Leonardo Pedro Donas-Boto de Vilhena Martins

Licenciatura em Ciências de Engenharia Biomédica

STOCHASTIC MODEL OF TRANSCRIPTION INITIATION OF CLOSELY SPACED PROMOTERS IN ESCHERICHIA COLI

Dissertação para obtenção do Grau de Mestre
em Engenharia Biomédica

Orientador: José Manuel Fonseca, Professor Auxiliar, FCT-UNL
Co-orientador: André Sanches Ribeiro, Professor Assistente, TUT, Finlândia

Júri:

Presidente:	Prof. Doutor Mário António Basto Forjaz Secca
Arguente:	Prof. Doutora Ilda Santos Sanches
Vogais:	Prof. Doutor José Manuel Fonseca
	Prof. Doutor André Sanches Ribeiro

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Dezembro 2011

Copyright

“Aut inveniam viam aut faciam”

"I shall either find a way or make one."

Hannibal

Resumo

Os mecanismos reguladores da transcrição permitem aos organismos uma rápida adaptação a mudanças no meio ambiente e actuam frequentemente na iniciação da transcrição. Nesta Tese é proposto um modelo estocástico da iniciação da transcrição ao nível dos nucleotídos para estudar a dinâmica da produção de ácidos ribonucleicos (RNAs) em promotores com um curto espaçamento e os seus mecanismos de regulação.

Neste estudo analisa-se como diferentes disposições (convergente e divergente), distância entre os locais de iniciação da transcrição (TSS) e diferentes parâmetros cinéticos afectam a dinâmica da produção de RNAs e como diferentes passos na iniciação da transcrição podem ser regulados variando os locais de ligação do repressor.

Através dos resultados, observa-se que os passos que limitam a produção podem ter uma grande influência na cinética da produção de RNAs nas duas disposições. Descobre-se que uma maior interferência entre as RNA polimerases nos promotores divergentes com sobreposição e nos convergentes, aumentam a média e desvio padrão da distribuição dos intervalos de tempo entre produção de RNAs, provocando uma maior oscilação nos níveis de RNA.

Observa-se também que pequenas mudanças na distância entre os TSS podem provocar transições abruptas na dinâmica da produção de RNAs, principalmente na transição entre promotores com e sem sobreposição.

Dos estudos da correlação mostra-se que através da afinação das distâncias entre os TSS nas diversas disposições se pode obter tanto uma correlação negativa como positiva, quer na direccionalidade de RNAs consecutivos como nas séries temporais. Também se mostra que mecanismos de repressão distintos do início da transcrição, em tais passos como a formação dos complexos fechados e abertos e a libertação do promotor, têm diferentes efeitos na dinâmica da produção de RNAs.

Este tipo de modelos podem ajudar a explicar como os circuitos genéticos evoluíram, podendo ainda ajudar a produzir circuitos genéticos com propriedades específicas.

Palavras-chave: simulação estocástica, iniciação da transcrição, expressão genética em procariontes, mecanismos de regulação, disposição dos promotores

Abstract

The regulatory mechanisms of transcription allow organisms to quickly adapt to changes in their environment and often act during transcription initiation. Here, a stochastic model of transcription initiation at the nucleotide level is proposed to study the dynamics of RNA production in closely spaced promoters and their regulatory mechanisms.

We study how different arrangements (convergent e divergent), distance between transcription start sites (TSS), and various kinetic parameters affect the dynamics of RNA production. Further, we analyze how the kinetics of various steps in transcription initiation can be regulated by varying locations of repressor binding sites.

From the results, we observe that the rate limiting steps have strong influence in the kinetics of RNA production. We find that interferences between RNA polymerases in divergent overlapped and convergent geometries causes the distribution of time intervals between the production of consecutive RNA molecules from each TSS to increase in mean and standard deviation, which leads to stronger fluctuations in the temporal levels of RNA molecules.

We observe that small changes in the distance between TSSs can lead to abrupt transitions in the dynamics of RNA production, particularly when this change changes the geometry from overlapped to non-overlapped promoters.

From the study of the correlation in the choices of directionality and on the time series of RNA productions we show that by tuning the distances and directions of the two TSS one can obtain both negative and positive correlations. We further show that distinct repression mechanisms of transcription initiation in steps such as the open and closed complex formation and promoter escape have different effects on the dynamics of RNA production.

The study of these models will help the study of how genetic circuits have evolved and assist in designing artificial genetic circuits with desired dynamics.

Keywords: stochastic simulation, transcription initiation, prokaryotic gene expression, regulation mechanisms, promoter arrangements

Acknowledgements

First, I would like to thank my supervisor Prof. Dr. José Manuel Fonseca, who offered me this great opportunity of doing my Master Thesis project and gave me a great support.

To Prof. Dr. André Sanches Ribeiro, who welcomed me in his group and provided a great assistance during my stay, both at supporting and providing me with all the tools necessary for this work.

To Jarno Mäkelä, who helped me a lot during this Thesis, especially in the section of the results, where we discussed what type of results we needed for this work.

To Jason Lloyd-Price and Antti Häkkinen who provided me with the tools necessary for this work. Jason gave me the SGNS simulator, where we made all ours simulations and gave me some important insights in stochastic simulations which I used in the Chapter 2.1. Antti gave me a template of a model in which I created this model for transcription initiation.

To Abhishekh Gupta, who were always present and occupying the “Grid”, but also helped me find a home in Finland and invited me to so many sport games with his friends.

To everyone in the LBD group, who were always there to help me but also provided for times of fun and relaxation, especially during Bomberman.

To all my friends and especially the ones I met 5 years ago, when we all started our studies in Biomedical Engineering, because they made this 5 year journey so much more fun and pleasant. Joaquim Horta, Hugo Pereira, Sérgio Mendes, Pedro Martins, Luís Mendes, Fernando Mota, Bernardo Azevedo, Nuno Fernandes, Filipe Oliveira, João Martins, Pedro Cascalho, João Santinha, Mafalda Fernanda, Ana Frazão, Sara Gil, Rita Rosa, Ana Marques, Ana Margarida, Susana Gaspar, Susana Martinho, Cátia Rocha, Milene Bação, and to everyone else I want to show my greatest gratitude for all the good times spent with you guys and hopefully we shall continue this great friendship after completing this journey together. A picture is worth a thousand words, but sometimes we just need 4 to express our feelings: I love you guys!!!

A special thanks to Joaquim and João Santinha for sharing their house with all their friends in the all night studies or just when we needed a warm couch.

Finally I thank my family, especially my parents and my brother who always gave me strength to overcome any problem and who gave me the support to make this important move for my education.

Abbreviations and Symbols

λ phage	Lambda bacteriophage
$a_\mu(x)$	Propensity function
APR	Abortive to productive ratio
bp	Base pairs
CME	Chemical master equation
CV	Coefficient of variation
DNA	Deoxyribonucleic acid
EC	Elongation complex
E. coli	*Escherichia coli*
FF	Fano Factor
FRET	Förster resonance energy transfer
In vivo	Latin for "within the living"
In vitro	Latin for "within glass"
ITC	Initial transcribing complex
$h_\mu(x)$	Possible reactant combinations in the reaction volume
ODE	Ordinary Differential Equation
RNA	Ribonucleic acid
RNAp	Ribonucleic acid polymerase
$\mathrm{RP_c}$	Closed complex of RNA polymerase and the DNA
$\mathrm{RP_i}$	Isomerized complex of RNA polymerase and the DNA
$\mathrm{RP_o}$	Open complex of RNA polymerase and the DNA
RRE	Reaction Rate Equation
SSA	Stochastic simulation algorithm
TFBS	Transcription factor binding site
TSS	Transcriptional start site
UV	Ultraviolet
x	Vector of the current concentrations of all the molecules

Contents

Figure contents

Table contents

Chapter 1. Introduction

The objective of this work is to design a stochastic model of transcription initiation at the nucleotide level and use it to study the dynamics of RNA production in *Escherichia coli* (*E. coli*). We study the dynamics between pairs of closely spaced promoters and also single unidirectional promoters for a comparison.

Cells are able to respond to environmental changes using regulatory mechanisms connected to their genetic circuits. This regulation often takes place at the stage of transcription initiation [1], which is the first step of transcription. Transcription is the process of synthesizing a Ribonucleic acid (RNA) transcript using the information contained in a determined region of the Deoxyribonucleic acid (DNA). This process is executed by the RNA polymerase (RNAp) and is present in both eukaryotes and prokaryotes, such as our model organism: *E. coli*.

The stochastic nature of gene expression leads to cell to cell variability in the number of RNA and protein molecules in monoclonal cell populations [2]. Fluctuations in the RNA and protein numbers over time have been observed at the single cell level [3, 4]. These fluctuations were detected in transcription using single-cell experiments [5] and have an important effect in the behavior of the cell.

Researchers have recently developed tools to model and simulate biological processes at the single event level using the stochastic simulation algorithm (SSA) [6]. These models [7, 8] have shown the ability to predict the statistics of such processes that have a stochastic nature, which would not be possible using deterministic kinetics.

The first stochastic model [7] of gene expression considered transcription to be an instantaneous process as a first approximation, but the execution of this process by the RNAp can take some time. To account for this, time delays were added to the model of transcription [9]. A delayed stochastic model of transcription at the single nucleotide level [8] was then proposed, which included dynamically pertinent events in elongation such as transcriptional pauses, error correction, arrests, premature termination and collisions between elongating RNAps. In that model transcription initiation was modeled as a delayed event whose duration followed a Gaussian distribution, to account for the rate-limiting steps inherent to this process [10].

In this work, using *E. coli* as a model organism, we follow the same strategy to model transcription initiation for bacteria [9]. This model includes explicitly the steps of promoter search [11, 12, 13, 14], transition of the RNAp and DNA to a closed complex, isomerization of the closed complex, which leads to the open complex formation [15, 16, 17]. When attempting to start a productive elongation process, the model accounts for the occurrence of abortive initiation events [18, 19, 20]. In this process, the RNAp releases a small and incomplete transcript and then cycles back to the open complex state. Also, it is noted that in this event the RNAp does not move but rather stays in the same position and "scrunches" the DNA [21, 22]. When the energy accumulated inside the RNAp is enough to break the promoter bond, the RNAp escapes the promoter and starts the elongation process.

Our model of transcription initiation at the nucleotide level, which includes all the reactions described earlier, has a great ability to study the dynamics of RNA production in promoters that are closely spaced, one event at a time. Further, we can use it to test the effects of varying the distance between transcription start sites (TSSs) or the influence of the directions of transcription of each promoter.

These closely spaced promoters can be organized with regard to the direction of transcription (which we will describe as "promoter arrangement"). In terms of arrangement, promoters can be tandem, convergent or divergent [1]. Closely spaced tandem promoters are oriented in the same direction, as they usually transcribe the same gene or operon. Convergent promoters are arranged in a face to face fashion and, thus, have a common region of elongation. Finally, divergent promoters have a common binding region for the RNAp and transcribe in opposite directions.

Closely spaced promoters have been found to be abundant in all simple organisms, ranging from viruses to chloroplasts and bacteria [23]. On the other hand, the eukaryotic genomes are generally less dense and less organized than prokaryotes, so the finding of such promoters in this organisms, as for example in the human PCNA locus [24], was less expected. It is now known that bidirectional promoters are also a common gene organization in higher order organisms, including humans [25, 26].

The use of a model at the nucleotide level enables the study transcription regulation at that level and the investigation of how the location of such regulators can affect the dynamics of RNA production [1, 27]. Using different transcription factor binding sites (TFBSs), it is possible to study the mechanisms of transcription regulation at various levels of transcription initiation [28]. Due to the interference between colliding RNAps we expect that closely spaced promoters have more complex gene expression patterns [29] and so can have a more complex regulation and so to understand such a complex system, using a detailed at a single nucleotide level is the best approach.

With this model, it is possible to simulate all possible arrangements in these types of promoters, it is also possible to modify the number of nucleotides in the binding region, the nucleotides between TSSs, among other parameters and study how the dynamics of RNA production changes as a function of these parameters. It is achievable to modify the features of the repression mechanism, such as size and location of the TFBS. Finally, it is also possible to modify the kinetic constants of the rate limiting steps for both promoters independently, such as the open complex formation, in an independent fashion.

Using all this different parameters we investigated how asymmetric features affect the relative expression levels of both genes separately. To study the dynamics of RNA production, we characterized it in terms of mean and noise in RNA numbers, distribution of time intervals between consecutive production events, the correlation between choices of direction of consecutive elongation events and correlation between time series of RNA numbers.

In Figure 1.1 it is presented a scheme with the location where this work was done (TUT in Finland). This work was done in collaboration with the LBD group, who studies the dynamics of prokaryotic gene expression and gene regulatory network. This work appeared because in previous models, transcription initiation was modeled as a single delayed event.

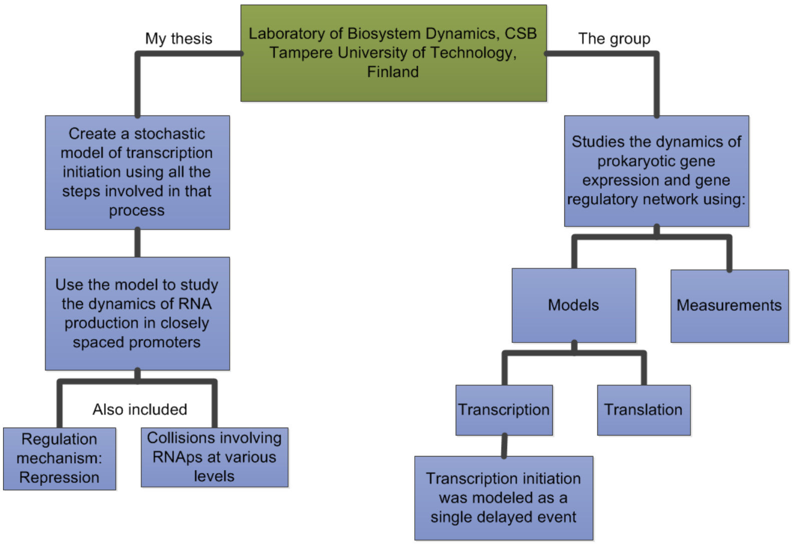
Laboratory of Biosystem Dynamics, CSB
Tampere University of Technology,
Finland
My thesis
The group
Create a stochastic model of transcription initiation using all the steps involved in that process
Use the model to study the dynamics of RNA production in closely spaced promoters
Also included
Regulation mechanism: Repression
Collisions involving RNAps at various levels
Studies the dynamics of prokaryotic gene expression and gene regulatory network using:
Models
Measurements
Transcription
Translation
Transcription initiation was modeled as a single delayed event

Figure 1.1 – Scheme with the objectives of this work (left side), the motivation behind this work (right side) and the location where this work was done (green box).

This thesis is divided in 5 more Chapters. Chapter 2 introduces a few notions behind the main topics of this thesis: stochastic simulation of chemical reactions, the execution of transcription initiation by the RNAp and regulation mechanisms of closely spaced promoters and we also present a distribution of the distance between the TSSs of such promoters. In Chapter 3 we present a description of the stochastic model of transcription initiation at the nucleotide level, including all the reactions that we used and their description but also their standard kinetic parameters used for this study. Chapter 4 contains the results on the study on the dynamics of RNA production as a function of different parameters, the study on the dynamics of RNA production in different types of arrangements and as a function of the distance between TSSs and the dynamics of RNA production of different repression mechanisms. Chapter 5 presents the conclusions of this work and also the future developments perspectives. Finally, Chapter 6 contains all the references used for this work.

Chapter 2. Theoretical framework

In this chapter we give a theoretical description of the framework behind the main topics of this work. Particularly we give some insights on how stochastic simulations appeared and evolved throughout the years and why this type of simulations is considered to be very useful in studies of biological processes such as gene expression. Then we describe the steps involved in transcription initiation, including a structural description of the existing RNAps and their interactions with the DNA during these steps. Finally we also have a topic on closely spaced promoters and how they can be organized in different arrangements and their influence in gene regulatory mechanisms.

2.1. Stochastic Simulation Algorithm

Modelling is an approach to characterize the state of the elements inside a system and the interactions between those elements. Models using the present knowledge of a system can help in testing if our understanding of that particular system match the data obtained in experimental procedures.

From the point of view of classical mechanics, systems of chemical reactions are considered as deterministic, because it is possible to predict the evolution of the system. The deterministic approach has slight problems in explaining such processes that have a stochastic nature, for example gene expression and their regulatory mechanism [3, 4, 5]. Thanks to the impossibility of calculating the exact moment that an event takes place, although we can estimate that probability.

Solving a system of coupled ordinary differential equations (ODEs) is the most conventional way of describing and simulating (using the deterministic approach) the behavior of molecules reacting in a homogeneous and thermally equilibrated mixture. This system has one equation for each of N active chemical species in the volume, where each equation describe changing rate of the concentration X_i of each chemical species S_i taking in consideration the concentration of the other species, stoichiometry and reaction constants of the R channels through which they interact. This set of equations forms the reaction rate equation (RRE) [30]:

$$d\boldsymbol{x} = \sum_{\mu=1}^{R} v_\mu r_\mu(\boldsymbol{x})dt \tag{2.1}$$

Here v_μ is the vector that describes the stoichiometry of reaction R_μ and r_μ is considered as the mean 'rate' at which the same reaction occurs as a function of the vector of the current concentrations of all the molecules: $x = \left(X_1(t), \dots, X_N(t)\right)$.

As stated before, some biological systems have a stochastic nature and so due to this reason a new approach is needed. The stochastic approach instead of calculating the exact moment that an event takes place, it is involved in estimating the probability $P(x,t|x_0,t_0)$ of having the given concentration x in the reaction volume at time t after the initial concentrations of all the molecules and the initial time of reactions: x_0 and t_0 respectively. Using the probability that the molecules x in the volume at time t react in the next infinitesimal time interval (*t*+*dt*) via reaction R_μ, which is defined as the propensity function or $a_\mu(x)$.

The stochastic approach can be expressed as a partial-differential equation also known as the chemical master equation (CME) [31], which is also known as the Kolmogorov forward equation for a stochastic kinetic process [32]:

$$\frac{\partial P(x,t|x_0,t_0)}{\partial t} = \sum_{\mu=1}^{R}\left[a_\mu(x-v_\mu)P(x-v_\mu,t|x_0,t_0) - a_\mu(x)P(x,t|x_0,t_0)\right] \qquad (2.1)$$

Here v_μ also corresponds to the absolute number of all the reactants that change when the reaction R_μ occurs.

This propensity function can be written as $a_\mu(x) = h_\mu(x)c_\mu$, where $h_\mu(x)$ is the number of possible reactant combinations in the reaction volume and $c_\mu$is a kinetic constant such that $c_\mu dt$ gives the probability that in the next infinitesimal time *dt* a determined molecule will spontaneously react via R_μ. Different types of reactions can have different $h_\mu(x)$ as can be seen in table 2.1.

Table 2.1 – Various values of $h_\mu(x)$ as a function of the type of reaction. Taken from [33]

Type of reactions	$h_\mu(x)$
$\rightarrow$ ***products***	1
$S_i \rightarrow$ ***products***	X_i
$S_i + S_j \rightarrow$ ***products***	X_iX_j
$2S_i \rightarrow$ ***products***	$\frac{X_i(X_i-1)}{2}$
$\sum_{i\in S_\mu} N_{(i,\mu)}S_i \rightarrow$ ***products***	$\prod_{i\in S_\mu}\prod_{q=1}^{N_{(i,\mu)}}\frac{X_i-q+1}{q}$

Numerical simulations of complex systems is an easier way to describe the time evolution of such systems, where solving analytically the CME or RREs is a very a very difficult task, so this is way analytical simulators started to be more utilized. The Stochastic Simulation Algorithm (SSA) [34] is a Monte Carlo method that simulates numerically the time evolution of well stirred reaction systems. The time goes forward in discrete steps and in each step a reaction is explicitly executed and the effect on the number of each molecule is settled. Probability distributions are used to determine the time of the next reaction.

The SSA produces in a single run one of the possible exact temporal trajectories of the CME and can be represented by the following steps:

1. Initialization: Define R reactions rates (k_1,…,k_R) and the initial molecule number x (x_1,…, $x_{N)}$ and then set time t = 0 and reaction counter n=0.
2. Calculate R propensities using the current population of molecules, $p_1=k_1 \bullet h_1, \ldots,$ $p_R=k_R \bullet h_R$ where h is the number of all possible distinct molecular interactions in the current state(see table 2.1 for different types of *h).* Calculate $p_0 = \sum_1^R p_i$ and store all the propensity values.
3. Calculate the pair (τ, μ) using two random numbers r_1 and r_2 (from a *U*(0,1) uniform distribution, using $\tau = \ln(1/r_1) \cdot (1/p_0)$ and μ has to satisfy the following: $\sum_1^{\mu-1} p_i < r_2 \cdot p_0 < \sum_1^{\mu} p_i$
4. Calculate the actual value of t using the pair (τ, μ) and adjust the reaction counter by one.
    a. If $t + \tau \geq t_{stop}$, end the simulation.
    b. If $t + \tau < t_{stop}$ then set t = t + τ, and update the molecular numbers according to the type of reaction that occurred using $x = x + v_\mu$.
5. Go back to step 2.

