^{E_1^L}} \emptyset \tag{4.58}$$

$$\emptyset \xrightarrow{k_p^{E_2^L}\, mE_2^L} E_2^L \tag{4.59}$$

$$E_2^L \xrightarrow{\delta_p^{E_2^L}} \emptyset \tag{4.60}$$

$$\emptyset \xrightarrow{k_p^{E_1\hat{\ }E_4^L}\, mE_1\hat{\ }E_4^L} E_1\hat{\ }E_4^L \tag{4.61}$$

$$E_1\hat{\ }E_4^L \xrightarrow{\delta_p^{E_1\hat{\ }E_4^L}} \emptyset \tag{4.62}$$

$$\emptyset \xrightarrow{k_p^{L1}\, mL_1} L_1 \tag{4.63}$$

$$L_1 \xrightarrow{\delta_p^{L1}} \emptyset \tag{4.64}$$

$$\emptyset \xrightarrow{k_p^{L2}\, mL_2} L_2 \tag{4.65}$$

$$L_2 \xrightarrow{\delta_p^{L_2}} \emptyset \tag{4.66}$$

where E_1^L, E_2^L, $E_1\hat{}E_4^L$, L_1 and L_2 are the proteins translated from the transcripts mE_1^L, mE_2^L, $mE_1\hat{}E_4^L$, mL_1 and mL_2, respectively. The translation rates for these latter proteins are proportional to their transcripts through the rate constants $k_p^{E_1^L}$, $k_p^{E_2^L}$, $k_p^{E_1\hat{}E_4}$, $k_p^{L_1}$ and $k_p^{L_2}$, respectively. $\delta_p^{E_1^L}$, $\delta_p^{E_2^L}$, $\delta_p^{E_1E_4^L}$, $\delta_p^{L_2}$ and $\delta_p^{L_1}$ are the degradation rates for E_1^L, E_2^L, $E_1E_4^L$, L_1 and L_2, respectively. The biochemical reactions regarding the formation of the DE_1 and DE_2 dimers and of the E_1E_2 tetramer are the same of that explained in the early promoter section, just with the proteins produced by the late promoter, and are the following.

$$E_1^L + E_1^L \underset{k_r^{DE_1}}{\overset{k_f^{DE_1}}{\rightleftharpoons}} DE_1^L \xrightarrow{\delta_{DE_1^L}} \emptyset \tag{4.67}$$

$$E_2^L + E_2^L \underset{k_r^{DE_2}}{\overset{k_f^{DE_2}}{\rightleftharpoons}} DE_2^L \xrightarrow{\delta_{DE_2^L}} \emptyset \tag{4.68}$$

$$DE_1^L + DE_2^L \underset{k_r^{E_1E_2}}{\overset{k_f^{E_1E_2}}{\rightleftharpoons}} E_1E_2^L \xrightarrow{\delta_{E_1E_2^L}} \emptyset \tag{4.69}$$

where $k_f^{DE_1}$, $k_r^{DE_1}$ are the forward and reverse rate constants regarding the formation of the dimer DE_2^L, respectively and $k_f^{DE_2}$, $k_r^{DE_2}$ are the forward and reverse rate constants regarding the formation of the dimer DE_1^L, respectively. $k_f^{E_1E_2}$ and $k_r^{E_1E_2}$ are the forward and reverse rate constants regarding the formation of the tetramer $E_1E_2^L$, respectively; $\delta_{E_1E_2^L}$, $\delta_{DE_2^L}$ and $\delta_{DE_1^L}$ are the degradation rates for the tetramer and the dimers, respectively. To take into account of the stabilization effect mediated by E_1^L on E_2^L, we have assumed the constraint $\delta_{E_1E_2} < \delta_{E_2}$.

Finally, we model the exogenous forcing function, to the viral system, depicted by the $C/EBP\beta$ transcription factor. It is very likely $C/EBP\beta$ has a low stochasticity due to its very low degradation, and further stabilization by other exogenous pathways. Moreover its differentiation-dependent regulation is not well characterized yet. We need it by slowly modulate the late promoter activity during the cell differentiation so, in the first place, we choose to model $C/EBP\beta$ as a simple birth and death process, without modeling its promoter dynamics. We also choose to do not distinguish its transcripts and proteins considering them as a unique entity.

The synthesis and degradation of $C/EBP\beta$ are

$$\emptyset \xrightarrow{s_{C/EBP\beta}} C/EBP\beta \tag{4.70}$$

$$C/EBP\beta \xrightarrow{\delta_p^{C/EBP\beta}} \emptyset \tag{4.71}$$

where $s_{C/EBP\beta}$ and $\delta_p^{C/EBP\beta}$ are the synthesis and degradation rates of $C/EBP\beta$, respectively.

We know that $C/EBP\beta$ undergoes post-translational modifications that generate different isoforms (LIP, LAP, LAP*) that can associate together and regulate the late promoter. We condense all these by modeling the dimerization of $C/EBP\beta$ and assuming it can only activate the late promoter, that is actually we are interested in when we want to describe the late viral life cycle.

The dimerization reactions are

$$C/EBP\beta + C/EBP\beta \underset{k_r^{C/EBP\beta}}{\overset{k_f^{C/EBP\beta}}{\rightleftharpoons}} DC/EBP\beta \xrightarrow{\delta_{DE_2^L}} \emptyset \tag{4.72}$$

where $DC/EBP\beta$ is the $C/EBP\beta$ dimer and when $k_f^{C/EBP\beta}$ and $k_r^{C/EBP\beta}$ are the forward and revers rate constants in forming the dimer.

We know from literature that $C/EBP\beta$ has a very low degradation. However it is strongly stabilized by different mechanisms in order to slowly increase during the cellular differentiation as well as other important factors involved in the differentiation, like p63. We do not model all the complex mechanisms that stabilize $C/EBP\beta$ because we want to maintain a model structure as simple as possible. A trade-off among these needs could be satisfied by condensing the numerous and complex mechanisms interacting with $C/EBP\beta$ with the insertion of an explicit delay between the $C/EBP\beta$ dimer and its binding to the promoter. To do this we can model a chain of delay reactions each of them with the synthesis rate equal to the degradation rate in order to have a "pure" delay (i.e. the steady state will be the same of the non delayed $DC/EBP\beta$). To account for a strong delay the rate constants can be made small enough. The delay chain is the following

$$\begin{aligned} DC/EBP\beta \xrightarrow{k_{delay}} DC/EBP\beta^{d_1} \xrightarrow{k_{delay}} DC/EBP\beta^{d_2} \xrightarrow{k_{delay}} DC/EBP\beta^{d_3} \\ \xrightarrow{k_{delay}} DC/EBP\beta^{d_4} \end{aligned} \tag{4.73}$$

where k_{delay} is the delay rate constant for the delay chain, and $DC/EBP\beta^{delayed_i}$ is $DC/EBP\beta$ after "i" delays.

5
Master Equation

All the chemical species of the system are random variables, because the chemical reactions that change the state of the system occur randomly in time. If we assume that the dwell time in any particular chemical state of the system is exponentially distributed, then the system satisfies the Markov property (Kepler and Elston (2001), Scott (2013)). Physically, this means that the time evolution of the system is determined solely by its current state and is independent of its past. The Markov property allows us to write down a Master Equation (ME) for the time evolution of the systems probabilities (VanKampen (2007), Gardiner (2009)).

The Master Equation of the whole system is quite complex and it is not possible to solve it analitically. The simulations we'll show in the results chapter 8, will be performed by using the Gillespie's algorithm: This latter reproduces the exact numerical solution of the Master Equation related to the whole system. However, to get some analytical insights, a treatable simulation of the promoters probability distribution and a quite good and manageable deterministic version of the stochastic model, we account for a simplified version of the master equation considering the stochasticity only associated to the promoters, splicing sites and ploy-adenylation sites, that are the major stochastic sources. This will be useful to design the promoters and properly tune their parameters in order to fit the available data set we will report in chapter 7. Moreover, by using the steady state probabilities of this simplified master equation we will easily get a quasi-equilibrium approximation to the system in order to get a possible deterministic formulation, as we will see in chapter 6.

5.1 Master Equation of the early promoter "regulatory core"

At any given time t, the state of the system, described by the biochemical reactions, reported in the previous chapter, is specified by the random vector

$$\mathbf{X} = \left[\begin{array}{c} P_i^E \\ P_i^{SRSF1} \\ SS_1^i \\ \hline pM \\ mE_1 \\ mE_2 \\ E_1 \\ E_2 \\ DE_1 \\ DE_2 \\ E_1E_2 \end{array}\right] = \left[\begin{array}{c} \mathbf{X^C} \\ \hline \mathbf{X^S} \end{array}\right] \tag{5.1}$$

where $\mathbf{X^C}$ is the sub-vector that corresponds to the states of the control markov chains regarding the promoters and the splicing sites, while $\mathbf{X^S}$ is the sub-vector which is inherent the "standard" chemical species.

The whole master equation would describe the time evolution of the probabilities of the entire state space

$$
\begin{aligned}
p_{\mathbf{X^S}}^{\mathbf{X^C}}(t) = p_{pM,\, mE_1,\, mE_2,\, E_1,\, E_2,\, E_1E_2,\, DE_2}^{P_i^E,\, P_i^{SRSF1},\, SS_i}(t) = \\
& Pr\Big[pM(t) = pM,\; mE_1(t) = mE_1,\; mE_2(t) = mE_2,\; E_1(t) = E_1 \\
& E_2(t) = E_2,\; E_1E_2(t) = E_1E_2,\; DE_2(t) = DE_2,\; P_i^E(t) = P_i^E \\
& P_i^{SRSF1}(t) = P_i^{SRSF1},\; SS_i(t) = SS_i\Big]
\end{aligned}
\tag{5.2}
$$

We will simulate the whole system with the aid of the Gillespie's syochastic algorithm, nevertheless, as we said in the chapter introduction, in this section we will consider a simplified master equation of the early promoter, the SRSF1 promoter and SS_1 splicing site.
The ME of the early promoter markov chain in fig. 4.5.1 is described by the following differential equations system

$$
\begin{aligned}
\frac{\mathrm{d}p_0^{EP}}{\mathrm{d}t} &= -(k_1\, DE_2 + k_5\, E_1E_2)\, p_0^{EP} + k_2\, p_1^{EP} + k_6\, p_3^{EP} \\
\frac{\mathrm{d}p_1^{EP}}{\mathrm{d}t} &= -(k_2 + k_3\, DE_2 + k_9\, E_1E_2)\, p_1^{EP} + k_1\, DE_2\, p_0^{EP} + k_4\, p_2^{EP} + k_{10}\, p_4^{EP} \\
\frac{\mathrm{d}p_2^{EP}}{\mathrm{d}t} &= -k_4\, p_2^{EP} + k_3\, DE_2\, p_1^{EP} \\
\frac{\mathrm{d}p_3^{EP}}{\mathrm{d}t} &= -(k_6 + k_7\, DE_2)\, p_3^{EP} + k_5\, E_1E_2\, p_0^{EP} + k_8\, p_4^{EP} \\
\frac{\mathrm{d}p_4^{EP}}{\mathrm{d}t} &= -(k_8 + k_{10})\, p_4^{EP} + k_7\, DE_2\, p_3^{EP} + k_9\, E_1E_2\, p_1^{EP}
\end{aligned}
\tag{5.3}
$$

where p_n^E denotes the probability of being in chemical state P_n^E at time t. The initial conditions are $p_0^E = 1$ and $p_i^E = 0$ for $i \neq 0$, so we start from the basal transcription, coherently with literature. Where, since the states are mutually exclusive, we can write, by applying the total probability theorem the constraint $Pr\left[\bigcup_{i=0}^{4} P_i^E(t)\right] = \sum_{i=0}^{4} p_i^E(t) = 1, \forall\, t \geq 0$, since the EP states are mutually exclusive ($P_i^E(t) \cap P_j^E(t) = \emptyset,\ i \neq j,\ \forall\, t \geq 0$) hence the probabilities of the intersection of two or more states are zero.

The ME of the SRSF1 promoter is

$$\begin{aligned} \frac{\mathrm{d}p_0^{SRSF1}}{\mathrm{d}t} &= -k_{on}^{SRSF1}(E_2)\ \cdot p_0^{SRSF1} + k_{off}^{SRSF1}\, p_1^{SRSF1} \\ \frac{\mathrm{d}p_1^{SRSF1}}{\mathrm{d}t} &= k_{on}^{SRSF1}(E_2)\ \cdot p_0^{SRSF1} - k_{off}^{SRSF1}\, p_1^{SRSF1} \end{aligned} \tag{5.4}$$

where p_n^{SRSF1} denotes the probability of being in chemical state P_n^{SRSF1} at time t. The initial conditions are $p_0^{SRSF1} = 1$ and $p_1^{SRFS1} = 0$, since at the beginning E_2 protein, positively regulating this promoter, is not present. The two states of $SRSF1$ promoter are mutually exclusive, hence $Pr\left[P_0^{SRSF1}(t) \cup P_1^{SRFS1}(t)\right] = p_0^{SRSF1}(t) + p_1^{SRSF1}(t) = 1, \forall\, t \geq 0.$

Finally, the ME of the SS splicing site is the following

$$\begin{aligned} \frac{\mathrm{d}p_0^{SS_1}}{\mathrm{d}t} &= -k_{on}^{SS_1}\, SRSF1\, p_0^{SS_1} + k_{off}^{SS_1}\, p_1^{SS_1} \\ \frac{\mathrm{d}p_1^{SS_1}}{\mathrm{d}t} &= k_{on}^{SS_1}\, SRSF1\, p_0^{SS_1} - k_{off}^{SS_1}\, p_1^{SS_1} \end{aligned} \tag{5.5}$$

where $p_n^{SS_1}$ denotes the probability of being in chemical state $P_n^{SS_1}$ at time t. The initial conditions are $p_0^{SS_1} = 1$ and $p_1^{SS_1} = 0$. This is coherent with literature because at the beginning we have the only production of E_1 and E_2.
As in the previous two cases, the two states of SS_1 splicing site are mutually exclusive, hence $Pr\left[P_0^{SS_1}(t) \cup P_1^{SS_1}(t)\right] = p_0^{SS_1}(t) + p_1^{SS_1}(t) = 1, \forall\, t \geq 0.$

5.1.1 Master equation with oncogenes

Considering the oncogenes the state of the system for any time is the following random vector

$$\mathbf{X} = \left[\begin{array}{c} P_i^E \\ P_i^{SRSF1} \\ SS_1^i \\ SS_2^i \\ \hline pM \\ mE1 \\ mE2 \\ E1 \\ E2 \\ DE1 \\ DE2 \\ E1E2 \\ mE6 \\ mE7 \\ E6 \\ E7 \end{array}\right] = \left[\begin{array}{c} \mathbf{X^C} \\ \hline \mathbf{X^S} \end{array}\right] \tag{5.6}$$

where we have added to the $\mathbf{X^C}$ sub-vector the SS_1 splicing site random variable.

We are interested in the only control part of the state. This latter is exactly the same of the early promoter regulatory core with the only addition of the splicing site SS_2. Then the ME will be the same of before, hence the equations 5.3, 5.4 and 5.5, with the addition of the master equation for SS_2 splicing site

$$\begin{aligned} \frac{\mathrm{d}p_0^{SS_2}}{\mathrm{d}t} &= -k_{on}^{SS_2} \cdot SRSF1 \cdot p_0^{SS_2} + k_{off}^{SS_2}\, p_1^{SS_2} \\ \frac{\mathrm{d}p_1^{SS_2}}{\mathrm{d}t} &= k_{on}^{SS_2} \cdot SRSF1\, p_0^{SS_2} - k_{off}^{SS_2} \cdot p_1^{SS_2} \end{aligned} \tag{5.7}$$

where $p_n^{SS_2}$ denotes the probability of being in chemical state $P_n^{SS_2}$ at time t. The initial conditions are quite irrelevant in this case because SS_2 just controls among E_6 and E_7 which of them has to be generated.

5.2 Late Promoter Master Equation

At any given time t, the state of the system described by the biochemical reactions modeling the late promoters (reported in chapter 4), is specified by the random vector

$$\mathbf{X} = \left[\begin{array}{c} P_i^L \\ P_i^{SRSF1} \\ SS_2^i \\ SS_3^i \\ pAE_i \\ P_i^{CEBP} \\ \hline pM^L \\ mE1^L \\ mE2^L \\ mE1\hat{\ }E4^L \\ mL1 \\ mL2 \\ E1^L \\ E2^L \\ E1\hat{\ }E4^L \\ L1 \\ L2 \\ DE1^L \\ DE2^L \\ E1E2^L \\ C/EBP\beta \\ DC/EBP\beta \\ DC/EBP\beta^{delay} \end{array}\right] = \left[\begin{array}{c} \mathbf{X^C} \\ \hline \mathbf{X^S} \end{array}\right] \tag{5.8}$$

where $\mathbf{X^C}$ is the sub-vector that corresponds to the states of the control markov chains regarding the promoters and the splicing sites, while $\mathbf{X^S}$ is the sub-vector which is inherent the standard chemical species.
The ME of the late promoter markov chain, 4.33, is described by the following differential equations system

$$\begin{aligned}
&\frac{\mathrm{d}p_0^L}{\mathrm{d}t} = -k_1^{LP} \cdot DC/EBP\beta^{delay} \cdot p_0^{LP} + k_2^{LP}\, p_1^{LP} \\
&\frac{\mathrm{d}p_1^L}{\mathrm{d}t} = -(k_2^{LP} + k_3^{LP}) \cdot DC/EBP\beta^{delay} \cdot p_1^{LP} + k_1^{LP} \cdot DC/EBP\beta^{delay} \cdot p_0^{LP} \\
&\qquad + k_4^{LP}\, p_2^{LP} \\
&\frac{\mathrm{d}p_2^L}{\mathrm{d}t} = -k_4^{LP}\, p_2^{LP} + k_3^{LP} \cdot DC/EBP\beta^{delay}\, p_1^{LP}
\end{aligned} \tag{5.9}$$

where p_n^{LP} denotes the probability of being in chemical state P_n^{LP} at time t. The initial conditions are $p_0^{LP} = 1$ and $p_i^{LP} = 0$ for $i \neq 0$. This is because at the beginning the late promoter is turned off.

Since the states are mutually exclusive, we can write, by applying the total probability theorem the constraint $Pr\left[\bigcup_{i=0}^{2} P_i^{LP}(t)\right] = \sum_{i=0}^{2} p_i^{LP}(t),\ \forall\, t \geq 0$, since the LP states are mutually exclusive $(P_i^{LP}(t) \cap P_j^{LP}(t) = \emptyset,\ i \neq j,\ \forall\, t \geq 0)$ hence the probabilities of the intersection of two or more states are zero.

The ME of the polyadenylation site biochemical equation, 4.36, is

$$\begin{aligned}
\frac{\mathrm{d}p_0^{pAE}}{\mathrm{d}t} &= -k_{on}^{pAE}(E2^{TOT})\, pAE_0 + k_{off}^{pAE}\, pAE_1 \\
\frac{\mathrm{d}p_1^{pAE}}{\mathrm{d}t} &= -k_{off}^{SS_2}\, pAE_1 + k_{on}^{pAE}(E2^{TOT})\, pAE_0
\end{aligned} \tag{5.10}$$

where p_n^{pAE} denotes the probability of being in chemical state P_n^{pAE} at time t. The initial conditions are $p_0^{pAE} = 1$ and $p_1^{pAE} = 0$. At the beginning the late genes are not produced.

As in the previous two cases, the two states of pAE polyadenylation site are mutually exclusive, hence $Pr\left[P_0^{pAE}(t) \cup P_1^{pAE}(t)\right] = p_0^{pAE}(t) + p_1^{pAE}(t) = 1, \forall\, t \geq 0$.
The remaining ME results in the master equation for the SS_1 splicing site, 5.5, we have already written down before. This is because the splicing regulation of E_1, E_2 and E_4 is the same for both the promoters.

5.3 Whole system Master Equation

The ME for the entire system, consisting of both the early and the late promoters, is exactly the union of all the master equations we have reported so far. The only thing to consider is the repression of the early promoter by the late one. To account of this regulation we have just to consider the early promoter markov chain ME to be dependent on the total $DE_2^{TOT} = DE_2^E + DE_2^L$ dimer and $E_1E_2^{TOT} = E_1E_2^E + E_1E_2^L$ concentrations produced by both the promoters. In the SRSF1 promoter ME we have to consider the dependence on $E_2^{TOT} = E_2^E + E_2^L$ and finally for the SS_1 splicing site we have to consider $E_1^{TOT} = E_1^E + E_1^L$ concentration.

6
Quasi Equilibrium Approximation

From the stochastic model we can get a deterministic version in terms of its quasi-equilibrium approximation. This latter is performed by applying the mass action kinetic law to the system of biochemical reactions describing the model. Before doing that, it is important to deterministically revisit the contribution of the stochastic sources of the system. The idea behind the quasi-equilibrium (QE) is to assume there exists two distinct time scales inside the system dynamics: one fast and the other slow. The hypothesis is to consider the time scale of the control part of the system (promoters, splicing sites, polyadenylation sites) fast with respect to the other processes (transcription, degradations, dimerization,...) and subsequently averaging over the fast variables. In particular, to satisfy the assumption of fast fluctuations we'll consider the steady state level of master equation describing the control part of the system (Kepler and Elston (2001), Gardiner (2009)).

The QE follows very well the average of the stochastic system when the control part of the system is under a fast fluctuations regimen. Otherwise (in the presence of slower promoters, splicing sites and polyadenylation sites fluctuations) the QE tends to fail in following the stochastic predictions as we'll see in the results chapter. This is good when we are interested to investigate the additive stochastic behavior a deterministic formulation cannot predict. In any case, QE is an easy approach to get a good deterministic version of a model starting from its stochastic formulation.

We will follow the same scheme of the previous chapters. At first we'll derive the QE for the early promoter regulatory core, secondly we'll consider the oncogenes and finally the late promoter.

6.1 Early Promoter quasi-equilibrium (QE)

For each promoter and splicing states we can build an indicator function

$$\chi_i^{EP}(t) = \begin{cases} 1 & if \quad P_j^E(t) = P_i^E(t) \\ 0 & if \quad P_j^E(t) \neq P_i^E(t) \end{cases} \tag{6.1}$$

$$\chi_i^{SRSF1}(t) = \begin{cases} 1 & if \quad P_j^{SRSF1}(t) = P_i^{SRSF1}(t) \\ 0 & if \quad P_j^{SRSF1}(t) \neq P_i^{SRSF1}(t) \end{cases} \tag{6.2}$$

$$\chi_i^{SS_1}(t) = \begin{cases} 1 & if \quad P_j^{SS_1}(t) = P_i^{SS_1}(t) \\ 0 & if \quad P_j^{SS_1}(t) \neq P_i^{SS_1}(t) \end{cases} \tag{6.3}$$

where $\chi_i^E(t)$, $\chi_i^{SRSF1}(t)$ and $\chi_i^{SS_1}(t)$ are the indicator functions for the early promoter, $SRSF1$ promoter and SS_1 splicing site. In this way we can rewrite the formulae 4.11, 4.15,4.22,4.21 as follows

$$S_{pM}(P_i^E(t)) = \sum_{i=0}^{4} s_i^{pM^E} \, \chi_i^{EP}(t) \tag{6.4}$$

$$S_{mSRSF1}(P_i^{SRSF1}(t)) = s_0^{SRSF1} \, \chi_0^{SRSF1}(t) + s_1^{SRSF1} \, \chi_1^{SRSF1}(t) \tag{6.5}$$

$$k_m^{E_2}(SS_1^i(t)) = \hat{k}_m^{E_2}\,\chi_0^{SS_1}(t) + 0\,\chi_1^{SS_1}(t) = \hat{k}_m^{E_2}\,\chi_0^{SS_1}(t) \tag{6.6}$$

$$k_m^{E_1}(SS_1^i(t)) = \hat{k}_m^{E_1}\,\chi_0^{SS_1}(t) + 0\,\chi_1^{SS_1}(t) = \hat{k}_m^{E_1}\,\chi_0^{SS_1}(t) \tag{6.7}$$

The previous formulae being linear combination of indicator functions are, in turn, stochastic variables. To get a deterministic approximation of these latter we can calculate their expectation. From probability (Billingsley (2012)) we can write

$$\mathbb{E}\left[\chi_i^{EP}(t)\right] = p_i^{EP}(t) \tag{6.8}$$

$$\mathbb{E}\left[\chi_i^{SRSF1}(t)\right] = p_i^{SRSF1}(t) \tag{6.9}$$

$$\mathbb{E}\left[\chi_i^{SS}(t)\right] = p_i^{SS_1}(t) \tag{6.10}$$

where $\mathbb{E}[\cdot]$ is the expectation operator and p_i^E, p_i^{SRSF1}, p_i^{SS} are the probabilities to be in the state i-th for the early promoter, the $SRSF1$ promoter and for SS_1 splicing site, respectively. To get the fast fluctuation assumption verified we need to consider the steady state probabilities of the ME. In what follows we'll directly assume to be in the steady state master equation regimen, in order to get the steady state probabilities when we apply the expectation operator.