Note that the procedure in step 3 is done by one of the original formulations of the SSA, the Direct Method [34] which is computationally less intensive then the other formulation: the First Reaction Method. For large systems, these methods can be computationally heavy and so other methods started to appear in order to that improve the computational performance without affecting its exactness, and so the Next Reaction Method [35] and the Logarithmic Direct Method [36] were proposed. The First Family Method [30] is a generalization of the above methods and has the advantage of being able to choose either the Direct Method or the First Reaction Method on step 3 of the SSA based on the total propensities of the reactions (which are grouped into “families”).

Finally as previously described, the need to account with the duration of processes that take non-negligible to complete, such as transcription, led to the modification of the SSA [35], where time delays started to be added to account for those durations. These delays can be implemented using a “wait list” or modifying step 4 of the SSA to account for the time of the release while updating the new molecular numbers.

The delayed SSA was used to model transcription at the nucleotide level [9] and was shown to match [37] the dynamics of RNA and protein production at the single RNA and protein level [38, 39].

2.2. RNA polymerase and the execution of transcription initiation

Transcription initiation is the first step in transcription and is executed by an important enzyme: the RNAp. Due to this, the structure of the RNAp and how that structure affects its function in the transcription process is an important feature to understand in this process.

Different organisms have different types of RNAps, which can be divided into single-unit multi-subunit RNAps. The single-unit RNAp is normally associated to virus and one of the most studied examples of this type comes from the T7 bacteriophage RNAp. Multi-subunit RNAps are associated to eukaryotes, bacteria and archaea, although there are major differences in the RNAps within those domains, even though there is evidence of a correspondence between structure and function of various subunits found in archaea and eukaryotes compared to bacteria [40, 41]. In this case, one of the most studied examples comes from the *E. coli* RNAp.

Both the T7 and *E. coli* RNAp have been the focus of single-cell studies both structurally and functionally. The core structure of the *E. coli* RNAp (formed by five subunits: two α, β, β' and ω) and its relevance to transcription initiation were studied at a resolution of 15 Å using cryo-electron microscopy and image processing of helical crystals [42].

Other bacterial organisms have been used to gather information on *E. coli* RNAp using higher resolutions, such as the *Thermus aquaticus* [43, 44] or the *Thermus thermophilus* [45]. On the other hand T7 structure has been studied at a resolution of 3.3 Å [46] and despite the structural differences, evidence of a similar functional mechanism of the RNAp was found in both organisms using structural and functional studies [47, 48, 49], making T7 RNAp also suitable for mechanistic comparisons with the *E. coli* RNAp in the process of transcription.

Transcription initiation in bacteria starts with the localization of the promoter-specific region (also known as promoter search) [11, 12, 13, 14], which in bacteria requires the binding a polypeptide (called σ factor) [50] to the core enzyme, forming the holoenzyme, reducing the affinity of the RNAp for nonspecific DNA and increasing the affinity to various promoters. T7 RNAp also locates promoters using a similar mechanism but doesn't need the binding of additional polypeptides due to its high affinity to specific T7 promoters [51].

The housekeeping sigma factor (σ^{70}), which is considered to be the most important σ factor, recognizes the core bacterial promoter. The core bacterial promoter has two hexameric motifs, normally centered close to the -10 and -35 positions (or at those positions) relative to the TSS [1]. In addition to these motifs, some promoters also have a sequence enriched with adenine and thymine, upstream of the -35 element (that is normally addressed as the "UP element") [52] . The "UP" element is recognized by both α subunits [53].

Other σ factors, which are normally related to genes that respond to stress situations like UV radiation or heat shock, recognize less common promoter motifs, as they are only needed in special conditions [50]. In figure 2.1 we show the interaction between the various subunits of RNAp and the promoter motifs.

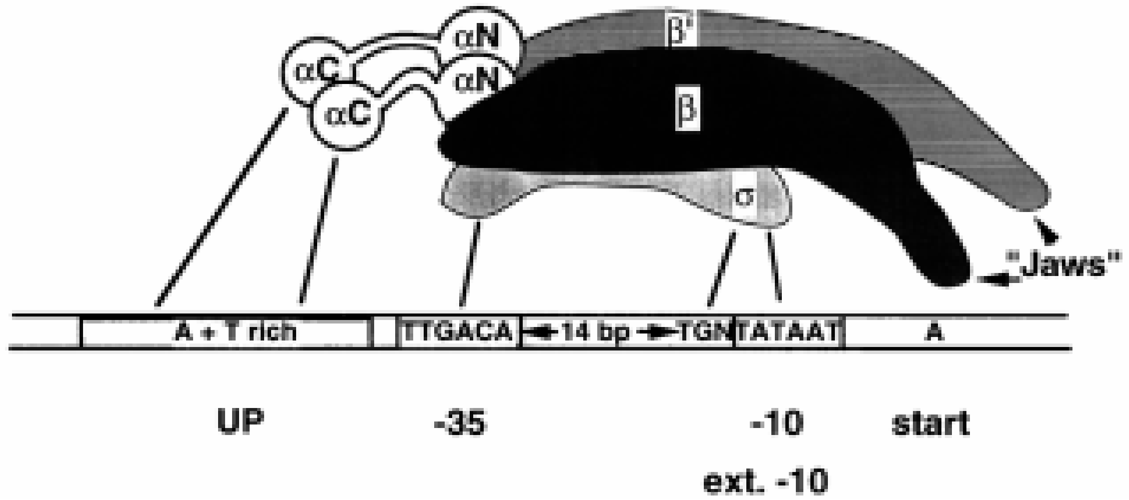

Figure 2.1 – The interaction between the subunits of RNAp and the core promoter. The consensus sequences for the -10 and -35 promoter motifs and the "UP element" are also shown. The contacts between RNAp and the promoter are shown in solid lines. Figure taken from [54].

Eukaryotes have several types of RNAps, depending on the type of the synthesized RNA. The RNAp II is the most studied one and as other eukaryotic RNAps, RNAp II alone doesn't recognize the eukaryotic promoter motifs, namely, TATA box, CCAAT-box and GC-box and others [55]. Due to this the binding of initiation factors is required before transcription initiation starts, which leads to a high level of control over transcription [56]. This control is also associated with the binding of other accessory factors, transcriptional activators and co-activators that regulate the rate of RNA production from each gene in response to different conditions [56]. Since in this paper we focus on bacterial transcription initiation, we will not enter into a detailed vision of the eukaryotic mechanisms.

The localization of the promoter in prokaryotes results in conformational changes in the DNA and the RNAp, which leads to the formation of a closed complex (RP_c). These changes include the DNA bending to wrap of upstream DNA and loading the downstream DNA in the active site of the RNAp [57, 58]. This mechanism proceeds with the RNAp loading of the template and non-template DNA positioned at the TSS, the unwinding of the double stranded DNA, the assembly and tightening of the RNAp clamp [58]. This stage is called isomerization and can be simplified into just one single isomerized complex (RP_i). This step can also be considered a rate limiting step, as observed in studies of the *lac*UV5 promoter of the *in vitro* kinetics of transcription

[15]. A real-characterization of these isomers in the T7A1 promoter, using *E. coli* RNAP [59] showed that the transition between the various intermediates is very fast, supporting this simplification. This mechanism finally proceeds with the formation of the open complex (RP_o).

Before starting elongation and escaping the promoter (which is the last step of transcription initiation) the initial transcribing complex (ITC) is involved in a repetitive cycling of the RNAp back to the open complex, and releasing short abortive RNA transcripts (that normally range from 2 to 16 nucleotides, but sometimes it is observed aborted transcripts up to 20 nucleotides) [60, 61], this process is called abortive initiation. Recent studies showed this process is involves a "Scrunching" mechanism [21, 22], where the RNAp doesn´t move forward, but "scrunches" the downstream duplex DNA in a through formed by subunit β′ and enclosed on top by the subunit β. This process accumulates energy, necessary to break the bond between the RNAp and the promoter leading the RNAp to escape the promoter and turning into an elongation complex (EC), capable of producing a full RNA transcript after termination of the transcription and the promoter available to receive another RNAp and starting the initiation process all over again.

In a recent study [62], abortive initiation was also detected *in vivo* using the bacteriophage T5 N25 promoter using the *E. coli* RNAP, which was an important find because until that, this process was only observable *in vitro*.

Note that the sigma factor has also an important role in abortive initiation [63], so after the RNAp starts elongating, the σ factor is released stochastically [64] and can bind to other RNAps helping in the location of other promoters (defined as the σ cycle). Mooney and colleagues found that sometimes the sigma factor is not released until termination and that it can be used as an elongation regulator [65]. We decided to not include specifically the σ action into our model due to the complexity associated with using all the different observations of the previous models.

To model transcription initiation we need not only spatial information about the process, but also temporal. The duration of some of the steps previously described can vary widely between different promoters [11], even when their sequences only differ by one or two nucleotides [66]. It has also been studied that for example just one change in the nucleotides base in the spacer region between both promoter motifs can also change dramatically the promoter activity [67]. For example the abortive initiation and promoter escape durations are dependent on the promoter interactions (and in this case, the stronger the promoter the slower is this process) but can also vary due to small changes in the downstream region (initial transcribing sequence) [60, 61]. Temperature can also affect the duration of these steps [15, 68] as well as the concentration of Mg^{2+} [68, 69] or the concentration of K^+ [68].

In figure 2.2 we show a representation of several steps that take place during transcription initiation.

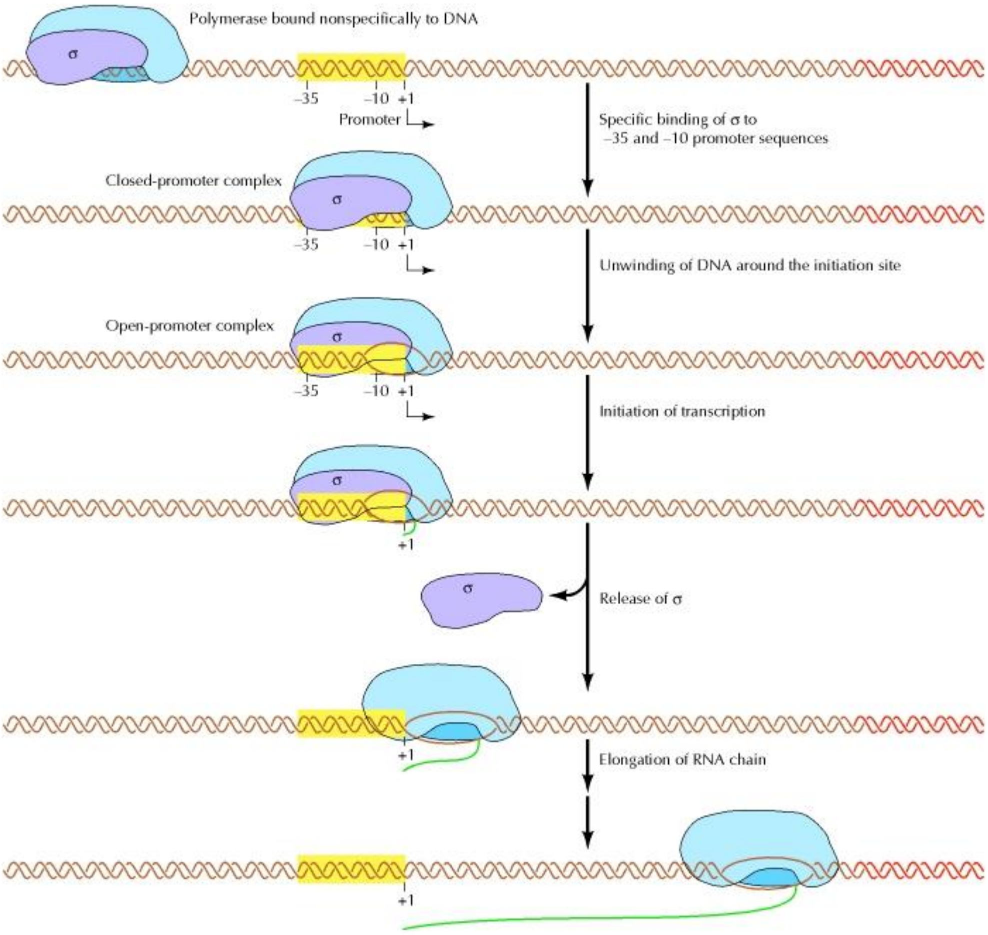
Polymerase bound nonspecifically to DNA
σ
−35
−10
+1
Promoter
Specific binding of σ to −35 and −10 promoter sequences
Closed-promoter complex
Unwinding of DNA around the initiation site
Open-promoter complex
Initiation of transcription
Release of σ
Elongation of RNA chain

Figure 2.2 – The RNAp can bind nonspecifically to DNA and searches for the promoter. The σ^{70} subunit recognizes to the -35 and -10 promoter motifs and forms a closed-promoter complex. The RNAp then unwinds DNA around the initiation site forming an open complex. Then transcription is initiated followed by the release of sigma and the elongation of the RNA chain. Figure taken from [70].

We note that in figure 2.2 the “UP element” and its interaction with subunits α (which we note that some promoters actually don’t have) is not represented and also some steps of transcription initiation like the abortive initiation are not included.

2.3. Closely spaced promoters and regulation mechanisms

In the previous section we talked about how transcription initiation starts with the localization of promoters and how the interaction between the RNAp and those promoters affect this process. In this section we will emphasize on the actual location of such promoters in the DNA template, as this location can also play an important role in regulatory mechanisms [1].

The compilation of several promoters in *E. coli* [71, 72] led to an organization of closely spaced promoters in terms of arrangement (direction of transcription), and as it was mentioned, there are three types of arrangements: tandem, convergent or divergent promoters [1]. The last two types of arrangements can also be designated as bidirectional promoters, since transcription is done in two different directions. In figure 2.3 we exhibit the three possible arrangements of closely spaced promoters. Note that in some cases, the promoters P1 and P2 can be overlapped.

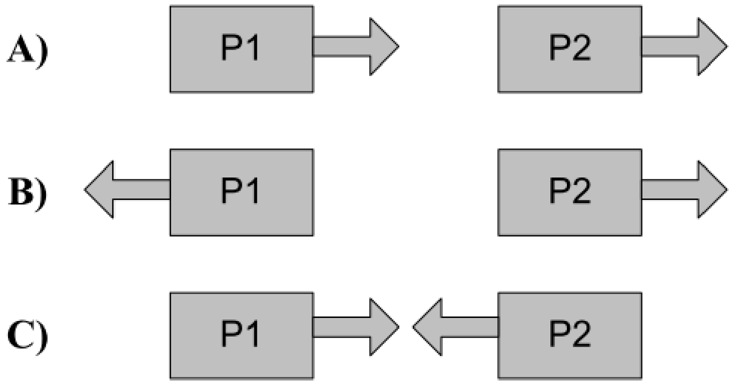

Figure 2.3 – Possible types of closely spaced promoter arrangements. There are three types of arrangements: tandem (A), divergent (B), convergent (C) promoters.

The first divergent promoters to be discovered were the promoters P_R and P_{RM}, in the lambda bacteriophage (λ phage) [73]. These promoters are one of the most studied cases of gene regulatory network [74, 75] and since the in the decision between two different paths in this network involves a random process, they were also the reason why stochastic models started to be used in such studies [7]. The studies in λ phage lead to the discovery of other divergent promoters present in *E. coli*, for example the arginine gene (arg_E and arg_{CBH} promoters) [76] or the P_C and P_{BAD} promoters of the L-arabinose operon, which have been extensively studied by Schleif and colleagues for more than 30 years [77, 78].

A compilation of closely spaced promoters found several promoters in various simple organisms such as bacteria and their viruses (for example the phages), mitochondria, chloroplasts and viruses of eukaryotes [23] showed that these types of promoters represent a general type of gene organization in this type of organisms. We should mention that Beck and colleague [23] used a different notation for the promoter arrangement, as they address the convergent arrangement as “face-to-face” promoters and divides the divergent promoters into “back-to-back” and “overlapping” promoters.

With the completion of the genome sequencing in *E. coli* K-12 (one of the most used strains) [79] most promoters and their functions were identified. The location of such promoters have a regulatory role in transcription, as a recent statistical analysis [80] revealed that both operons that regulate each other and operons that are co regulated tend to be in close distance of each other and showed a tendency of divergent promoters in this type of regulation.

The genome in simple organisms is considered to be highly organized and dense, as for an instance the *E. coli* genome with a 4.6 Mbp sequence has around 4000 genes [79]. With an average distance of 118 bp between genes, the genome is made mostly of coding DNA. The human gene, which is less dense then the *E. coli* genome, contains a 2.9 Gbp sequence with an estimation of around 39000 genes [81]. This means that for a mean spacing of around 75 kbp for each human gene there is an average gene size of 27 kbp.

This implies that there is a large amount of noncoding DNA present in the sequence. Since the eukaryotic genome isn't considered to be as organized as in bacteria, it was a surprising discovery when Adachi and Lieber identified bidirectional genes in the human chromosomes 21 and 22 [25]. A subsequent study [26] showed a prevalence of bidirectional genes in the human genome and among mouse orthologs, which means that this prevalence is often conserved along the species evolution.

We should notice again, that another notation for the promoter arrangement was used by Adachi and Lieber [25]. They address tandem orientation as "head-to-tail", convergent as "tail-to-tail" and divergent as "head-to-head". To avoid the confusion of having various notations, we decided to use the "tandem, convergent and divergent" notation in the rest of the work, and also use divergent overlap, when the distance between TSS only allows one RNAp to transcribe at a given time, which we consider to be of 110 or less nucleotides This is because we consider the size of the RNAp as 55 nucleotides when in diffusion process (note that we use a different size for elongation, which will be focused on more detail in the model section) based on footprint studies [52, 82, 83, 84, 85], where the RNAp protects around 50 to 60 nucleotides of the DNA template in the upstream region (there is also a protection of the downstream region with the opening of the DNA strand leading to the open complex formation).

In this thesis we focus on *E. coli* bidirectional promoters, considering both the divergent and convergent arrangements, hence we extracted 897 known and 4010 predicted promoters (specifically recognized by σ^{70})in *E. coli* from RegulonDB database (version 7.0) [86].

From the predicted promoters, there were several promoters of the same gene so we counted only the first promoter (p1 in the database) of each gene to avoid repetition of the same genes. From this, 1671 predicted promoters were extracted for this study, for a total of 2568 promoters. Using this number of promoter we found 258 pairs of divergent promoters and 186 pairs of convergent promoters with a distance between their TSSs lower than 800 nucleotides.

The distributions of nucleotide distance between the TSSs for both geometries are shown in figure 2.4. We observe that the bulk of the distribution of distance between adjacent TSSs is below 200 nucleotides (88.8% for convergent and 61.8% for divergent). The mean distance (in nucleotides) for convergent is 108.4 and 225.7 for divergent, therefore, this range of was used as a reference for our models.

We can also see from figure 2.4, that there is a good match between measured and predicted lengths distributions.

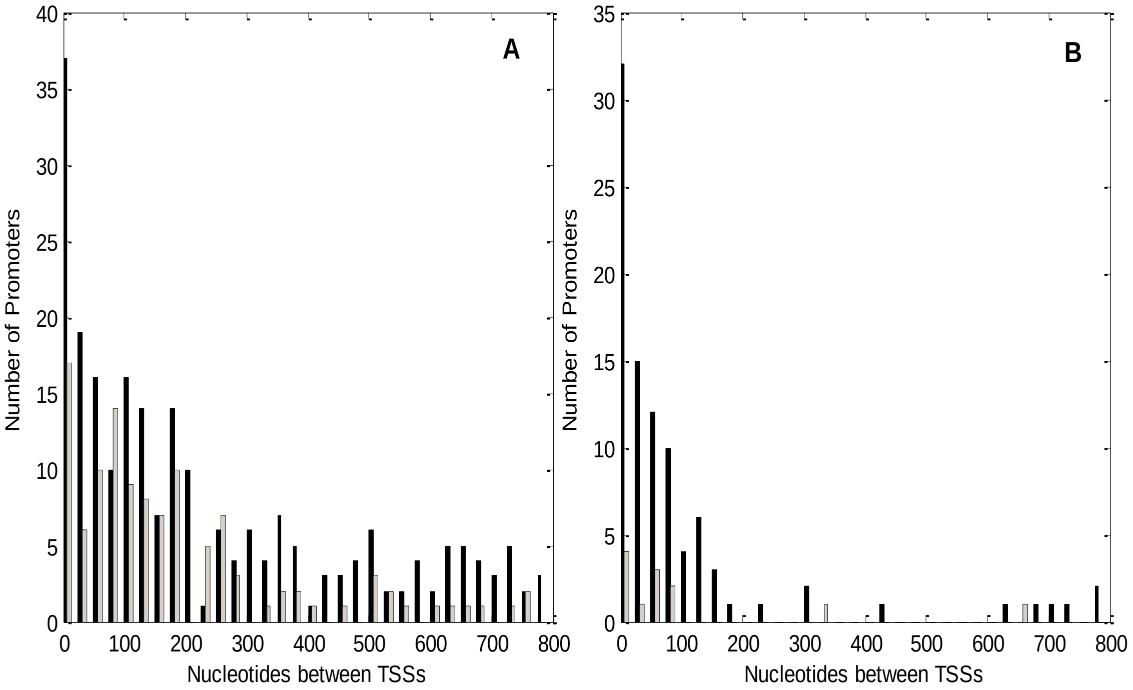

Figure 2.4 – Numbers of promoters pairs with a given number of nucleotides between the TSSs for (A) divergent promoters, and (B) convergent promoters. Gray bars are used for the known promoters and black bars are used for the predicted promoters.

This type of promoters can also be classified depending on the function of the gene products. Beck and colleague [23] classifies them as S-S if both transcripts determine structural polypeptides, R-S if one transcript determines a regulatory molecule and the other a structural polypeptide, or R-R where both transcripts determine regulatory molecules. These regulatory mechanisms [27] can be positive (activation), negative (repression), or depending on the circumstance can be of both types.

In this work we focus on repression mechanisms, as repressing the gene expression at the transcription level is probably the most used method of controlling the RNA production. Steric occlusion is probably the most common and simple method of repression, as in this method the repression molecule binds to the DNA and prevents the RNAp to start transcription [87, 88]. This repression can prevent specific steps in transcription depending on their TFBSs and size (such as the binding to the promoter, the transition from closed to open complex, or the promoter escape) [28].

It has been shown that the most common TFBS position is near the TSS (both in downstream and upstream regions) [88]. Our model allows us to choose the TFBS location and repressor size to be able to study the effect of repressor location and repression of specific steps, not only in single promoters but also in bidirectional promoters. Multiple TFBSs on the same DNA template are also allowed.

Chapter 3. Materials and Methods

In this section we present the description of the reactions used to create our model, the respective kinetic rate constants and delays. The calculations used to characterize RNA production is also presented in this section.

3.1. Creating the model

Modeling transcription initiation traditionally consist of three overall steps: binding of RNAp to promoter, open complex formation and the promoter escape [1]. Some of those steps present in that type of model are undoubtedly oversimplified as the numerous small steps involved in each of them are integrated into single reaction.

We use the proposed strategy [9] to model the dynamics of transcription initiation in bidirectional genes, considering both convergent and divergent promoters [1]. In figure 3.1 we show two model representations of bidirectional promoters and some of the reactions that used in this model. Locations within the promoter region are designated by the position relative to the TSS at the right side. The position of this TSS is set to be +1, positions to the left are negative and to the right are positive (note that by convention in Biology the first upstream nucleotide before the TSS is at position -1 and the TSS is the position +1, meaning that there is no position 0).

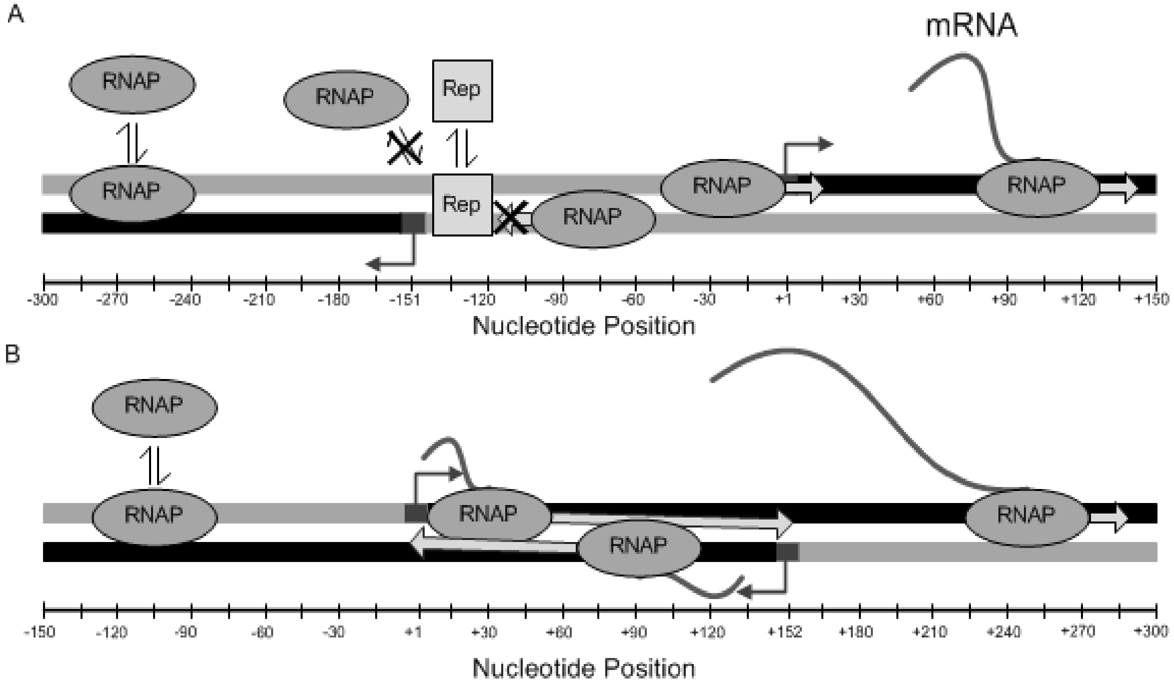

Figure 3.1 – Structure of (A) divergent promoters, and (B) convergent promoters, where binding region is gray, elongation region is black and the angled arrows presents transcriptional start sites (TSSs). Harpoons represent the binding and unbinding of RNAps and repressor molecules, the x-axis represents the nucleotide position in relation to the TSS of the “right” promoter.

In figure 3.1 the regions of elongation (which the RNAp can also percolate by diffusion) are represented in black, while regions where only diffusion can take place are represented in gray. Elongating RNAps are represented with an elongating RNA chain, which are not present in diffusing RNAps. RNAps bound to the DNA template have an arrow that represents the current direction of movement.

If a repressor is bound to the DNA template (at the TFBS), it blocks the movement of both diffusing and elongating RNAps. In figure 3.1 (A) the TFBS is represented between position -140 and -120 on the template. A bound repressor also prevents the binding of RNAps to that region, and that region alone. This repression is represented using a cross over the movement arrows of the RNAp.

In figure 3.1 (A), it is illustrated two divergent promoters, whose TSSs are located at -151 and +1. The “left” gene can only be transcribed by RNAps diffusing in that direction and the gene to the right can only be transcribed by RNAps diffusing to the corresponding direction. In figure 3.1 (B) it is illustrated two convergent promoters, whose TSSs are located at +152 and +1 and also differs from figure 3.1 (A) since it contains an overlapping elongation region. . The elongating RNAp removes the diffusing RNAps and RNAps bound to TSS. In case of two elongating RNAps colliding in the overlapped region, either one of them randomly dissociates or both dissociate.

The reactions, the standard rate constants and delays used to model transcription initiation are presented in table 3.1. In this table we present the binding and unbinding of the RNAp to the DNA template, the search for the promoter using linear diffusion, the closed and open complex formation, isomerization, and collisions between diffusing RNAps.

Note that in the reactions presented in tables 3.1, 3.2 and 3.3 we don’t express specifically the differences between different promoters present in our model. This means that we don´t specify if the reactions are from the left or the right side (see figure 3.1) or even if they are from convergent or from divergent promoters. The only difference between these situations will be on the indexes, so because of this we decided to only include one reaction for each event, which is much clearer.

Most reactions in the model are instantaneous, that is, once the two reacting molecules meet and react, the product is produced instantaneously and its amount updated. Instantaneous reactions are exemplified as: $A + B \xrightarrow{k} C$. In this reaction, when both A and B meet according to the rules of the SSA [6], molecule C is instantaneously produced. The expected time for A and B to meet is determined by the propensity of this reaction at each moment function (as described in section 2.1). Looking at table 2.1 (see third reaction) we can see that the propensity function of this reaction is calculated by the product of k with the molecular concentrations of A and B [6].

Some reactions need to account for the time that the processes take to occur, once initiated. Such delays in the release of products are exemplified as follows: $A + B \xrightarrow{k} C(\tau)$. When this reaction occurs, molecule C is placed on a waitlist and is only made available for reactions, after τ seconds have elapsed. We can generate τ randomly from any desired distribution each time the reaction is chosen to occur. Such delayed events are only introduced when the time that the process takes to occur is sufficiently long to affect the kinetics of the system. Substrates that are not consumed in the reaction are indicated with (*).

Table 3.1 - Chemical reactions, rate constants (in s^{-1}), and time delays (in s) used to model transcription initiation.

Identifier	Reaction	Rate constants and delays	Ref.
(1)	$RNAp + U_{[n-\Delta_D,\, n+\Delta_D]} \xrightarrow{k_b} O_n$	$k_b = 0.000075$	[12]
(2)	$O_n \xrightarrow{k_f} RNAp + U_{[n-\Delta_D,\, n+\Delta_D]}$	$k_f = 0.3$	[12]
(3)	$O_n + U_{n+\Delta_D+1} \xrightarrow{k_m} O_{n+1} + U_{n-\Delta_D}$	$k_m = 660$	[12]
(4)	$O_{TSS+\Delta_D} \xrightarrow{k_c} RP_c$	$k_c = 0.5$	[12]
(5)	$RP_c + U_{[TSS+1,\, TSS+19]} \xrightarrow{k_i} RP_i$	$k_i = 0.095$	[15]
(6)	$RP_i \xrightarrow{k_o} RP_o$	$k_o = 2$	[15]
(7)	$*O_{n+2\Delta_D+1} + O_n \xrightarrow{k_m}$ $U_{[n-\Delta_D,\, n+\Delta_D]} + RNAp$	$k_m = 660$	[12, 101]

In the reactions presented in table 3.1 RNAp stands for the RNA polymerase, U_n stand for the nth unoccupied nucleotide. Ranges of nucleotides are denoted such as $U_{[start,\ end]}$, denoting a stretch of unoccupied nucleotides from indexes *start* to *end*. As reported in footprint studies [52, 82, 83, 84, 85] the bound RNAp protects around 50 nucleotides on the template. In our model each RNAp occupies 55 nucleotides, and we name it as O_n, to account only for the active center (this means that the diffusing RNAp is occupying the range $[n-\Delta_D, n+\Delta_D]$, where $\Delta_D = 27$). Here RP_c, RP_i and RP_o correspond respectively to the closed complex, the isomerization complex and the open complex, which are all just conformation changes of the bonds between the RNAp and the DNA template.

The reactions used to model the steps that occur after the open complex formation, such as the abortive initiation, promoter clearance, the initial transcribing complex, the elongation complex formation and the collision between the elongation complexes and different RNAps at various steps is presented in table 3.2.

Table 3.2 - Chemical reactions, rate constants (in s^{-1}), and time delays (in *s*) used to model the steps that occur after the open complex formation.

Identifier	Reaction	Rate constants and delays	Ref.
(8)	$*E_{n+2\Delta_E+1} + O_n \xrightarrow{k_m}$ $U_{[n-\Delta_D, n+\Delta_D]} + RNAp$	$k_m = 660$	[12, 101]
(9)	$*E_{n+2\Delta_E+1} + E_n \xrightarrow{k_{el}}$ $U_{[n-\Delta_E, n+\Delta_E]} + RNAp$	$k_{el} = 42$	[101]
(10)	$*E_{TSS-\Delta_E} + RP_c \xrightarrow{k_{el}} U_{[TSS, TSS-2\Delta_D]}$ $*E_{TSS} + RP_i/RP_o/E_{TSS-12} \xrightarrow{k_{el}}$ $U_{[TSS+\Delta_E, TSS-2\Delta_D]}$	$k_{el} = 42$	[101]
(11)	$RP_o \xrightarrow{k_{el}} E_{TSS}$	$k_{el} = 42$	[21, 22, 61]
(12)	$E_{TSS+n} \xrightarrow{k_{el}/4} E_{TSS+n+1}$	$k_{el} = 42$ $n \leq 12$	[21, 22, 61]
(13)	$E_{TSS+n} \xrightarrow{k_a} RP_o$	$k_a = 4.2$	[21, 22, 61]
(14)	$E_{TSS+12} + U_{TSS+\Delta_E+12} \xrightarrow{k_{el}}$ $E_{TSS+13} + U_{[TSS+12, TSS+2\Delta_D+12]}$	$k_{el} = 42$	[21, 22, 61]
(15)	$E_n + U_{n+\Delta_E} \xrightarrow{k_{el}} E_{n+1} + U_{n-\Delta_E}$	$k_{el} = 42$ $n \geq 13$	[100]
(16)	$E_{n_{last}} \xrightarrow{k_{el}}$ $RNA(\tau_{el}) + RNAp(\tau_{el}) + U_{[n_{last}-2\Delta_E, n_{last}]}$	$k_{el} = 42$ $\tau el \sim G(k, 1/k_{el})$	[100]
(17)	$RNA \xrightarrow{k_d} \varnothing$	$k_d = 0.006$	[105]

In the reactions presented in table 3.2, the Ribonucleic acid is designed as RNA for and E_n stands for Elongation Complex (EC) and so during elongation the EC occupies normally around 30 to 23 nucleotides [83], we decided to use 25 nucleotides as it is also reported by [89], so because of this it occupies the range [n-Δ_E, n+ Δ_E], where Δ_E=12.

The reactions used for repression mechanism, as the mechanism of transcription regulation used for this model is presented in table 3.3.

Table 3.3 - Chemical reactions, rate constants (in s^{-1}), and time delays (in *s*) used to model repression and dissociation of repressor.

Identifier	Reaction	Rate constants and delays	Ref.
(18)	$Rep + U_{[n-\Delta_{rep},\, n+\Delta_{rep}]} \xrightarrow{k_r} R_n$	$k_r = 0.0167$	[113]
(19)	$R_n \xrightarrow{k_u} Rep + U_{[n-\Delta_{rep},\, n+\Delta_{rep}]}$	$k_u = 0.004$	[113]

In the reactions presented in table 3.3, a repressor not bound to the DNA template is designed as Rep and the repressor bound to the DNA template is designed as R_n and occupies the range [n-Δ_{rep}, n+Δ_{rep}]. Each different repressor molecule can occupy a different number of nucleotides. In this work we use a standard value of Δ_{rep} = 10 as it is the value for the principal binding site of the lac repressor [87], which is one of the most studied repressors in *E. coli*, and it is also close to other repressor footprints measured in *E. coli* [90]. We also varied the values of repressor size to study its effects in transcription regulation.

A description of the specific reactions used in our model is given below (using the respective identifiers from tables 3.1, 3.2 and 3.3). Parameter values used in the reactions were obtained from measurements in *E. coli*, directly or indirectly from the respective references column.

Once the RNAp binds to the template via reaction (1) it diffuses linearly (also known as “sliding”) [12, 91] on the DNA template (3). We point out that the binding of RNAp happens to one strand at a time thus the direction chosen by the RNAp can´t change during the sliding unless the RNAp unbinds. We didn´t include the three-dimensional diffusion in solution and intersegment transfer [92] as they are mostly means to make long transfers in the DNA template. If the RNAp is blocked by another RNAp or by a repressor, and doesn´t find a TSS, the RNAp eventually dissociates from the DNA strand (2) [12]. In our simulations we used a value for the initial concentration of free RNAp to be 28 molecules per cell [93].

If the RNAp finds the specific TSS (in theory the RNAp should find the promoter motifs, but in our model we don´t use specific nucleobases, so we use the TSS for this search), we use a set of non-delayed, consecutive, chained reactions [1, 14] to model the closed complex formation (4).

We note that when referring to closed complex formation, we define it as the conformational changes that occurs after the finding of the TSS [12, 14], and not as all steps including finding of the promoter and diffusion [1].

As stated before, because of the size occupied by the RNAp, we consider promoters with a distance between TSS of 110 or less as divergent overlap so we consider that for this case, only one RNAp can transcribe at a given time and that once it starts transcription in one direction it cannot be redirected to the other direction [94]. We point out that we don't consider special cases that have been studied in λ phage [95], where the interference between the occupancy of an RNAp at one of the promoters (P_R) and the other promoter (P_{RM}) was studied. It was found that for a very specific distance between the TSSs this interference was greatly diminished, allowing RNAps at both promoter sites to start transcription at the same time.

The next step in this process is isomerization (5), where the RNAp structure changes, and occupies around 20 more nucleotides (from +1 to +20) as reported by DNA footprints [82, 83, 84, 85] and finally the formation of the open complex is done (6). At this step the RNAp occupies around 75 nucleotides (from -55 to +20).

We didn't included the reverse reactions in the open complex formation because on strong promoters, open complexes are much more energetically favorable, and the transition from closed to open complex is essentially irreversible [85], but we point out that these reactions can be easily added to our model if we actually needed. There are also other models of transcription initiation that don´t use the reverse reactions after the formation of the closed complex [96]

Following the formation of the ITC via reaction (11), the RNAp doesn´t move forward, but "scrunches" the DNA (12) [21, 22]. This process continues until the energy inside the RNAp is enough for it to be able to break the bonds with the promoter (14). We added the abortive path (13) which competes with the scrunching path (12), returning to the state of open complex.

A model of the abortive initiation step in transcription initiation steps, which includes the abortive path, scrunching path but also an escape path was proposed by Xue and colleagues [97]. This is a detailed model, where for every different path at different positions have specific kinetic parameters. Unfortunately the kinetic parameters were calculated for T5N25 and T7A1 (bacteriophage) promoters, which are considered to be much faster initiators then for example the *lac*UV5 promoters [18] even though the interaction with such different promoters share similar mechanisms in the abortive initiation [19]. For simplicity we used the same kinetic parameters at every nucleotide and we only allow the escape after the scrunching of the 12th nucleotide.

Note that in this model we don´t include a path for unproductive ITC´s, where the RNAp escapes the promoter very slowly, if at all. This process was observed in few cases, thus it doesn't have a profound effect to the RNA production. [97, 98]

We tested an abortive rate ranging from 1 to 10.5 s^{-1} and decided to use 4.2 s^{-1} as our standard value, which have a value of Abortive to Productive (APR) ratio that was within the values reported in abortive initiation studies [61]. With this value, the promoter can also be considered a rate limiting step (as it takes an average of 20 s for the RNAp to escape the promoter). This complies with observations made with the lacUV5 promoter [15, 99]. These results are presented in Chapter 4, section 4.2.

Note that as soon as the RNAp escapes the promoters and starts the elongation (15), another free RNAps in the cell may occupy that promoter and start a new cycle of transcription initiation.

We add a final reaction for the elongation (16), where the elongation complex exits our region of interest. We add a delay for the release of the RNAp and the mRNA using a Gamma Distribution (the function G in the reaction), where k equals to the number nucleotides that still needs to be elongated the respective gene and θ equals to 1/42 s (coming from the elongation rate) [100].

Considering that we wanted to mainly study the effects of the transcription initiation on the production of RNA and since interference is done in our zone of interest, normally we use a value of k equal to 150 nucleotides. This means that the elongation region is normally the same as the binding region, but we sometimes change this value to account for the increase in the binding region. We point out that since the elongation rate is so high, that changes up to 200 nucleotides, just makes an average increase of 5 s to the production time.

We include in our model collisions between ECs (9), where one of the two colliding RNAps is released from the template (randomly chosen among the two), or both of them are released from the template.

We also use the same system as in (9) to model collisions between 2 diffusing RNAps (7), where one of the RNAps or both of them randomly dissociates from the template.

In convergent promoters, it is possible for elongating and diffusing RNAps (8) to collide, and since the binding between an elongating RNAp and the template is so strong that it can remove a complex at a promoter [101], these can force, when collisions occur, diffusing RNAps, which have a weaker binding affinity then complexes at a promoter [102]. This means that the EC remains in the template and the diffusing RNAp dissociates from the template.

We also add the “Sitting duck” mechanism as reported by [101] where an EC collide with promoter complexes (closed, isomerization, open or ITC) at a converging promoter (10). Degradation of mRNA via reaction (17) is modeled as a single step reaction.

Repression is modeled via reaction (18) as competition for the RNAp binding or blocking the RNAp movement and dissociation of the repressor molecule from the template (19). As said before the repressor footprint occupies the range [n-Δ_{rep}, n+Δ_{rep}], and as a standard value we used $\Delta_{rep} = 10$.

With our model, we can easily modify the distance between TSSs and the binding region to study dynamics of RNA production, the repressor size and its TFBSs can also be changed. This together with the variability of kinetic constants extracted from *in vitro* and *in vivo* single-cell studies in *E. coli* allows us to study the dynamics of transcription initiation in various conditions. We also have to use *in vitro* studies to choose our parameters as for some steps, the *in vivo* kinetics of the steps described before still remain unknown. These simulations can be used for comparison with the measurements in single-cell studies that have been done in the last years [14].

All simulations in this work were executed with SGNS Simulator [103].

3.2. Calculations

Unless stated otherwise, we normally use 50 independent simulations with duration of 10^5 s each, sampled at every 10 s for every value that we want to simulate. We do this to save more time, because we can run multiple simulations in parallel and since the simulations can take some time to be obtained, which can range from just a few seconds up to almost an hour (depending rates and the conditions used for the simulation), this proved to be a very useful technique.