To get a deterministic approximation of 6.4, 6.5 and 6.26 we apply the expectation operator

$$\begin{aligned}\bar{S}_{pM}(P_i^{EP}) &= \mathbb{E}\left[S_{pM^E}(P_i^{EP})\right] = \mathbb{E}\left[\sum_{i=0}^{4} s_i^{pM^E}\,\chi_i^{EP}(t)\right] = \sum_{i=0}^{4} s_i^{pM^E}\,\mathbb{E}\left[\chi_i^{EP}(t)\right]\\ &= \sum_{i=0}^{4} s_i^{pM^E}\,\hat{p}_i^{E}\\ &= \sum_{i=0}^{4} \mathbb{E}[S_{pM^E}(P_i^{EP}|P_j^{EP} = P_i^{EP}]\,\hat{p}_i^{EP}\end{aligned} \tag{6.11}$$

$$\begin{aligned}\bar{S}_{mSRSF1}(P_i^{SRSF1}) &= \mathbb{E}\left[S_{mSRSF1}(P_i^{SRSF1})\right]\\ &= \mathbb{E}\left[s_0^{mSRSF1}\,\chi_0^{SRSF1} + s_1^{mSRSF1}\,\chi_1^{SRSF1}\right]\\ &= s_0^{mSRSF1}\,\mathbb{E}\left[\chi_0^{SRSF1}\right] + s_1^{mSRSF1}\,\mathbb{E}\left[\chi_1^{SRSF1}\right]\\ &= s_0^{mSRSF1}\,\hat{p}_0^{SRSF1} + s_1^{mSRSF1}\,\hat{p}_1^{SRSF1}\\ &= \mathbb{E}[S_{mRSF1}(P_i^{SRSF1}|P_i^{SRSF1} = P_0^{SRSF1}]\,\hat{p}_0^{SRSF1} +\\ &\mathbb{E}[S_{mRSF1}(P_i^{SRSF1}|P_i^{SRSF1} = P_1^{SRSF1}]\,\hat{p}_1^{SRSF1}\end{aligned} \tag{6.12}$$

$$\begin{aligned}\bar{k}_m^{E_2}(SS_1^i) &= \mathbb{E}\left[k_m^{E_2}(SS_1^i)\right] = \mathbb{E}\left[\hat{k}_m^{E2}\,\chi_0^{SS_1} + 0\,\chi_1^{SS_1}\right]\\ &= \hat{k}_m^{E2}\,\mathbb{E}\left[\chi_0^{SS_1}\right]\\ &= k_m^{E2}\,\hat{p}_0^{SS_1} = \mathbb{E}[k_m^{E2}(SS_1^i)|SS_1^i = SS_1^o]\,\hat{p}_0^{SS_1}\end{aligned} \tag{6.13}$$

where $\hat{p}_i^E$, $\hat{p}_i^{SRSF1}$ and $\hat{p}_i^{SS_1}$ are the steady state probabilities of the approximated master equation presented in the previous paragraph. The steady state probabilities are function of DE_2 and E_1E_2 but they are not explicitly dependent on time anymore. Hence, we can interpret $\bar{S}_{pM}(P_i^{EP})$, $\bar{S}_{mSRSF1}(P_i^{SRSF1})$, $\bar{k}_m^{E_2^E}(SS_1^i)$, $\bar{k}_m^{E_1^E}(SS_1^i)$ as weighted sums of the probability to stay in each of the markov chain state they refer to, where the weights are the rate constants associated to each state. Actually these latter are the conditional expectations of the stochastic variable we defined before as linear combination of indicator functions, conditioned to the correspondent state of the markov chain.

Now, applying the mass action kinetics law and and using the formulae 6.11, 6.12, 6.13 we can write down the quasi-equilibrium approximation

$$\frac{\mathrm{d}pM^E}{\mathrm{d}t} = \bar{S}_{pM}(P_i^{EP}) - (k_s^E + \bar{k}_m^{E_2^E}(SS_1^i) + \bar{k}_m^{E_1^E}(SS_1^i))\, pM^E$$

$$\frac{\mathrm{d}mE_2^E}{\mathrm{d}t} = \bar{k}_m^{E_2^E}(SS_1^i)\, pM^E - \delta_m^{E_2}\, mE_2^E$$

$$\frac{\mathrm{d}mE_1^E}{\mathrm{d}t} = k_m^{E_1^E}(SS_1^i)\, pM^E - \delta_m^{E1}\, mE_1^E$$

$$\frac{\mathrm{d}E_2^E}{\mathrm{d}t} = k_p^{E_2^E}\, mE_2^E - \delta_p^{E_2^E}\, E_2^E - 2\, k_f^{DE2}\, (E_2^E)^2 + 2\, k_r^{DE2}\, DE_2^E$$

$$\frac{\mathrm{d}E_1^E}{\mathrm{d}t} = k_p^{E_1^E}\, mE_1^E - \delta_p^{E1}\, E_1^E - 2\, k_f^{DE2}\, (E_1^E)^2 + 2\, k_r^{DE1}\, DE_1^E$$

$$\frac{\mathrm{d}DE_2^E}{\mathrm{d}t} = k_f^{DE2}\, (E_2^E)^2 - \left(k_r^{DE2} + \delta_{DE2}\right)\, DE_2^E - k_f^{E1E2}\, DE_1^E\, \cdot DE_2^E + k_r^{E1E2}\, E_1E_2^E$$

$$\frac{\mathrm{d}DE_1^E}{\mathrm{d}t} = k_f^{DE1}\, (E_2^E)^2 - \left(k_r^{DE1} + \delta_{DE1}\right)\, DE_1^E - k_f^{E1E2}\, DE_1^E\, \cdot DE_2^E + k_r^{E1E2}\, E_1E_2^E$$

$$
\begin{aligned}
\frac{\mathrm{d}E_1E_2^E}{\mathrm{d}t} &= k_f^{E1E2}\, DE_1^E \cdot DE_2^E - \left(k_r^{E1E2} + \delta_{E1E2}\right)\, E_1E_2^E \\
\frac{\mathrm{d}mSRSF1}{\mathrm{d}t} &= \bar{S}_{mSRSF1}(P_i^{mSRSF1}) + k_{rms}\, mSRSF1/SRSF1 + \\
&\quad - k_{fms}\, mSRSF1\, SRFS1 - \delta_{SRSF1}\, SRSF1 \\
\frac{\mathrm{d}SRSF1}{\mathrm{d}t} &= k_p^{SRSF1}\, mSRSF1 + k_{rms}\, mSRSF1\, SRSF1 + \\
&\quad - k_{fms}\, mSRSF1\, SRSF1 - \delta_{SRSF1}\, SRSF1 \\
\frac{\mathrm{d}mSRSF1/SRSF1}{\mathrm{d}t} &= k_{fms}\, mSRSF1\, SRFS1 + k_{rms}\, mSRSF1/SRSF1 + \\
&\quad - \delta_{mSRSF1/SRSF1}\, mSRSF1/SRSF1
\end{aligned}
\tag{6.14}
$$

6.1.1 Oncogenes QE extension

To perform the oncogenes quasi equilibrium approximation we have to calculate the expectation of the oncogenes conversion rates by means of the steady state probabilities of the simplified master equation, as we have done so far.

Regarding E_6, we can observe from the definition 4.31 that the E_6 mRNA conversion rate is non-zero only when the stochastic event $\{SS_1^i = SS_1^1\} \wedge \{SS_2^i = SS_2^0\}$ happens. Therefore we can define just an indicator function for this latter event, as

$$
\chi_{SS_1^1 \wedge SS_2^0}(t) = \begin{cases} 1 \quad if \quad \{SS_1^i(t) = SS_1^1\} \wedge \{SS_2^i(t) = SS_2^0\} \\ 0 \qquad\qquad\qquad otherwise \end{cases} \tag{6.15}
$$

Similarly, E7 mRNA conversion rate is non-zero only when the stochastic event $\{SS_1^i = SS_1^1\} \wedge \{SS_2^i = SS_2^1\}$ ir realized. Therefore we can define the indicator function

$$\chi_{SS_1^1 \wedge SS_2^1}(t) = \begin{cases} 1 \quad if \quad \{SS_1^i(t) = SS_1^1\} \wedge \{SS_2^i(t) = SS_2^1\} \\ 0 \qquad\qquad\qquad\qquad otherwise \end{cases} \tag{6.16}$$

From the probability we can write the expectation of the indicator functions as follows

$$\begin{aligned} \mathbb{E}[\chi_{SS_1^1(t) \cap SS_2^0}(t)] &= Pr\big[\{SS_1^i(t) = SS_1^1\} \cap \{SS_2^i(t) = SS_2^0\}\big] \\ &= Pr\big[\{SS_1^i(t) = SS_1^1\}\big] \cdot Pr\big[\{SS_2^i(t) = SS_2^0\}\big] \\ &= p_1^{SS_1}(t) \cdot p_0^{SS_2}(t) \quad , \forall\, t \geq 0 \end{aligned} \tag{6.17}$$

$$\begin{aligned} \mathbb{E}[\chi_{SS_1^1(t) \cap SS_2^1}(t)] &= Pr\big[\{SS_1^1(t) = SS_1^1\} \cap \{SS_2^i(t) = SS_2^1\}\big] \\ &= Pr\big[\{SS_1^i(t) = SS_1^1\}\big] \cdot Pr\big[\{SS_2^i(t) = SS_2^1\}\big] \\ &= p_1^{SS_1}(t) \cdot p_1^{SS_2}(t) \quad , \forall\, t \geq 0 \end{aligned} \tag{6.18}$$

where $Pr[\cdot]$ is the symbol we use for calculating the probability of an event, while lowercase p is the probability nomenclature we have decided to write down the master equation probabilities. In the previous formulae, we have written the probability of the intersection events (i.e. the realizations of the stochastic processes $SS_1^i(t)$ and $SS_2^i(t)$ for a fixed time t) as the product of of each event probabilities, thanks to the statistical independence of the two splicing sites binding dynamics.

As we have done so far, we can write the expectation of the conversion rates for the oncogenes mRNAs

$$\begin{aligned}
\bar{k}_m^{E_6}(SS_1^i(t), SS_2^i(t)) = \mathbb{E}\left[k_m^{E_6}(SS_1^i(t), SS_2^i(t))\right] &= \mathbb{E}\left[\hat{k}_m^{E6}\,\chi_{SS_1^1(t)\cap SS_2^0(t)}\right] \\
&= \hat{k}_m^{E6}\,\mathbb{E}\left[\chi_{SS_1^1(t)\cap SS_2^0(t)}\right] \\
&= \hat{k}_m^{E6}\,\hat{p}_1^{SS_1}\cdot\hat{p}_0^{SS_2} \\
&= \mathbb{E}[k_m^{E_6}(SS_i(t))|\{SS_1^i(t) = SS_1^1\}\cap\{SS_2^i(t) = SS_2^0\}]\,\hat{p}_1^{SS_1}\cdot\hat{p}_0^{SS_2}
\end{aligned} \tag{6.19}$$

$$\begin{aligned}
\bar{k}_m^{E_7}(SS_1^i(t), SS_2^i(t)) = \mathbb{E}\left[k_m^{E_7}(SS_1^i(t), SS_2^i(t))\right] &= \mathbb{E}\left[\hat{k}_m^{E7}\,\chi_{SS_1^1(t)\cap SS_2^1(t)}\right] \\
&= \hat{k}_m^{E7}\,\mathbb{E}\left[\chi_{SS_1^1(t)\cap SS_2^1(t)}\right] \\
&= \hat{k}_m^{E7}\,\hat{p}_1^{SS_1}\cdot\hat{p}_1^{SS_2} \\
&= \mathbb{E}[k_m^{E_7}(SS_i(t))|\{SS_1^i(t) = SS_1^1\}\cap\{SS_2^i(t) = SS_2^1\}]\,\hat{p}_1^{SS_1}\cdot\hat{p}_1^{SS_2}
\end{aligned} \tag{6.20}$$

where we have considered the steady state probabilities, of the simplified master equation, in order to satisfy the fast fluctuation to get the quasi-equilibrium approximation. We observe we haven't indicated the steady state probabilities as function of time.

The QE extended system accounting the oncogenes is pretty the same we have already derived. WE just need to update the pre-mRNA equation to account for the oncogenes conversion and to add the E_6 and E_7 transcripts and proteins equation by using the mass action kinetic law and the formulae 6.19 and 6.20, we have just derived.

The QE update of the pre-mRNA is

$$
\begin{aligned}
\frac{\mathrm{d}pM^E}{\mathrm{d}t} =& \bar{S}_{pM^E}(P_i^{EP}) - (k_s^E + \bar{k}_m^{E2}(SS_1^i) \\
&+ k_m^{E1}(SS_1^i) + \bar{k}_m^{E6}(SS_1^i, SS_2^i) + \bar{k}_m^{E7}(SS_1^i, SS_2^i))\, pM^E \\
\frac{\mathrm{d}mE_6}{\mathrm{d}t} &= \bar{k}_m^{E_6}(SS_1^i, SS_2^i)\, pM^E - \delta_m^{E_6}\, mE_6^E \\
\frac{\mathrm{d}mE_7}{\mathrm{d}t} &= \bar{k}_m^{E_7}(SS_1^i, SS_2^i)\, pM^E - \delta_m^{E_7}\, mE_7 \\
\frac{\mathrm{d}E_6}{\mathrm{d}t} &= k_p^{E_6}\, mE_6 - \delta_p^{E_6}\, E_6 \\
\frac{\mathrm{d}E_7}{\mathrm{d}t} &= k_p^{E_7}\, mE_7 - \delta_p^{E_7}\, E_7
\end{aligned}
\tag{6.21}
$$

6.2 Late promoter Quasi Equilibrium Approximation

Following the same procedure of the previous sections, considering the stochasticity associated to the only promoters, splicing sites and polyadenylation sites, we can derive a deterministic formulation for the late promoter pre-mRNA synthesis, and of the conversion rates of the transcripts.

Let's define the indicator functions of the late promoter, splicing sites and polyadenylation site markov chains defined in chapter 4.

The indicator function for the late promoter states is

$$
\chi_i^{LP}(t) = \begin{cases} 1 & if \quad P_j^L(t) = P_i^L \\ 0 & if \quad P_j^L(t) \neq P_i^L \end{cases} \tag{6.22}
$$

where uppercase P_i^L is the i-th state of the late promoter markov chain. We can observe that the indicator function for the i-th state of the late promoter assumes the value 1 if and only if the promoter is exactly in that state.

Late promoter needs of SS_1 splicing site but we have already written the indicator function for SS_1 splicing site in 6.3.

Regarding L_1 and L_2 transcription, from the definition of their conversion rates (4.54, 4.55) it is non-zero if only if the stochastic event $\{pAE_i = pAE_1\}$ happens. In other words when pAE is inhibited. Therefore we define an indicator function for this latter stochastic event

$$\chi_{pAE_i=pAE_1}(t) = \begin{cases} 1 & if \quad pAE_i = pAE_1 \\ 0 & if \quad otherwise \end{cases} \tag{6.23}$$

As in the previous section we rewrite the formulae 4.39, 4.51,4.52, 4.53, as the following stochastic variables

$$S_{pM^L}(P_i^L(t)) = \sum_{i=0}^{2} s_i^{pM^L} \, \chi_i^{LP}(t) \tag{6.24}$$

$$k_m^{E1^L}(SS_1^i(t)) = \hat{k}_m^{E1^L} \, \chi_0^{SS_1}(t) = \hat{k}_m^{E1^L} \, \chi_0^{SS_1}(t) \tag{6.25}$$

$$k_m^{E2^L}(SS_1^i(t)) = \hat{k}_m^{E2^L} \, \chi_0^{SS_1}(t) \tag{6.26}$$

$$k_m^{E4^L}(SS_1^i(t)) = \hat{k}_m^{E4^L} \, \chi_1^{SS_1}(t) \tag{6.27}$$

$$k_m^{L1}(pAE_i(t)) = \hat{k}_m^{L1} \, \chi_{pAE_i=pAE_1}(t) \tag{6.28}$$

$$k_m^{L2}(pAE_i(t)) = \hat{k}_m^{L2} \, \chi_{pAE_i=pAE_1}(t) \tag{6.29}$$

The previous formulas, being linear combination of indicator functions are, in turn, stochastic variables. To get a deterministic approximation of these latter we can calculate their expectations.

Before doing that let's calculate the expectations of their indicator functions. Actually, we have already done that with the exception of the late genes L_1 and L_2

$$\begin{aligned}\mathbb{E}[\chi_{pAE_i=pAE_1}(t)] &= Pr\left[\chi_{pAE_i=pAE_1}(t)\right] \\ &= Pr\left[\chi_{pAE_i=pAE_1}\right] \\ &= p_1^{pAE}(t) \quad ,\forall\, t \geq 0\end{aligned} \tag{6.30}$$

$$\begin{aligned}\mathbb{E}[\chi_{pAE_i=pAE_1}(t)] &= Pr\left[\chi_{pAE_i=pAE_1}(t)\right] \\ &= Pr\left[\chi_{pAE_i=pAE_1}\right] \\ &= p_1^{pAE}(t) \quad ,\forall\, t \geq 0\end{aligned} \tag{6.31}$$

where p_1^{pAE} is the probability to have the early polyadenylation site inhibited (by E2).

By inserting in the formulas 6.24, 6.25, 6.26, 6.27, 6.28 and 6.29 the previous expectations 6.30 and 6.31 we can get a deterministic formulation for the late promoter pre-mRNA synthesis and for the transcript conversion rates

$$S_{pM^L}(P_i^{LP}(t)) = \sum_{i=0}^{2} s_i^{pM^L}\, \chi_i^{LP}(t) \tag{6.32}$$

$$\begin{aligned}\bar{k}_m^{E_1^L}(SS_1^i(t)) &= \mathbb{E}\left[k_m^{E_1^L}(SS_1^i)\right] = \mathbb{E}\left[\hat{k}_m^{E1^L}\, \chi_0^{SS_1}\right] \\ &= \hat{k}_m^{E1^L}\, \mathbb{E}\left[\chi_0^{SS_1}\right] \\ &= \hat{k}_m^{E1^L}\, \hat{p}_0^{SS_1}\end{aligned} \tag{6.33}$$

$$\begin{aligned}\bar{k}_m^{E_2^L}(SS_1^i(t)) &= \mathbb{E}\left[k_m^{E_2^L}(SS_1^i)\right] = \mathbb{E}\left[\hat{k}_m^{E2^L}\,\chi_0^{SS_1}\right]\\ &= \hat{k}_m^{E2^L}\,\mathbb{E}\left[\chi_0^{SS_1}\right]\\ &= \hat{k}_m^{E2^L}\,\hat{p}_0^{SS_1}\end{aligned} \tag{6.34}$$

$$\begin{aligned}\bar{k}_m^{E_4^L}(SS_1^i(t)) &= \mathbb{E}\left[k_m^{E_4^L}(SS_1^i)\right] = \mathbb{E}\left[\hat{k}_m^{E4^L}\,\chi_1^{SS_1}\right]\\ &= \hat{k}_m^{E4^L}\,\mathbb{E}\left[\chi_1^{SS_1}\right]\\ &= \hat{k}_m^{E4^L}\,\hat{p}_1^{SS_1}\end{aligned} \tag{6.35}$$

$$\begin{aligned}&\bar{k}_m^{L1}(pAE_i(t)) = \mathbb{E}\left[k_m^{L2}(pAE_i(t))\right] = \mathbb{E}\left[\hat{k}_m^{L1}\,\chi_{pAE_i=pAE_1}(t)\right]\\ &= \hat{k}_m^{L1}\,p_1^{pAE}\\ &= \mathbb{E}[k_m^{L1}(pAE_i(t))|pAE_i(t) = pAE_1]\ \cdot \hat{p}_1^{pAE}\end{aligned} \tag{6.36}$$

$$\begin{aligned}&\bar{k}_m^{L2}(pAE_i(t)) = \mathbb{E}\left[k_m^{L2}(pAE_i(t))\right] = \mathbb{E}\left[\hat{k}_m^{L1}\,\chi_{pAE_i=pAE_1}(t)\right]\\ &= \hat{k}_m^{L1}\,p_1^{pAE}(t)\\ &= \mathbb{E}[k_m^{L1}(pAE_i(t))|pAE_i(t) = pAE_1]\ \cdot \hat{p}_1^{pAE}\end{aligned} \tag{6.37}$$

where, as usual $\hat{p}_i$ are the steady state probabilities of the master equation.

Using the mass action kinetics law and the synthesis and conversion rates formulation we have just derived we can write the quasi equilibrium of the late promoter

$$\begin{aligned}\frac{\mathrm{d}pM^L}{\mathrm{d}t} = \bar{S}_{pM^L}(P_i^{LP}) - (&k_s + \bar{k}_m^{E2^L}(SS_1^i) + \bar{k}_m^{E1^L}(SS_1^i) + \bar{k}_m^{E4}(SS_1^i) \\ &+ \bar{k}_m^{L1}(pAE_i) + \bar{k}_m^{L2}(pAE_i))\, pM^L\end{aligned}$$

$$\frac{\mathrm{d}mE2^L}{\mathrm{d}t} = \bar{k}_m^{E2^L}(SS_1^i)\, pM^L - \delta_m^{E2}\, mE2^L$$

$$\frac{\mathrm{d}mE1^L}{\mathrm{d}t} = k_m^{E1^L}(SS_1^i)\, pM^L - \delta_m^{E1}\, mE1^L$$

$$\frac{\mathrm{d}mE4^L}{\mathrm{d}t} = k_m^{E4^L}(SS_1^i)\, pM^L - \delta_m^{E4}\, mE4^L$$

$$\frac{\mathrm{d}mL1}{\mathrm{d}t} = k_m^{L1}(pAE_i)\, pM^L - \delta_m^{L1}\, mL1$$

$$\frac{\mathrm{d}mL2}{\mathrm{d}t} = k_m^{L2}(pAE_i)\, pM^L - \delta_m^{L2}\, mL2$$

$$\frac{\mathrm{d}E1^L}{\mathrm{d}t} = k_p^{E1^L}\, mE1^L - \delta_p^{E1^L}\, E1^L - 2\, k_f^{DE2^L}\, (E1^L)^2 + 2\, k_r^{DE1^L}\, DE1^L$$

$$\frac{\mathrm{d}E2^L}{\mathrm{d}t} = k_p^{E2^L}\, mE2^L - \delta_p^{E2}\, E2^L - 2\, k_f^{DE2^L}\, (E2^L)^2 + 2\, k_r^{DE2^L}\, DE2^L$$

$$\frac{\mathrm{d}E4^L}{\mathrm{d}t} = k_p^{E4^L}\, mE4^L - \delta_p^{E4}\, E4^L$$

$$\frac{\mathrm{d}L1}{\mathrm{d}t} = k_p^{L1}\, mL1 - \delta_p^{L1}\, L1$$

$$
\begin{aligned}
\frac{\mathrm{d}L2}{\mathrm{d}t} &= k_p^{L2}\, mL2 - \delta_p^{L2}\, L2 \\
\frac{\mathrm{d}DE2^L}{\mathrm{d}t} &= k_f^{DE2^L}\,(E2^L)^2 - \left(k_r^{DE2^L}\delta_{DE2^L}\right)\, DE2^L - k_f^{E1E2^L}\, DE1^L\, DE2^L \\
&\quad + k_r^{E1E2^L}\, E1E2^L \\
\frac{\mathrm{d}DE1^L}{\mathrm{d}t} &= k_f^{DE1^L}\,(E2^L)^2 - \left(k_r^{DE1^L} + \delta_{DE1^L}\right)\, DE1^L - k_f^{E1E2^L}\, DE1^L\, DE2^L \\
&\quad + k_r^{E1E2^L}\, E1E2^L \\
\frac{\mathrm{d}E1E2^L}{\mathrm{d}t} &= k_f^{E1E2^L}\, DE1^L\, DE2^L - \left(k_r^{E1E2^L} + \delta_{E1E2^L}\right)\, E1E2^L
\end{aligned}
\tag{6.38}
$$

6.3 WHOLE SYSTEM QUASI EQUILIBRIUM APPROXIMATION

The QE of the whole system is described by the same equations we have written down so far in this chapter. The few thing we have to account handle the two promoters together are

- The early promoter regulation depends on the total E_2 dimer $DE_2^{TOT} = DE_2^E + DE_2^L$ and tetramer $E_1E_2^{TOT} = E_1E_2^E + E_1E_2^L$ produced by both the promoters.

- The SRSF1 promoter activation and the eraly polyadenylation signal, pAE, depend on the total $E_2^{TOT} = E_2^E + E_2^L$ produced by both the promoters.

- The dissociation rate constant of SRSF1, from SS_1 splicing site, depends on the total $E_1^{TOT} = E_1^E + E_1^L$ produced by both the promoters.

7

Dataset, parameters and design of *in silico* experiments

In this chapter we report the parameters we have gathered/inferred from literature and that we have assumed to perform the model predictions.

We describe an available dataset gathered from literature about the early promoter activity and the qualitative measurements about the temporal evolution of E_1 and E_2 transcripts during the infected cell differentiation.

Finally we'll design some *in silico* experiments to investigate the dynamical behavior of the developed model and test its capability to qualitative reproduce the expected biology. The predictions will be performed in the next chapter.

7.1 Dataset

7.1.1 Early promoter activity

A dataset about the early promoter activity for HPV16 and 18 is available in literature (Hou et al. (2002)). In this dataset there are considered early promoters with all the four binding sites for E_2 dimer and the promoter activities was tested under different E_2 concentration levels. Only the negative feedback effect was evaluated.

In tables 7.1.3 and 7.1.2 we report the experimental data about the promoter activity. In fig. 8.1.1 and 8.1.2 we report the plot of the data.

Table 7.1.1: HPV16 early promoter activity

E_2 [nM]	promoter activity
0	1
1.67	0.6
8.33	0.08
41.67	0.02
166.67	0.02

Table 7.1.2: HPV18 early promoter activity

E_2 [nM]	promoter activity
0	1
1.67	0.98
8.33	0.55
41.67	0.08
166.67	0.04

In the next chapter we'll show the fit of both the promoter activities performed by using the early promoter markov chain model we have developed.

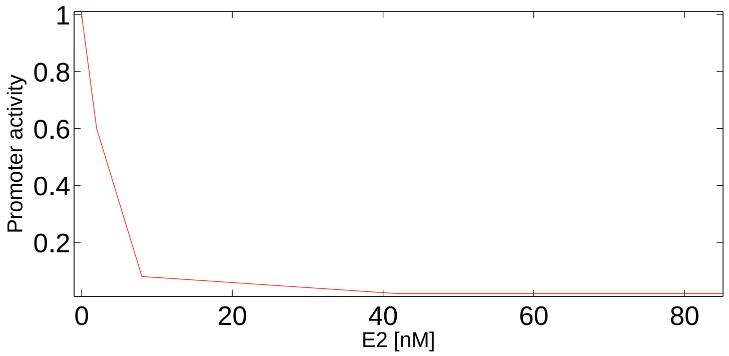

Figure 7.1.1: HPV16 early promoter activity.

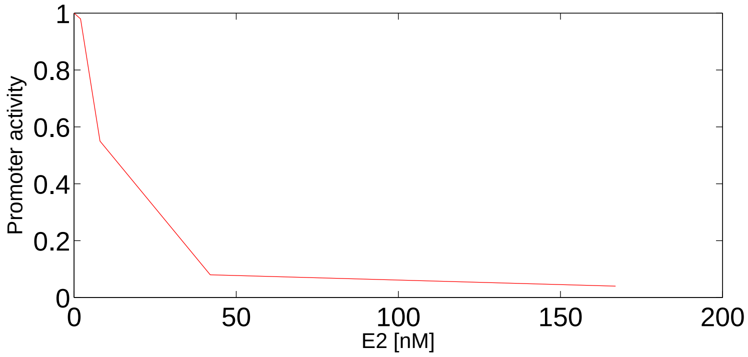

Figure 7.1.2: HPV18 early promoter activity.