To have a better temporal resolution we calculate the distribution of time intervals between productions we use a sampling of 1 s. To calculate the mean and noise in RNA numbers, we concatenate all the simulated time series. On the other hand for time intervals we calculate the values for each independent simulation and for each promoter and then we concatenate all the values to make the histogram.

Note that for the calculation of time intervals, using the time series of RNA levels we can´t discern the case where at the same second there was a production event and a degradation event on the same side of productions, because if this happens the RNA numbers on that side doesn´t change. This can be avoided using a counter just for the production events, but this special case is very rare, so it doesn´t influence if we use enough simulations to have a significant sample of values.

To calculate the correlation between consecutive choices of production, we first produced a vector collecting the data from which side produced the next RNAp. Then we used MATLAB® *corr* function to compute Pearson's linear autocorrelation coefficient for that vector at different lags.

For the calculation of correlation between two unidirectional independent promoters, we put the corresponding time series side by side, and we produce the same vector of choices. There can be an event at the same second (due to the sampling) where both promoters produced a RNA transcript and for this we created a special case that consider as well that there was a change in choice of production. To calculate the correlation between time series, we do the same procedure but with a vector containing the time series of RNA numbers.

We calculate the Fano Factor (FF), the noise (CV^2) in RNA production using the equation (3.1) using the time series of RNA production we can calculate the variance and the mean. We can use this value to compare with the Fano Factor of a Poisson process (which is equal to 1)

$$\text{Fano Factor} = \frac{\text{Var(RNA)}}{\langle\text{RNA}\rangle} \tag{3.1}$$

We also calculate the noise (CV^2) in RNA production using the equation (3.2) but using the value of the squared mean.

$$\text{CV}^2 = \frac{\text{Var(RNA)}}{\langle\text{RNA}\rangle^2} \tag{3.2}$$

Chapter 4. Results and Discussion

We study the dynamics of transcription initiation as a function of the binding affinity, arrangements of the promoters, distance between TSSs and the effects of repression. In all models of bidirectional promoters, regardless of the arrangement, unless stated otherwise, the two TSSs and the dynamics of the rate limiting steps are in all identical, thus, unless there is some external factor causing an asymmetry (for example a repressor molecule bound close to one of the TSS), the mean RNA numbers expressed by the two promoters will be in all identical, provided a sufficiently large sampling.

4.1. Dynamics of RNA production as function of the binding affinity

In this section we first analyze the mean RNA numbers as a function of the binding affinity of the RNAp to the promoter region in divergent, convergent promoter with of 150 nucleotides between TSSs. We also use and unidirectional promoters and for all the arrangements we set the binding region with a length of 200 nucleotides for both genes. As said before, for these simulations we performed 50 independent simulations, 10^5 s each, sampled every 10 s for each different value of k_{bind} and promoter arrangement, and used all the values that were reported in table 3.1 and 3.2. Specifically we used the values for the closed complex formation to 0.5 s^{-1} [12], k_{open} to 2 s^{-1} and k_{isom} to 0.095 s^{-1} [15].

Given those parameter values, and knowing that there are genome wide measurements for the number of RNA molecules at steady state of most genes [104] and using a value for the RNA degradation rate to 0.36 min^{-1} (corresponding to a half-life ~2 min) [105], it is possible to test what intervals of values for the rate of binding to the DNA template (k_{bind}) in the model leads to realistic mean RNA numbers at near-equilibrium [104]. We will then compare them with the value of k_{bind} extracted from the simulations to the one measured by indirect means [12]. Note that we used a value for the rate of dissociation (k_{unbind}) that value by Singer and colleagues by the same indirect means.

In figure 4.1 we exhibit the values for the RNA mean numbers over time for each value of k_{bind} and each arrangement. In figure 4.2 we exhibit the square of coefficient of variation (CV^2) of RNA numbers over time. For the calculation of the mean and CV^2 we concatenated the 50 times series for each value of k_{bind} and for each arrangement. Given the length of the time series, while we initialize the simulated with no RNA molecules, the time to reach near-equilibrium is negligible for the calculations of mean and CV^2 of the RNA numbers.

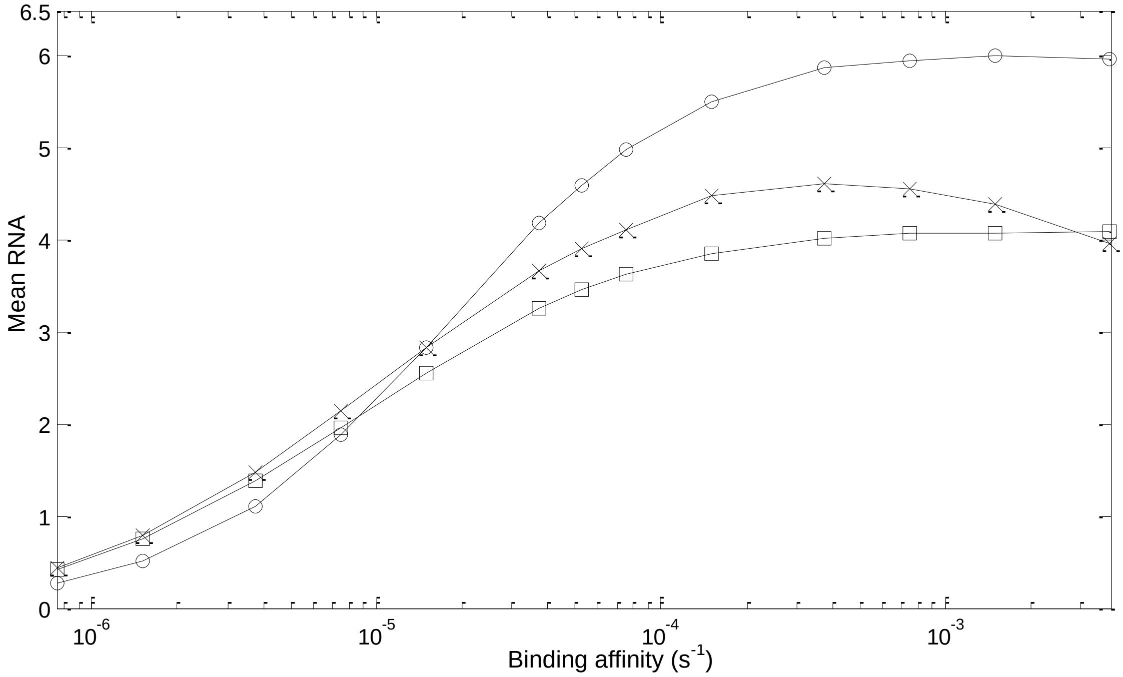

Figure 4.1 – Mean RNA numbers as a function of k_{bind} rate. Here we represent (○) for unidirectional promoters, (x) for divergent promoters and (□) for convergent promoters.

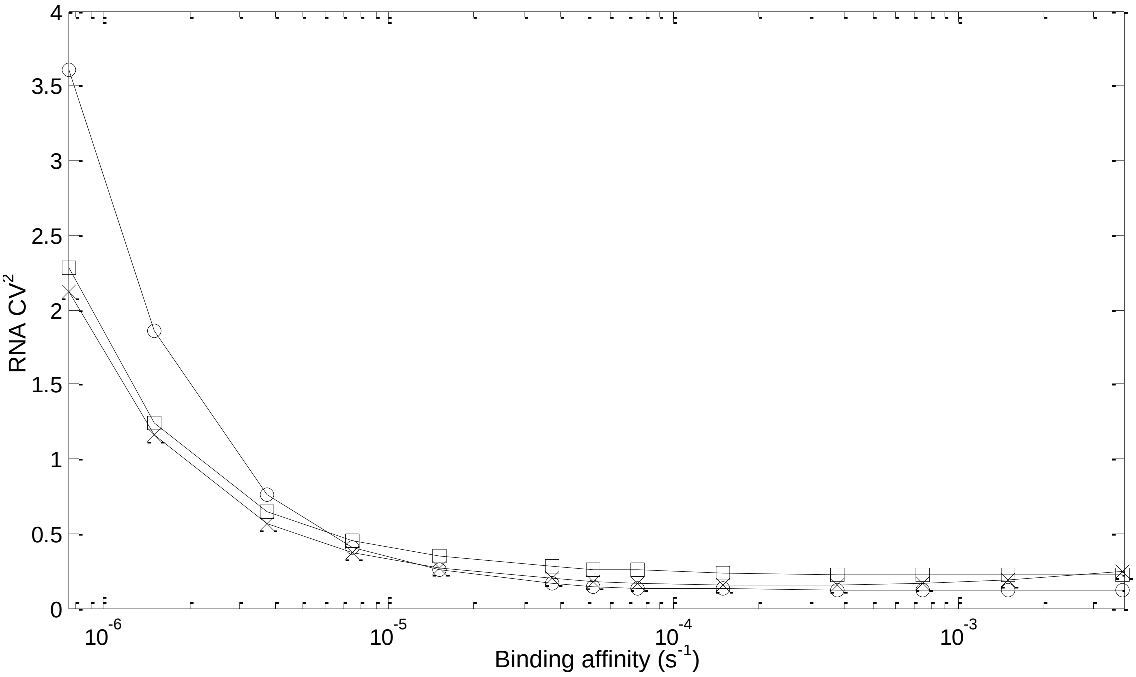

Figure 4.2 – CV^2 of the RNA numbers as a function of the binding affinity (k_{bind}) rate over time (at near-equilibrium). Here we represent (○) for unidirectional promoters, (x) for divergent promoters and (□) for convergent promoters.

The measurements *in vitro* from Singer and colleagues [12] suggest that the RNAp can bind to any nucleotide in template and that the binding rate to the promoter region is higher than for other regions, so we use the model to first investigate what is the saturation rate for k_{bind} (namely, the rate for which further increases will not increase mean RNA levels).

From the figure 4.1, we observe that for all promoter structures, the resulting mean RNA numbers at near-equilibrium are within realistic intervals for the entire interval of values of k_{bind} used for this simulation [104]. Lower values correspond to weakly expressed genes, while higher values correspond to more strongly expressed genes.

However, we note that the genes are not subject to repression in this specific simulation (we will use the repression mechanisms in later sections), while in *E. coli*, weakly expressing genes usually have such behavior due to the presence of repressor molecules. Due to this, we consider the most realistic interval of values of k_{bind} to be the ones above 10^{-5} s^{-1}. Also, we can observe that the rate of "saturation" is approximately k_{bind} ~ 10^{-4} s^{-1}, given the kinetics of the closed and open complex formation. Thus, realistic values for this case are likely between 10^{-5} and 10^{-4} s^{-1}.

Singer and co-workers [12] estimated the *in vitro* rate of non-specific association of RNAp to circular DNA to be 4.6 × $10^4M^{-1}s^{-1}$ (per nucleotide). From this, it is possible to estimate the corresponding value for k_{bind} in this case by dividing the measured value by the expected volume of the *E. coli* (10^{-15}L, taken from the CyberCell Database [106]) and the Avogadro constant. This results in a value for k_{bind} of 0.75 × 10^{-4} s^{-1} per nucleotide, which is within the interval resulting from our estimation. Due to this, here onwards we set the value of k_{bind} to 0.75 × 10^{-4} s^{-1}. Interestingly, according to the results in Figure 4.1, this value for k_{bind}, for all the promoter arrangements with 150 nucleotides between their TSSs, is close to the saturating rate of RNA production.

As said before there are differences in the dynamics of RNA production between different promoter arrangements due to the collisions between RNAps in different states which can cause interference in those dynamics. Because of this the unidirectional naturally has the highest production rate while both the divergent and convergent cases the RNA production rate is considerable lower. At the saturation level, the resulting near-equilibrium mean RNA numbers for the unidirectional case is ~6, while for the divergent and convergent case is ~4 which is well within realistic intervals for strongly expressing genes [104]

In the divergent arrangement we observe that the mean value doesn´t saturate but starts decreasing in production as the k_{bind} rate go higher. This can be explained by the increase in traffic in the DNA template by the diffusing RNAps. On the other hand, in the convergent arrangement, the converging elongating RNAp can remove the diffusing RNAps, which partly helps the traffic problem.

From figure 4.2 we can see that for values of k_{bind} above 10^{-5} s^{-1} the CV^2 remains constant and within the same range for all the arrangements. From this and with the increase of the mean observed in figure 4.1 we can see that the that with the increase of the k_{bind} the production of RNA approximates to a Poisson process, as can be seen in table 4.1 where we present the values of the Fano Factor (as calculated by equation 3.1) for all the simulated values of k_{bind}.

Table 4.1 - The Fano factor of RNA production for different arrangements as a function of the isomerization rate.

k_{bind} /s^{-1}	Unidirectional	Divergent	Convergent
$7,5 \times 10^{-7}$	0,95	0,94	0,95
$1,5 \times 10^{-6}$	0,93	0,90	0,92
$3,8 \times 10^{-6}$	0,85	0,85	0,89
$7,5 \times 10^{-6}$	0,78	0,79	0,87
$1,5 \times 10^{-5}$	0,72	0,75	0,88
$3,8 \times 10^{-5}$	0,68	0,72	0,91
$5,3 \times 10^{-5}$	0,67	0,71	0,92
$7,5 \times 10^{-5}$	0,68	0,71	0,92
$1,5 \times 10^{-4}$	0,70	0,70	0,92
$3,8 \times 10^{-4}$	0,71	0,72	0,91
$7,5 \times 10^{-4}$	0,71	0,75	0,91
$1,5 \times 10^{-3}$	0,72	0,81	0,90
$3,8 \times 10^{-3}$	0,72	0,99	0,91

Comparing the values with a pure Poisson process, which has a Fano Factor of 1, we can observe that all our processes have an inferior value then 1. This could mean that delays in chemical reactions have less variance than a pure Poissonian process and due to that can act as noise filters. The values closer to a Poissonian process are for the lowest values of k_{bind} and also for the highest values. The values with the lowest Fano Factor are the ones near the 10^{-5} s^{-1} order of magnitude. Interestingly, according to the results in table 4.1, this range of values for k_{bind} (which are in the interval of the ones estimated by Singer and co-workers [12]) are the ones that reduce the noise in the production of RNA. This is an example of a noise regulation mechanism in gene expression. [4].

4.2. Dynamics of RNA production as function of the abortive ratio

In the last section we studied the dynamics of transcription initiation as a function of the binding parameter, now in this section we study the effects of changing the abortive ratio on the RNA production. For this we calculate the abortive to productive ratio (APR), the noise (CV^2) and the Fano Factor. For these simulations we maintained a mean RNA number of around 5.5 for all the simulations and do this we had to change the RNA degradation rate [105] (a higher abortive rate leads to a higher smaller degradation rate, or a higher RNA half-life) in order to tune the mean.

The values of APR and Fano Factor are presented in Table 4.2 for each of the values of the chosen k_{abort}. The noise (CV^2) and the mean RNA numbers (in the inset graphic) for the RNA production is presented in figure 4.3. These simulations were done for the divergent arrangement using promoters with a distance of 200 nucleotides between the TSS.

Table 4.2 - Fano factor of RNA production and Abortive to Productive ratio (APR) as a function of the abortive ratio (k_{abort}).

k_{abort} (s^{-1})	1	1.9	2.3	2.6	3	3.5	4.2	5.3	7	10.5
Fano Factor	0.72	0.71	0.70	0.69	0.68	0.68	0.69	0.75	0.87	1.04
APR	2	7	10	14	19	31	55	128	445	4108

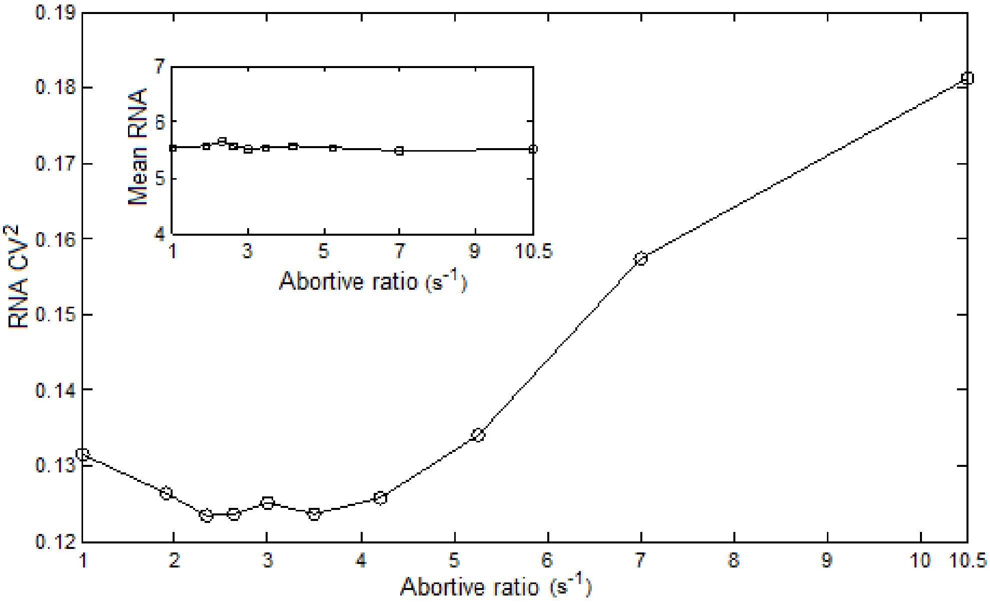

Figure 4.3 – The CV^2 of RNA production as a function of the abortive ratio (k_{abort}). As can be seen in the inset graphic we maintained the RNA number constant around 5.5 varying the degradation rate. These simulations were all done on a divergent promoter with a distance of 200 nucleotides between the TSS.

From the results in table 4.2 we observed an APR ranging from 2 to 4108 using a range of abortive ratio respectively from 1 to 10.5 s^{-1}. We also observe that the lower Fano Factor values were obtained in the range from 2.3 to 4.2 s^{-1}. At 10 s^{-1} the process has a noise higher than a pure Poissonian process (Fano Factor is higher than 1).

Studies using the same promoter sequence but with different initial transcribing sequence (the downstream region) have observed values of APR ranging from around 13 to 386 [61]. Another study observed values ranging from around 6 to 100 [107]. These values are within the values that we obtained with range of k_{abort} used in the simulations. This means that the value used for our standard parameter should be in this range, so we decided to use 4.2 s^{-1} as a standard parameter in the rest of the work (with an APR = 55), which is a value in the middle of that specific range.

From the figure 4.3 we observed that k_{abort} also allows regulating noise in RNA for values higher than 5.3 s^{-1}. In all cases the mean RNA number was maintained at around 5.5 (see inset graphic from figure 4.3) to remove the influence of the mean in the noise values.

From the studies using the *E. coli* RNAp interacting with T7 promoters [97] and from a recent study [108] used a single-molecule Förster resonance energy transfer (FRET) and stopped-flow FRET to monitor T7 RNAp transition from initiation to elongation we can see that in the process of abortive initiation, each path of “scrunching”, “abort” or “escape” can have their own kinetic parameters depending on the position of the nucleotide regarding the TSS and depending on the DNA sequence. As said before we assume, for simplicity proposes, a uniform value of the abortive ratio for every nucleotide and we also only allow the escape at exactly the 12th nucleotide (meaning that we don’t have an escape path for every nucleotide) [97]. We also decided to maintain this simple mechanism because there are still no studies using the *E. coli* RNAp on regular *E. coli* promoters and as we said in the theoretical framework, although the mechanisms in T7 promoters and the T7 RNAps are similar, their kinetic parameters are considered to be much different.

We should note that we already constructed a model that includes an escape path and with all the different kinetic parameters for each nucleotide. Since with this modification we allow that for every abortive initiation, the RNAp escapes the promoter at a different position leading to different values for the maximum size of abortive transcripts, which in this work is always 12, but with this modification could go up to 20 as reported in abortive initiation studies [60,61]. This modification leads to process that is more random process, however we decided not to include it in this work due to lack of kinetic parameters and the complexity of studying the effects of each value.

This modification in our model could also be included in a future work, where new information from single-cell studies [14] could give valuable information about the kinetic parameters.

4.3. Time Series of RNA production of different promoter arrangements

Using the standard values defined in the previous sections we now present a time series of the RNA production for each of the different promoter arrangements. For this time series we used just one simulation of 10^6 s sampled every 1 s (for a better temporal resolution) and extracted the first 5000 s for each arrangement. In figure 4.4 we present the time series for the divergent (with a distance of 150 nucleotides between the TSSs) case and in figure 4.5 we present the time series for the convergent (also with a distance of 150 nucleotides) case. In figure 4.6 we also present the time series for a divergent arrangement with overlapped promoters (with a distance of 65 nucleotides between the TSSs). In this section the binding region has 300 nucleotides for all the arrangements. The line in red present in figures 4.4, 4.5 and 4.6 corresponds to the right promoter (the one with the TSS at position +1) for all the arrangements (see figure 3.1)

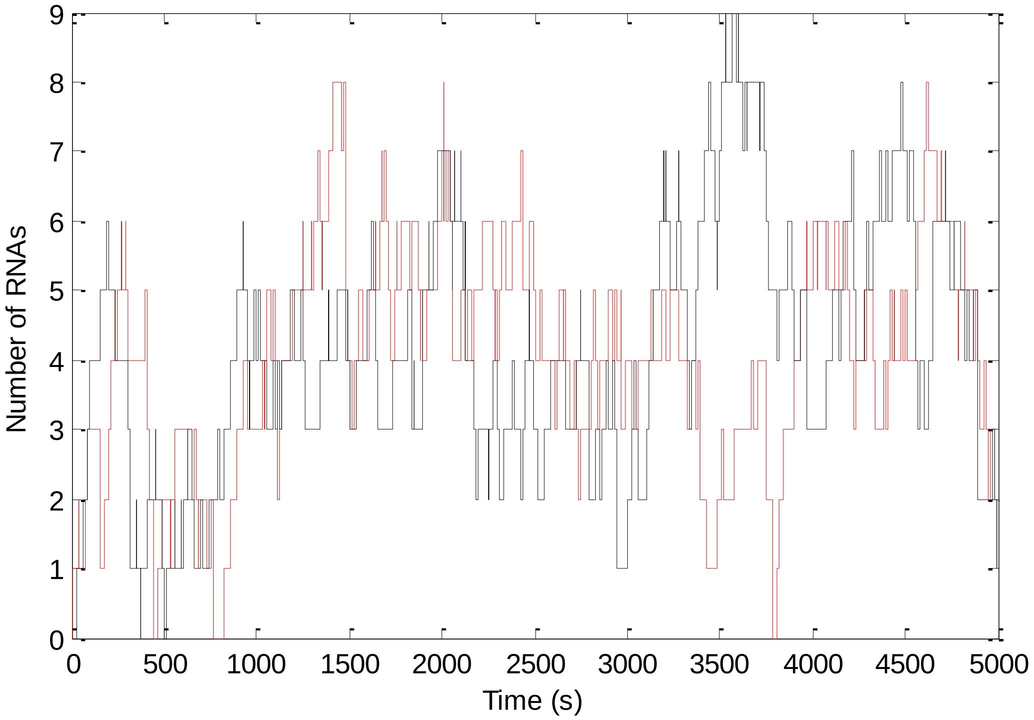

Figure 4.4 – Time series of RNA production for the divergent case.

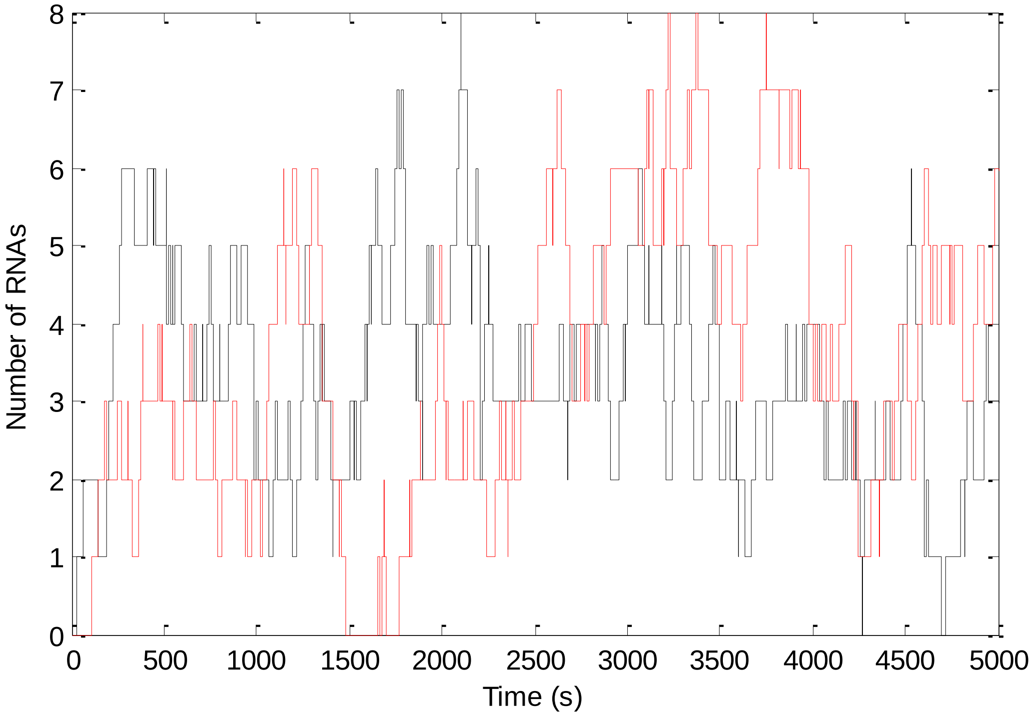

Figure 4.5 – Time series of RNA production for the convergent case.

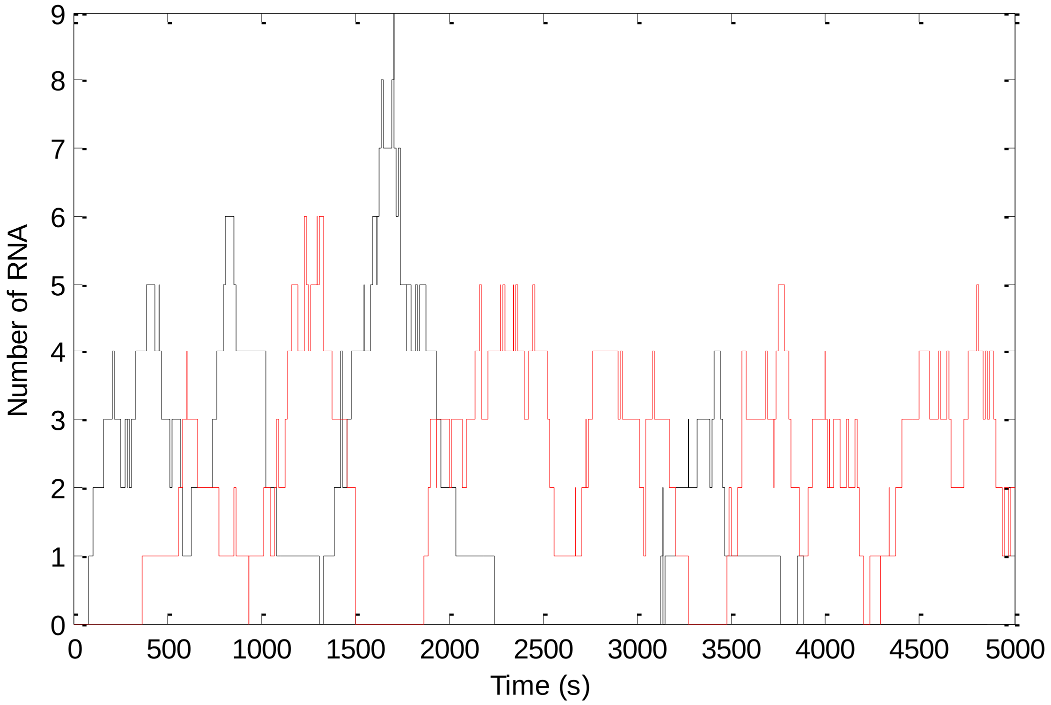

Figure 4.6 – Time series of RNA production for the divergent overlap case.

From the figures 4.4, 4.5 and 4.6 we can observe that all the three cases presented here have a different behavior in the production of RNA, but we can observe that in all of them, there are some big fluctuations caused by the stochasticity of this processes.

As expected we observe from figure 4.4 that in the divergent arrangement both promoters can transcribe at the same time, but due to the stochasticity of the gene expression, there are times where one promoter is producing and the other isn´t, or times where both genes are producing RNAs. From figure 4.5 we can see that in the convergent case when one promoter is producing RNAs, the promoter can´t produce any RNA. From figure 4.6 we observe that in the divergent overlap, this situation is even more visible and we can see that in this case the time where one promoter is actively producing a RNA transcript is bigger than in the convergent arrangement.

This could mean that process of producing one RNA transcript can influence production of the next RNA transcript. We will study these effects in a latter section using correlation studies of directionality and time series.

4.4. Asymmetric rate limiting steps in different promoter arrangements

For these simulations we use the same arrangements as in the previous section, so the divergent promoters have a distance of 300 nucleotides between the TSSs, the convergent promoters have a distance of 150 nucleotides and the divergent arrangement with overlapped promoters have a distance of 65 nucleotides, and all the arrangements have a binding region of 300 nucleotides for all the arrangements.

In this section we study the effects of introducing asymmetric rate limiting steps in different promoter arrangements. These asymmetric rates can be achieved for example with an activator that can affect various steps in transcription initiation. For example in the lac gene, the CRP protein can active the binding step, while at the gal gene the CRP protein affects the isomerization steps In the repression case, we can have the LacR protein repressing the binding step or the Arc protein can repress the isomerization step [54]. . In this section we don´t use a specific activators or repressors, but we change directly the rate constants of the chemical reactions to see the effect of asymmetry on such parameters.

In table 4.3 we present the mean, CV^2 and Fano Factor of RNA production for different promoter arrangements as a function of the isomerization rate. Note that this rate was only changed on the promoter with the TSS at +1 (see figure 3.1), which we call the “right” promoter.

Table 4.3 - The mean and CV^2 of RNA production for different arrangements as a function of the isomerization rate. This rate was changed on the “right” promoter. Here k_{isom} stands for the standard isomerization rate used in this work: 0.095 s^{-1}.

		Mean		CV2		Fano Factor	
		Left	Right	Left	Right	Left	Right
Convergent	k_{isom}	3.76	3.69	0.26	0.27	0.98	1.00
	$k_{isom}/2$	4.41	2.29	0.21	0.44	0.93	1.01
	$k_{isom}/5$	5.15	1.04	0.16	0.97	0.82	1.01
	$k_{isom}/10$	5.44	0.54	0.14	1.85	0.76	1.00
Divergent overlap	k_{isom}	1.92	1.94	0.88	0.88	1.69	1.71
	$k_{isom}/2$	1.46	1.45	1.41	0.88	2.06	1.28
	$k_{isom}/5$	0.86	0.85	2.96	1.21	2.55	1.03
	$k_{isom}/10$	0.52	0.49	5.53	1.96	2.88	0.96
Divergent	k_{isom}	5.60	5.57	0.13	0.12	0.73	0.67
	$k_{isom}/2$	5.57	4.15	0.13	0.18	0.72	0.75
	$k_{isom}/5$	5.56	2.32	0.12	0.35	0.67	0.81
	$k_{isom}/10$	5.53	1.33	0.13	0.66	0.72	0.88

In the case of UV5 isomerization is the strongest rate limiting step leading to the open complex formation [15], so it is natural that we change this value in this study.

From the results in table 4.3 we observe that in the convergent arrangement, the promoter with a higher k_{isom} has an increase in the mean RNA numbers (and decrease in the CV^2) while the other promoter has a decrease in the mean RNA numbers (and decrease in the CV^2). This is because the increase in the the number of ECs (firing rate) that are produced by one promoter leads to a higher number of collisions between those ECs and the ones produced by the other promoter and also between the ECs and complexes at the slower promoter (slower firing rate). Sneppen and colleagues [109] saw the interference due to collisions from 2 convergent promoters and observed that if both promoters had the same strength (the normalized case), they had a logarithmic interference with the increase of distance between TSSs. If one promoter was stronger and one was weaker (then the normalized case), then the stronger suffers less interference (lines lower than the K^A/K^S) and the weaker suffers a higher interference (lines higher than the K^A/K^S). We also observed a similar pattern, which means that our model is in accordance with their model of interference.

Looking at the “sitting duck” interference is not that simple, because with a slower promoter, the aspect ratio goes higher, but the K^A/K^S also goes higher, so we might not see an increase in this type of interference.

In the divergent overlapped promoters, it is observed that with one weaker promoter, at both promoters the mean RNA production decreases within the same order of magnitude. Interestingly the CV^2 is higher for the promoter with a slower isomerization. As we can see from figure 4.6, when one promoter is producing an RNA, it stays producing for some time and only then it changes to the other side (we will study this effect in a latter section using correlation between production choices). So this results means that the isomerization step doesn´t affect this choice of production (because the mean is equal on both sides), and only affects the duration of the production. But since this production is alternate we will have around the same number of RNA´s produced at each side but with a slower production with the slower isomerization, which leads to a higher noise.

In the divergent case we can observe that only the affected promoter has a significant change in the RNA production (the other promoter has a 1.2% decrease in the mean value, while the affected promoter has 76% decrease with a 10 times decrease on the isomerization rate). This is because when one promoter is producing, the other promoter can still produce but has a smaller binding region (less available nucleotides), so he gets a small decrease in promoter activity, thus a lower mean RNA levels.

From the values of the Fano Factor (FF) we can observe that on the convergent case, the affected promoter (lower isomerization) has the same noise as a Poisson process (FF =1), while the other promoter has a reduction on the noise (FF < 1). In the divergent overlap case, both promoters start with a higher noise than a Poisson process (FF > 1) but the affected promoter starts to lower the noise until it is less than 1, while the other promoter actually grows the noise (FF >>1). In the divergent case, both promoters have a noise lower than a Poisson process, but the affected promoter has an increase in noise with the decrease in isomerization.

In table 4.4 we present the mean, CV^2 and Fano Factor of RNA production for different promoter arrangements as a function of the binding rate. In this case we actually changed the binding rate of some nucleotides in order to achieve asymmetry.

In the convergent promoter, all the binding region of the "right" promoter (see gray region in figure 3.1) was affected, while in the divergent promoters, we divided the binding region in half, and only one half was affected. So we affected with the binding rate a region from +1 to -150 (note that here the distance between the TSSs is 300 nucleotides, so the other promoter is at position -301. On the divergent overlapped promoters, we had to make a decision on where to divide the binding region: one that included both TSSs and one that would affect mainly just one TSS (so the affected region would be at the middle or at the end of the binding region respectively). We simulated both cases, and saw that when we affected just the middle, the mean, CV^2 and FF stayed almost unchanged for both promoters (less than 1% change). Because of this we only show the values for the other simulation in table 4.4.

Table 4.4 - The mean and CV^2 of RNA production for different arrangements as a function of the binding rate. This rate was changed in a particular region of the template. Here k_{bind} stands for the standard binding rate used in this work: 0.000075 s^{-1}.

		Mean		CV2		Fano Factor	
		Left	Right	Left	Right	Left	Right
Convergent	k_{bind}	3.71	3.71	0.27	0.27	1.00	1.00
	$k_{bind}/2$	3.93	3.22	0.25	0.31	0.98	1.00
	$k_{bind}/5$	4.45	2.21	0.20	0.46	0.89	1.02
	$k_{bind}/10$	4.93	1.35	0.17	0.75	0.84	1.01
Divergent overlap	k_{bind}	1.92	1.92	0.88	0.88	1.69	1.69
	$k_{bind}/2$	2.73	1.48	0.53	1.13	1.45	1.67
	$k_{bind}/5$	3.34	1.18	0.38	1.32	1.27	1.56
	$k_{bind}/10$	3.55	1.09	0.34	1.41	1.21	1.54
Divergent	k_{bind}	5.56	5.59	0.12	0.12	0.67	0.67
	$k_{bind}/2$	5.49	5.33	0.13	0.13	0.71	0.69
	$k_{bind}/5$	5.46	5.08	0.13	0.14	0.71	0.71
	$k_{bind}/10$	5.40	4.94	0.13	0.14	0.70	0.69

From the results in table 4.4, we can observe the same effect as the previous simulation on the convergent case, where the mean of the affected promoter (lower binding rate) decreased, while the other increases the mean. In this case, the CV^2 of the affected promoter increased by almost 3 times, while the other promoter had a reduction on the CV^2. We also observe that the affected promoter also has a noise similar to a Poisson process (FF =1), while the other decreases the noise. In this case we are reducing the amount of RNAps that bind to the DNA template on the region of the affect promoter, so the promoter will have less converging EC´s to collide into, so it means that they can clean the converging region of the diffusing and "sitting ducks", so it has less interference than the case where both promoter have the same binding affinity.

On the divergent overlap case, we can see the same effect, that the affected promoter has a reduction on the mean, while the other increases. We can see an increase of 60% on the CV^2 of the affected promoter and a decrease in 61% on the other promoter and from the Fano Factor we can see that while both promoters have a FF higher than 1, and both promoters have a reduction on the FF, the affected promoter have a smaller reduction than the other promoter (9% compared to 28% respectively).

Finally on the divergent case, we can see that both promoters have a slight decrease on the mean (the affected promoter has a decrease of 12% and the other of 3%). Both the CV^2 and the Fano Factor have very small variations (they seem to have a tendency of increasing both values at both promoters, but there are some values that had a small decrease). This is because both promoters share a big binding region, so even if we divide in half the binding affinity, the other half still has lots of RNAps binding in both directions, so this effect is less pronounced in the divergent case than in the other cases.

With these results we observed the various effects on the mean, CV^2 and Fano Factor using asymmetric rate constants, on chemical reactions before and after the binding of the RNAp on the different arrangements.

4.5. Binding kinetics of RNA polymerases to promoter regions

In this section we first study the binding kinetics of RNAps to the DNA template using divergent promoters with a distance between TSSs of 200 nucleotides. For this simulation we measured from 50 independent simulations each 10^5 s long the number of times each nucleotide on the DNA template is bound by an RNAp. In figure 4.7 we show the distribution of probabilities with shows the fraction of times that each of the nucleotides was the one to which the center of the RNAp first binds to in the template, thus, there is no recorded binding in the first and in the last 27 nucleotides of the binding region.

We also tested the effect of varying the binding affinity (k_{bind}), which was first set to the standard value of $0.75.10^{-4}$ s^{-1} [12] and then changed to 10 and 100 times smaller. Note that when the formation of the closed complex take place at the TSSs, the RNAp occupies 55 nucleotides, which means that they occupy the position +1 to -54 ("right" promoters) or -147 to -201 ("left" promoter), since for this simulations the promoters had a distance of 200 nucleotides between their TSSs.

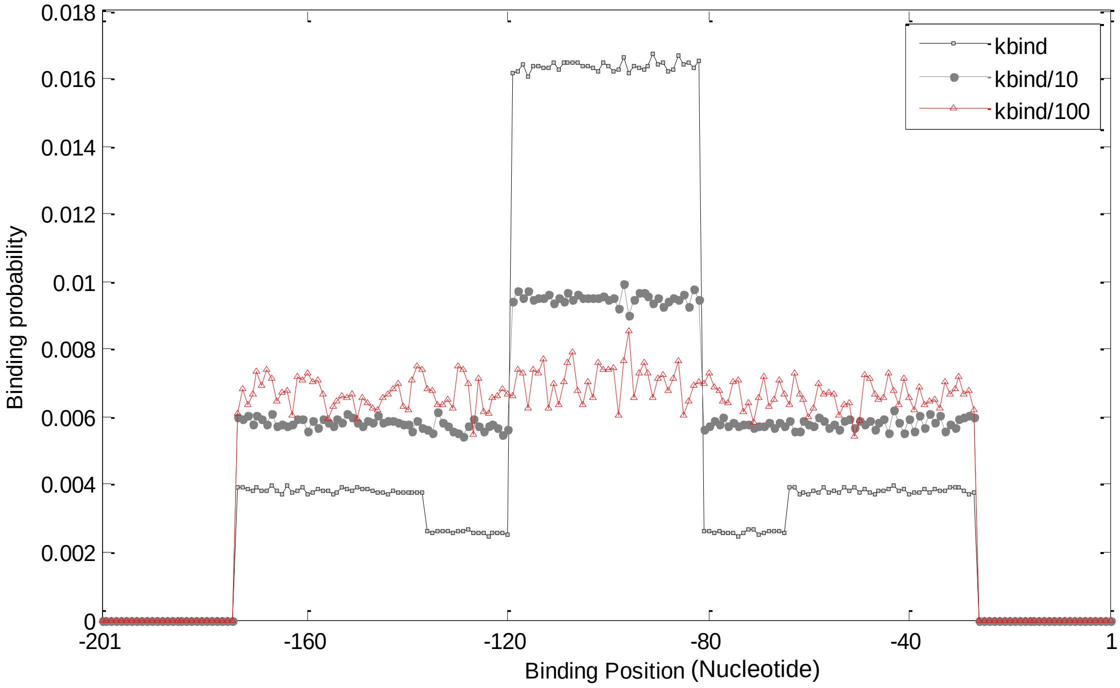

Figure 4.7 – Probabilities for each nucleotide that an RNAp will bind to the DNA template of two divergent promoters. The binding region is located between both TSSs with 200 nucleotides between them. We tested three different values of k_{bind} (standard value, 10 and 100 smaller values than the standard).

In figure 4.7 we show the distributions of the probability that each nucleotide which was bound by an RNAp for different values of k_{bind}. We should also note again that that since in Biology there is no position 0 we had to shift one position in the x-axis (from figures 4.7 and 4.8), so that although there are 200 nucleotides between the TSS, their positions are at +1 and -201.