7.1.2 E_1 AND E_1 MRNAS DURING DIFFERENTIATION

The only available time series about HPV transcripts is a qualitative measurements about E_1 and E_2 mRNAs (measured as relative change in band intensity from a Northen blot analysis) upon induced differentiation and a total experiment duration of 16 days (Ozbun and Meyers (1998)). Most likely these transcripts are produced by the late promoter.

In table... we report the qualitative data and inf fig..... we plot them. We specify that the last temporal data has to be neglected as it accounts for an experimental set up inaccuracy.

Table 7.1.3: E_1 and E_2 mRNAs during differentiation

time [days]	mE_1 [Relative change]	mE_2 [Relative change]
0	1.4	2
4	0.67	0.34
8	2.13	0.81
12	5.72	1.91
16	2.19	1.72

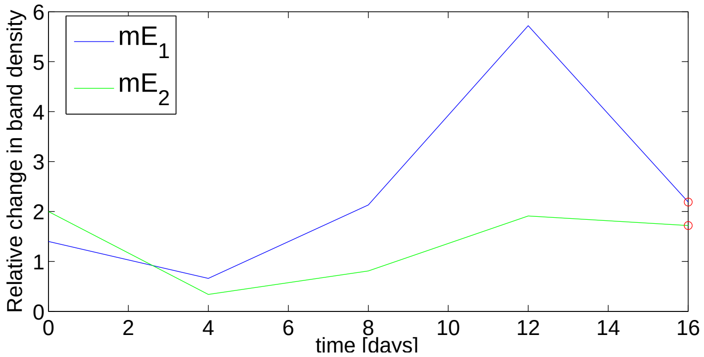

Figure 7.1.3: E_1 and E_2 mRNAs produced by the late promoter during differentiation.In red are indicated the last temporal data to neglect.

7.2 Parameters

In this section we report the parameters of the model in different tables (to distinguish between rate constants, degradations, splicing control,...) Most of them are gathered or inferred from literature. Some of them are inferred during the first qualitative assessment of the model. In fact a first validation is always performed during model formulation in order to reproduce the qualitative behavior we expect from the biological knowledge present in literature (Cobelli Claudio (2008)). Some inferred parameters were fixed in order to satisfy a fast fluctuation regimen. This is the case for the markov chain rate constants of $SRSF1$ promoter and SS_1 splicing site. We have made this choice in order to have, at first, a good agreement with the quasi equilibrium approximation. However, these parameters will be varied in different *in silico* experiments (reported in the next section) to account for stronger stochastic noise contribution.

Table 7.2.1: Rate constants Early promoter markov chain HPV16

parameter	value	unit of measure	reference
k_1^{EP}	1.89 e-2	$nM^{-1}\,min^{-1}$	Mok et al. (1996), Demeret et al. (1997), Hartley and Alexander (2002), Hou et al. (2002)
k_2^{EP}	1.92 e-2	min^{-1}	Demeret et al. (1997)
k_3^{EP}	2.00 e-2	$nM^{-1}\,min^{-1}$	Mok et al. (1996), Demeret et al. (1997), Hartley and Alexander (2002), Hou et al. (2002)
k_4^{EP}	1.15 e-1	min^{-1}	Demeret et al. (1997)
k_5^{EP}	2.58 e-2	$nM^{-1}\,min^{-1}$	Demeret et al. (1997), Chao et al. (1999)
k_6^{EP}	5.0 e-2	min^{-1}	Demeret et al. (1997), Chao et al. (1999)
k_7^{EP}	2.00 e-2	$nM^{-1}\,min^{-1}$	Mok et al. (1996), Demeret et al. (1997), Hartley and Alexander (2002), Hou et al. (2002)
k_8^{EP}	1.15 e-1	min^{-1}	Mok et al. (1996), Demeret et al. (1997), Hartley and Alexander (2002), Hou et al. (2002)
k_9^{EP}	4.54 e-3	$nM^{-1}\,min^{-1}$	Demeret et al. (1997), Chao et al. (1999),
k_{10}^{EP}	5.0 e-2	min^{-1}	Demeret et al. (1997), Chao et al. (1999)

Table 7.2.2: Rate constants Early promoter markov chain HPV18

parameter	value	unit of measure	reference
k_1^{EP}	1.89 e-2	$nM^{-1}\,min^{-1}$	Mok et al. (1996), Demeret et al. (1997), Hartley and Alexander (2002), Hou et al. (2002)
k_2^{EP}	1.92 e-2	min^{-1}	Demeret et al. (1997)
k_3^{EP}	2.98 e-2	$nM^{-1}\,min^{-1}$	Mok et al. (1996), Demeret et al. (1997), Hartley and Alexander (2002), Hou et al. (2002)
k_4^{EP}	1.15 e-1	min^{-1}	Demeret et al. (1997)
k_5^{EP}	2.58 e-2	$nM^{-1}\,min^{-1}$	Demeret et al. (1997), Chao et al. (1999)
k_6^{EP}	5.7 e-2	min^{-1}	Demeret et al. (1997), Chao et al. (1999)
k_7^{EP}	2.98 e-2	$nM^{-1}\,min^{-1}$	Mok et al. (1996), Demeret et al. (1997), Hartley and Alexander (2002), Hou et al. (2002)
k_8^{EP}	1.15 e-1	min^{-1}	Mok et al. (1996), Demeret et al. (1997), Hartley and Alexander (2002), Hou et al. (2002)
k_9^{EP}	6.79 e-3	$nM^{-1}\,min^{-1}$	Demeret et al. (1997), Chao et al. (1999)
k_{10}^{EP}	5.7 e-2	min^{-1}	Demeret et al. (1997), Chao et al. (1999)

Table 7.2.3: Degradations

parameter	value	unit of measure	reference
δ_{E1}	3.3 e-3	min^{-1}	Ozbun and Meyers (1998)
δ_{mE1}	1.65 e-2	min^{-1}	Ozbun (2002)
δ_{E2}	3.3 e-3	min^{-1}	King et al. (2011),McBride (2013), Bellanger et al. (2001), Taylor et al. (2003)
δ_{mE2}	1.65 e-2	min^{-1}	Ozbun (2002)
$\delta_{E1\hat{}E4}$	1.9 e-3	min^{-1}	Assumed
$\delta_{mE1\hat{}E4}$	1.65 e-2	min^{-1}	Assumed
δ_{E6}	1.54 e-2	min^{-1}	Ajiro Masahiko (2015)
δ_{mE6}	3.08 e-2	min^{-1}	Ozbun (2002)
δ_{E7}	1.24 e-2	min^{-1}	Ajiro Masahiko (2015)
δ_{mE7}	2.48 e-2	min^{-1}	Ozbun (2002)
δ_{L1}	1.9 e-3	min^{-1}	Collier et al. (2002)
δ_{mL1}	1.93 e-2	min^{-1}	Assumed
δ_{L2}	3.9 e-3	min^{-1}	Finnen et al. (2003)
δ_{mL2}	3.85 e-2	min^{-1}	Assumed
δ_{DE1}	6.6 e-4	min^{-1}	Hou et al. (2002)
δ_{DE2}	6.6 e-4	min^{-1}	Hou et al. (2002),Mok et al. (1996), Demeret et al. (1997)
δ_{E1E2}	3.3 e-4	min^{-1}	King et al. (2011), Hartley and Alexander (2002)
δ_{mSRSF1}	7.7 e-3	min^{-1}	Moulton et al. (2014)
δ_{SRSF1}	7.7 e-4	min^{-1}	Assumed
$\delta_{mSRSF1/SRSF1}$	1.54 e-2	min^{-1}	Sun et al. (2010)
$\delta_{C/EBP\beta}$	4.81 e-4	min^{-1}	Maytin and Habener (1998), Huber et al. (2015)

Table 7.2.4: Rate constants multimeric complexes

parameter	value	unit of measure	reference
k_f^{E1E2}	4.2 e-4	$nM^{-1}\,min^{-1}$	Chao et al. (1999), Demeret et al. (1997),Hou et al. (2002)
k_r^{E1E2}	5.00 e-2	min^{-1}	Chao et al. (1999), Demeret et al. (1997),Hou et al. (2002)
k_f^{DE1}	3.97 e-4	$nM^{-1}\,min^{-1}$	Graham (2012)
k_r^{DE1}	1.00 e-2	min^{-1}	Graham (2012)
k_f^{DE2}	3.97 e-4	$nM^{-1}\,min^{-1}$	Demeret et al. (1997),Hou et al. (2002)
k_r^{DE2}	1.0 e-2	min^{-1}	Demeret et al. (1997),Hou et al. (2002)
k_f^{ms}	9.94 e-3	$nM^{-1}\,min^{-1}$	Somberg and Schwartz (2010), Sun et al. (2010)
k_r^{ms}	1.00 e-3	min^{-1}	Somberg and Schwartz (2010), Sun et al. (2010)
$k_f^{C/EBP\beta}$	3.00 e-4	$nM^{-1}\,min^{-1}$	Assumed
$k_r^{C/EBP\beta}$	2.00 e-2	min^{-1}	Assumed

Table 7.2.5: Synthesis and conversion rates

parameter	value	unit of measure	reference
s_0^{pME}	7.55 e-1	$nM\,min^{-1}$	Hou et al. (2002)
s_1^{pME}	2.00 e-0	$nM\,min^{-1}$	Hou et al. (2002)
s_2^{pME}	2.52 e-2	$nM\,min^{-1}$	Hou et al. (2002)
s_3^{pME}	4.03 e-1	$nM\,min^{-1}$	Hou et al. (2002)
$\hat{k}_m^{E1}$	4.50 e-1	min^{-1}	Ozbun and Meyers (1998)
$\hat{k}_m^{E2}$	5.00 e-1	min^{-1}	Ozbun and Meyers (1998)
$\hat{k}_m^{E6}$	2.00 e-1	min^{-1}	Assumed
$\hat{k}_m^{E7}$	2.00 e-1	min^{-1}	Assumed
$\hat{k}_p^{E1}$	3.00 e-2	min^{-1}	Assumed
$\hat{k}_p^{E2}$	3.50 e-2	min^{-1}	Hou et al. (2002)
$\hat{k}_p^{E6}$	8.00 e-3	min^{-1}	Assumed
$\hat{k}_p^{E7}$	8.00 e-3	min^{-1}	Assumed
$\hat{k}_s^{E}$	6.93 e-2	min^{-1}	Audibert et al. (2002)
$\hat{k}_p^{SRFS1}$	3.00 e-3	min^{-1}	Somberg and Schwartz (2010), Sun et al. (2010)
s_0^{SRSF1}	1.01 e-2	$nM^{-1}\,min^{-1}$	Assumed
s_1^{SRSF1}	5.03 e-2	$nM^{-1}\,min^{-1}$	Assumed

Table 7.2.6: Splicing sites parameters and rate constants

parameter	value	unit of measure	reference
$k_{on}^{SS_1}$	2.00 e-0	$nM^{-1}\,min^{-1}$	Somberg and Schwartz (2010), Sun et al. (2010),Johansson and Schwartz (2013)
$\hat{k}_{off}^{SS_1}$	1.70 e-2	min^{-1}	Somberg and Schwartz (2010), Sun et al. (2010),Johansson and Schwartz (2013)
$\tilde{k}_{off}^{SS_1}$	1.00 e-5	min^{-1}	Assumed
λ_{SS_1}	3.50 e-0	nM	Somberg and Schwartz (2010), Sun et al. (2010),Johansson and Schwartz (2013)
n_{SS_1}	2.00 e-0	$-$	Assumed
$\hat{k}_{on}^{SRSF1}$	9.93 e-0	$nM^{-1}\,min^{-1}$	Sun et al. (2010)
k_{off}^{SRSF1}	7.0 e-1	min^{-1}	Sun et al. (2010)
λ_{SRSF1}	5.00 e-0	$nM^{-1}\,min^{-1}$	Johansson and Schwartz (2013)
n_{SRSF1}	6.00 e-0	$-$	Assumed
$k_{on}^{SS_2}$	0.1 e-0	min^{-1}	Assumed
$k_{off}^{SS_2}$	0.3 e-0	min^{-1}	Assumed

Table 7.2.7: Rate constants Late promoter markov chain

parameter	value	unit of measure	reference
k_1^{LP}	1.98 e-1	$nM^{-1}\,min^{-1}$	Assumed
k_2^{LP}	2.00 e-1	min^{-1}	Assumed
k_3^{LP}	1.98 e-2	$nM^{-1}\,min^{-1}$	Assumed
k_4^{LP}	1.00 e-1	min^{-1}	Assumed

Table 7.2.8: Additive splicing sites parameters and rate constants

parameter	value	unit of measure	reference
$\hat{k}_{on}^{pAE}$	3.00 e-1	$nM^{-1}\,min^{-1}$	Johansson and Schwartz (2013)
k_{off}^{pAE}	5.00 e-3	min^{-1}	Johansson and Schwartz (2013)
n_{pAE}	6.00 e-0	–	Assumed
λ_{pAE}	1.25 e+1	nM	Johansson and Schwartz (2013)

Table 7.2.9: Synthesis and conversion rates Late promoter

parameter	value	unit of measure	reference
$s_0^{pM^L}$	2.5 e-1	$nM\,min^{-1}$	Assumed
$s_1^{pM^L}$	1.25 e-0	$nM\,min^{-1}$	Assumed
$s_2^{pM^L}$	2.51 e-0	$nM\,min^{-1}$	Assumed
$\hat{k}_m^{E1^L}$	5.00 e-1	min^{-1}	Ozbun and Meyers (1998)
$\hat{k}_m^{E2^L}$	2.00 e-1	min^{-1}	Ozbun and Meyers (1998)
$\hat{k}_m^{E1E4^L}$	3.00 e-1	min^{-1}	Johansson and Schwartz (2013)
$\hat{k}_m^{L1}$	3.00 e-1	min^{-1}	Finnen et al. (2003)
$\hat{k}_m^{L2}$	3.50 e-1	min^{-1}	Finnen et al. (2003)
$\hat{k}_s^{L}$	6.93 e-2	min^{-1}	Audibert et al. (2002)
$\hat{k}_p^{E1^L}$	5.00 e-3	min^{-1}	Assumed
$\hat{k}_p^{E2^L}$	1.00 e-2	min^{-1}	Assumed
$\hat{k}_p^{E1E4^L}$	5.00 e-3	min^{-1}	Assumed
$\hat{k}_p^{L1}$	3.00 e-2	min^{-1}	Finnen et al. (2003)
$\hat{k}_p^{L2}$	5.00 e-2	min^{-1}	Finnen et al. (2003)
$s^{C/EBP\beta}$	2.3 e-2	$nM\,min^{-1}$	Assumed

7.3 *in silico* EXPERIMENTS DESIGN

We design five *in silico* experiments in order to investigate the behavior of the model. We'll consider four experiments about the early promoter since it presents a more complex structure with an autoregulated promoter and the splicing regulation too. The last experiment will test the late promoter functioning.

All the experiment have been designed by assuming a keratynocyte cell diameter between 15 to 35 μm and a nuclear diameter of 8.6 μm (Gareau (2011)).

In all the experiments, with the exception of Experiment 1, will show the chemical species in copy number, suitable with a stochastic approach.

7.3.1 EARLY PROMOTER

The early promoter *in silico* experiments can be useful to describe different scenarios: a basal infection situation; a precancerous lesion where the early promoter can stay active for most of the differentiation time before being strongly repressed; HPV integration into the human DNA where it over expresses the oncogenes to sustain the cancer condition.

Even if the late promoter should strongly repress the early promoter, there is no evidence the early promoter is completely turned off.

We know from literature that the copy number of different viral proteins should be around some thousands per cell. However, these numbers should consider the entire amount of proteins and transcripts from three contributions: basal infected cells, cell that committed the differentiation and integrated HPV. We know that the late promoter can generate higher amount of E_1 and E_2 and of their multimeric complexes; we also know that the integrated HPV can overexpress the oncogenes. Actually, for a qualitative investigation it is not very important to respect the real amount of molecules, unless to consider very high copy number where the stochastic contributions become negligible. We have chosen to consider a steady state copy number around some hundreds for the transcripts and few thousands for the proteins. This choice, on the one hand could be consistent with the real copy number magnitude the early promoter should produce, and on the other hand is a good choice to investigate the stochastic effects coming from a slow promoters and splicing sites fluctuations, and a finite number effect, as well.

The predictions will be simulated for an *in silico* experiment duration of a week. We are interested in investigating the qualitative dynamical behavior of the early promoter and if the purpose is to investigate the stochastic dynamics, the effective duration of the experiment is not so crucial. Anyway, a biological consistency with a simulation period of a week could be consistent with the situation with the virus in the basal cells that not differentiate; another case could be relative to cells in a precancerous stage where the effect of the late promoter can be strongly delayed.

Since in this section we are interested in the investigation of the early promoter dynamics we do not account for the repression effect mediated by the late promoter. Actually the late promoter repression on the first one is not so interesting dynamically speaking, since there isn't a co-regulation between the two promoters.

7.3.1.1 Experiment 1

In this experiment we consider the early promoter heuristically modeled in (Giaretta et al. (2015)). The main purpose is to verify the E_1 coregulation mediated on E_2 and the possible repression of the early promoter that in this first attempt we have assumed condensed as a forcing function modulating E_1 mRNA. In this chapter we have made the choice to report only the parameters for the new model version. The parameters fixed for this simulation are reported in (Giaretta et al. (2015)). Moreover in this in silico experiment we will show the results expressed in concentration, with a synthesis rate assumed to an indicative value (consistent with the upper bound copy number limit we have inferred from literature, considering the average volume of a keratynocyte).

7.3.1.2 Experiment 2

In this experiment we consider the early promoter markov chain with the parameters in table 7.2.1, in order to perform an experiment with the promoter structure and parameters we have used to fit the experimental data about its activity, as we'll show in chapter 7. In particular, we consider a eak positive feedback and a medium strong negative feedback. The oncogenes are not produced in a very hogh concentration deterministically, in order to show a first interesting stochastic behavior. The promoters and splicing sites fluctuations are imposed to be quite fast in order to have a good agreement with the quasi-equilibrium approximation, hence with the deterministic behavior of the system.

The *in silico* experiments will account for a total duration of a week.

7.3.1.3 Experiment 3

In this experiment we investigate the stochastic behavior of the early promoter in the context of slow fluctuations of both SRSF1 promoter and SS_1 splicing site. Hence investigating a slow dynamics of the splicing control part of the system. The SRSF1 and SS_1 markov chains parameters will be fixed two order of magnitude lower than respect to the previous experiment to be sure the slower fluctuations make the difference. All the other parameters are the same we have reported in the tables from 7.2.1 to 7.2.6.

7.3.1.4 Experiment 4

In this experiment we investigate the stochastic behavior of the early promoter in the context of medium slow fluctuations (i.e., on/off rates one order of magnitude lower than the values reported on tables) of both SRSF1 promoter and SS_1 splicing site.

The early promoter feedback are the same of the previous experiments.

In partcular we put to zero the parameter $\tilde{k}_{off}^{SS_1}$ in equation 4.7. In this way, the SS_1 splicing site detachment rate constant, described in equation

4.7, is a "pure" Hill function depending on E_1 concentration, accounting for the enhancement regulation of E_2 transcription. We'll see that in enough slow fluctuations a very interesting complex pattern will arise.

7.3.1.5 Experiment 5

In this experiment we consider the same promoter feedback strength but we make the early promoter markov chain rate constants slower of one magnitude order with respect the parameters inferred from literature.

Then, we consider a quite fast SRSF1 promoter fluctuations (i.e., on/off rates one order of magnitude lower than the values on the tables) and a quite slow SS_2 splicing fluctuations (i.e., on/off rates two order of magnitude lower than the values on tables). Moreover, we have considered a lower total conversion rate of the oncogenes (i.e., about one third the values on tables) with respect to that of the regulatory genes E_1 and E_2.

This experiment has the purpose to study a case with slower fluctuations dynamics of the early promoter regulatory module together with a slower dynamics of the splicing regulation.

7.4 Late Promoter

The early promoter *in silico* experiments can be useful to describe different scenarios upon differentiation: normal infection with the differentiation commitment; precancerous stages CIN-1,-2,-3 with a progressive strong delay in the activation of the late promoter.

We won't show an experiment for each of the previous cases since, as far as we know the only difference in the dynamics are different delays in the promoter activity and among the transcripts but qualitatively the behaviors should be the same.

7.4.0.6 Experiment 6

In this experiment we want to show the capability of the late promoter model to reproduce the qualitative behavior we expect from the literature. We consider an *in silico* experiment of 16 days of duration and a quite strong delay to slowly activate the late promoter.

The fixed parameters are reported in the previous tables.

8
Results: in silico experiments

In this chapter we will show the in silico experiments designed in the previous chapter. The predictions are perfomed in MATLAB by fixing the parameters indicated in chapter 7 for all the experiments, unless differently specified in the figures caption or in the text.

The stochastic simulations were predicted by using the Gillespie's stochastic algorithm (Gillespie (2002), Gillespie (2007), Gillespie (2013)).

8.1 Early promoter fit

In this section we show the predicted early promoter activity, modeled in terms of the early promoter markov chain reported in chapter 4, can fit very well the experimental data for both HPV16 and HPV18, actually by keeping the early promoter parameters very close to what we could directly inferred from literature.

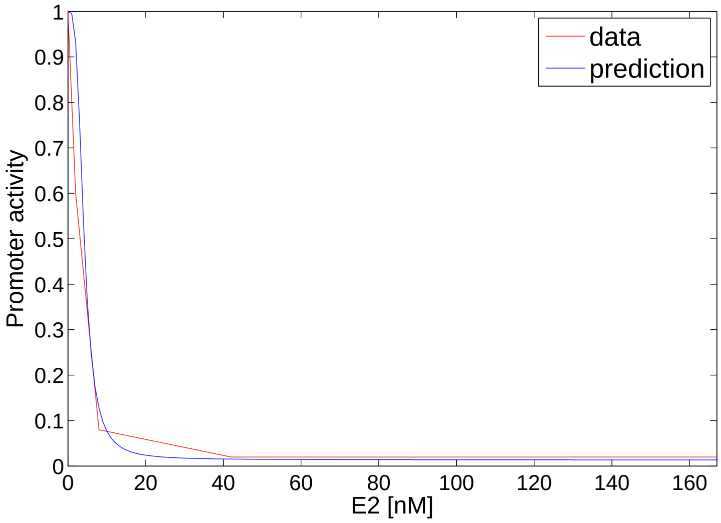

Figure 8.1.1: HPV16 early promoter activity fit.

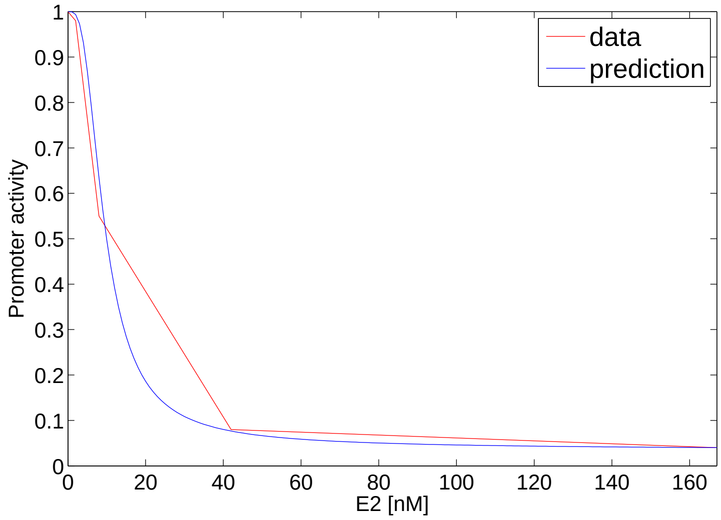

Figure 8.1.2: HPV18 early promoter activity fit.

8.2 Experiment 1

In this section we report the qualitative behavior of the early promoter repression predicted by the first heuristic model we have developed (Giaretta et al. (2015)).

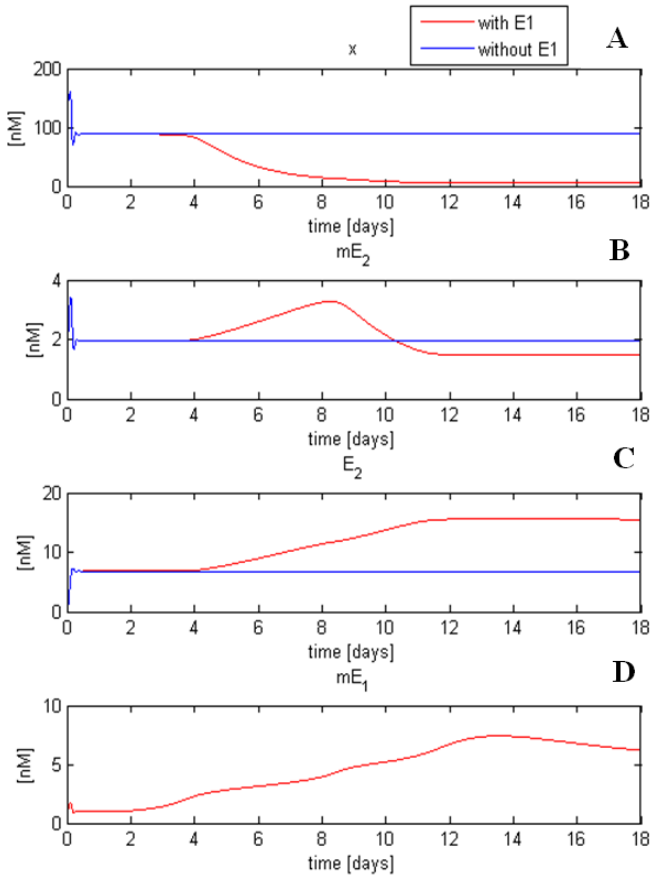

Figure 8.2.1: Model state variables with and without E_1 regulation during cellular differentiation. A. Early promoter pre-mRNA x. B. mE_2 transcript. C. E_2 protein. D. mE_1 transcript.

The dynamic behaviour of state variables with and without E1 co-regulation is shown in Fig. 8.2.1. In both cases the initial phase is very fast and all transcripts reach a steady state nearly 6 hours after infection. This is consistent with experimental evidences that have shown the presence of the early transcripts as early as 4-10 hours post infection.

We can see that both E_1 effect to enhance E_2 transcript and to knock down E_2 degradation are necessary for turning off the primary transcript during the early stage.

In this first modeling attempt we remember that we have condensed the late promoter effect in a forcing function explaining the high increase of E_1 mRNA.

8.3 Experiment 2

Here we show the experiment about the novel stochastic model of the early promoter by fixing the parameters in order to maintain a fast fluctuations regimen. In what follows we firstly show the deterministic predictions in terms of the QE.