In figure 4.7 we can observe that the spatial distributions of binding probabilities are not uniform, except for the case k_{bind}/100. This is due to the rate limiting steps that occur at the TSSs, namely, the closed complex formation, isomerization and the open complex formation. For *lac*UV5 the step of RP_i formation through RP_c is the slowest of the three (see table 3.1 for the values used) [15]. Since the duration of those steps, along with the abortive initiation [18, 19] is not negligible, this means that RNAps primarily occupy the regions the TSSs. After, the regions more occupied are those just adjacent to the first ones as diffusing RNAps wait for the TSS to be cleared. The non-negligible nature of the size of the RNAp compared to the promoter length cause the distributions to be discontinuous.

To verify this, we did a new simulation where k_{bind} is set to its standard value (see table 3.1), but the rates of the four rate limiting steps are set so that they no longer were rate limiting steps of the RNAp movement along the DNA template. Specifically, k_{close}, k_{isom}, k_{open} and k_{elong} are set equal to k_{move} and k_{abort} = 0. When doing so, the distribution became identical to the one for k_{bind}/100 (red line in figure 4.7) as observed in figure 4.8.

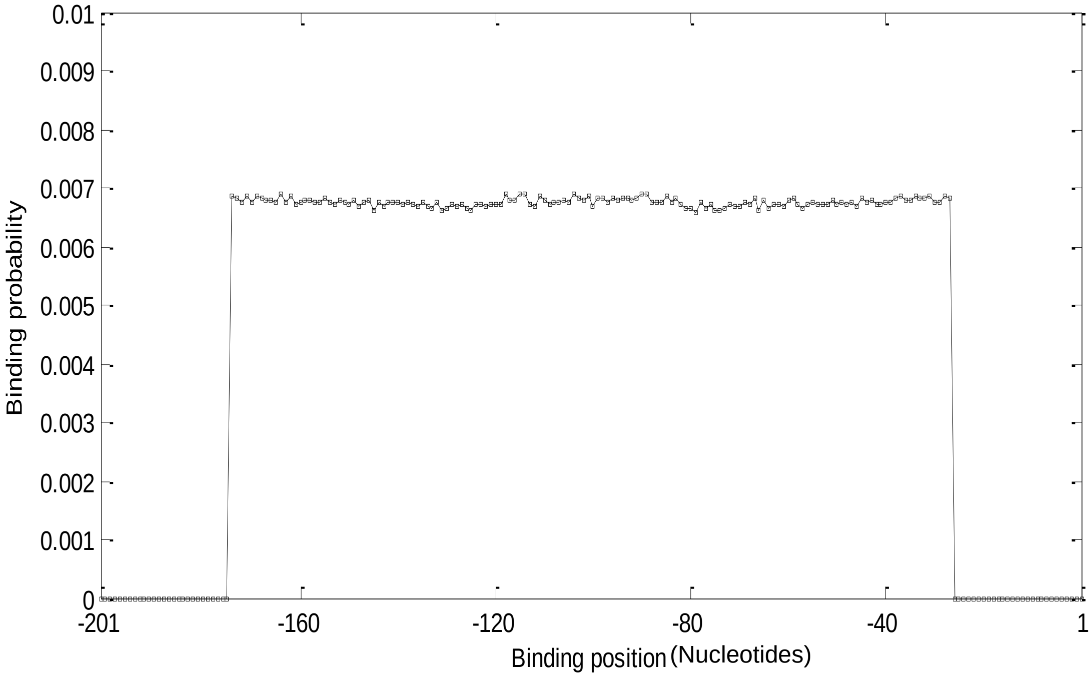

Figure 4.8 – Probabilities for each nucleotide that an RNAp will bind to the DNA template of two divergent promoters. The binding region is located between both TSSs with 200 nucleotides between them. For this simulation we set the rate constants so that they are no longer rate limiting of the RNAp movement along the DNA template.

From figure 4.8 it is possible to see that the distribution became uniform and identical to the one for k_{bind}/100 because there were no rate limiting steps of RNAp movement in this simulation. We note that this simulation has less noise, because it has more samples (since the rates were much faster, there were more binding events). Since there are 148 available nucleotide positions that the reading head of the RNAp can bind to, the value at the y-axis is ~0.68.

It is of interest to note that all these four steps have similar effects on the distribution because all those steps take place in the same positions (although they may differ in other effects such as noise in RNA levels).

As each of those will cause the distribution to become non uniform similarly to what is shown in figure 4.7 for the standard k_{bind} case. The slower the rates of the chemical reactions (that take place at the TSS), the more non-uniform will be the distribution.

The shape of the distribution shown in figure 4.7 is expected to be dependent on the distance between TSSs. To study the effect we show distributions of binding probability for divergent promoters with different distances between TSSs in figure 4.9. For these simulations, we only use the standard value of k_{bind} (see table 3.1) to see what is the effect on the non-uniform distribution. Note that for the simulations in figure 4.9 we had to do just 1 simulation 10^6 s duration, because it was simpler to just count the binding events in 1 simulation, and this duration is enough to understand the shape of the distribution.

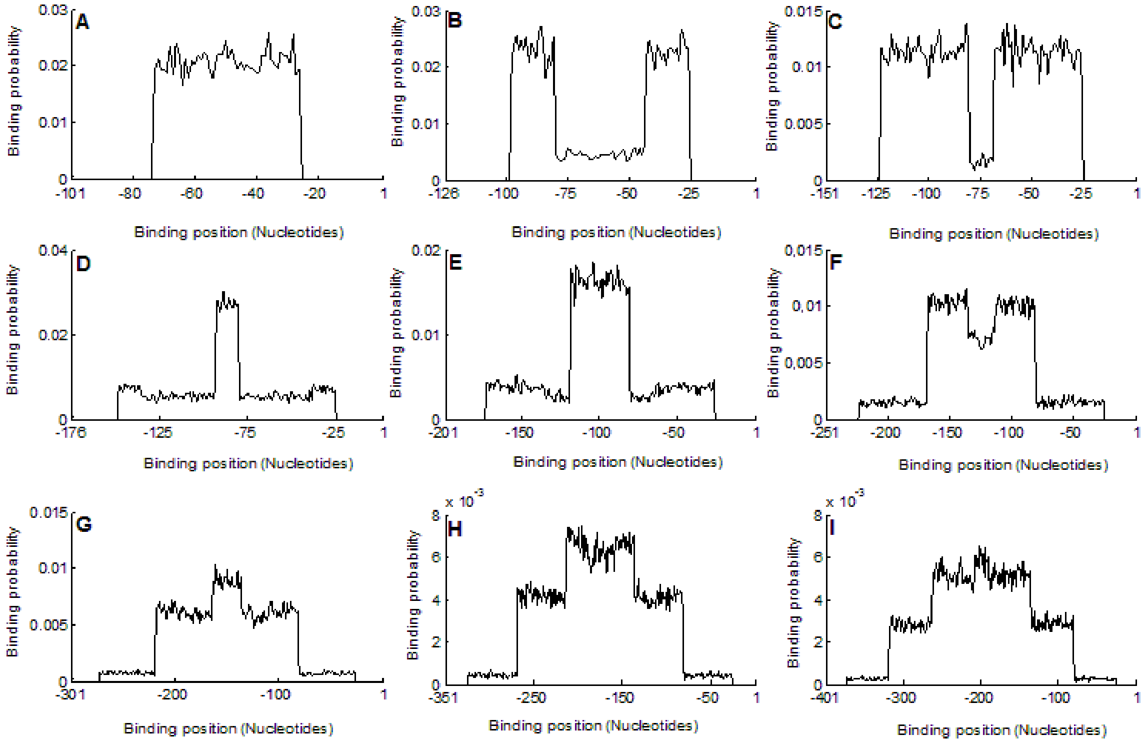

Figure 4.9 – Probabilities for each nucleotide that an RNAp will bind to the DNA template of two divergent promoters. The binding region is located between both TSSs differing in number of nucleotides (N) between both TSSs. (A) N=100 (B) N=125 (C) N=150 (D) N=175 (E) N=200 (F) N=250 (G) N=300 (H) N=350 (I) N=400. In the y-axis we have the binding probability and in the x-axis we have the binding position.

From figure 4.9 it is possible to conclude that the shapes of the of the positions where the binding occurs depend on the ratio between the number of nucleotides that an RNAp occupies when diffusing on the template and the size of the binding region

This distribution will only be spatially uniform in two cases. First, as shown in figure 4.8 (A), the distribution is uniform if the distance between TSSs region is too small to have more than one RNAp bound to it at any moment (divergent overlapping promoters). The second case would be for a distance between TSSs so large that the number of free RNAps in the cell would not be sufficient to fully occupy it (with 28 free RNAps, each occupying 55 nucleotides we would need a sequence with 1540 nucleotides). This can be inferred from observing how the distribution changes (and how the spikes become less pronounced) with increasing distance.

In figure 4.10 we present the results of the study of the binding kinetics of RNAps to the DNA template using convergent promoters with a distance between TSSs of 150 nucleotides and a binding region of 150 nucleotides for both promoters.

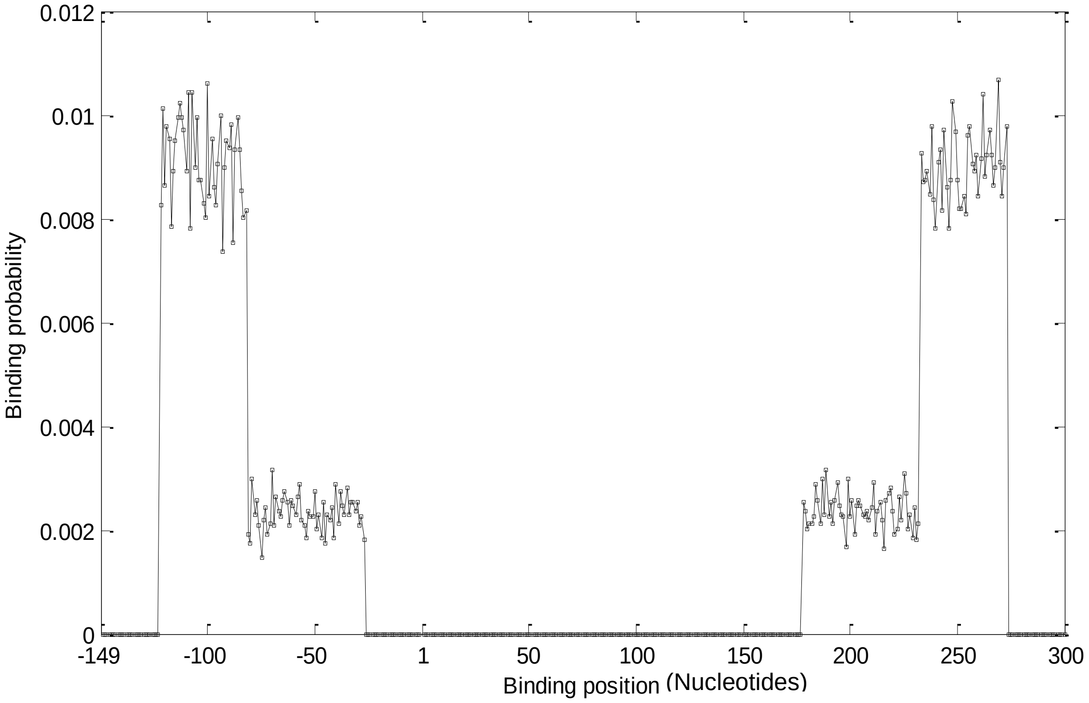

Figure 4.10 – Probabilities for each nucleotide that an RNAp will bind to it, when binding to the promoter region. This simulation was done with convergent promoters with a binding region of 150 nucleotides and a distance between TSSs of 150 nucleotides.

From figure 4.10 we can observe that the convergent promoter behaves similarly in case of the distribution shape but the binding regions are not overlapping. It is possible to get the distribution of convergent promoter by taking the half distribution of the overlapping area, turning it around and separating the space with the distance between TSSs.

4.6. Dynamics of RNA production in different arrangements

We next investigate how the geometry of a promoter affects the dynamics of RNA production. We compare the distributions of intervals between the productions of consecutive RNAs from one TSS (since this distribution is identical for both TSSs in all models, we just join the values obtained separately for each TSS into a single vector to make the distribution). For these simulations we used the same 50 simulations with 10^5s each, but with a sampling of 1s for a better temporal resolution.

In all models, the binding region behind each the TSS is 200 nucleotides long which implies that differences in the means of the distributions are solely due to the differences in geometry and not, for example, due to different rates of binding of RNAps to the templates. The tails of the distributions are shown in inset for each case (except F, where it is within the range of 120 s) and differ significantly in length.

Models A to C are divergent promoters, differing in the distance between both TSSs. In A the distance is 200, in B is 150 and C is 65 (divergent overlapped promoters). As this distance decreases, the mean of the intervals decreases, and so does the standard deviation because of the decrease in number of collisions between elongating and diffusing RNAps, and thus the width of the distribution decreases.

Model D is also a divergent promoter, identical in structure to A, but without the rate limiting steps at both TSSs. Due to that, in comparison to A, the mean of the intervals is much smaller and the distribution becomes exponential-like, given the absence of the rate limiting steps [110].

Model E is a convergent promoter with 100 nucleotides between both TSSs. As expected, in comparison to model A, the kinetics of RNA production is reduced and noisier. Namely, the distribution of intervals between production events increases in mean and standard deviation due to the interference between elongating RNAps from different TSSs.

Finally, model F, a unidirectional promoter, behaves similarly to a divergent promoter where there are no collisions between elongating and diffusing RNAps (similarly to model C).

We present the mean and standard deviation of those time intervals in table 4.5. In figure 4.11 we also present the distributions of time intervals between the productions of consecutive RNAs, using 6 different models described in this section.

Table 4.5 - The mean and the standard deviation of the time intervals between the productions of consecutive RNAs from just one promoter. We joined the data obtained separately for each promoter to have more samples.

Models	Mean /s	σ /s
A	34.30	20.11
B	41.06	25.17
C	88.75	176.84
D	5.12	4.58
E	46.70	42.84
F	34.51	19.97

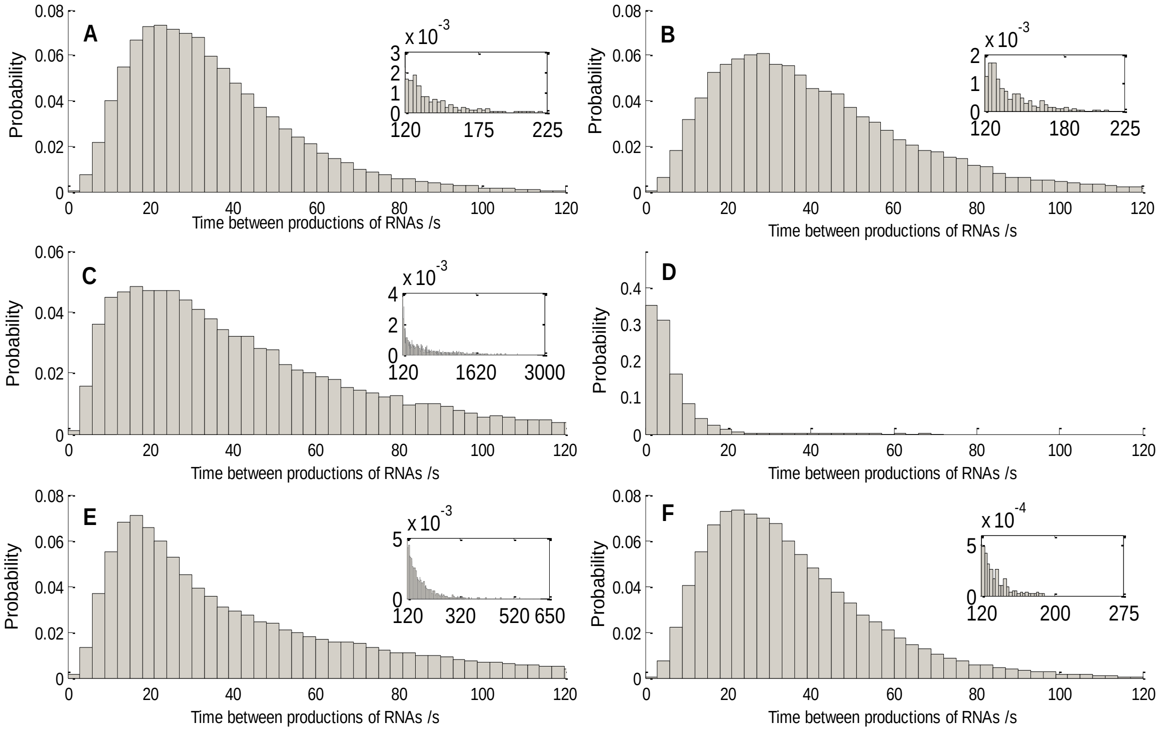

Figure 4.11 – Probability distribution of time intervals between the productions of consecutive RNAs on one side. The models vary in their distance between the TSSs (in nucleotides) but all of them have a binding region of 200 nucleotides and were previously described. The figures were cut from 120 s and the long tail is presented in small subfigure. The x-axis is divided in bins of 3 s while the y-axis represents the probability that the production events occurs within the bin duration.

In general, the kinetics of transcript production is similar in all bidirectional promoters, giving rise to distributions of similar shape. The steps that most contribute to the shape of the distributions are the rate limiting steps at both TSSs. However, the results show that mean and standard deviation of the distribution of intervals between productions of consecutive RNA molecules can be tuned, to some extent, by the relative positions of both TSSs and by the geometry of the promoter.

Divergent promoters with binding region between TSSs (figure 4.11 A) has approximately the same mean time between consequent produced RNAs on one side as the single unidirectional promoter (figure 4.11 F) and also the standard deviation is almost the same. However the other divergent promoters (with binding region of 50 nucleotides outside of the TSSs – see figure 4.11 B) has a slower production rate and slightly higher standard deviation than the other divergent model (figure 4.11 A).

The overlapping divergent promoters (figure 4.11 C) has a time interval distribution with a long tail that emerges from the interference between occluding RNAps and thus more than double the mean production time while the standard deviation is almost 9 times higher than the divergent promoter in model A. This mean production value is explained because only one RNAp is able to initiate at a time and the elongating RNAp removes the diffusing RNAps from the other side completely.

Convergent promoter (figure 4.11 E) has only slightly higher mean production time than divergent and unidirectional ones but the interference causes the standard deviation to become higher (two times the value from model A) and the time distribution to have a long tail as well. For example, in a convergent promoter, when an RNAp starts the elongation process, it will, generally, encounter one to a few diffusing RNAps in the opposite direction. When it does so, and according to reaction (8) in table 3.2, it will cause the diffusing RNAp to be released from the DNA template. Consequently, it is more likely that the next diffusing RNAp that will successfully reach the TSS and begin the elongation process will be one travelling in the same direction as the previous elongating RNAp, than one diffusing in the opposite direction.

If the above is true, than one should expect that the bidirectional promoters with distinct arrangements differ in dynamics of production of the two types of RNA, specifically, they should to differ in the degree of correlation of choices between the production of one or the other RNA. If the autocorrelation is null, there is no effect of the interferences between RNAps on the production of transcripts. If the autocorrelation is positive, it implies that once one of both types of RNAs are produced, the promoter is biased by the interferences to produce the same type of RNA in the next event. If the autocorrelation is negative, the opposite is more likely. However, there may be additional sources of correlation, aside geometry.

In figure 4.12 we present the correlation (this calculations are explained in the Materials and Methods section) of these choices for several lags for the models described before. For the F model we constructed a vector of two, non-interacting, unidirectional promoters and also calculate the degree of correlation between them.

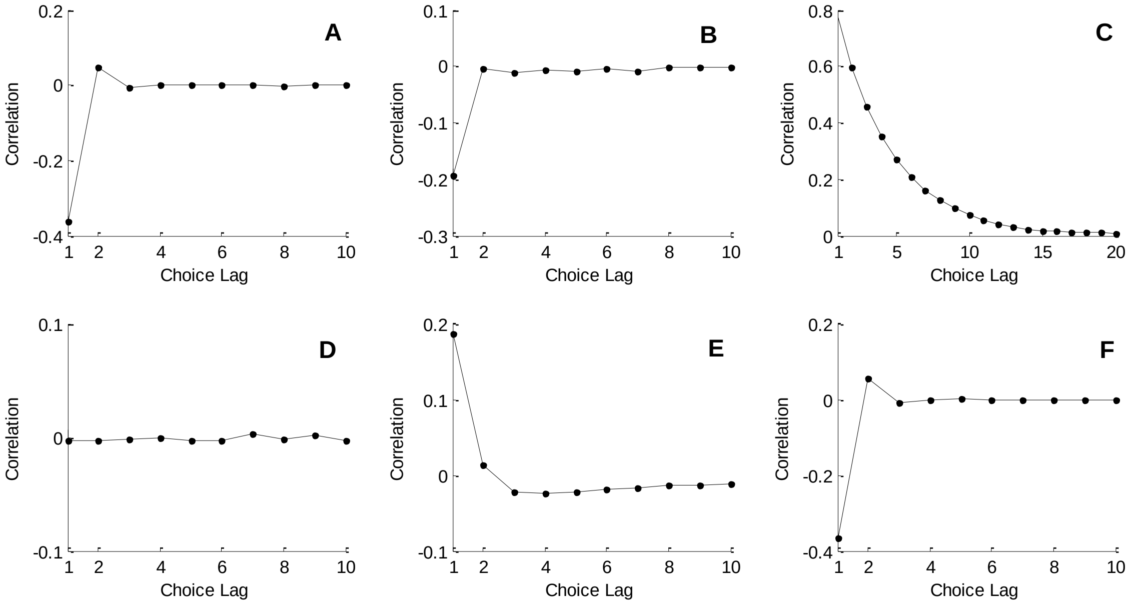

Figure 4.12 – The Pearson correlation between sequences of choices of elongation directions. All the models are the same as in figure 4.11. Here, Choice Lag does not have units, since we use vectors of choices (0 for the "left" promoter and 1 for the "right" promoter").