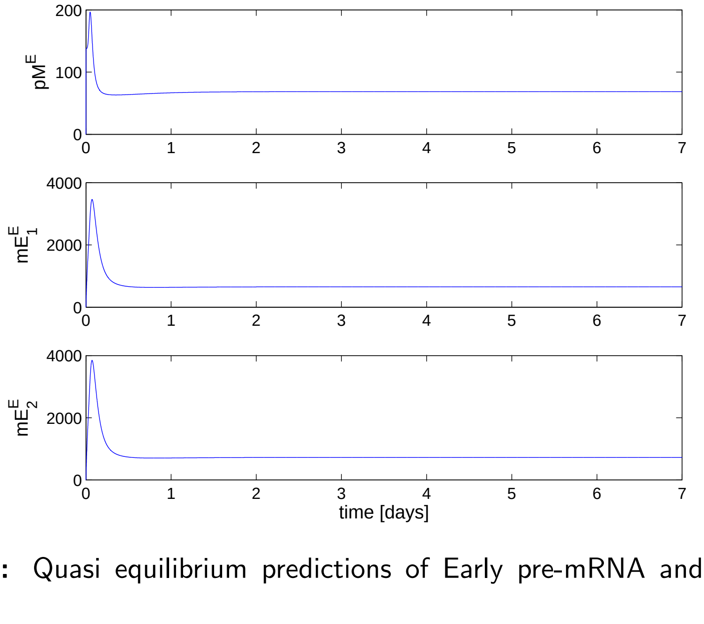

Figure 8.3.1: Quasi equilibrium predictions of Early pre-mRNA and E_1 and E_2 transcripts.

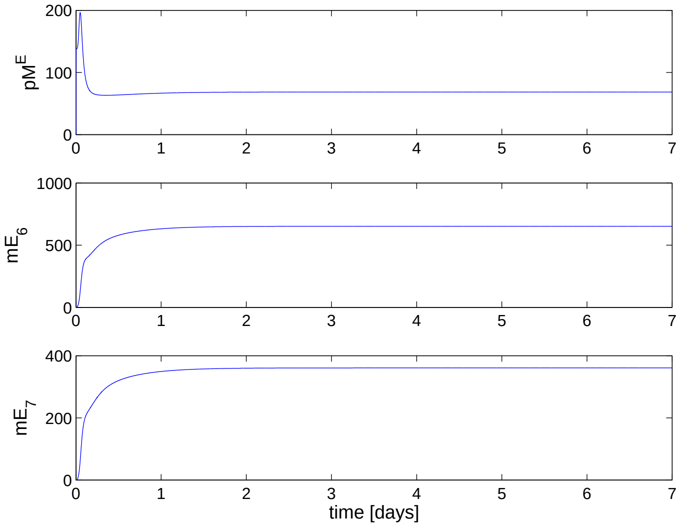

Figure 8.3.2: Quasi equilibrium predictions of Early pre-mRNA and E_6 and E_7 transcripts.

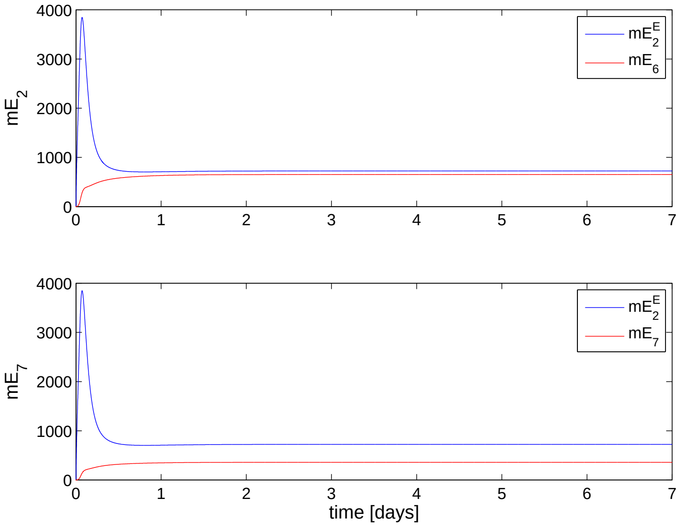

Figure 8.3.3: Deterministic comparison between E_2 transcripts and the oncogenes transcripts.

From this deterministic *in silico* experiment we can see an initial peak in the production of the pre-mRNA and of E_1 and E_2 transcripts. This is due to the transient response and the very small degradation of the pre-mRNA consistent with literature, that makes the transient very fast. The peak is also higher thanks to the presence of the positive feedback which

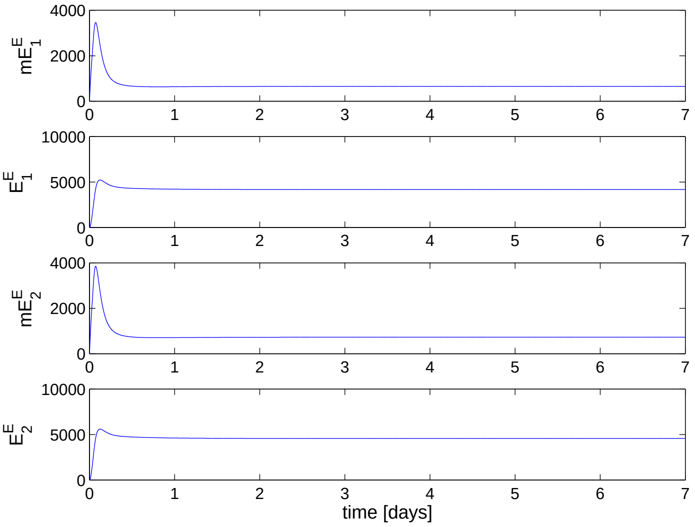

Figure 8.3.4: Quasi equilibrium predictions of E_1 and E_2 proteins and comparison with their transcripts.

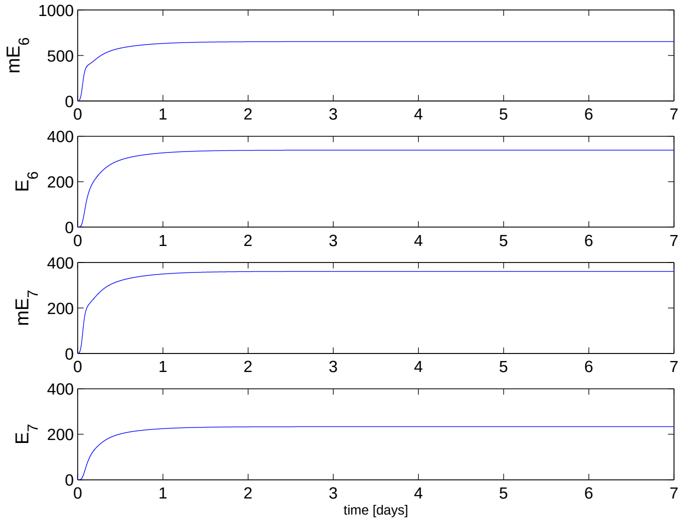

Figure 8.3.5: Quasi equilibrium predictions of E_6 and E_7 oncoproteins and comparison with their transcripts.

acts in low DE_2 and E_1E_2 concentrations, hence in particular at the beginning of the viral infection. The transcripts are very fast, as well. This is because they are a direct conversion of the very fast pre-mRNA and their degradations are of the order of few hours. This quick response is consistent with literature evidences (see chapter 2) that confirm a higher

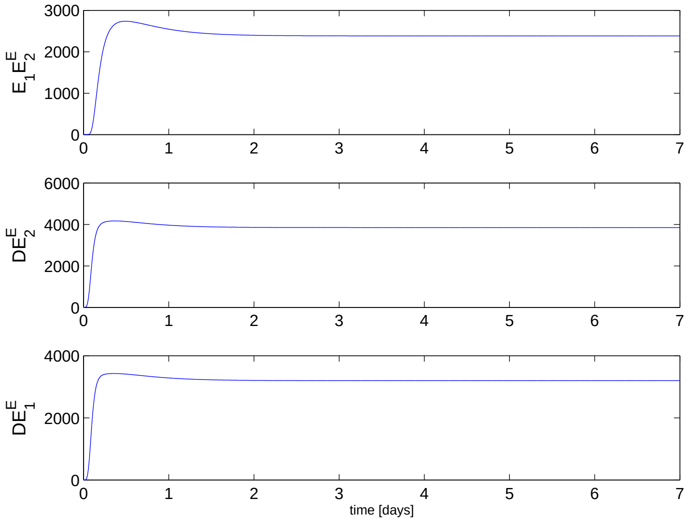

Figure 8.3.6: Quasi equilibrium predictions of DE_1 and DE_2 dimers and E_1E_2 tetramer.

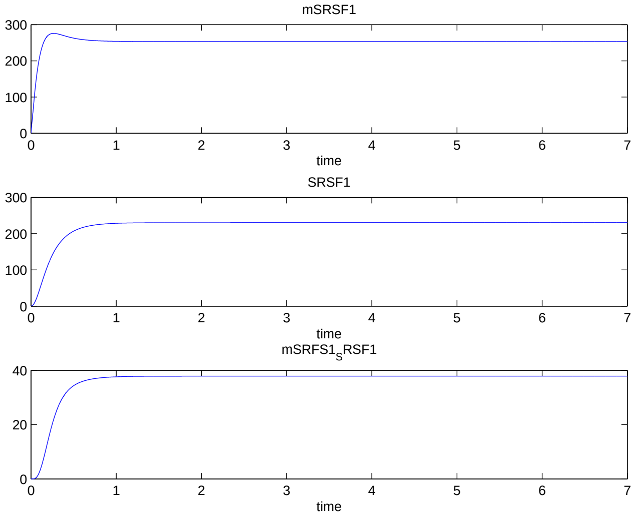

Figure 8.3.7: Quasi equilibrium predictions of $SRSF1$ transcript, protein and the heterodimer resulting from their association.

production of E_1 and E_2 just after the infection to guarantee a minimal replication level and consequent decreasing in their expression due to the negative feedback control. The oncogenes E_6 and E_7 are delayed with respect to E_1 and E_2 and they eventually win these latter production. This is consistent with literature evidences (see chapter 2). In our

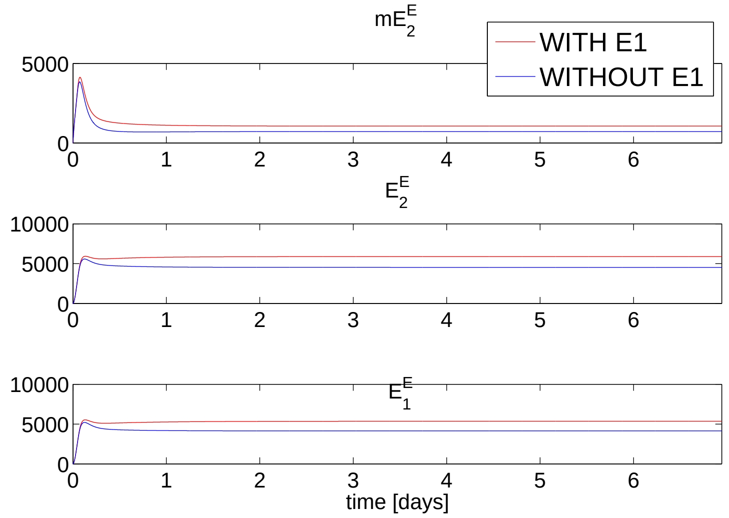

Figure 8.3.8: Quasi equilibrium predictions of E_2 transcript and protein and E_1 proteins to investigate the positive regulation of E_1 on E_2.

simulations, in the comparison between E_2 and the oncogenes transcript we can clearly see the delay in the production of the oncogenes that can be made even stronger by acting on the SS_1 splicing site (formally condesing splicing SA3358 and SA2709 sites), on E_1 enhancement or acting through SRSF1 dynamics. Actually from Fig. 8.3.3 it does not seem, looking at the steady states, the oncogenes win over E_1 and E_2. Nevertheless, as we will see later in the stochastic simulations it is not really true. It does not necessarily mean that if the deterministic steady state levels of the oncogens is lower than the steady state of E_1 and E_2, then the oncogenes do not win over these latter in the proceeding of the viral life cycle, as we'll see in what follows, when we'll show the stochastic predictions. The E_1 and E_2 proteins do not present the strong peak as their mRNAs. This is because there is a delay in producing proteins during the translation and especially because of their much lower degradation rates (five times lower in this case).

Then, we can observe SRSF1 is predicted in order to be around few hundreds copy number as inferred from literature. It can be modulated by the heterodimer $mSRSF1/SRSF1$ that induces a negative feedback on SRSF1 mRNA in terms of translational efficiency. In particular, SRSF1 takes around one day and half to completely reach its steady state, in our predictions. This is consistent with literature (it can have a slow dynamics up to 3 days due to its negative feedback regulation (as reported in chapter 2, SRSF1 section) and can be modulated by acting on $mSRSF1/SRSF1$ heterodimer dynamics.

Finally we can observe from fig. 8.3.8 that E_1 can positively enhance E_2 mRNA and proteins, and this is in agreement with the qualitative behavior the first heuristic model has predicted. The difference is that the positive regulation is implemented here at the splicing site control and the E_1-mediated enhancement happens immediately and not after some days as in the first experiment. This is because, in this latter we considered the E_1 positive regulation only modulated by the activation of the late promoter. Most likely it happens immediately. We also observe that, E_1 positively regulate itself as we have assumed in the new version consistently with literature (see chapter 2).

We won't show the repression of the early pre-mRNA mediated by the late promoter since the qualitative behavior would be the same as in experiment 1 and wouldn't add any new insights in the early promoter dynamics.

Before proceeding with the stochastic simulations we investigate the early promoter markov chain functioning with the aid of its steady state probability distribution, reported in Fig. 8.3.9.

In fig. 8.3.9, the steady state probabilities are dependent on DE_2 and E_1E_2 concentrations, being the regulators of the promoter. We have chosen to show the only dependence on DE_2, due to its main importance, and since in literature is usual to do that. Looking at the probabilities in function of DE_2 copy number we can see that the probability accounting for the basal transcription (state p_0^E) is initially the highest (we started with initial condition in the basal state, with probability 1) and then it converges fast to zero when DE_2 gets higher. Subsequently the positive feedback (state p_1^E) wins over the other and finally we can see that, for

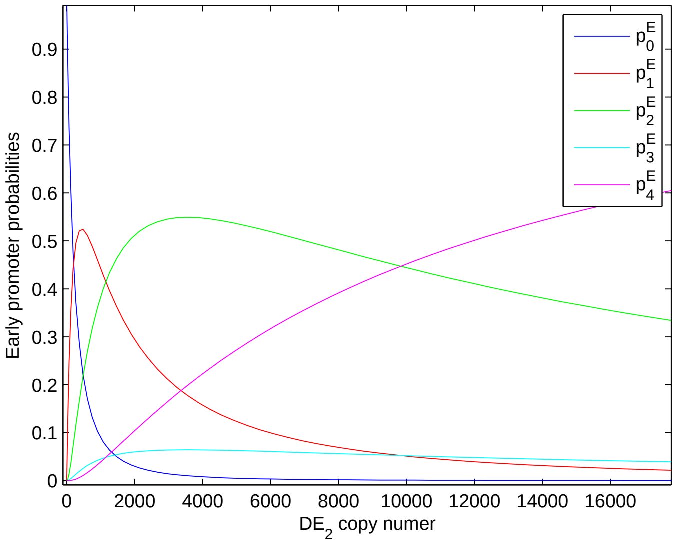

Figure 8.3.9: Steady state Early promotr markov chains probabilities in function of DE_2 dimer copy number.

higher concentrations, the negative feedbacks dominate the scene (states p_2^E, p_3^E and p_4^E). This also corresponds to a time dependence evolution in the feedback control change, since the dimers and tetramers take time to grow up to their steady state condition.

In what follows we present and discuss the stochastic simulations relative to the deterministic predictions we have discussed so far. For each stochastic variables we have also plotted its probability distribution to get insights in the state levels the variable can assume in its stochastic switching dynamics, due to the multiple levels control coming from the promoters and the splicing sites markov chains (that can change the synthesis and conversion rates of pre-mRNA and transcripts conversions, respectively).

These stochastic predictions were performed in a quite fast promoters (early and SRSF1 promoters) and splicing sites (SS_1 and SS_2 splicing sites) fluctuations regimen. In fact, we can see (from all the figures we have plotted) the quasi-equilibrium approximation follows quite well the average of the stochastic processes. In the limit of speed fluctuations tend to the infinity we would exactly get the quasi-equilibrium. Anyway, even in a fast fluctuations regimen we can understand the unpredictable behavior, due to the intrinsic noise, and clearly the deterministic approach cannot afford such a complex pattern predictions. Just to start, look at the pre-mRNA dynamics in Fig. 8.3.10. We clearly see the very fast fluctuations of the pre-mRNA due to its very fast conversion rate, into its component transcripts. The quasi-equilibrium cannot follow such fast variations and just predicts a weighted average among the different synthesis values of the pre-mRNA, with respect to the permanence probabilities in each of early promoter states. We can also see that, by looking at the multi-modal pre-mRNA distribution (it seems bimodal or maybe trimodal from the simulations; but we can't see very well the other peaks due to fast fluctuations).

As we anticipated before, observing the stochastic amplitude and qualitative frequency in Fig. 8.3.11 in comparison with Fig. 8.3.10 that the oncogenes win over E_1 and E_2 production (thanks to SRSF1 control on SS_2 splicing site). It is true that this stronger bursts are also due to the higher oncogenes degradations, but anyway the probability to get their expression is stronger than to get E_1 and E_2 expressions as we can see from the SS_2 splicing site steady state probabilities depicted in Fig 8.3.18.

In fact, in this figure we can see that the higher SRSF1 copy number (growing during its transient), the higher the probability (of state SS_1^1) to transcribe the oncogenes. This is anyway qualitatively captured with the only stochastic simulations, even if it would be clearer in a slower fluctuations regimen.

About the other state variables we can see the stochastic fluctuations are not so strong, and they follow quite well the quasi-equilibrium, as we can also depict from their monomodal probability distribution (excluding the peak, in the distribution, due to the transient response of the state variables), centered in the quasi-equilibrium steady state. The only variable showing a hint of bimodality is, in this particular in silico experiment

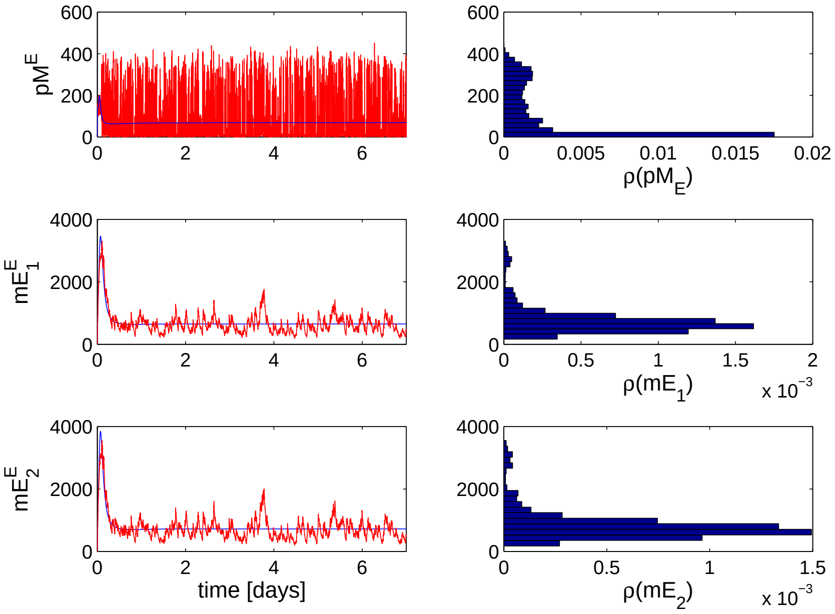

Figure 8.3.10: In the first coloumn: comparison between deterministic and stochastic predictions of Early pre-mRNA and E_1 and E_2 transcripts. In the second coloumn: probability distributions of the state variables

realization, the E_7 protein as shown in Fig. 8.3.15.

In the next sections we will show some in silico predictions with slower fluctuations of the promoters and the splicing sites. In this case the quasi-equilibrium fails in following the stochastic trajectories as exemplified in Fig. 8.3.19,8.3.20. In fact, in the first we can see the quasi-equilibrium predict a constant steady state for both E_2 protein and transcript, while the stochastic predictions show a kind of excitable "bursty" behavior for the occurrence of E_2 mRNA and huge variations of E_2 proteins over predicted by the quasi-equilibrium. In the second figure we can see large variations of the dimers and tetramer again overpredicted by a constant

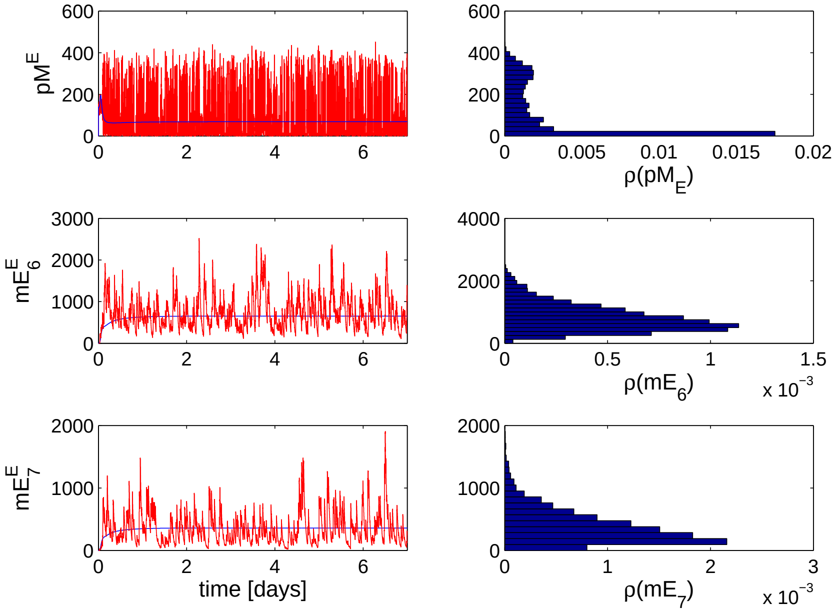

Figure 8.3.11: In the first coloumn: comparison between deterministic and stochastic predictions of Early pre-mRNA and E_6 and E_7 transcripts. In the second coloumn: probability distributions of the state variables

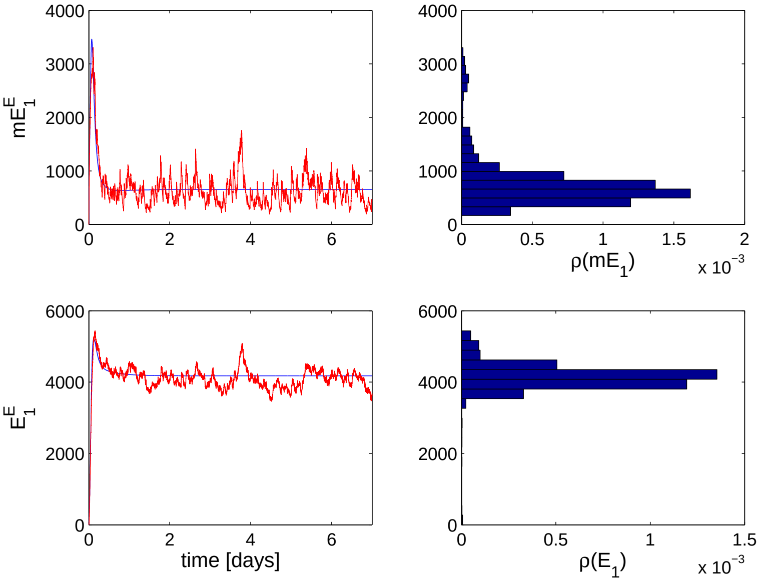

Figure 8.3.12: In the first coloumn: comparison between deterministic and stochastic predictions of E_1 protein and comparison with its transcript.In the second coloumn: probability distributions of the state variable.

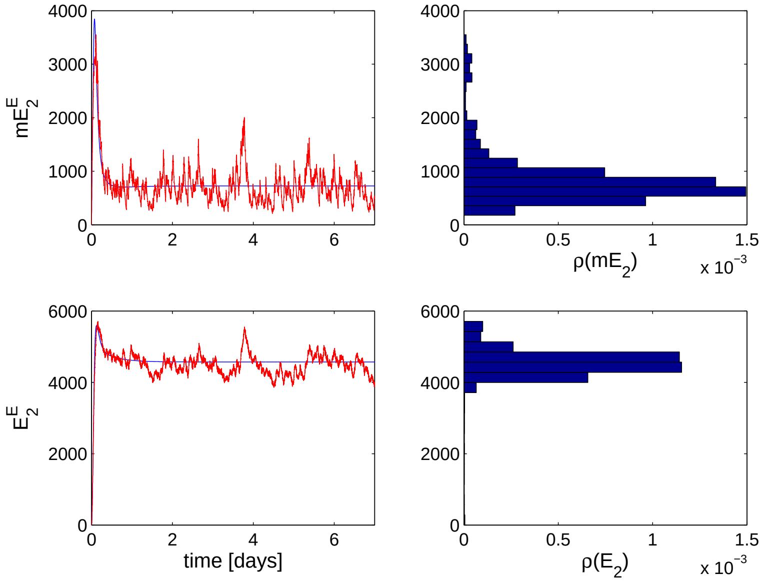

Figure 8.3.13: In the first coloumn: comparison between deterministic and stochastic predictions of E_2 protein and comparison with its transcript.In the second coloumn: probability distributions of the state variable

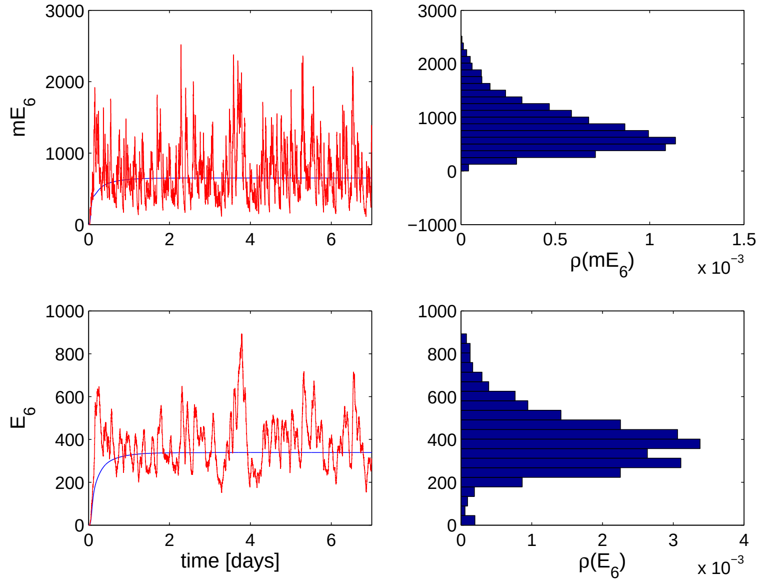

Figure 8.3.14: In the first coloumn: comparison between deterministic and stochastic predictions of E_6 oncoprotein and comparison with its transcript.In the second coloumn: probability distributions of the state variable.