Results from figure 4.12 confirm that the dynamics of RNA production from both TSSs are dynamically correlated because of the interference between the RNAps at various levels, and add to this, the role that the rate limiting steps play on the correlation between the kinetics of RNA production by both promoters.

First, we can observe in figure 4.12 A, that there is negative correlation between consecutive production events, but this correlation is not propagated for longer lags. This correlation arises from the existence of rate limiting steps, as seen by comparing with figure 4.12 D, whose model has no rate limiting steps. When a TSS is occupied by RNAp, it will remain in that state for a period of time (whose duration is determined mostly by the open complex formation). Due to that, it is more likely that the next RNA produced will be from the other TSS, since for a production event to occur in the first TSS, it is necessary to wait not only for the clearance of the promoter, but also for the start and completion of the next open formation event. It is also due to this, that there is a small positive correlation between choices for lag of 2.

Decreasing the nucleotide distance between both TSSs of a divergent promoter weakens this correlation and as this distance decreases, the “loading capacity” of the promoter also decreases, that is, the number of RNAps that can be bound to it at any moment is smaller. It is this loading capacity that allows one successful transcription event at one TSS to be followed by another at the other TSS in a correlated fashion.

By further decreasing the distance between both TSSs, there is an abrupt change in the kinetics of transcription, which might not be very clear in figure 4.11 C, but is clear in figure 4.12 C. This can be explained because if one of both TSSs is loaded with RNAp, going through the open complex formation. The other one is likely to not be loaded by an RNAp, since the region between both TSSs is smaller and because the RNAp at the TSS impedes diffusing RNAps to reach the TSS. Once the open complex formation is completed, the RNAp moves along the template, and as it does so, it collides to any RNAp diffusing in the opposite direction, therefore not allowing them to reach the other TSS.

Only RNAps diffusing in the same direction can reach a TSS, specifically, the one from which the elongating RNAp departed from. Due to that, a very strong positive correlation of consecutive choices emerges in the divergent overlapped arrangement (see figure 4.12 C).

In figure 4.12 E, we observe a similar phenomenon. Due to the convergent arrangement, it is more likely that one elongation event from one TSS is followed by another elongation from the same TSS than from the other. The effect is not as strong as in the previous case because in this case, there is a strong possibility that while one RNAp is elongating, another elongating event can still start at the other TSS, and when two elongating RNAps collide, they have identical probabilities of being removed from the DNA template, which prevents the RNA production from that fallen RNAp. Note that the probability of two elongating events from each of the TSS occurring simultaneously in the previous case is dependable on the distance between TSS (as will be confirmed in the next section).

Finally, in figure 4.12 F we show the correlations between consecutive transcripts production from two independent unidirectional promoters. The similarity between this figure and Figure 7A also confirms shows how negative correlation emerges due to the rate limiting steps.

This is similar to the toggle switch. In that case, the positive correlation is due to the protein levels not decaying instantaneously. In general it stays in the same state but since there is a probability of each moment of switching, as the lag increases, the correlation decreases, implying that for some lag, they are no longer correlated.

The correlation choosing a direction of transcription is also depending on the noise levels of promoter, for example how much variance is inherent in transcription initiation. Open complex formation, isomerization, closed complex formation and promoter escape being inefficient causes the time interval distribution to have larger mean and variance. This in turn reduces the occlusion or interference inherent correlation. Less noisy process can be coordinated more efficiently in accordance to time interval distributions and from there on to correlation distances.

In the next section, we show how the distance between both TSSs affects both the correlation between choices and the mean RNA levels for divergent and convergent promoters.

4.7. Dynamics of RNA production as a function of the distance between TSSs

In this section we study the effects of changing the distance between TSSs to dynamics of RNA production in the sense of correlation between the directions choosing of the promoters and of mean RNA levels. In order to do so we simulated divergent and convergent promoter dynamics as a function of the distance between TSSs. The binding region of both promoters in these simulations is set to 300 nucleotides.

For these simulations we did 50 independent simulations with 10^5s each, and calculated the Pearson correlation just for lag 1. Using lag 1, we can observe how the distance between TSSs affect exactly the next choice of production. In figure 4.13 we present the results from those simulations for convergent (○) and divergent (□) promoters.

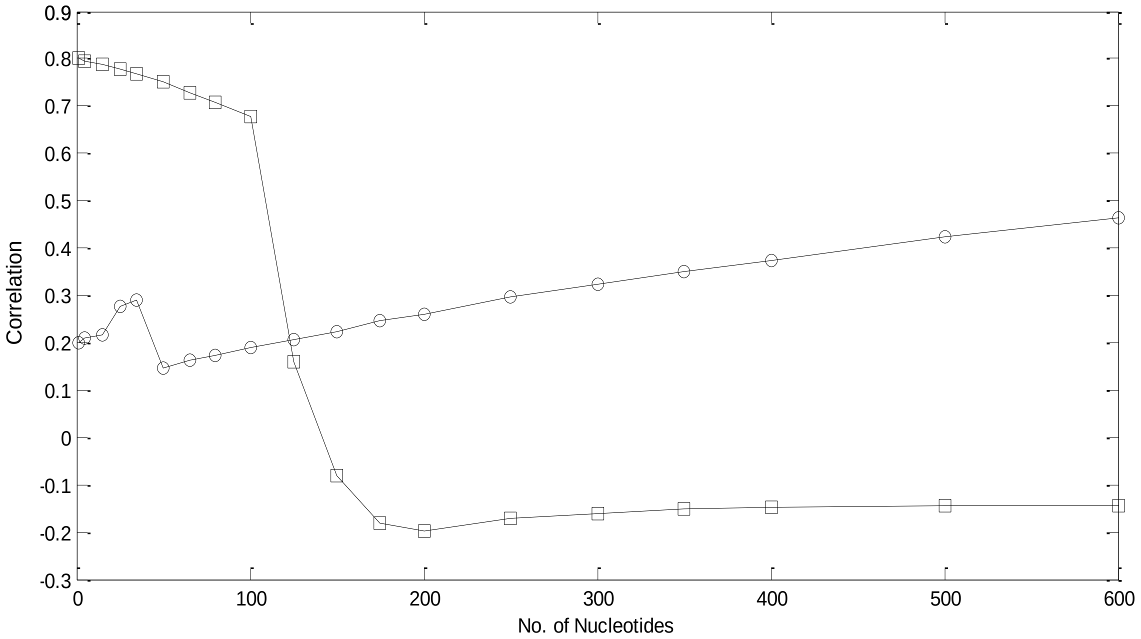

Figure 4.13 – Pearson correlation at lag 1 between the choices of directionality of consecutive RNAs as a function of the distance between both TSSs for convergent (○) and divergent (□) promoters.

From figure 4.13, we can observe that the distance between both TSSs is a critical variable in the degree of correlation of consecutive choices of direction of transcription (especially in divergent promoters). For distances smaller than ~110 nucleotides there is strong positive correlation between consecutive choices. At this point, as the distance is increased, there is an abrupt change, and the choices become anti-correlated. This transition corresponds to the change in the structure from overlapping to not overlapping promoters. When overlapping (<110 nucleotides), the RNAp at a TSS, when elongating, clears the other side from any diffusing RNAp, making it more likely that the next elongation event will be from the same TSS as the previous. When this distance is larger, the correlation becomes negative because of the rate limiting steps at the TSS and because the activity at one TSS is less affected by the activity at the other TSS (as can be seen by the increase in the mean RNA levels in figure 4.14).

In the case of convergent promoters, there is interference between the activities of both promoters for all distances, as elongating RNAps clear the other promoter from any RNAp. This interference increases with distance because the longer is the time that it takes for the elongating RNAp to pass by the other TSS, the longer will be the time interval during which no successful transcription event can arise from this other TSS. The small peak at the 35 nucleotides between TSSs is due to the fact that at such a distance, the RNAp at one TSS impedes even the formation of an open complex at the other TSS. Once the distance is larger, the correlation suffers a decrease as the other TSS can also form an open complex simultaneously, and (while rarely) this complex can end up firing before and thus clear the first TSS as well.

The temporal correlations shown in figure 4.13 have a significant effect on the mean RNA numbers at near equilibrium shown in figure 4.14. In divergent promoters, in general, the stronger is the positive correlation, the smaller is the mean number of RNA at near-equilibrium. In the case of convergent promoters, the relationship between correlation and mean RNA numbers is the opposite. In both geometries, the stronger is k_{bind}, the stronger are these correlations (both the positive and the negative ones). Finally, note that, beyond a certain distance between TSSs, further increases in distances no longer change the mean RNA levels significantly. This is due to other rate limiting steps, such as the open complex formation, that limit further increases in the rate of RNA production.

In convergent promoters, as the distances between the two TSSs increase, first there is strong increase in mean RNA numbers, as the distance becomes large enough for having both TSSs simultaneously occupied in the downstream region by an RNAp (same point as the decrease in correlation in figure 4.13). After that, further increases in the distance leads to a decrease in mean RNA numbers as the number of interferences and collisions between elongating RNAps increase (as show in figure 4.14).

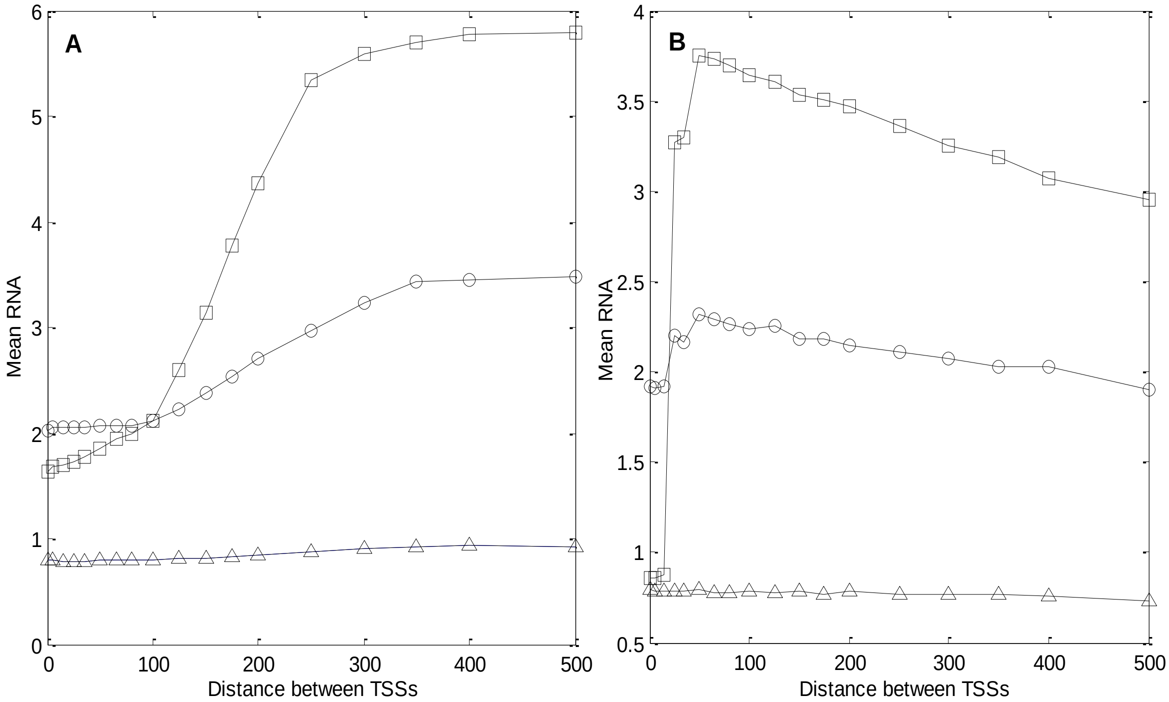

Figure 4.14 –Mean RNA numbers in the (A) divergent (B) convergent promoters as a function of the distance between the TSSs. In these models, both genes are identical and, thus, so are their mean RNA levels. The standard k_{bind} is represented in (□), k_{bind} / 10 is represented in (○) and k_{bind} / 100 is represented in (△).

In figure 4.14 we show how the mean RNA levels change as a function of the distance between TSSs with 3 different k_{bind} values. Overall increment of the distance between TSSs decreases the occlusion and collisions between elongating and diffusing RNAps in the divergent promoter (see figure 4.14 A). The consequences to the mean RNA number is more tangible in the regime where TSSs are partly overlapped, that is, RNAp binding and initiating transcription causes occlusion in the other TSSs. The low rate of k_{bind} does not seem to significantly have a change in

RNA numbers as promoter region changes. This is mainly because of non saturated RNAp binding rate as in opposition the standard k_{bind} rate has dramatic effects as the promoter region expands.

Beyond a certain length, further increases in length no longer increase significantly the mean RNA level. This is due to other rate limiting steps such as abortive initiation, open and closed complex formations, limiting further increases in the rate of RNA production. The size of the length for which further increases in length no longer cause an increase in RNA production rate depends on the kinetic parameters of these various rate limiting steps.

The convergent promoter dynamics (see figure 4.14 B) in other hand follows a different trend when it comes to mean RNA levels. The mean levels at overlapping TSSs (distance 1) are very close to divergent case but the mean level then increases very quickly reaching the optimal distance for RNA production at 50 nucleotides between TSSs. After this the mean starts to slowly decrease as the distance between TSSs grows and the traffic increases. With two higher binding rates this trend is clearly visible but as before the slowest binding rate produces nearly uniform production rate with all distances between TSSs. This can be seen as a consequence of having low population of diffusing RNAps.

4.8. Repression dynamics

4.8.1. Size of the repressor and the binding kinetics of a repressor

In this section we also used 50 simulations each with 10^5 s of duration on a divergent promoter with a distance of 200 nucleotides between the TSSs and a binding region of 200 nucleotides.

For these simulations we used just one repressor with a kinetic constant for its binding of $k_{rep} = 0.01\ s^{-1}$ and of its dissociation of $k_{unrep} = 0.01\ s^{-1}$. The repressor also has a reading head in the center, so that’s why it is easier to choose an odd number of occupied nucleotides (where they have the same amount of nucleotides divided in both halves). If we needed to have an even repressor, we would just need to choose one of the sides to remove occupied nucleotides until we reached the amount that we wanted. The center of the repressor was always at the position -100.

In table 4.6 we show the effects of varying the repressor size from 10 nucleotides to 100 nucleotides in the dynamics of RNA production: mean RNA levels, CV^2 and Fano Factor. We also test this effect with a lower binding affinity (100 times lower than the standard value [12]).

Table 4.6 - The mean, CV2 and Fano Factor with the variation of the repressor size. For these simulations we used two binding kinetics of the RNAp (the standard value and 100 smaller).

k_{bind}	Mean		CV^2		Fano Factor	
	Right	Left	Right	Left	Right	Left
Size 11	4.36	4.39	0.16	0.16	0.70	0.70
Size 51	3.97	4.03	0.18	0.18	0.71	0.73
Size 101	4.91	4.90	0.15	0.15	0.74	0.74
k_{bind}/100	Mean		CV^2		Fano Factor	
	Right	Left	Right	Right	Left	Right
Size 11	0,32	0,33	2,99	2,92	0.96	0.96
Size 51	0,27	0,27	3,63	3,64	0.98	0.98
Size 101	0,22	0,22	4,63	4,65	1.02	1.02

The repressor size can vary from small molecules to large distances (looping the DNA) [90,111,112] to gain different effects on the promoter dynamics. DNA looping is one way of achieving longer repression distances by preventing the RNAp binding to DNA at that region. Single molecules can also bind directly to TSS to prevent the transcription initiation.

From the results in table 4.6 we can see that the most effective size, with the same repression kinetic would around 50 nucleotides, which is around the same size of the RNAp. This is because the repressor competes with the RNAp in the binding to the DNA template, so if the size is bigger than the RNAp, the RNAp will have a bigger propensity to bind to the DNA than the repressor. If the repressor size is smaller than the RNAp then when bound the repressor occupies less nucleotides, which means that the RNAp will have more space to bind to. This is only true for a k_{bind} of the RNAp on the standard value, because with 100 times smaller, the RNAp will not bind enough times to be able to compete with the repressor, so the bigger the size, the smaller the mean. In the standard value, we don´t observe significant changes in the CV^2 and the Fano Factor, while with the value 100 times smaller both the CV^2 (with a 50% increase with the higher size) and and Fano Factor increases (the FF goes higher than 1).

In table 4.7 we exhibit the effects of varying the dissociation kinetic constant in the dynamics of RNA production: mean RNA levels, CV^2 and Fano Factor. For this simulation we used a repressor with 101 nucleotides and the standard binding value for the RNAp.

Table 4.7 - The mean, CV2 and Fano Factor with the variation of the dissociation kinetic constant. For these we used a repressor with 101 nucleotides and the standard value for the RNAp binding.

k_{unrep}	Mean		CV^2		Fano Factor	
	Right	Left	Right	Left	Right	Left
k_{unrep}	4.93	4.91	0.15	0.15	0.74	0.74
k_{unrep} /5	4.50	4.49	0.23	0.24	1.04	1.08
k_{unrep} /10	4.13	4.10	0.35	0.35	1.45	1.44
k_{unrep} /15	3.75	3.77	0.47	0.47	1.76	1.77
k_{unrep} /20	3.48	3.49	0.58	0.58	2.02	2.02

From table 4.6 we saw that for the standard value, the repressor with 101 nucleotides could not compete with the binding of the RNAP, but lowering the dissociation constant, we were able to repress the production (with the constant 20 times smaller, we obtained a decrease in the mean RNA levels up to 29%). The CV^2 and the Fano Factor increased with the decrease of the constant, and we were able to observe a higher noise then a pure Poisson process with the addition of the repressor.

In the next section we will study how different binding positions can affect different mechanisms, and how the repression affects the correlation between Time series of RNA production.

4.8.2. Different repression mechanisms

The most common mechanism of repression of transcription is steric occlusion, where the repressor blocks the access of the RNAp to a specific region of the promoter [1]. Depending on where the binding site is located, the blocking can affect different steps in transcription initiation, going from the binding to the DNA template to the closed complex formation, open complex formation or preventing the promoter escape [87, 88].

This mechanism allows, in theory, to block transcription completely since, provided a very large number of repressor molecules, the expected time for a repressor to bind to the DNA is virtually zero, hampering any transcription event. The only case where repression would not be complete is if there was sufficient space between the region occupied by the repressor and the TSS for an RNAp to bind. In this scenario, as the number of repressor molecules increase, the rate of RNA production would decrease until reaching a plateau of minimum expression rate that could not be further decreased.

Here we first investigate the kinetics of transcription of a unidirectional promoter subject to a repressor as a function of the number of repressor molecules and the position of their binding site.

Namely, we model promoters with the binding site for the repressor centered at positions +1, +12, and +37. In all cases, the repressor occupies 10 nucleotides in each side of this position.

To model repression we need to introduce in the model realistic rate constants controlling the binding and the dissociation of the repressor molecules to the DNA template. The ratio between these rates has been estimated for several genes and repressor molecules [113]. On average, one repressor molecule is bound to its binding site approximately 80% of the time. We set the rate constants of binding and unbinding such that the ratio between them is 4.167 (see table 3.3). For all the simulations in this section the rates of binding and dissociation of the repressors are always identical.

It is noted that the model assumes that RNAps cannot, by collision or other means, dislodge the repressors from their binding sites, but stays paused at the same place [114]. If an RNAp is occupying the binding region of the repressor, then the repressor molecule cannot bind to the DNA, which means that the kinetics of dissociation of the repressor from the DNA template only depends on the kinetic rate of dissociation.

Table 4.8 - Repression of unidirectional promoter at various steps of transcription initiation. Mean RNA levels as a function of the number of repressors present.

Number of repressors	Repressor of promoter escape	Repressor of open complex formation	Repressor of closed complex formation
1	4.26	4.88	5.78
2	3.47	4.25	5.71
3	2.85	3.72	5.64
5	2.23	2.99	5.56
7	1.85	2.52	5.41
10	1.54	2.01	5.30
20	1.07	1.2	4.86
50	0.76	0.57	3.98
100	0.66	0.31	3.14
200	0.62	0.17	2.40
500	0.59	0.07	1.73
1000	0.57	0.03	1.45
10000	0.43	0.003	1.14
50000	0.21	0.0007	1.01

Finally, to assess the strength of repression we define a “repression factor” as the ratio between mean RNA numbers at near-equilibrium when no repressors are present (equal to 5.85) and the mean RNA numbers when a certain number of repressors are present. In figure 4.15 we present how this quantity varies with the position of the binding site and with the number of repressors in the cell.