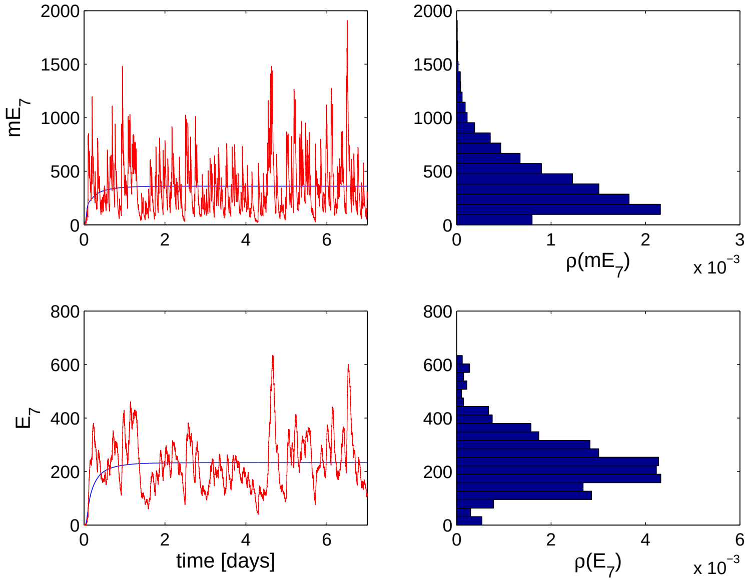

Figure 8.3.15: In the first coloumn: comparison between deterministic and stochastic predictions of E_7 oncoprotein and comparison with its transcript.In the second coloumn: probability distributions of the state variable.

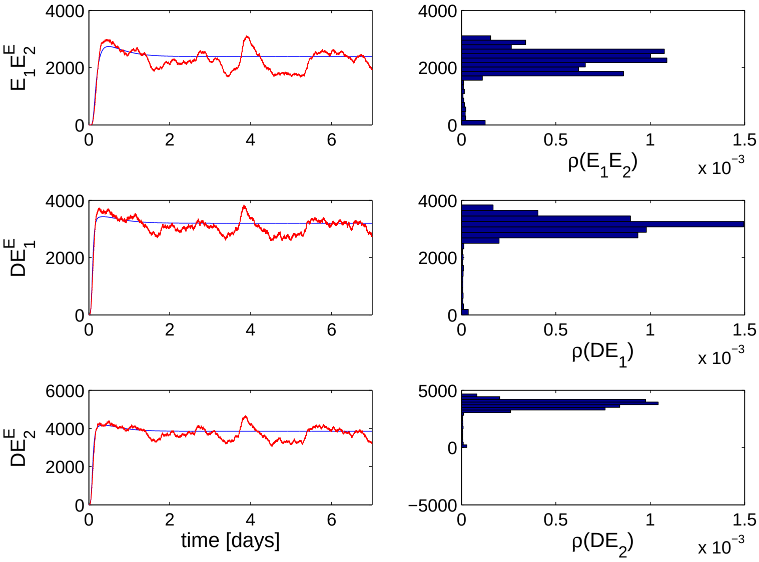

Figure 8.3.16: In the first coloumn: comparison between deterministic and stochastic predictions of DE_1 and DE_2 dimers and E_1E_2 tetramer.In the second coloumn: probability distributions of the state variables.

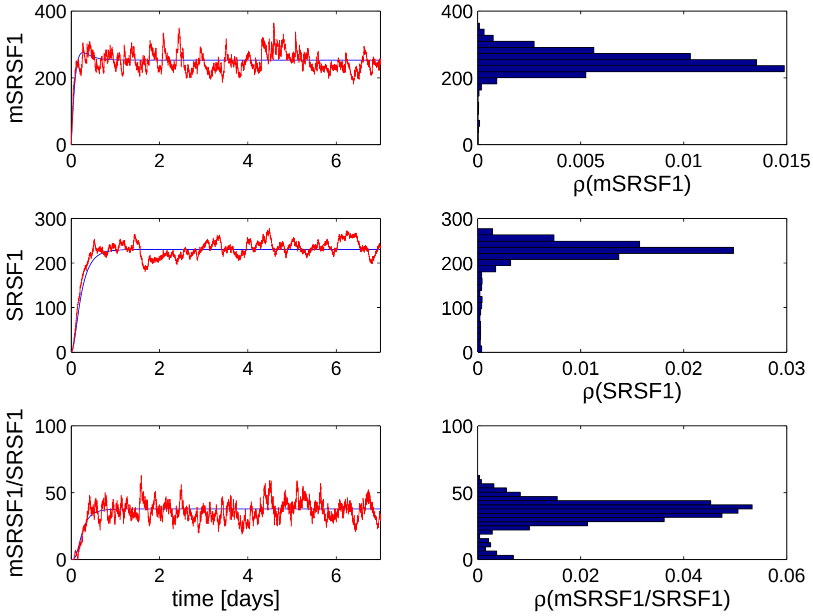

Figure 8.3.17: In the first coloumn: comparison between deterministic and stochastic predictions of $SRSF1$ transcritp, protein and the heterodimer resulting from their association. In the second coloumn: probability distributions of the state variables.

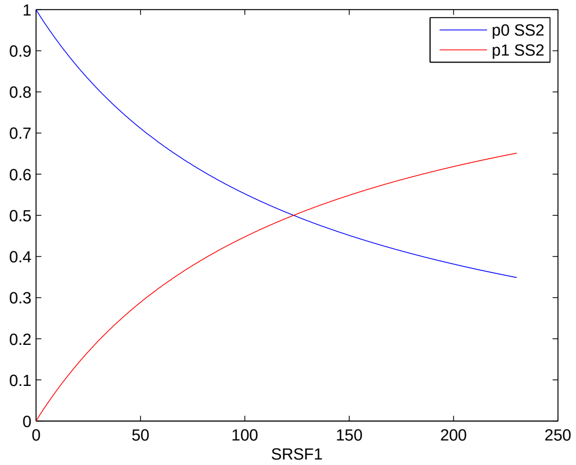

Figure 8.3.18: Probabilities of the SS_1 splicing site (formally condensing SA3358 and SA2709 splicing sites) in function of SRSF1 copy number. The probability for SS_1 to be occupied wins when SRSF1 is in higher concentration, thus driving the oncogenes , eventually diminishing E_2 and E_1 transcription.

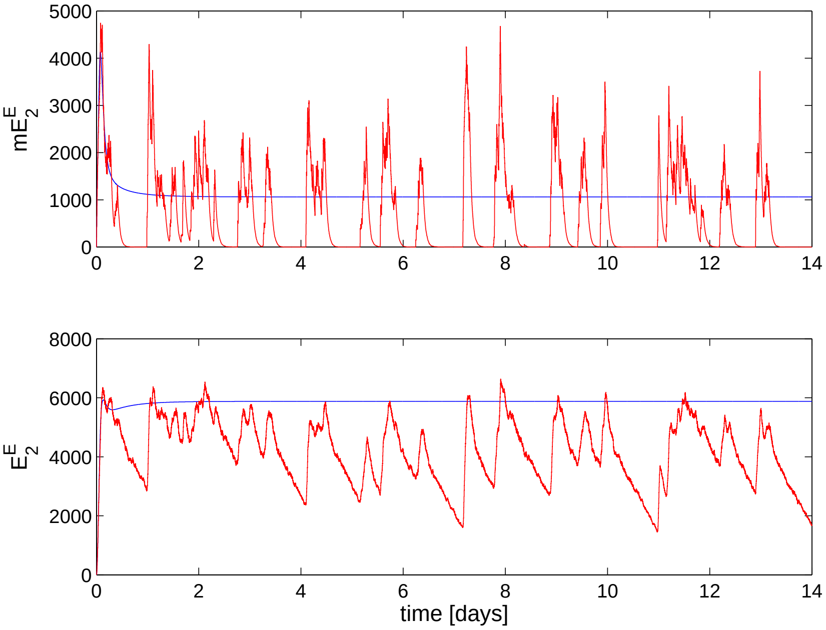

Figure 8.3.19: Comparison between quasi-equilibrium and stochastic predictions of E_2 protein and transcripts in a slow splicing fluctuations regimen.

quasi-equilibrium steady-state.

We could have a lot of interesting stochastic behaviors due to slow fluctuations. In the next section we will show some interesting behaviors that could be associated to interesting biological explanation. In what follows we will show the only stochastic predictions since we already know the quasi-equilibrium will fail.

8.4 Experiment 3

In this experiment we can interestingly observe that E_1 and E_2 transcripts are produced in a sort of excitable manner, showing sudden strong bursts. The same happens for the oncogenes in a mutually exclusive manner (i.e., when E_1 and E_2 are produced the oncogens are not). Moreover, the dimers and tretamer dynamics shows strong stochastic variation be-

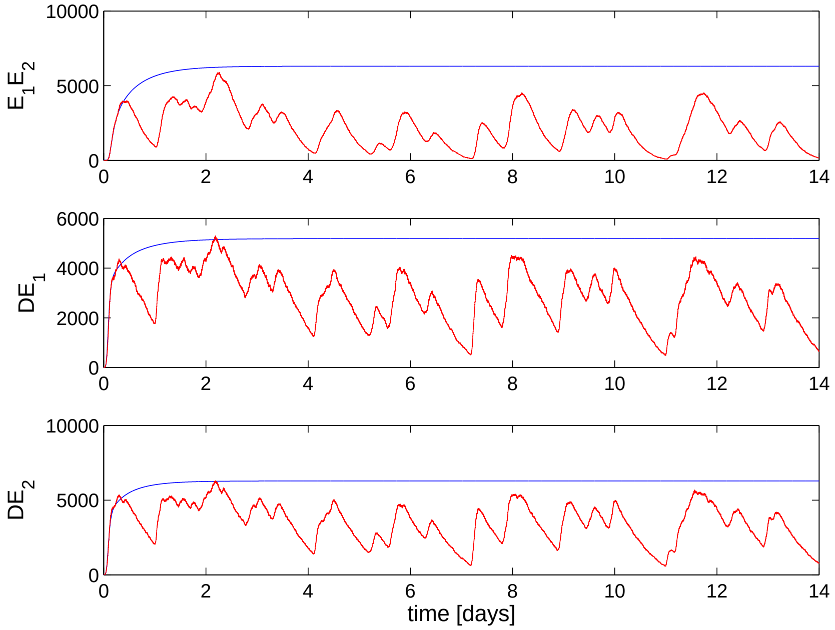

Figure 8.3.20: Comparison between quasi-equilibrium and stochastic predictions of DE_1, DE_2 dimers and E_1E_2 tetramer in a slow splicing fluctuations regimen.

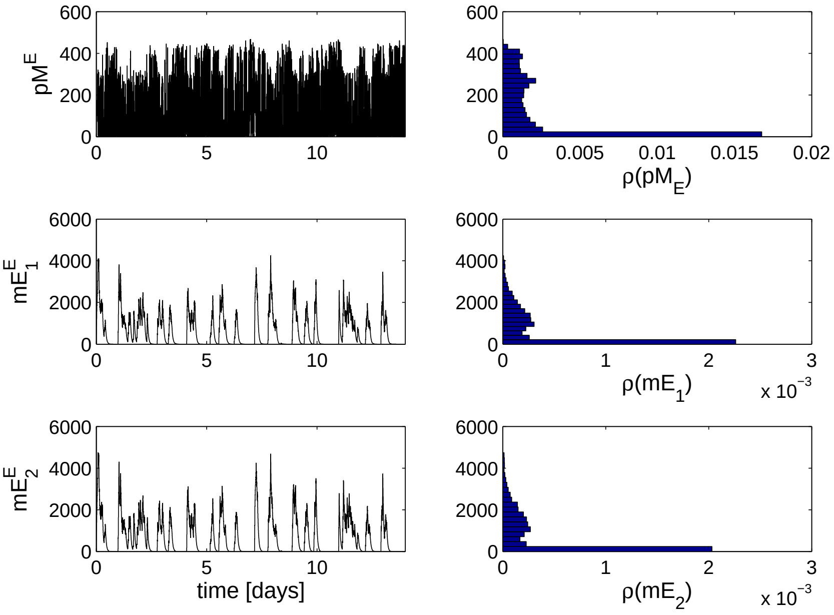

Figure 8.4.1: Comparison between quasi-equilibrium and stochastic predictions of E_2 protein and transcripts in a slow splicing fluctuations regimen.

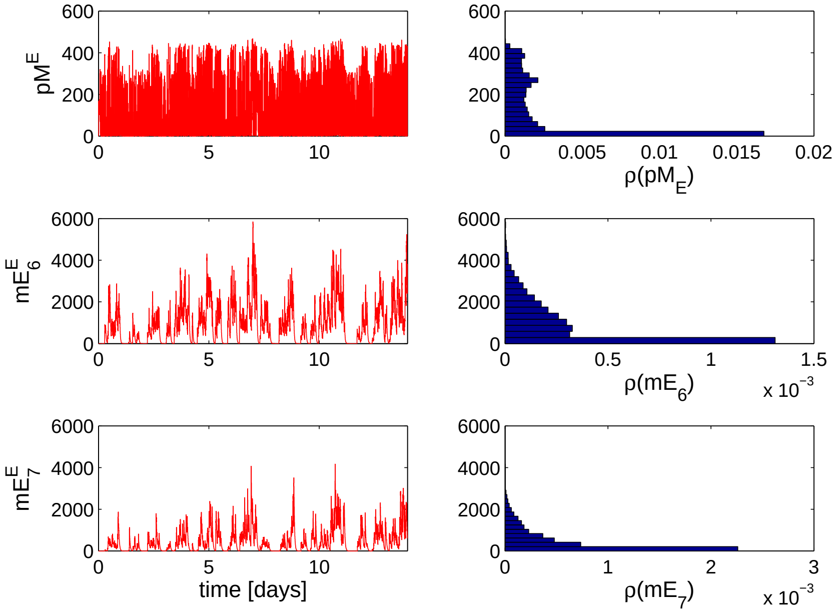

Figure 8.4.2: Comparison between quasi-equilibrium and stochastic predictions of DE_1, DE_2 dimers and E_1E_2 tetramer in a slow splicing fluctuations regimen.

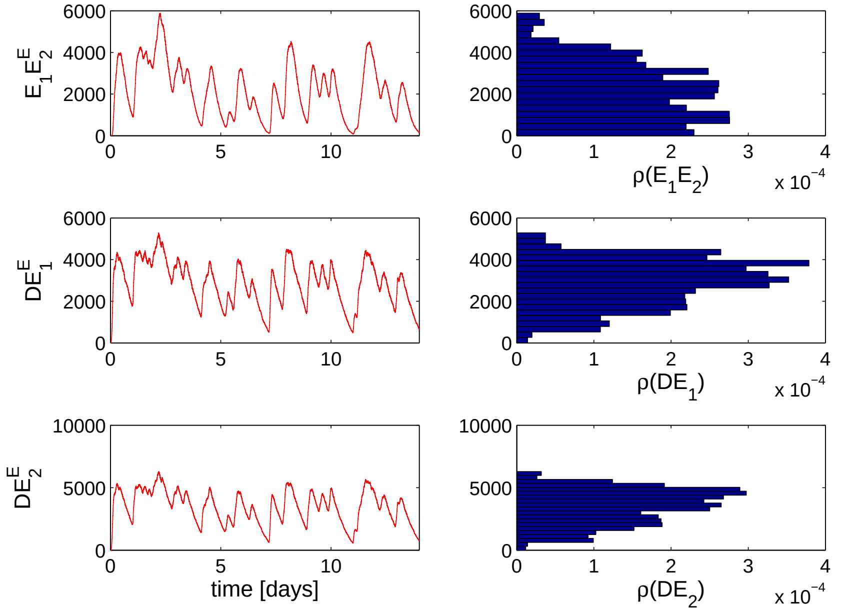

Figure 8.4.3: Comparison between quasi-equilibrium and stochastic predictions of DE_1, DE_2 dimers and E_1E_2 tetramer in a slow splicing fluctuations regimen.

cause of the sudden ″bursty″ dynamics of E_1 and E_2. What is interesting to observe is that, with the fixed parameters, looking at the steady state probabilities (on which the quasi-equilibrium is based) of the SS_1 splicing site we would notice that the oncogenes should be generated with a very low probability with respect to E_1 and E_2. However, the steady state probabilities of the control markov chains are not reliable in a slow fluctuations regimen. In fact, as we can see from the stochastic predictions in figures 8.4.1 and 8.4.2 the oncogenes seem to have more or less the same probabilities of E_1 and E_2 to be produced, even a bit more. That is why the quasi-equilibrium cannot predict this situation.

We can also see that the oncogenes are produced in very strong bursts, even stronger than the regulatory transcripts (E_1 and E_2). This is due in part to their higher degradation rates than the regulatory transcripts. The other reason has to be found in the regulatory transcripts (E_1, E_2) bursts. In fact they induce, as argued before, strong variation of the dimers and tetramer regulating the early promoter. In particular, these latter are strongly decreased when there is no E_1 and E_2 production, resulting in a weaker negative feedback on the early promoter and a consequent strong burst of the oncogenes.

In a biological context this pulsatile ″bursty″ expression of the regulatory transcripts and oncogenes, if it was experimentally verified in future, could be a very interesting mechanism the virus could usein order to efficiently control the eukaryotic cell DNA replication machinery by E_1 and E_2, in order to properly replicate the viral DNA; or, alternatively, to efficiently control the restart, mediated by the oncogenes, of the cell cycle after the infection. In fact, it was recently proven that a frequency pulsatile control could efficiently modulate gene expression (Cai et al. (2008), Dalal et al. (2014)).

8.5 EXPERIMENT 4

In this experiment, by placing to zero the parameter $\tilde{k}_{off}^{SS_1}$, we create the possibility to have an "absorbing" state (also called a "limbo" state) in the SS_1 two state markov chain. If E_1 stochastically reaches the zero copy number, there is no way to produce neither E_1 nor E_2 anymore and the only oncogenes will be produced. Moreover, since the regulatory proteins cannot be produced anymore, their dimers and tetramer will disappear by degrading in time. This will produce the effect to create a reverse action on the early promoter transactivation feedback. The negative feedback will became a positive feedback till to reach a basal transcription that will last forever. This will, in turn, over express the oncogenes (the basal transcription is higher than the negative feedback transcription).

Anyway, all this is reached thanks to a proper balance between an enough strong negative feedback and intrinsic noise strength to make this complex dynamical regulation possible. Hence, this is one of the complex pattern we look for when we consider the entire behavior of a system composed by interconnected sub-modules. This is in line with the actual holistic paradigm of systems biology.

This is a sort of bistable property for the system. The interesting fact is that it is not due to the bistability of the early promoter but to the downstream splicing dynamics. This qualitative behavior the system can exhibit can be very interesting biologically speaking. It could explain, in an autoregulatory manner, the HPV integration event in the human DNA. As reported in chapter 2, some HPVs, after infection, can integrate into the human DNA with the consequent over expression of the oncogenes and no more production of E_1 and E_2. Actually this is performed thanks to the lost of splicing mechanisms and not to a change in the early promoter structure. Finally, we observe that to dynamically exhibit this behavior we needed to eliminate $\tilde{k}_{off}^{SS_1}$, resulting in a change of the splicing model we had performed. This is acceptable since on the one hand the virus upon integration loses part of its genome, resulting in a different regulatory network for the splicing control; on the other hand, because our model is condensing more than one splicing mechanisms all together.

From Fig. 8.5.1 the pre-mRNA is initially, after its transient response, subjected to the dominant negative feedback, but subsequently the early promoter converges indefinitely to the basal transcription. This is because, as argued before, the E_1 protein can stochastically reach a zero copy number. When this happens no more E_1 and E_2 are produced as their dimers, as we can see in the other figures 8.5.3, 8.5.4, 8.5.7. The only genes that will be expressed (actually over expressed) are the oncogenes as we can see in the Fig.8.5.2.

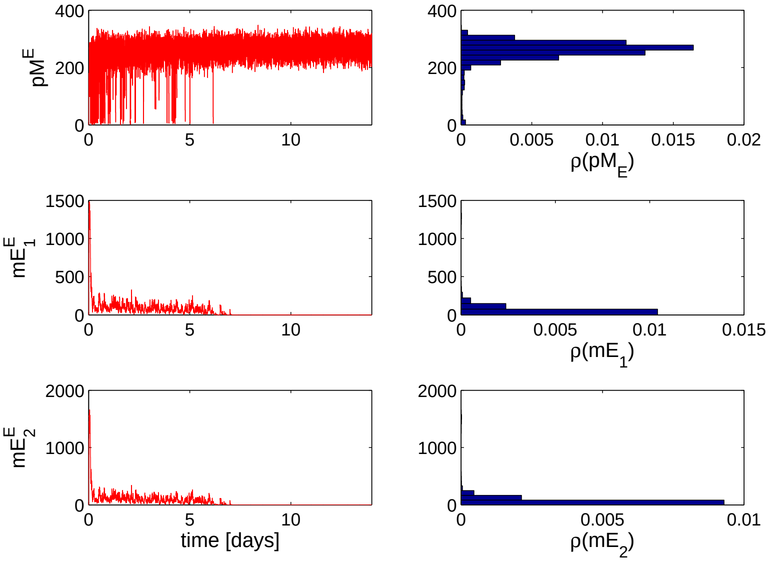

Figure 8.5.1: In the first coloumn: stochastic predictions of Early pre-mRNA, E_1 and E_2 transcripts. In the second coloumn: probability distributions of the state variables

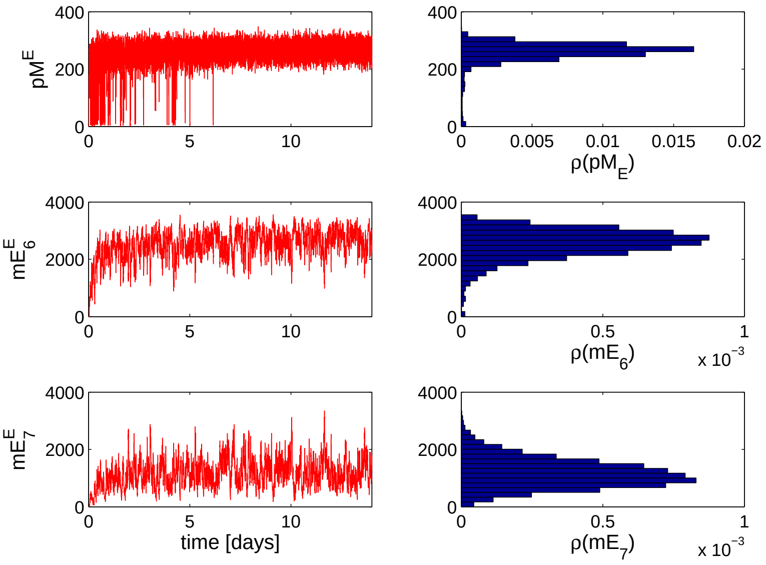

Figure 8.5.2: In the first coloumn: stochastic predictions of Early pre-mRNA, E_6 and E_7 transcripts. In the second coloumn: probability distributions of the state variables

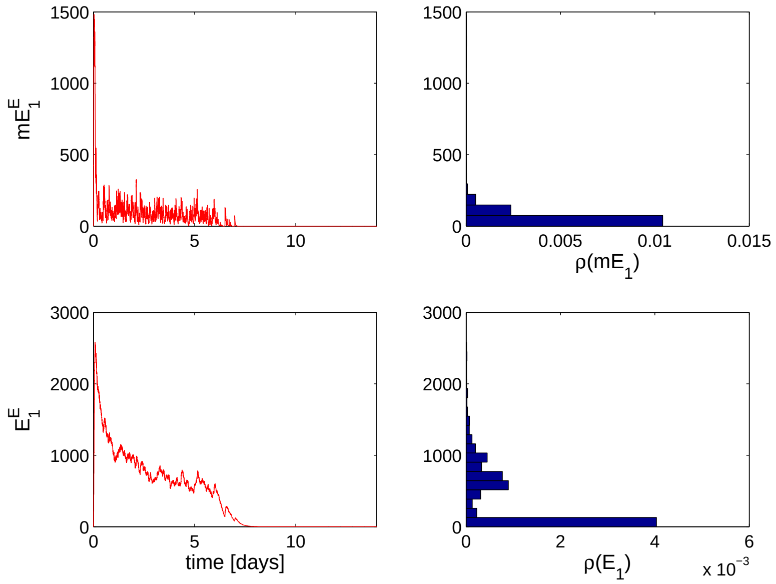

Figure 8.5.3: In the first coloumn: comparison between predictions of E_1 protein and its transcript. In the second coloumn: probability distributions of the state variables.

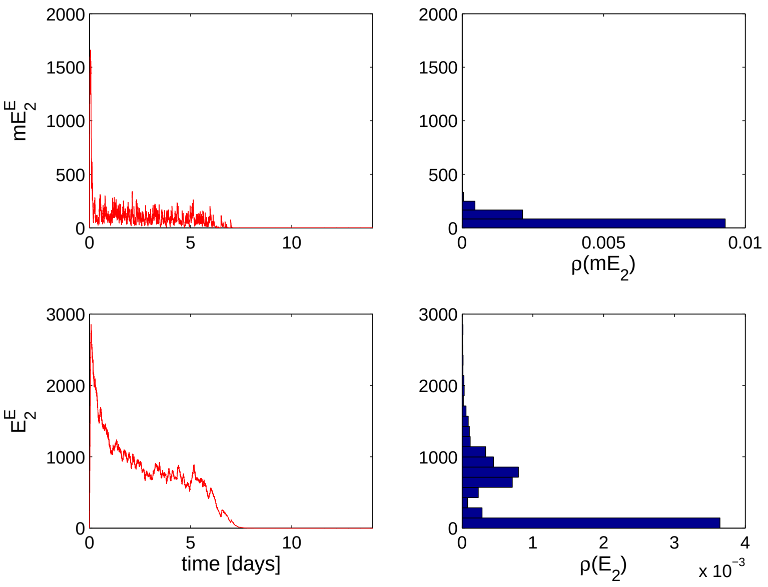

Figure 8.5.4: In the first coloumn: comparison between predictions of E_2 protein and its transcript. In the second coloumn: probability distributions of the state variables.

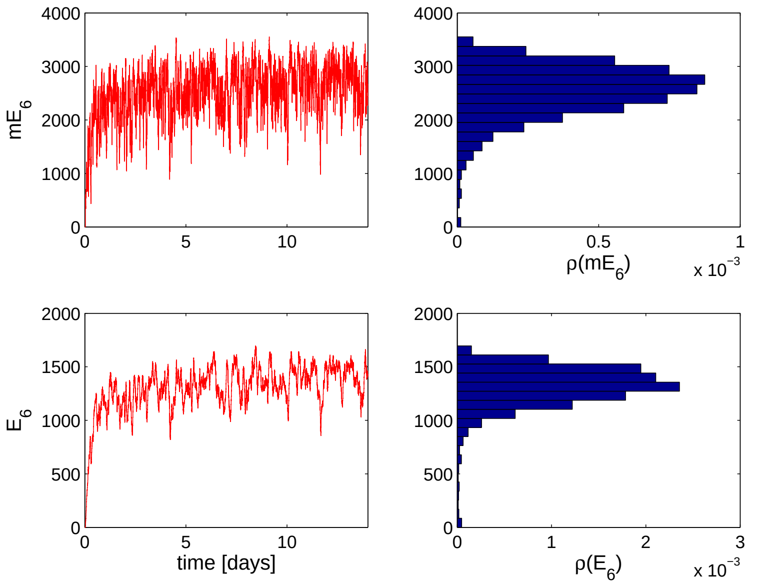

Figure 8.5.5: In the first coloumn: comparison between predictions of E_6 protein and its transcript. In the second coloumn: probability distributions of the state variables.

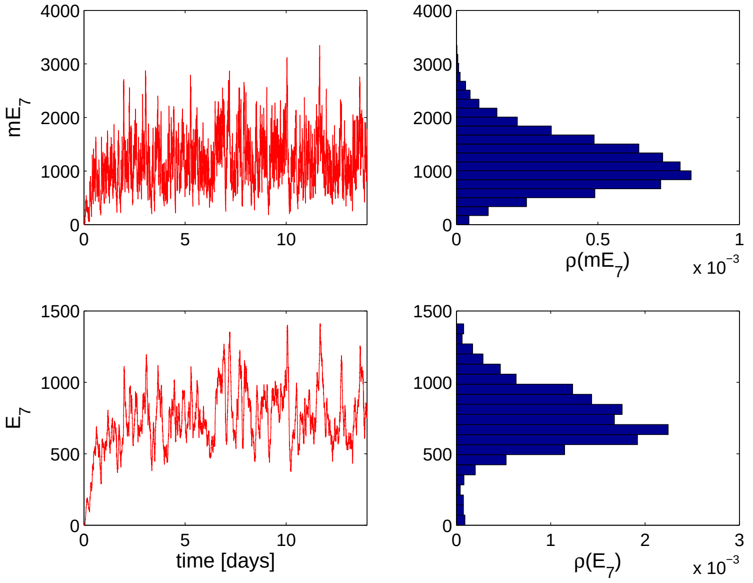

Figure 8.5.6: In the first coloumn: comparison between predictions of E_7 protein and its transcript. In the second coloumn: probability distributions of the state variables.

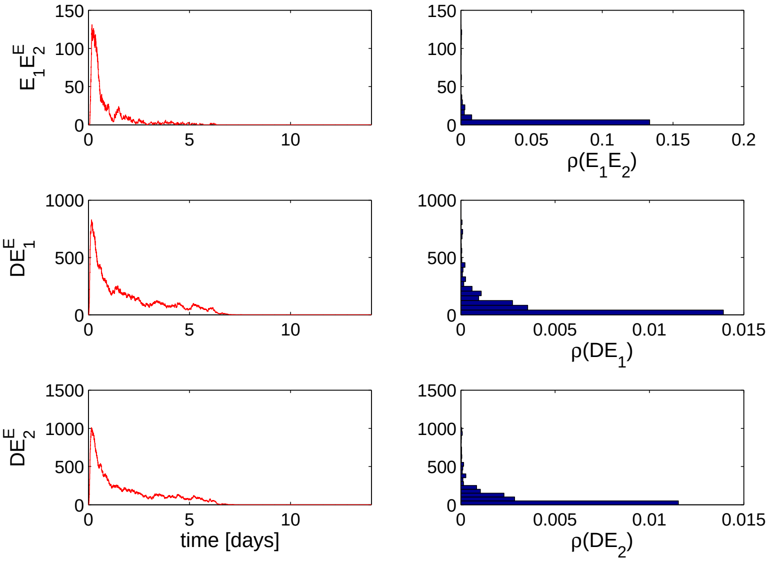

Figure 8.5.7: In the first coloumn: stochastic predictions of DE_1 and DE_2 dimers and E_1E_2 tetramer.In the second coloumn: probability distributions of the state variables.

8.6 Experiment 5

This experiment was performed in order to have an insight in the slower dynamics of both the early promoter, the SRSF1 promoter and the splicing sites.

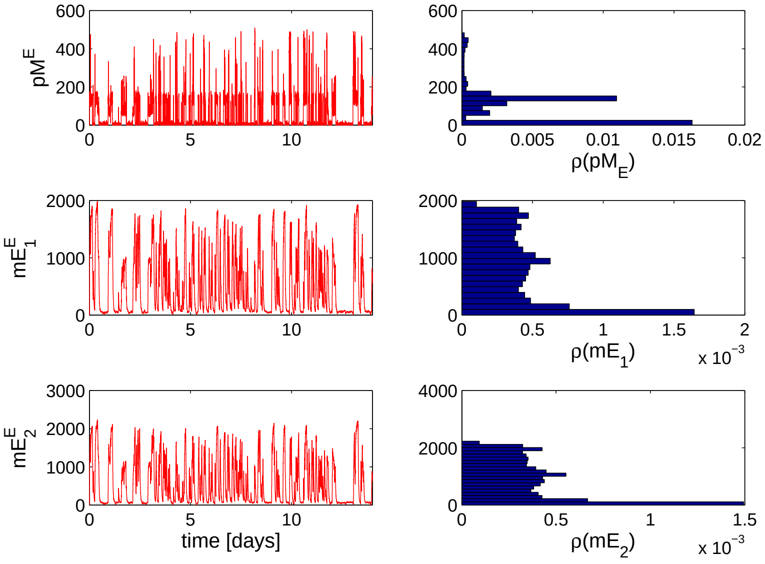

Figure 8.6.1: In the first coloumn: stochastic predictions of Early pre-mRNA, E_1 and E_2 transcripts. In the second coloumn: probability distributions of the state variables

We can see from Fig. 8.6.1 that the early promoter can exhibit lower frequent and strong bursts. With slower fluctuations it is also more demarcated its multimodal probability distribution. E_1 and E_2 transcripts exhibit strong bursts and we can see their protein and expecially the dimers and tetramer can exhibit stronger and more lasting busrts, as shown in Fig. 8.6.3, 8.6.4 and especially in Fig. 8.6.7.

About the oncogenes in this situation they are produced in low quantity and frequency as we can see from their isolated strong like-excitable bursts behavior. It is interesting to note that when the oncogenes occur

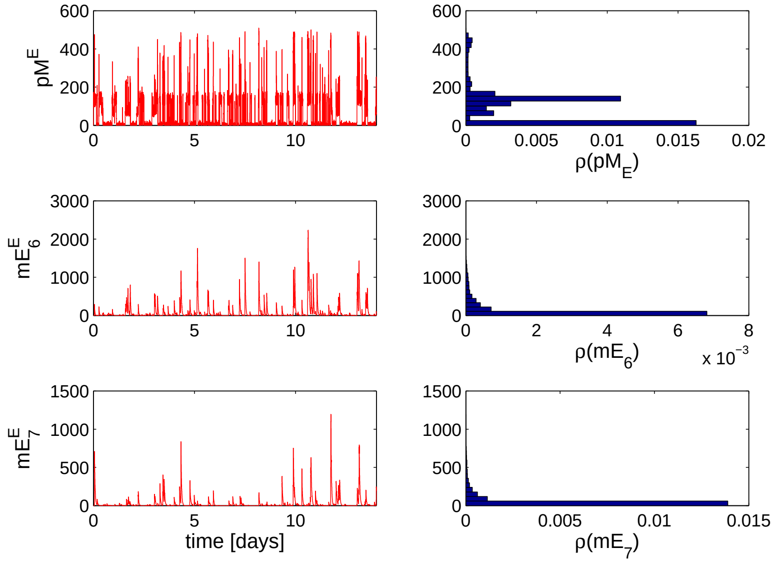

Figure 8.6.2: In the first coloumn: stochastic predictions of Early pre-mRNA, E_6 and E_7 transcripts. In the second coloumn: probability distributions of the state variables

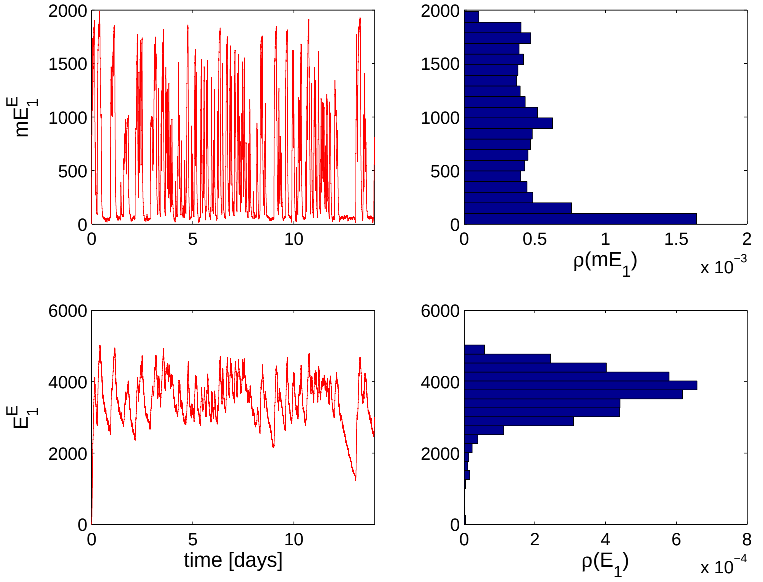

Figure 8.6.3: In the first coloumn: comparison between predictions of E_1 protein and its transcript. In the second coloumn: probability distributions of the state variables.

we can observe a stronger "bursty" spike in the pre-mRNA. In this case is more evident than in the other experiments, because we made the total

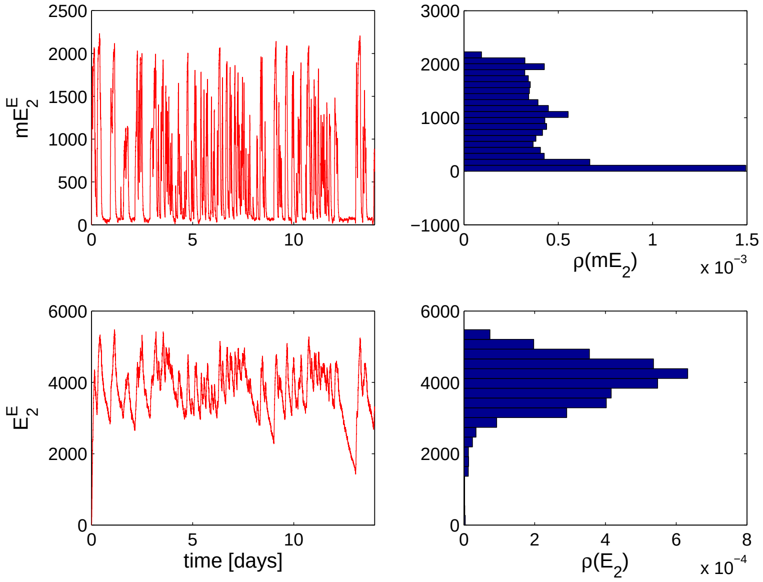

Figure 8.6.4: In the first coloumn: comparison between predictions of E_2 protein and its transcript. In the second coloumn: probability distributions of the state variables.

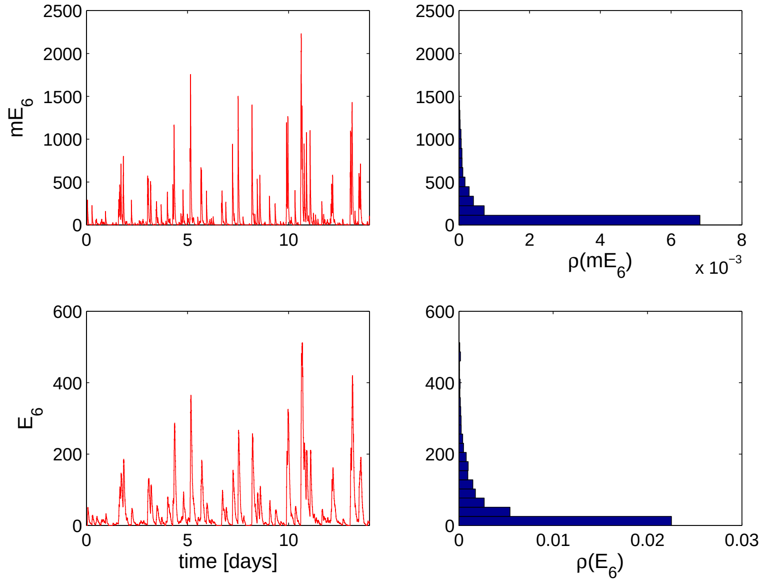

Figure 8.6.5: In the first coloumn: comparison between predictions of E_6 protein and its transcript. In the second coloumn: probability distributions of the state variables.

conversion rate of the oncogenes lower than the total conversion of the regulatory proteins (E_1 and E_2). In this manner, when the oncogenes are

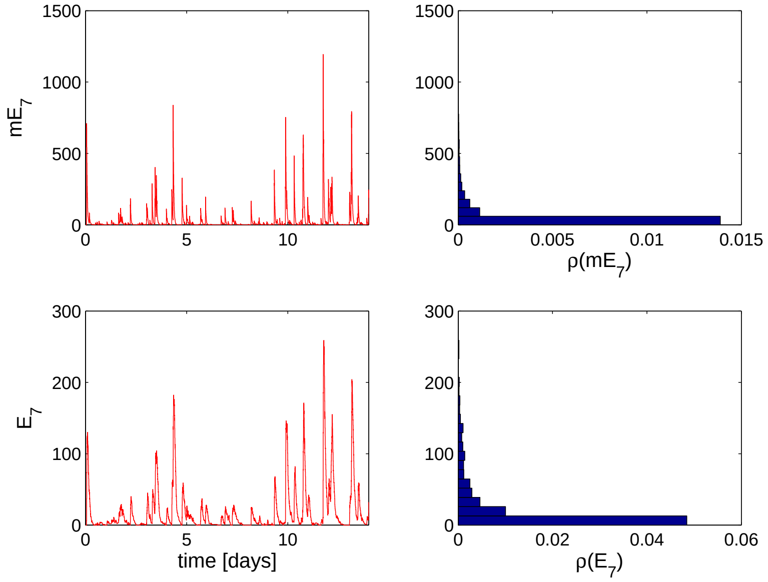

Figure 8.6.6: In the first coloumn: comparison between predictions of E_7 protein and its transcript. In the second coloumn: probability distributions of the state variables.

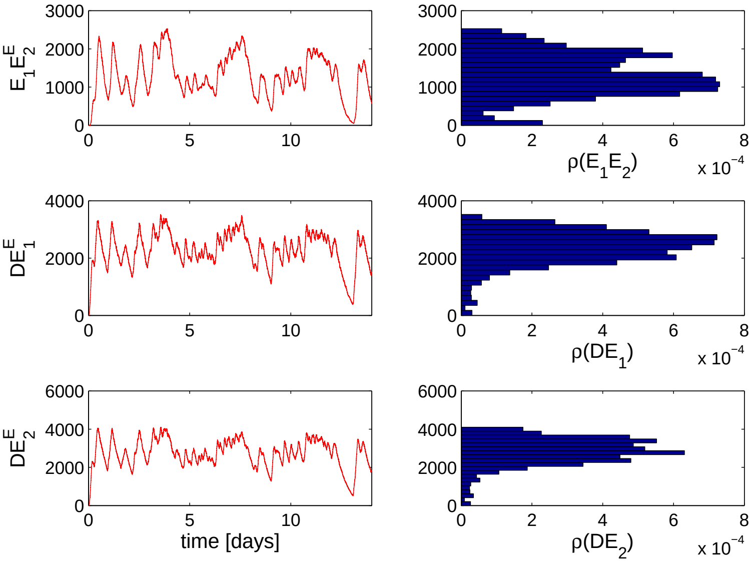

Figure 8.6.7: In the first coloumn: stochastic predictions of DE_1 and DE_2 dimers and E_1E_2 tetramer.In the second coloumn: probability distributions of the state variables.

produced the pre-mRNA has a lower "degradation" (i.e., sum of conversion rates) than in the case when the regulatory genes are transcribed, re-

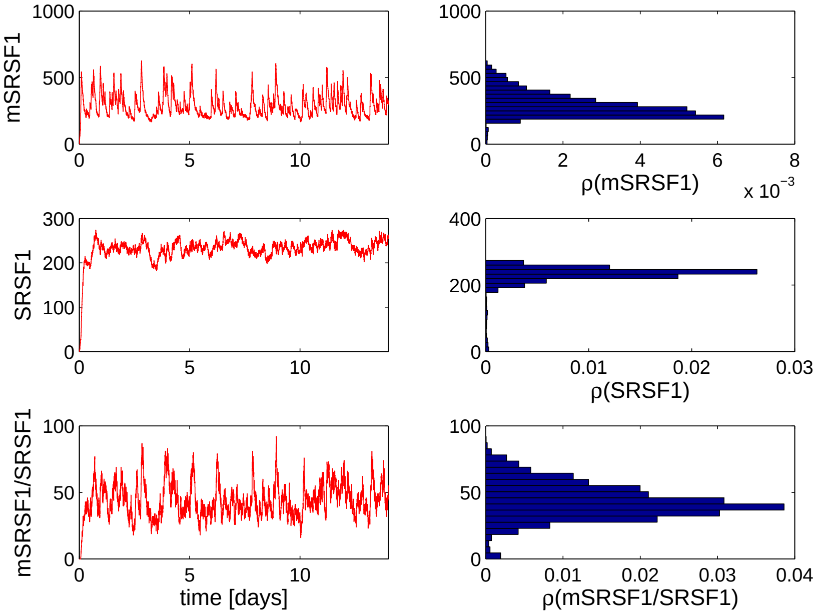

Figure 8.6.8: In the first coloumn: stochastic predictions of $SRSF1$ transcritp, protein and the heterodimer resulting from their association. In the second coloumn: probability distributions of the state variables.

sulting in a stronger burst response, given the very fast conversion rates. We can see in this case that the joint regulation of the early promoter and the splicing control layer add a sort of another state to the early promoter that the early promoter alone couldn't exhibit, from its designed markov chain alone.

To find a possible interesting biological context of this dynamical behvior, we could think about the basal infected cells, where the main purpose of the virus is to maintain a viral DNA reservoir through a major regulation of the only E_1 and E_2. Their pulsatile like excitable response could be interesting to effciently control in a pulsatile manner the cellular replication machinery.

8.7 Experiment 6

In this experiment we will investigate the late promoter functioning.

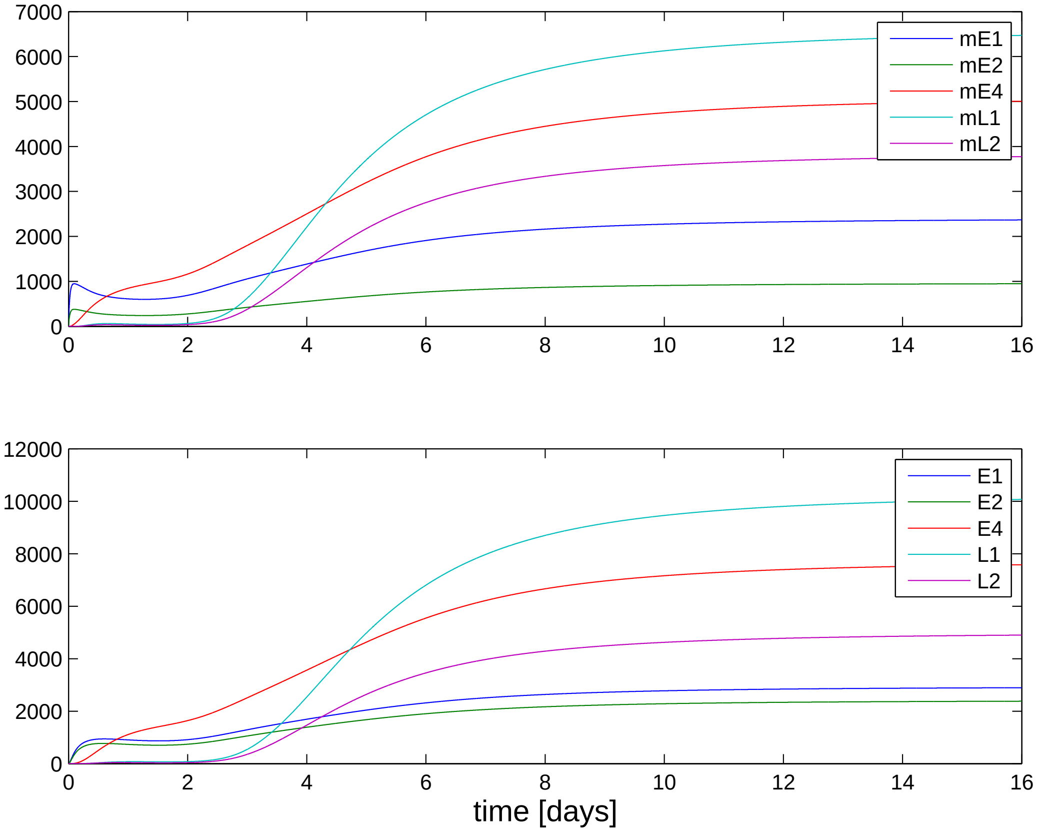

Figure 8.7.1: Deterministic quasi-equilibrium prediction of the HPV transcripts and proteins copy numbers produced by the late promoter.

As we can see from QE predictions in fig. 8.7.1 the replication proteins E_1^L and E_2^L are the first to be produced and this is consistent with the necessity to get an immediate and higher replication of the viral DNA during the differentiation. Moreover we can observe that for E_1 and E_2 mRNAs we can get the same qualitative pattern (they initially decrease and then strongly increase) in the dataset reported in the previous chapter. E_4^L is produced just after E_1^L and E_2^L since it is dependent on higher values of SRSF1 splicing factor. Since this latter is activated by E_2 the system needs more time to grow up E_2^L to enough level in order to produce E_4^L; moreover we have to remember that E_2^L and E_4^L are mutually exclusive and at the beginning the production of the replication proteins has the priority.

L_1 and L_2 transcripts and proteins are produced with a stronger delay with respect to the early genes. This happens when E_2 reaches enough high concentrations to inhibit the early polyadenylation site, as we can see from fig. 8.7.2, paving the way to the late genes transcription. This is consistent with the biological knowledge. In fact the late genes, are produced in the upper epithelium strata. In fig. 8.7.3 we can see the pAE site is inhibited (state pAE_1) when E_2 gets higher concentration.

All these reciprocal delays among the viral transcripts and proteins are possible thanks to a complex regulation at post-transcriptional level in terms of splicing and polyadenylation regulations.

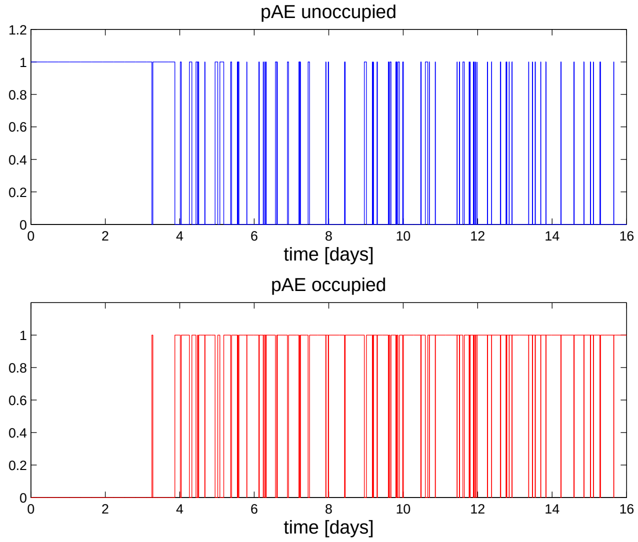

Figure 8.7.2: Time evolution of pAE site.

However, the slow modulation increase of all the transcripts is obtained at the late promoter level. The late promoter was designed to reproduce the modulation by $C/EBP\beta$ and to consider other possible intrinsic delays such as the nucleosome recruitment mediated by $C/EBP\beta$,

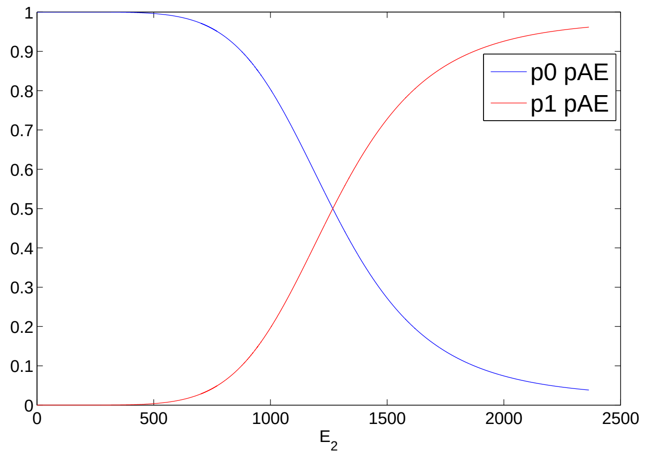

Figure 8.7.3: Probabilities distribution of polyadenylation site pAE in function of E_2 copy number.

as argued in chapter 2. To account for these delays we have assumed three different increasing levels of the late pre-mRNA production for each state of the late promoter markov chain. Moreover, the synthesis parameters and the markov chain parameters were tuned in order to make a quite good intrinsic stochastic delay, before reaching its full regimen functioning. We can see this dynamical features in the late promoter probability distribution, in function of $C/EBP\beta$ copy number, as shown in fig. 8.7.5. From the distributions we observe the basal transcription (state p_0^{LP}) starts at the beginning in low $C/EBP\beta$ concentration. The intermediate transcripton level occurs when $C/EBP\beta$ reaches medium concentrations and finally, the regimen condition accounting for a strong transcription has the highest probability to occur when $C/EBP\beta$ is in high concentration. The temporal transition between the states of the promoter markov chain will occur very slowly in time, as shown in fig. 8.7.5. This is because $C/EBP\beta$ grows very slowly because of its very high half life and because of its strong stabilization, at post-translational levels we are accounting with the insertion of a delay, as explained in Chapter 4.

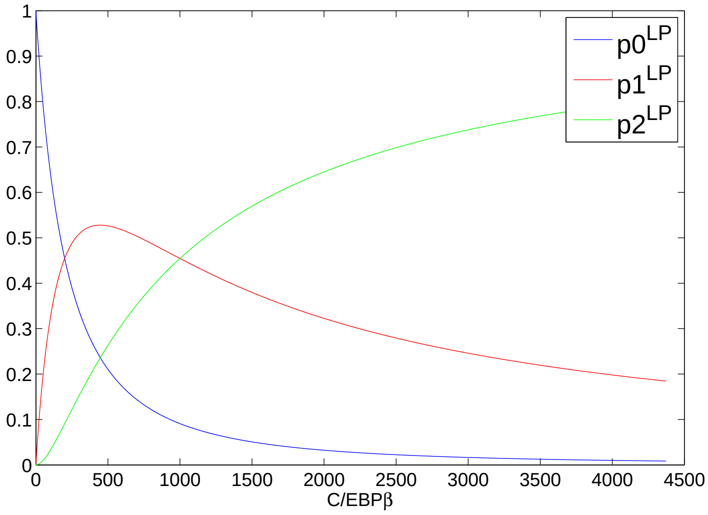

Figure 8.7.4: Late promoter markov chain probability distribution in function of $C/EBP\beta$ copy number concentration. p_0^{LP} is the probability of basal transcription, p_1^{LP} when $C/EBP\beta$ has medium concentration and the transcription is not so strong yet. p_2^{LP} is the probability accounting late promoter regimen functioning.

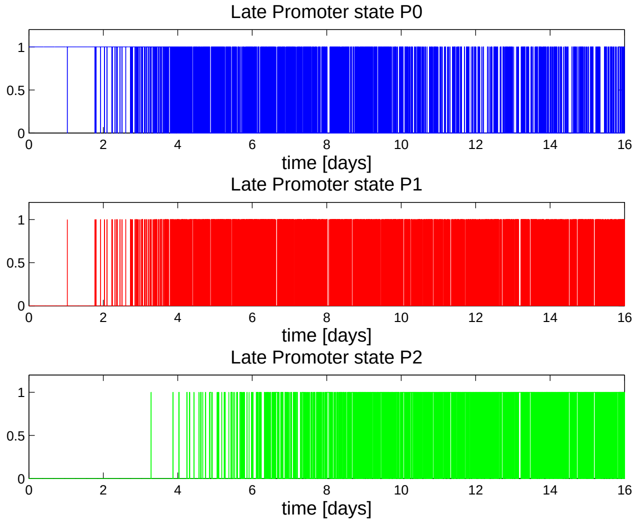

Figure 8.7.5: Time evolution of the late promoter markov chain states in function of the $C/EBP\beta$ slow modulation. At the beginning only p_0 state is active, then P_1 state is activated and finally P_3 state. Each state temporal evolution is a dichotomous random process describing when the state is active and when it is not.

In what follows we show the stochastic simulations of the most interesting state variables.

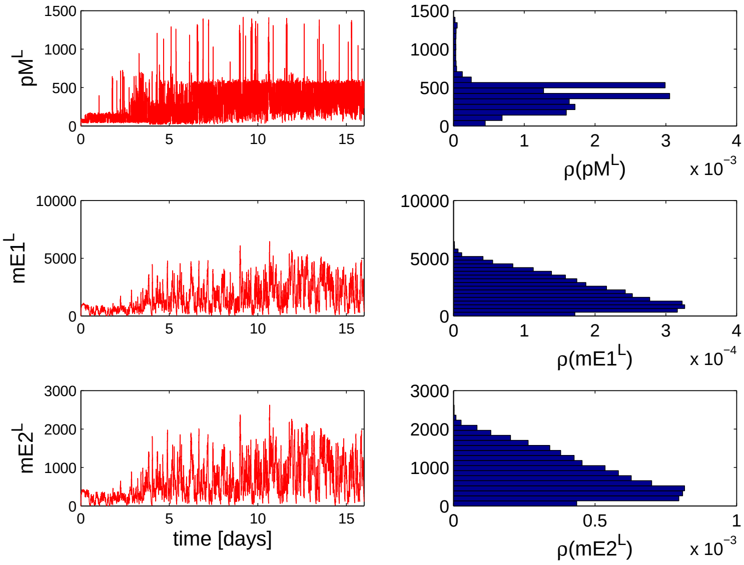

Figure 8.7.6: In the first coloumn: stochastic predictions of Late pre-mRNA, E_1^L and E_2^L transcripts. In the second coloumn: probability distributions of the state variables

In fig. 8.7.6 the late pre-mRNA is slowly modulated to it higher transcription regimen by a combined effect due to an intrinsic delay in the late promoter markov chain and a strong delay in $C/EBP\beta$ dynamics. The stronger spike bursts we can observe in the late phase are due, as was for the early promoter, to a dynamical change in time of the total conversion rate of the pre-mRNA. This results in the presence of bursts-like spikes. In some way the splicing modulation insert an additive state, as we can see from the multimodal distribution of the pre-mRNA, to the late promoter transcription that cannot be produced by the promoter alone. This is another example of complex pattern that can arise from the interconnection of more control layers, in line with the current paradigm of systems biology.

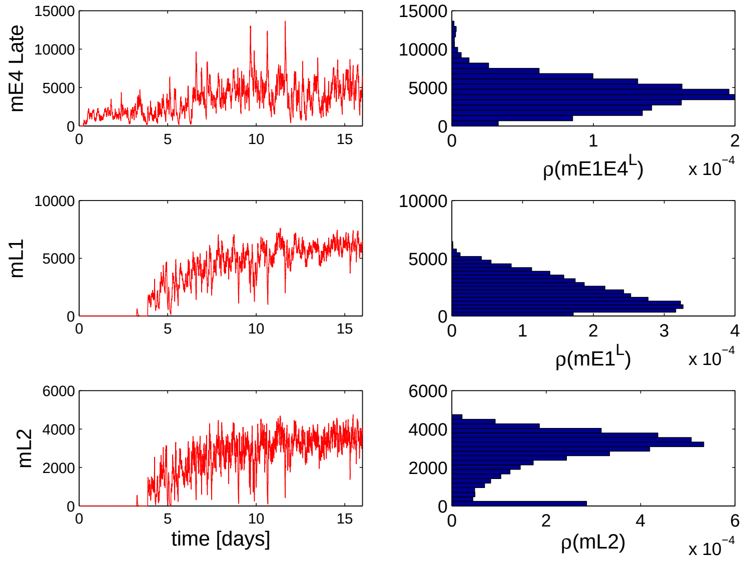

Figure 8.7.7: In the first coloumn: stochastic predictions of $E1E_4^L$, L_1 and L_2 transcripts. In the second coloumn: probability distributions of the state variables

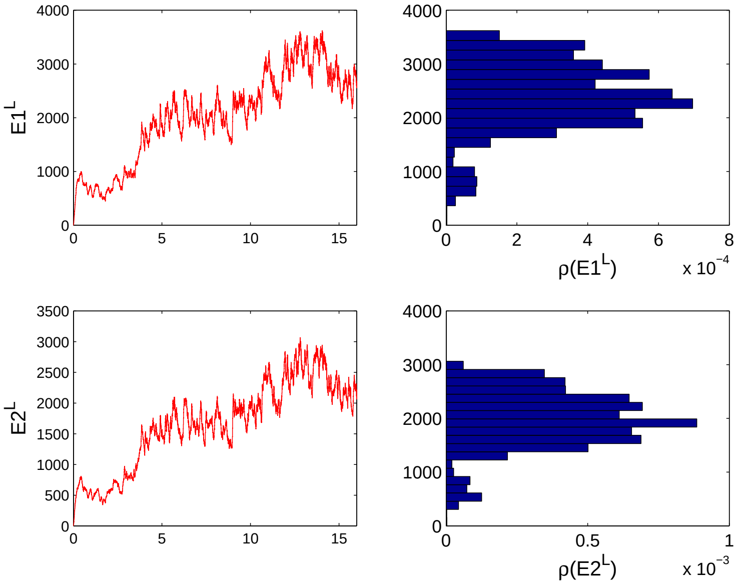

Figure 8.7.8: In the first coloumn: comparison between predictions of E_1 and E_2 proteins. In the second coloumn: probability distributions of the state variables.

Transcripts and proteins follows the induced delay arising from the pre-mRNA. In particular the late mRNAs are strongly delayed in the stochastic predictions too, thanks to the E_2-mediated control of the early

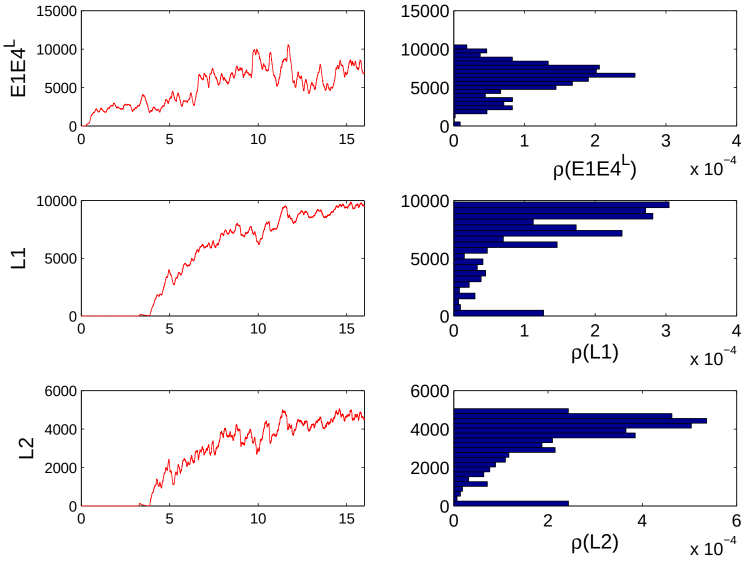

Figure 8.7.9: In the first coloumn: comparison between predictions of $E1E_4^L$, L_1 and L_2 proteins. In the second coloumn: probability distributions of the state variables.

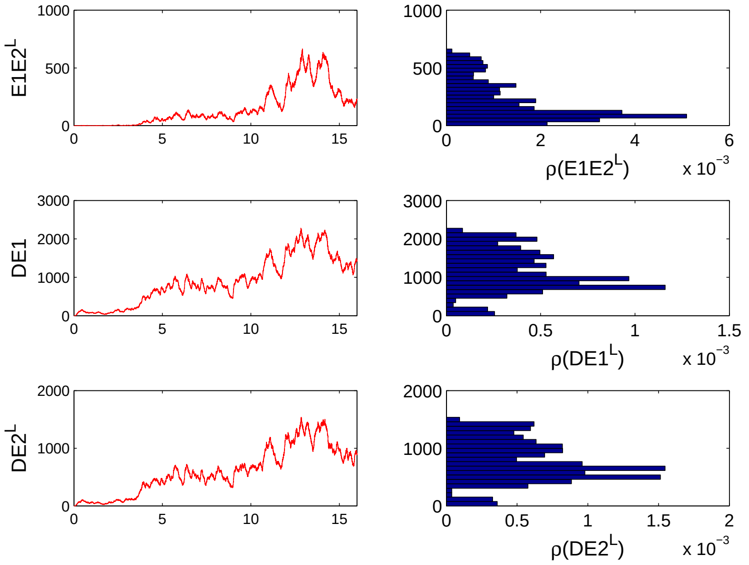

Figure 8.7.10: In the first coloumn: stochastic predictions of DE_1 and DE_2 dimers and E_1E_2 tetramer.In the second coloumn: probability distributions of the state variables.

polyadenylation signal, pAE.

We can observe that both the stochastic and the deterministic formulation can predict the late promoter behavior in good agreement with the current biological knowledge.

9
Discussion

9.1 Novelties

9.1.1 Models of specific mechanisms

The novelty of this project is both on methodological and biological /clinical site. The former is in line with the current challenge in recent years to have a holistic view of the basics regulatory mechanisms interconnected to form a complex machinery where complex patterns can arise only form the interconnection of basics modules. The purpose is thus the development and the study of a complex regulatory system made up of most of the main individual processes both at a deterministic and stochastic description level. In fact, the developed model accounts for different control layers, having a general validity, such as transcriptional regulation by means of the promoters regulation activity, post-transcriptional regulation by modeling alternative splicing and polyadenylation, translational regulation by modeling the protein synthesis from the transcripts and post-translational regulation by modeling proteins stabilization mechanisms and translational efficiency. These modules were interconnected together in order to develop a complex control system capable to predict complex dynamical patterns that can arise only from the interconnection of the above-mentioned modules.

As far as we know, it is the first time a stochastic model accounts for the complex post-transcriptional control, modeling the splicing and polyadenylation sites regulation, and connect this latter to the transcriptional control layer mediated by the promoters activities in order to explore complex patterns that can arise only from the modeling and the interconnection of different control layers.

9.1.2 Novel HPV model

In this thesis we have developed a novel stochastic and deterministic mathematical model of HPV gene regulatory network, in collaboration with the Elston Lab at Chapel Hill, University of North Carolina (UNC), and with the Department of Molecular Medicine (DMM) at University of Padova.

HPV offers a case of study of great interest in molecular systems biology and clinical studies. It involves a number of relevant regulatory mechanisms (e.g. transcription, translation, promoter modulation, polyadenylation regulation, splicing,) connected together to form a complex network, albeit its genome is relatively simple, thus suitable for an accurate deterministic and even stochastic modeling. Modeling its promoters activities and its post-transcriptional regulation gave use the possibility to investigate complex patterns formation that on the one hand have a general validity in dynamically explain the behavior of gene expression in its entireness and on the other allowed us to explore interesting qualitative behavior of the HPV gene network. In particular, a stochastic modeling gave us the perspective to understand how the intrinsic noise can affect and modify the behavior of each single modules and how it can generate complex patterns due to the interconnection of different modules and that cannot be predicted by a deterministic formulation.

As regards the biological/clinical aspects, we have developed a new synthesis of the HPV molecular biology with especial regard to gather/infer from literature the parameters useful for designing a dynamical model, and to shed light in what is still lacking in the biological literature, to complete and optimize a model of HPV gene regulatory network. In particular, it was done an effort in trying to get together the poor knowledge about the late promoter regulation, in order to present the minimal prerequisites necessary to design a dynamical model of this still elusive but of great importance in the context of the viral life cycle, upon differentiation.

The model shows a good agreement with the current biological knowledge and, as far as we know, there is no model of HPV gene regulation available in literature, apart a first heuristic/deterministic model on the early promoter regulation we have developed (Giaretta et al. (2015)).

The model is also able to predict complex patterns by connecting together the promoter activity regulation and the post-transcriptional regulation. These complex patterns were also very interesting in a biological perspective in finding out novel control mechanisms the virus could exhibit, such as an efficient and pulsatile control of viral replication or cell cycle restoration. Among these patterns it was possible to discover the capability of the developed model to predict the qualitative behavior of HPV integration and making the model structure interesting in a clinical perspective.

9.2 Experimental design and model validation

Validation is integral to the overall modeling process. It needs to be performed both during model building, and upon completion of the model.

9.2.1 Validation during model formulation

Model validation involves assessing the model in order to check that is well-founded and fulfils the purpose for which it was intended. Validation needs to be considered all the way through the modeling process. It begins when specifying the modeling activity and continues right on to the completed model.

The building of a model normally proceeds by postulating a conceptual model and then proceeding to a mathematical realization of this conceptual form. In putting up a conceptual framework, care needs to be taken to ensure that all the relevant physical and chemical concepts are properly described. This is important when it comes to producing the mathematical realization.

So far we followed and accomplished all these tasks throughout the the HPV model development process. We have stated the main purpose of the model was to reproduce the HPV gene regulatory network dynamics and we built the model structure consistently with the known molecular biology gathering/inferring all the available parameters from literature. During the building of the model we took care in reproducing the available knowledge and data we had available. In fact we were able to fit the experimental data about the early promoter activity and to qualitatively reproduce the dynamical evolution of E_1 and E_2 transcripts produced by the late promoter during differentiation.

9.2.2 Validation of the completed model

The validation process is dependent on the model purpose. In other words, the task is problems-specific. Validity is a general concept reflecting model purpose, current theories, and experimental data relating to the particular biological system of interest. Thus, as new theories are developed and new data become available, the requirements for a model to be deemed valid can change although its validity is still assessed in terms of the same criteria. A model to be valid needs to satisfy one or more of the criteria of: theoretical, empirical, pragmatic and heuristic validity (Cobelli Claudio (2008)).

Theoretical validity is concerned with ensuring that the model is consistent with accepted biological theories; Empirical validity is assessed by examining how well the model corresponds to available data; Pragmatic validity is used in supporting clinical decision making and heuristic validity is important when a model is to be used to test biological hypothesis.

So far we have theoretically validated the model being its structure has been designed by translating the major regulatory mechanisms behind its gene regulatory networks. Moreover it was developed by using a stochastic formalism representing the most accurate to describe molecular systems.

We have also in part accomplished the empirical validation being able to fit the early promoter activity and to qualitatively reproduce the E_1 and E_2 mRNAs pattern generated by the late promoter during differentiation. Obviously this part is still under development and we need new and complete data to completely and accurately accomplish this validation.

9.2.3 Experimental design

The model allowed us to properly design an experiment in order to get time courses of the viral transcripts during the viral life cycle. The model was useful to select the proper chemical species to measure, in what conditions and to set up an appropriate sample times grid.

The experiment will start in the next months and will take account for the analysis of HPV transcriptome and host gene expression during keratinocyte differentiation. The analysis will be performed at different time points after induction of keratinocyte differentiation (mediated by Ca^{2+}) thickening the sampling grid at the beginning of the infection to account for the splicing regulation modulating the oncogenes, then it will be reduced in the middle of the experiment and will be thickened again to account for the possible delayed production of E_1, E_2, E_4, L_1 and L_2 mR-

NAs. The analysis will be carried out for HPV16 genome by means of quantitative real-time PCR in order to obtain time series of all the viral transcripts (both early and late) and of the splicing factor SRSF1 and the transcription factors p63 and $C/EBP\beta$. This will be the first time that all the viral transcripts will be evaluated in time series together with temporal series of the major splicing factor and the two major late promoter regulators during the whole HPV life cycle.

The experimental data will then be useful to empirically validate the whole model.

9.3 HPV MODEL USAGES

Definitely this model will help in achieving a deep understanding of HPV gene expression in order to shed light into both the early and late promoter regulation but also in the post-transcriptional and post-translational modifications to reach a better comprehension of the viral life cycle and the delayed expression of the viral transcripts and proteins, especially upon differentiation. Moreover it will be useful to investigate the dynamical properties of post-transcriptional regulation and its dynamical cross-talk with the promoter regulation.

Once empirically validated, the model will represent an in silico simulator able to predict the HPV gene expression in different biological conditions. Updating the model with new experimental evidences even on the clinical side could be of potential usage in predict the evolution of pre-cancerous and cancerous stages.

Another very interesting and actual problem is regarding the development of antiviral therapies. In literature there are several antiviral therapy proposals but every of them tends to analyze just a single molecular actor at a time, speculating on new therapies proposals. Nevertheless, nowadays there isn't any working antiviral therapy. The problem could precisely reside in not keeping a holistic perspective, we have repeatedly stressed during this entire discussion. Most likely, the winner approach could be to consider the system in its entireness by analyzing it at different functioning levels (e.g. transcriptional and post-transcriptional regulation). The model, once empirically validated, would be of great help in deigning and/or optimizing an antiviral therapy. Actually, antiviral therapy is none other than imposing control to the viral system. In other terms it can be viewed as a modification of the HPV model structure. If the HPV model has been validated than it can be used in order to test some hypothesis and predict its modified behavior after its modification with the antiviral proposal therapy. In this way we could understand by means of in silico experiments if the antiviral proposal can be valid, before carrying out an experiment to verify the goodness of the idea. Subsequently we can use the model to test proper hypothesis in order to design a good an functioning antiviral therapy. In this way we will be able to accomplish an heuristic validation of the model, too.

9.4 Future development

- We will empirically validate the deterministic model (RT-PCR measurements are good to identify a deterministic model but not a stochastic one) by using the experimental measurements soon available. At first we will tune the parameters of the model in order to fit the experimental data, with the constraint to keep the parameters within acceptable biological values. Secondly we will seek to make the model identifiable, univocally identifying the parameters or combinations of them by means of standard identification algorithms such as weighted least squares (Cobelli Claudio (2008)).

  Even if these data will not suitable for validating the stochastic model, by identifying the deterministic version we will get a more robust set of parameters to subsequently better investigate the stochastic behavior of the model

- Once we will be sure the model is able to reproduce the experimental results we will perform an analysis of the statistical moments (mean, variance,...) of the chemical species.

- After the validation of the deterministic model we will design a single cell experiment in order to validate the stochastic formalism. This will be interesting in order to evaluate the strength of the stochastic sources inside the system, such as transcription, degradation, translation, promoter regulation, ... and find out what are the principal parameters constraints behind the noise strength modulation. Another interesting application could reside in the parameters identification of the stochastic system by identifying the deterministic system of its statistical moments (El Samad et al. (2005)).

  We will make use of the stochastic simulations, by fixing the parameters identified in the deterministic validation of the model, to analyze the probability distributions of the state variables. This will be useful to the best statistical moments order to stop in order to maintain a good statistical information. The statistical moments will be described by means of a system of ordinary differential equations, derived by the ME. One way to perform such analysis, given the complex structure of the syste, could be to approximate the whole master equation of the system in terms of its diffusion limit Fokker-Planck Equation (for each state combination of the control markov chains) (Gardiner (2009), VanKampen (2007), Scott (2013)), subsequently deriving the statistical moments by applying their definition to the FPE and the closure of moments to make the system closed (moments system won't be closed because of the model nonlinearities) (Pirone and Elston (2004)).

- In this thesis we have performed the stochastic simulations numerically solving the Master Equation by means of the Gillespie Algorithm. This is because on the one hand the computational effort in simulating the model was acceptable with the fixed parameters we inferred form literature and with the upper bound of copy number estimated. For this reason we preferred to investigate the stochastic model predictions in terms of the exact numerical solutions we can achieve by using the Gillespie's algorithm.

However, we will approximate the model in terms of stochastic differential equations (SDE) that are much more computationally efficient. This will be performed by following a hybrid scheme to the stochastic simulation. This is necessary in order to maintain the jump control processes intrinsically associated to the promoters, splicing sites and polyadenylation sites that are essential to explain strong stochastic events such as bursts in gene expression. The stochastic variables accounting for the control logic of the system (i.e., promoters, splicing and polyadenylation state variables) will be maintained discrete and simulated by means of an efficient version of Gillespie's algorithm. For the remaining chemical species a FPE will be derived by applying the diffusion limit of the Master Equation. Subsequently, the FPE will be converted into a system of stochastic differential equations (SDE) by applying the Ito's Lemma (Øksendal (2013), Gardiner (2009), Adalsteinsson et al. (2004)).

- The model will be extended in terms of a spatial diffusion formulation in order to describe the distribution of the viral proteins throughout the epithelium during cellular differentiation.

Bibliography

David Adalsteinsson, David McMillen, and Timothy C Elston. Biochemical Network Stochastic Simulator (BioNetS): software for stochastic modeling of biochemical networks. *BMC bioinformatics*, 5:24, 2004. ISSN 1471-2105. doi: 10.1186/1471-2105-5-24.

Zeng Zhi-Ming Ajiro Masahiko. E6ˆE7, a Novel Splice Isoform Protein of Human Papillomavirus 16, Stabilizes Viral E6 and E7 Oncoproteins via HSP90 and GRP78. 6(1):1–15, 2015. doi: 10.1128/mBio.02068-14.Editor.

A Audibert, D Weil, and F Dautry. In Vivo Kinetics of mRNA Splicing and Transport in Mammalian Cells. 22(19):6706–6718, 2002. doi: 10.1128/MCB.22.19.6706.

Theodore Hupp Ayeda Ayed. *Molecular Biology Intelligence Unit; p53*. Springer, 2010.

N S Banerjee, H K Wang, T R Broker, and L T Chow. Human papillomavirus (HPV) E7 induces prolonged G2 following S phase reentry in differentiated human keratinocytes. *J Biol Chem*, 286(17):15473–15482, 2011. ISSN 1083-351X. doi: 10.1074/jbc.M110.197574. URL http://www.ncbi.nlm.nih.gov/pubmed/21321122.

S Bellanger, C Demeret, S Goyat, and F Thierry. Stability of the human papillomavirus type 18 E2 protein is regulated by a proteasome degradation pathway through its amino-terminal transactivation domain. *J Virol*, 75(16):7244–7251, 2001. ISSN 0022-538X. doi: 10.1128/JVI.75.16.7244. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{&}db=PubMed{&}dopt=Citation{&}list{_}uids=11461997.

Hans-Ulrich Bernard. Gene expression of genital human papillomaviruses and considerations on potential antiviral approaches. *Antiviral therapy*, 7(4):219–37, dec 2002. ISSN 1359-6535. URL http://europepmc.org/abstract/med/12553476.

Hans-Ulrich Bernard. Regulatory elements in the viral genome. *Virology*, 445(1-2):197–204, 2013. ISSN 1096-0341. doi: 10.1016/j.virol.2013.04.035. URL http://www.ncbi.nlm.nih.gov/pubmed/23725692.

Patrick Billingsley. *Probability and Measure, Anniversary Edition*. Wiley, 4th edition, 2012.

Dukagjin M Blakaj, Narcis Fernandez-Fuentes, Zigui Chen, Rashmi Hegde, Andras Fiser, Robert D Burk, and Michael Brenowitz. Evolutionary and biophysical relationships among the papillomavirus E2 proteins. *Frontiers in bioscience (Landmark edition)*, 14:900–17, 2009. ISSN 1093-4715. doi: 3285[pii]. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3705561{&}tool=pmcentrez{&}rendertype=abstract.

Darron R Brown, Douglas Kitchin, Brahim Qadadri, Nicole Neptune, Teresa Batteiger, and Aaron Ermel. The human papillomavirus type 11 E1–E4 protein is a transglutaminase 3 substrate and induces abnormalities of the cornified cell envelope. *Virology*, 345(1):290–8, 2006. ISSN 0042-6822. doi: 10.1016/j.virol.2005.09.048. URL http://www.sciencedirect.com/science/article/pii/S0042682205006045.

Christopher B Buck, Naiqian Cheng, Cynthia D Thompson, Douglas R Lowy, Alasdair C Steven, John T Schiller, and Benes L Trus. Arrangement of L2 within the papillomavirus capsid. *Journal of virology*, 82(11):5190–5197, 2008. ISSN 0022-538X. doi: 10.1128/JVI.02726-07.

Long Cai, Chiraj K. Dalal, and Michael B. Elowitz. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature*, 455 (7212):485–490, 2008. ISSN 0028-0836. doi: 10.1038/nature07292. URL http://www.nature.com/doifinder/10.1038/nature07292.

Shih-fong Chao, Warren J Rocque, Selwyn Daniel, Linda E Czyzyk, William C Phelps, and Kenneth A Alexander. Subunit Affinities and Stoichiometries of the Human Papillomavirus Type 11. pages 4586–4594, 1999.

Louise T Chow, Thomas R Broker, and Bettie M Steinberg. The natural history of human papillomavirus infections of the mucosal epithelia. *Apmis*, 118(6-7):422–49, 2010. ISSN 1600-0463. doi: 10.1111/j.1600-0463.2010.02625.x. URL http://www.ncbi.nlm.nih.gov/pubmed/20553526.

Carson Ewart Cobelli Claudio. *Introduction to Modeling in Physiology and Medicine*. Academic Press, 2nd edition, 2008.

F Cobo. *Human Papillomavirus Infections*. Woodhead Publishing, 2012.

Brian Collier, Daniel Oberg, Xiaomin Zhao, and Stefan Schwartz. Specific inactivation of inhibitory sequences in the 5′ end of the human papillomavirus type 16 L1 open reading frame results in production of high levels of L1 protein in human epithelial cells. *Journal of virology*, 76(6):2739–52, 2002. ISSN 0022-538X. doi: 10.1128/JVI.76.6.2739. URL http://www.ncbi.nlm.nih.gov/pubmed/?term=11861841.

Alberto Corradin, Barbara DI Camillo, Francesca Rende, Vincenzo Ciminale, Gianna Maria Toffolo, and Claudio Cobelli. Retrovirus HTLV-1 gene circuit: a potential oscillator for eukaryotes. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 432:421–432, 2010. ISSN 2335-6936.

Alberto Corradin, Barbara Di Camillo, Vincenzo Ciminale, Gianna Toffolo, and Claudio Cobelli. Sensitivity analysis of retrovirus HTLV-1 transactivation. *Journal of computational biology : a journal of computational molecular cell biology*, 18(2):183–93, 2011. ISSN 1557-8666. doi: 10.1089/cmb.2010.0219. URL http://online.liebertpub.com/doi/abs/10.1089/cmb.2010.0219.

Kerstin Crusius, Isabel Rodriguez, and Angel Alonso. The human papillomavirus type 16 E5 protein modulates ERK1/2 and p38 MAP kinase activation by an EGFR-independent process in stressed human keratinocytes. *Virus Genes*, 20(1):65–69, 2000. ISSN 09208569. doi: 10.1023/A:1008112207824.

Chiraj K. Dalal, Long Cai, Yihan Lin, Kasra Rahbar, and Michael B. Elowitz. Pulsatile Dynamics in the Yeast Proteome. *Current Biology*, 24(18):2189–2194, 2014. ISSN 09609822. doi: 10.1016/j.cub.2014.07.076. URL http://linkinghub.elsevier.com/retrieve/pii/S0960982214009737.

Caroline Demeret, Christian Desaintes, and Moshe Yaniv. Different Mechanisms Contribute to the E2-Mediated Transcriptional Repression of Human Papillomavirus Type 18 Viral Oncogenes. 71(12):9343–9349, 1997.

John Doorbar. The papillomavirus life cycle. *Journal of Clinical Virology*, 32:7–15, 2005. ISSN 13866532. doi: 10.1016/j.jcv.2004.12.006. URL http://linkinghub.elsevier.com/retrieve/pii/S1386653204003671.

John Doorbar. Molecular biology of human papillomavirus infection and cervical cancer. *Clinical Science*, 110(5):525–541, 2006. ISSN 0143-5221. doi: 10.1042/CS20050369. URL http://clinsci.org/lookup/doi/10.1042/CS20050369.

John Doorbar, Wim Quint, Lawrence Banks, Ignacio G. Bravo, Mark Stoler, Tom R. Broker, and Margaret a. Stanley. The biology and life-cycle of human papillomaviruses. *Vaccine*, 30(SUPPL.5):F55–F70, 2012. ISSN 0264410X. doi: 10.1016/j.vaccine.2012.06.083. URL http://dx.doi.org/10.1016/j.vaccine.2012.06.083.

G Paolo Dotto. Crosstalk of Notch with p53 and p63 in cancer growth control. *Nature reviews. Cancer*, 9(8):587–595, 2009. ISSN 1474-175X. doi: 10.1038/nrc2675. URL http://dx.doi.org/10.1038/nrc2675.

F.S. Dukhovich. DRUG SYNTHESIS METHODS AND MANUFACTURING TECHNOLOGY RELATIONSHIP BETWEEN THE DISSOCIATION CONSTANT AND THE LIFETIME FOR COMPLEXES OF BIOLOGICALLY ACTIVE SUBSTANCES WITH RECEPTORS AND ENZYMES. *Pharmaceutical Chemistry Journal*, 36(5):248–254, 2002.

Hana El Samad, Mustafa Khammash, Linda Petzold, and Dan Gillespie. Stochastic modelling of gene regulatory networks. *International Journal of Robust and Nonlinear Control*, 15(15):691–711, 2005. ISSN 10498923. doi: 10.1002/rnc.1018.

M B Elowitz, A J Levine, E D Siggia, and P S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002. ISSN 1095-9203. doi: 10.1126/science.1070919. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve{&}db=PubMed{&}dopt=Citation{&}list{_}uids=12183631.

Frauke Fehrmann, David J Klumpp, and Laimonis A Laimins. Human papillomavirus type 31 E5 protein supports cell cycle progression and activates late viral functions upon epithelial differentiation. *Journal of virology*, 77(5):2819–31, 2003. ISSN 0022-538X. doi: 10.1128/JVI.77.5.2819. URL http://www.ncbi.nlm.nih.gov/pubmed/12584305.

L Finnen, Kimberly D Erickson, Xiaojiang S Chen, and Robert L Garcea. Interactions between Papillomavirus L1 and L2 Capsid Proteins. 77(8):4818–4826, 2003. doi: 10.1128/JVI.77.8.4818.

Crispin Gardiner. *Stochastic Methods. A Handbook for the Natural and Social Sciences*. Springer Complexity, 4th edition, 2009.

Dan Gareau. Automated identification of epidermal keratinocytes in reflectance confocal microscopy. *Journal of biomedical optics*, 16:030502, 2011. ISSN 10833668. doi: 10.1117/1.3552639.

Sybil M Genther, Stephanie Sterling, Stefan Duensing, Karl Münger, Carol Sattler, and Paul F Lambert. Quantitative role of the human papillomavirus type 16 E5 gene during the productive stage of the viral life cycle. *Journal of virology*, 77 (5):2832–42, 2003. ISSN 0022-538X. doi: 10.1128/JVI.77.5.2832. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=149772{&}tool=pmcentrez{&}rendertype=abstract.

A Giaretta, B Di Camillo, L Barzon, and G M Toffolo. Modeling HPV Early Promoter Regulation. 2:6493–6496, 2015.

Daniel T. Gillespie. The chemical Langevin and Fokker-Planck equations for the reversible isomerization reaction. *Journal of Physical Chemistry A*, 106(20):5063–5071, 2002. ISSN 10895639. doi: 10.1021/jp0128832.

Daniel T Gillespie. The chemical Langevin equation The chemical Langevin equation. 297(2000):297–306, 2013. doi: 10.1063/1.481811.

Daniel T DT Gillespie. Stochastic simulation of chemical kinetics. *Annual review of physical chemistry*, 58:35–55, 2007. ISSN 0066-426X. doi: 10.1146/annurev.physchem.58.032806.104637. URL http://www.ncbi.nlm.nih.gov/pubmed/17037977$\delimiter"026E30F$nhttp://www.annualreviews.org/doi/abs/10.1146/annurev.physchem.58.032806.104637.

Daniel T Gillesple. Exact Stochastic Simulation of couple chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977. ISSN 0022-3654. doi: 10.1021/j100540a008.

S V Graham. Europe PMC Funders Group Human papillomavirus : gene expression , regulation and prospects for novel diagnostic methods and antiviral therapies. 5(10):1493–1506, 2012. doi: 10.2217/fmb.10.107. Human.

Vignesh Gunasekharan, Guylaine Haché, and Laimonis Laimins. Differentiation-dependent changes in levels of C/EBPβ repressors and activators regulate human papillomavirus type 31 late gene expression. *Journal of virology*, 86(9):5393–8, 2012. ISSN 1098-5514. doi: 10.1128/JVI.07239-11. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3347327{&}tool=pmcentrez{&}rendertype=abstract.

Kelly A Hartley and Kenneth A Alexander. Human TATA Binding Protein Inhibits Human Papillomavirus Type 11 DNA Replication by Antagonizing E1-E2 Protein Complex Formation on the Viral Origin of Replication. 76(10):5014–5023, 2002. doi: 10.1128/JVI.76.10.5014.

Martin J Hicks, Bianca J Lam, and Klemens J Hertel. Analyzing mechanisms of alternative pre-mRNA splicing using in vitro splicing assays. 37:306–313, 2005. doi: 10.1016/j.ymeth.2005.07.012.

Samuel Y Hou, Shwu-yuan Wu, and Cheng-ming Chiang. Transcriptional Activity among High and Low Risk Human Papillomavirus E2 Proteins Correlates with E2 DNA Binding *. 277(47):45619–45629, 2002. doi: 10.1074/jbc.M206829200.

René Huber, Thomas Panterodt, Bastian Welz, Martin Christmann, Judith Friesenhagen, Andreas Westphal, Daniel Pietsch, and Korbinian Brand. C / EBP β -LAP Ã / LAP Expression Is Mediated by RSK / eIF4B-Dependent Signalling and Boosted by Increased Protein Stability in Models of Monocytic Differentiation. pages 1–23, 2015. doi: 10.1371/journal.pone.0144338.

C Johansson, M Somberg, X Li, E Backstrom Winquist, J Fay, F Ryan, D Pim, L Banks, and S Schwartz. HPV-16 E2 contributes to induction of HPV-16 late gene expression by inhibiting early polyadenylation. *Embo J*, 31(14):3212–3227, 2012. ISSN 0261-4189. doi: 10.1038/emboj.2012.147. URL http://www.ncbi.nlm.nih.gov/pubmed/22617423.

Cecilia Johansson and Stefan Schwartz. Regulation of human papillomavirus gene expression by splicing and polyadenylation. *Nature reviews. Microbiology*, 11(4):239–51, 2013. ISSN 1740-1534. doi: 10.1038/nrmicro2984. URL http://www.ncbi.nlm.nih.gov/pubmed/23474685.

Mads Kaern, Timothy C Elston, William J Blake, and James J Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature reviews. Genetics*, 6(6):451–464, 2005. ISSN 1471-0056. doi: 10.1038/nrg1615.

T B Kepler and T C Elston. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophysical journal*, 81(6):3116–3136, 2001. ISSN 00063495. doi: 10.1016/S0006-3495(01)75949-8. URL http://dx.doi.org/10.1016/S0006-3495(01)75949-8$\delimiter"026E30F$nhttp:

//www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1301773{&}tool=pmcentrez{&}rendertype=abstract.

Lauren E. King, Edward S. Dornan, Mary M. Donaldson, and Iain M. Morgan. Human papillomavirus 16 E2 stability and transcriptional activation is enhanced by E1 via a direct protein-protein interaction. *Virology*, 414(1):26–33, 2011. ISSN 00426822. doi: 10.1016/j.virol.2011.03.002. URL http://dx.doi.org/10.1016/j.virol.2011.03.002.

E Kowenz-Leutz and a Leutz. A C/EBP beta isoform recruits the SWI/SNF complex to activate myeloid genes. *Molecular cell*, 4(5):735–743, 1999. ISSN 10972765. doi: 10.1016/S1097-2765(00)80384-6.

Ulla Krüppel, Andreas Müller-Schiffmann, Stephan E. Baldus, Sigrun Smola-Hess, and Gertrud Steger. E2 and the co-activator p300 can cooperate in activation of the human papillomavirus type 16 early promoter. *Virology*, 377(1):151–159, 2008. ISSN 00426822. doi: 10.1016/j.virol.2008.04.006. URL http://linkinghub.elsevier.com/retrieve/pii/S004268220800233X.

I Kukimoto, T Takeuchi, and T Kanda. CCAAT/enhancer binding protein beta binds to and activates the P670 promoter of human papillomavirus type 16. *Virology*, 346(1):98–107, 2006. ISSN 00426822. doi: 10.1016/j.virol.2005.10.025. URL http://www.ncbi.nlm.nih.gov/pubmed/16307770.

Fausto Nelson Mitchell Richard N Kumar Vinay, Abbas Abul K. *Robbins Basic Pathology*. Saunders Elsevier, 4th edition, 2007.

Reet Kurg. The Role of E2 Proteins in. *intechopen*, 2009. doi: 10.5772/19609.

Michelle S Longworth and Laimonis a Laimins. Pathogenesis of Human Papillomaviruses in Differentiating Epithelia Pathogenesis of Human Papillomaviruses in Differentiating Epithelia. *Microbiology and molecular biology reviews*, 68(2):362–372, 2004. doi: 10.1128/MMBR.68.2.362.

E V Maytin and J F Habener. Transcription factors C/EBP alpha, C/EBP beta, and CHOP (Gadd153) expressed during the differentiation program of keratinocytes in vitro and in vivo. *The Journal of investigative dermatology*, 110(3):238–46, 1998. ISSN 0022-202X. doi: 10.1046/j.1523-1747.1998.00123.x. URL http://www.ncbi.nlm.nih.gov/pubmed/9506442.

Alison a. McBride. The Papillomavirus E2 proteins. *Virology*, 445(1-2): 57–79, 2013. ISSN 00426822. doi: 10.1016/j.virol.2013.06.006. URL http://dx.doi.org/10.1016/j.virol.2013.06.006.

Pauline B McIntosh, Stephen R Martin, Deborah J Jackson, Jameela Khan, Erin R Isaacson, Lesley Calder, Kenneth Raj, Heather M Griffin, Qian Wang, Peter Laskey, John F Eccleston, and John Doorbar. Structural analysis reveals an amyloid form of the human papillomavirus type 16 E1–E4 protein and provides a molecular basis for its accumulation. *Journal of virology*, 82(16):8196–8203, 2008. ISSN 1098-5514. doi: 10.1128/JVI.00509-08.

Caleb McKinney, Katherine Hussmann, and Alison McBride. The Role of the DNA Damage Response throughout the Papillomavirus Life Cycle. *Viruses*, 7(5):2450–2469, 2015. ISSN 1999-4915. doi: 10.3390/v7052450. URL http://www.mdpi.com/1999-4915/7/5/2450/.

Kristen K Mighty and Laimonis a Laimins. P63 Is Necessary for the Activation of Human Papillomavirus Late Viral Functions Upon Epithelial Differentiation. *Journal of virology*, 85(17): 8863–9, 2011. ISSN 1098-5514. doi: 10.1128/JVI.00750-11. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3165790{&}tool=pmcentrez{&}rendertype=abstract.

Yu-keung Mok, Gonzalo D E Prat Gay, P Jonathan Butler, and Mark Bycroft. Equilibrium dissociation and unfolding. pages 310–319, 1996.

Sarah Mole, Steven G Milligan, and Sheila V Graham. Human papillomavirus type 16 E2 protein transcriptionally activates the promoter of a key cellular splicing factor, SF2/ASF. *J Virol*, 83(1):357–367, 2009. ISSN 1098-5514. doi: JVI.01414-08[pii]10.1128/JVI.01414-08. URL http://www.ncbi.nlm.nih.gov/pubmed/18945764$\delimiter"026E30F$nhttp://dx.doi.org/10.1128/JVI.01414-08$\delimiter"026E30F$nhttp://jvi.asm.org/content/83/1/357.full.pdf$\delimiter"026E30F$nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2612322{&}tool=pmcentrez{&}rendertype=abstract.

Vaishali R Moulton, Andrew R Gillooly, and George C Tsokos. Ubiquitination Regulates Expression of the Serine / Arginine- rich Splicing Factor 1 (SRSF1) in Normal and Systemic Lupus Erythematosus (SLE) T Cells *. 289(7):4126–4134, 2014. doi: 10.1074/jbc.M113.518662.

Mandy Muller and Caroline Demeret. The HPV E2-Host Protein-Protein Interactions: A Complex Hijacking of the Cellular Network. *The open virology journal*, 6:173–89, 2012. ISSN 1874-3579. doi: 10.2174/1874357901206010173. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3547520{&}tool=pmcentrez{&}rendertype=abstract.

Bach-cuc Nguyen, Karine Lefort, Anna Mandinova, Dario Antonini, Vikram Devgan, Giusy Della Gatta, Maranke I Koster, Zhuo Zhang, Jian Wang, Alice Tommasi, Jan Kitajewski, Giovanna Chiorino, Dennis R Roop, Caterina Missero, and G Paolo Dotto. Cross-regulation between Notch and p63 in keratinocyte commitment to differentiation. pages 1028–1042, 2006. doi: 10.1101/gad.1406006.whereas.

Bernt Øksendal. *Stochastic Differential Equations. An Introduction with Applications*. Springer, 6th edition, 2013.

Michelle a Ozbun. Human papillomavirus type 31b infection of human keratinocytes and the onset of early transcription. *Journal of virology*, 76(22):11291–11300, 2002. ISSN 0022-538X. doi: 10.1128/JVI.76.22.11291-11300.2002.

Michelle A Ozbun and Craig Meyers. Human Papillomavirus Type 31b E1 and E2 Transcript Expression Correlates with Vegetative Viral Genome Amplification. 230:218–230, 1998.

Jason R. Pirone and Timothy C. Elston. Fluctuations in transcription factor binding can explain the graded and binary responses observed in inducible gene expression. *Journal of Theoretical Biology*, 226(1):111–112, 2004. ISSN 00225193. doi: 10.1016/j.jtbi.2003.08.008.

Rachel Raybould. Human Papillomavirus Integration and its Role in Cervical Malignant Progression. *The Open Clinical Cancer Journal*, 5(1):1–7, 2011. ISSN 18741894. doi: 10.2174/1874189401105010001.

Daniel DiMaio Robert L. Garcea. *The Papillomaviruses*. Springer, 2007.

Martina Schmitz, Corina Driesch, Lars Jansen, Ingo B. Runnebaum, and Matthias Dürst. Non-Random Integration of the HPV Genome in Cervical Cancer. *PLoS ONE*, 7(6):e39632, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0039632. URL http://dx.plos.org/10.1371/journal.pone.0039632.

Stefan Schwartz. Papillomavirus transcripts and posttranscriptional regulation. *Virology*, 445(1-2):187–196, 2013. ISSN 0042-6822. doi: 10.1016/j.virol.2013.04.034. URL http://dx.doi.org/10.1016/j.virol.2013.04.034.

M. Scott. *Applied Stochastic Processes in science and engineering*. University of Waterloo, 2013.

R C Sears and L Sealy. Multiple forms of C/EBP beta bind the EFII enhancer sequence in the Rous sarcoma virus long terminal repeat. *Molecular and cellular biology*, 14(7):4855–71, 1994. ISSN 0270-7306. doi: 10.1128/MCB.14.7.4855.Updated. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=358858{&}tool=pmcentrez{&}rendertype=abstract.

a. Seth. Pub Med Central CANADA. *Disabil Rehabil.*, 49(2):743–750, 2008. ISSN 1535-7228. doi: 10.1167/iovs.07-1072.Complement-Associated.

Hye-Jin Shin, Jungnam Joo, Ji Hyun Yoon, Chong Woo Yoo, and Joo-Young Kim. Physical status of human papillomavirus integration in cervical cancer is associated with treatment outcome of the patients treated with radiotherapy. *PloS one*, 9(1):e78995, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0078995.

Abhyudai Singh and Leor S Weinberger. Stochastic gene expression as a molecular switch for viral latency. *Current opinion in microbiology*, 12(4):460–6, 2009. ISSN 1879-0364. doi: 10.1016/j.mib.2009.06.016. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2760832{&}tool=pmcentrez{&}rendertype=abstract.

Abhyudai Singh, Brandon Razooky, Chris D. Cox, Michael L. Simpson, and Leor S. Weinberger. Transcriptional bursting from the HIV-1 promoter is a significant source of stochastic noise in HIV-1 gene expression. *Biophysical Journal*, 98(8):L32–L34, 2010. ISSN 00063495. doi: 10.1016/j.bpj.2010.03.001. URL http://dx.doi.org/10.1016/j.bpj.2010.03.001.

Emiko Soeda, Maureen C Ferran, Carl C Baker, and Alison a McBride. Repression of HPV16 early region transcription by the E2 protein. *Virology*, 351(1):29–41, 2006. ISSN 0042-6822. doi: 10.1016/j.virol.2006.03.016. URL http://www.ncbi.nlm.nih.gov/pubmed/16624362.

Monika Somberg and Stefan Schwartz. Multiple ASF/SF2 sites in the human papillomavirus type 16 (HPV-16) E4-coding region promote splicing to the most commonly used 3′-splice site on the HPV-16 genome. *Journal of virology*, 84(16):8219–8230, 2010. ISSN 0022-538X. doi: 10.1128/JVI.00462-10.

S W Straight, P M Hinkle, R J Jewers, and D J McCance. The E5 oncoprotein of human papillomavirus type 16 transforms fibroblasts and effects the downregulation of the epidermal growth factor receptor in keratinocytes. *Journal of virology*, 67(8):4521–4532, 1993. ISSN 0022-538X.

Shuying Sun, Zuo Zhang, Rahul Sinha, Rotem Karni, and Adrian R Krainer. SF2 / ASF autoregulation involves multiple layers of post-transcriptional and translational control. *Nature Publishing Group*, 17 (3):306–312, 2010. ISSN 1545-9993. doi: 10.1038/nsmb.1750. URL http://dx.doi.org/10.1038/nsmb.1750.

T Szekely Jr. and K Burrage. Stochastic simulation in systems biology. *Comput Struct Biotechnol J*, 12(20-21):14–25, 2014. ISSN 20010370. doi: 10.1016/j.csbj.2014.10.003. URL http://www.ncbi.nlm.nih.gov/pubmed/25505503$\delimiter"026E30F$nhttp://ac.els-cdn.com/S2001037014000403/1-s2.0-S2001037014000403-main.pdf?{_}tid=a3a5d306-ac86-11e4-9cb2-00000aab0f6c{&}acdnat=1423065679{_}86336ed95e790cb634e5584ed5a2e722.

S H Tan, L E Leong, P A Walker, and H U Bernard. The human papillomavirus type 16 E2 transcription factor binds with low cooperativity to two flanking sites and represses the E6 promoter through displacement of Sp1 and TFIID. *Journal of virology*, 68(10):6411–6420, 1994. ISSN 0022-538X. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=237061{&}tool=pmcentrez{&}rendertype=abstract$\delimiter"026E30F$nhttp://jvi.asm.org/content/68/10/6411.short.

Ewan R Taylor, Winifred Boner, Edward S Dornan, Eilidh M Corr, and Iain M Morgan. UVB irradiation reduces the half-life and transactivation potential of the human papillomavirus 16 E2 protein. *Oncogene*, 22 (29):4469–4477, 2003. ISSN 09509232. doi: 10.1038/sj.onc.1206746.

Françoise Thierry. Transcriptional regulation of the papillomavirus oncogenes by cellular and viral transcription factors in cervical carcinoma. *Virology*, 384(2):375–379, 2009. ISSN 0042-6822. doi: 10.1016/j.virol.2008.11.014. URL http://dx.doi.org/10.1016/j.virol.2008.11.014.

H Valdovinos-Torres, M Orozco-Morales, a Pedroza-Saavedra, L Padilla-Noriega, F Esquivel-Guadarrama, and L Gutierrez-Xicotencatl. Different Isoforms of HPV-16 E7 Protein are Present in Cytoplasm and Nucleus. *The open virology journal*, 2:15–23, 2008. ISSN 1874-3579. doi: 10.2174/1874357900802010015. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2678815{&}tool=pmcentrez{&}rendertype=abstract.

N. G. VanKampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, 3rd edition, 2007.

H.-K. Wang, A. A. Duffy, T. R. Broker, and L. T. Chow. Robust production and passaging of infectious HPV in squamous epithelium of primary human keratinocytes. *Genes & Development*, 23(2):181–194, 2009. ISSN 0890-9369. doi: 10.1101/gad.1735109. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2648537{&}tool=pmcentrez{&}rendertype=abstract.

Leor S Weinberger, Roy D Dar, and Michael L Simpson. Transient-mediated fate determination in a transcriptional circuit of HIV. *Nature Genetics*, 40(4):466–470, 2008. ISSN 1061-4036. doi: 10.1038/ng.116. URL http://www.nature.com/doifinder/10.1038/ng.116.

Maria Werner, Ingemar Ernberg, Jiezhi Zou, Jenny Almqvist, and Erik Aurell. Epstein-Barr virus latency switch in human B-cells: a physico-chemical model. *BMC systems biology*, 1(1):40, 2007a. ISSN 1752-0509. doi: 10.1186/1752-0509-1-40. URL http://www.biomedcentral.com/1752-0509/1/40.

Maria Werner, LiZhe Zhu, and Erik Aurell. Cooperative action in eukaryotic gene regulation: Physical properties of a viral example. *Phys. Rev. E*, 76:061909, Dec 2007b. doi: 10.1103/PhysRevE.76.061909. URL http://link.aps.org/doi/10.1103/PhysRevE.76.061909.

Acknowledgments

I would like to state my thankfulness to my advisor Prof.ssa Gianna Maria Toffolo for her guidance, assistance and support during these three years.

Thanks to Prof.ssa Barbara Di Camillo for her aid and support during these three years.

Thanks to Prof.ssa Luisa Barzon for her aid and numerous suggestions on the HPV project.

Thanks to Professor Morten Gram Pedersen for his aid and useful suggestions.

Thanks to Professor Timothy Elston for the opportunity of working in his group and for his teachings.