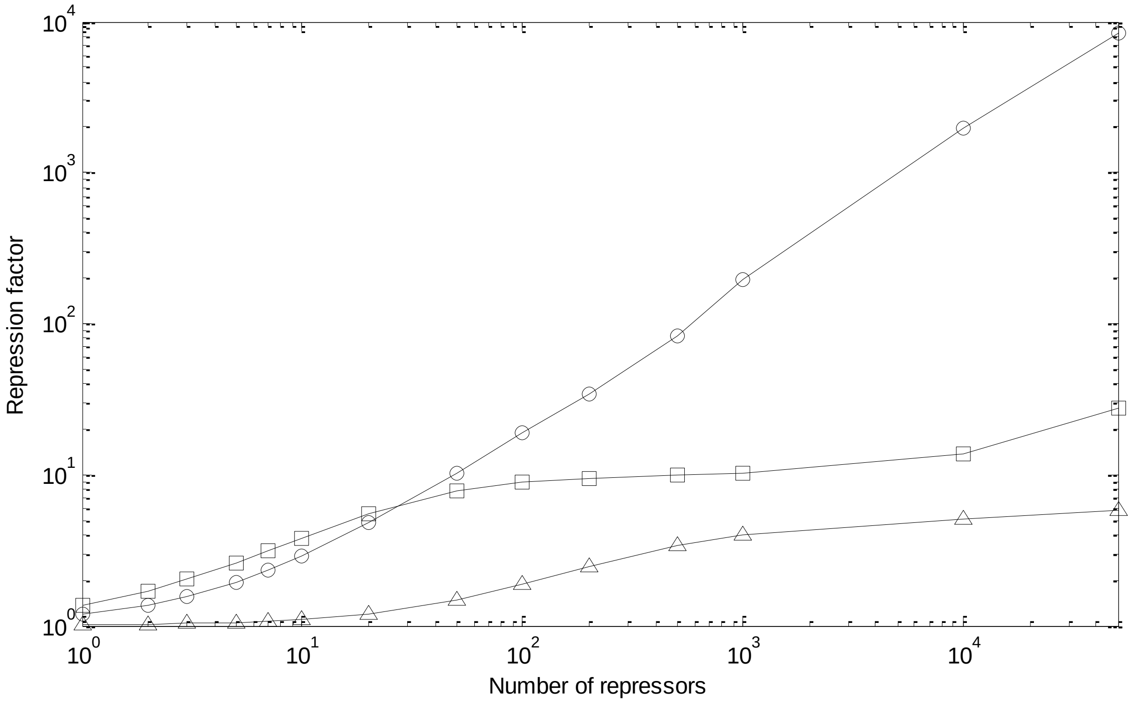

Figure 4.15 – Repression of unidirectional promoter at various steps of transcription initiation. Y-axis is the repression factor and x-axis is the number of repressor molecules. The repressed steps are, closed complex formation (△), open complex formation (○) and is the promoter escape (□).

To repress specific steps in transcription initiation we have to take a look at the RNAp footprint: so the RNAp is diffusing it occupies 55 nucleotides, but as the rate limiting leading to open complex step are taking place at the TSS the footprint of the RNAp increases to 75 in the downstream direction (during isomerization) implying that it now occupies the 20 nucleotides following the TSS. Following the promoter escape, the release of the sigma factor reduces the RNAp footprint to 25 nucleotides.

Using this information we used three different binding sites for the repressors such that at +1 it blocks the closed complex formation, at +12 it allows the closed complex but it blocks the open complex formation, and at +37, it allows initiation to be completed, but it blocks elongation. Note that in the section we are using a repressor with a size of 21 nucleotides as it is the value for the principal binding site of the lac repressor [87].

From figure 4.15 we can observe that in all cases the increase in the repressors numbers increased the repression factor, which also depends on the location of the binding site. A binding site at the TSS of right after it (at +12 and +37) provides equally efficient repression for small number of repressors. For large number of repressors, the efficiencies of these positions differ. Repression is stronger if the repressor blocks the open complex formation (at +12). Repressing the closed complex formation is the least efficient repression since binding and diffusing of RNAps on the template is a fast process, and thus able to compete with the kinetics of binding and unbinding of repressors. It of interest to note also that, for small number of repressors, blocking elongation is the most efficient mechanism since only elongating RNAps or repressors occupy this region (less competition than in the regions of diffusing RNAps).

When blocking the open complex formation at +12, repressors only compete with isomerization (thus also the open complex since they form very rapidly after isomerization), thus increasing the number of repressors causes linear increases on the rate of RNA production. On the other hand, increasing the number of repressors blocking elongation causes non linear changes in RNA production kinetics. At first, repressors only delay the movement of elongating RNAps but do not actually prevent elongation. This only has a limited effect in decreasing RNA production as for a certain range further increases in number of repressors cause little effect. Only when the speed of binding of repressors overcomes the speed of elongation do further increases in repressors numbers lead to additional decreases in the production of RNA molecules. In this regime, RNAps are prevented from leaving the TSS as the template is virtually always occupied by a repressor. Our results are in good agreement with the model made by Sanchez and colleagues [87] and his observations of the repression mechanisms.

In bidirectional promoters, the location of the repressor binding site, aside from affecting the overall production of RNA, it can also bias the kinetics of production of both RNA molecules. We tested different positions for the binding site in convergent and divergent promoters. Results are shown in tables 4.9 and 4.10 respectively for the convergent and divergent promoters.

Table 4.9 - Repression mechanism for convergent arrangement with 150 nucleotides between TSSs. Binding regions is consisted of 300 nucleotides. The repressor size is 21 nucleotides.

Convergent	Right	Left	Total
No repressor	3.56	3.54	7.1
-35	2.21 (closed)	3.37 (elongation)	5.58
+1	2.15 (closed)	3.15 (elongation)	5.3
+15	0.89 (open)	2.78 (elongation)	3.67
+36	0.59 (escape)	1.94 (elongation)	2.53
+75	0.92 (elongation)	0.93 (elongation)	1.85

Table 4.10 - Repression mechanism for the divergent arrangement with 150 nucleotides between TSSs. Binding regions is consisted of 300 nucleotides. The repressor size is 21 nucleotides.

Divergent	Right	Left	Total
No repressor	3.12	3.12	6.24
+36	1.15 (escape)	2.48 (diffusion)	3.63
+15	1.21 (open)	2.51 (diffusion)	3.72
-35	2.05 (closed)	2.88 (diffusion)	4.93
-50	2.04 (closed)	2.76 (diffusion)	4.8
-75	1.72 (diffusion)	1.75 (diffusion)	3.47
-35, -115	1.20 (closed)	1.22 (closed)	2.42

From the table 4.9, we can observe that the kinetics of RNA production from both TSSs can be biased in the convergent arrangement. The bias is a function of the relative positions of repressor binding site and TSSs. For example, placing a repressor at -65 in a convergent promoter (TSSs at +1 and +150) will affect only the expression of the TSS at +1 (right), thus biasing the otherwise unbiased production from both TSSs. (see table 4.9).

For example, placing the repressor at +36 in convergent promoters or at -35 in divergent promoters decreases the overall expression rate by approximately 60% and by 40%, respectively. Decreasing the overall expression without biasing the production of both TSSs is also possible for both arrangements (see tables 4.9 and 4.10), and to make it possible we need the repressor binding site to be located at the middle position between both TSSs, and using two repressors placed at symmetric positions from one another (see table 4.10).

It has been suggested that the relatively small distance between TSSs may facilitate the co-regulation of gene expression in both directions [1]. This would imply, for example, facilitating the simultaneous repression or activation of transcription from both genes. Due to this, we now study how repression may correlate the time series of RNA production from both TSSs.

We first used model with two identical unidirectional genes in the same cell and under the control of a repressor molecule with the same chemical kinetics. Aside this, they are independent from one another since the number of repressor molecules in this simulation is set to 100. In this case, the correlation was found to be null for all lags (this data is not shown), and so this model can be used as a null model of possible correlations between time series of RNA production. It is noted that if the number of repressor molecules was very small (from 1 to a few) a spurious anti correlation would be possible, as the repression of one of the genes would diminish the change of repression of the other.

We present in figures 4.16, 4.17 and 4.18 the correlation of the time series of RNA production at different lags for different arrangements with and without a repressor mechanism (the values in red are without repression and the values in black are with repression). In figure 4.16 we present a divergent overlapped promoter with a repressor molecule in the middle position between their TSSs, which blocks both promoters at the same time. In figure 4.17 we present two divergent promoters with 150 nucleotides between their TSSs and a repressor molecule also in the middle. This repressor only blocks the diffusion. We also use another repression mechanism, where there are two binding sites, and one binding sites blocks the other and blocks the production from the nearest TSS, this is represented in figure 4.17 with the filled balls. In figure 4.18 we present two convergent promoters also with a distance of 150 nucleotides and also a repressor in the middle of them.

To make this correlation studies, we compiled vectors of RNA time series (like the ones in figures 4.4, 4.5 and 4.6) and then we applied the correlation function as explained in the Materials and Methods chapter.

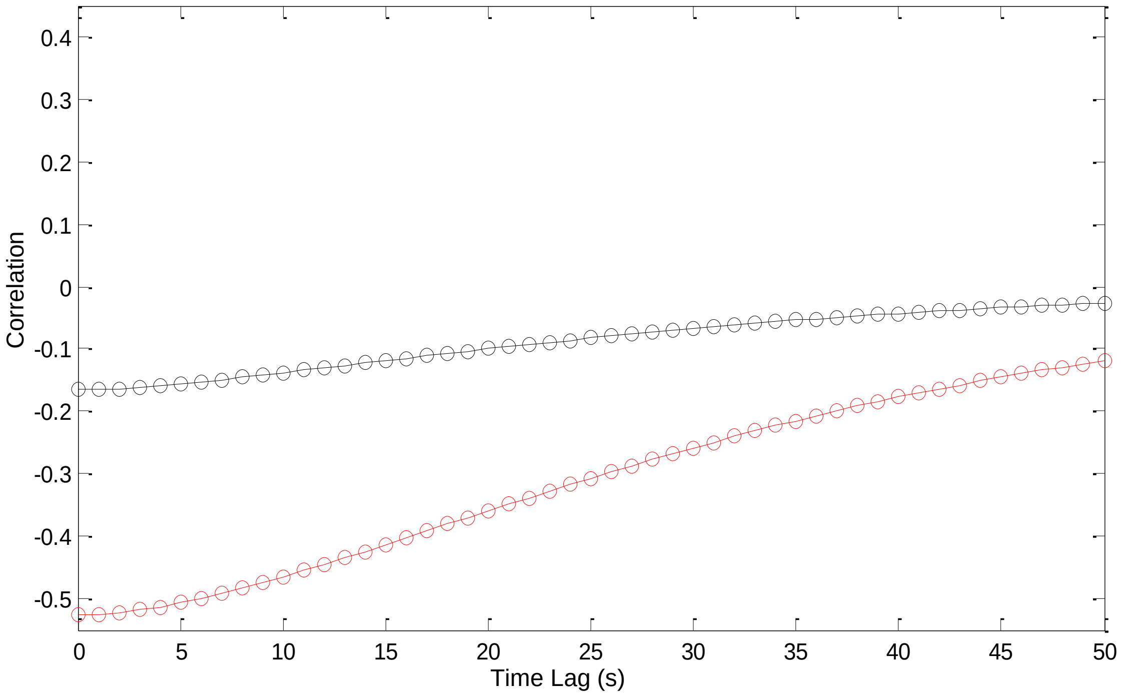

Figure 4.16 – Correlation of RNA production in time at different lags for different geometries: (○) divergent with 65 nucleotides between TSSs. The values in red correspond to the correlation without repression and the values in black to the correlation with repression.

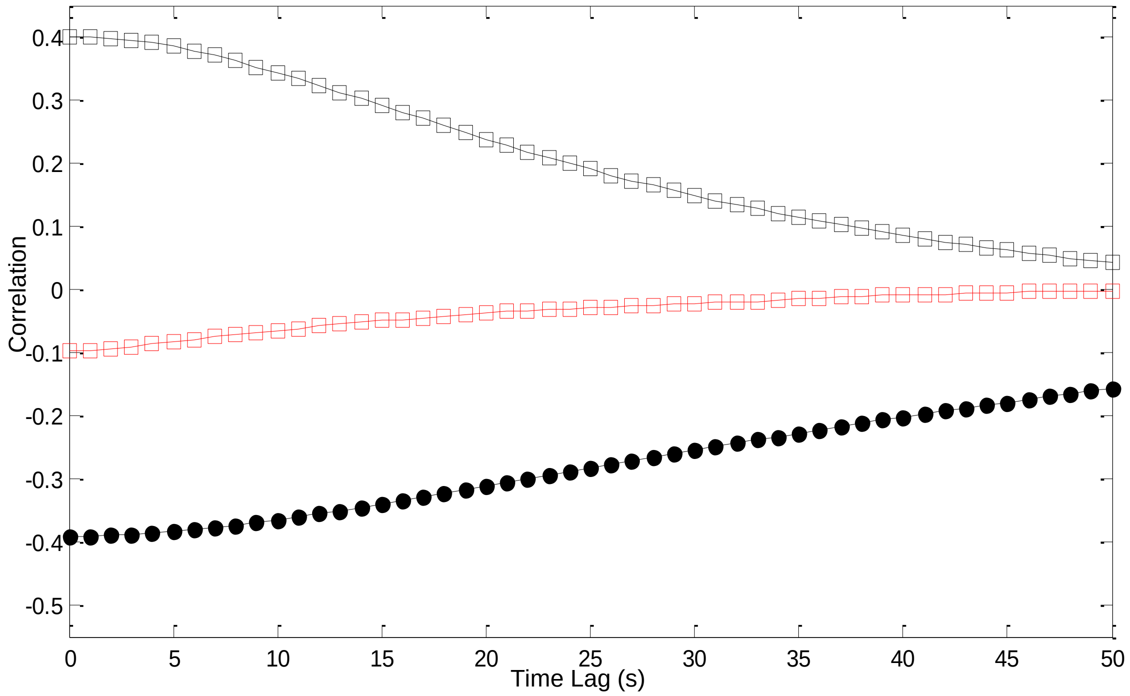

Figure 4.17 – Correlation of RNA production in time at different lags for different geometries: (□) divergent with 150 nucleotides between TSSs. Values in red correspond to correlation without repression and the values in black to correlation with repression, we also did a divergent promoter (•) in the same arrangement but with two repressor binding sites where one repressor site blocks the other and allows just one side to produce.

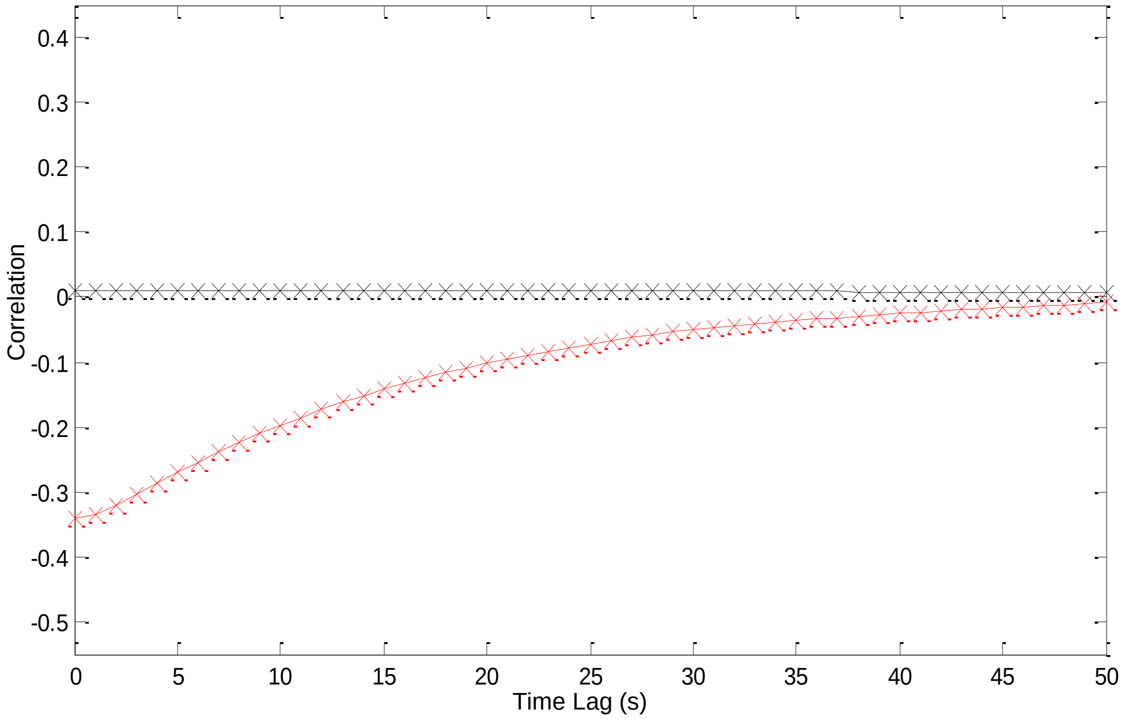

Figure 4.18 – Correlation of RNAp production in time at different lags for different geometries: (x) convergent with 150 nucleotides between TSSs. The values in red correspond to correlation without repression and the values in black to correlation with repression.

From figures 4.16, 4.17 and 4.18 and comparing the models with and without repressors (red and black lines respectively) is visible that, for all the arrangements (and for different distances between TSSs), the mutual repression mechanism correlates, in a positively the time series of RNA production productions of RNAs. With the addition of a repression mechanism in the convergent promoters (figure 4.18) the correlation goes from negative to a slight positive (almost null), while in overlapping divergent (figure 4.16) promoters it changes the inherent negative correlation between time series to far weaker negative correlation.

In the case of non-overlapped divergent promoters (figure 4.17) the repression went from almost null to a strong positive correlation. But in this promoter we can also make the correlation between the time series of the two genes more negative (filled balls in figure 4.17). This is possible if there are two binding sites and two distinct repressors (one for each TSS), and if the binding sites overlap, causing the repression of one of the two genes to hamper the repression of the other. This example suggests that complex repression mechanisms may allow correlations of any nature, positive or negative, and with any desired strength.

In figure 4.19 we exhibit a RNA time series and the time series of repression binding and dissociation in a convergent arrangement using the same model as the one described for figure 4.18, but using just 1 repressor.

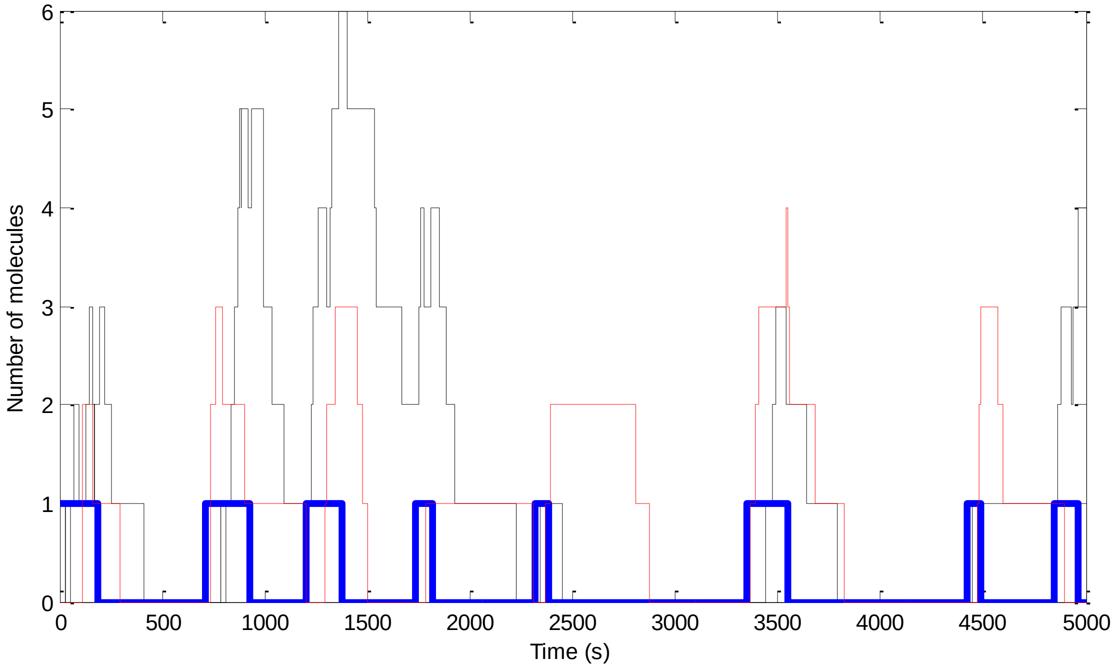

Figure 4.19 – Time series of RNA production for the convergent case (red and black line) and the time series of repression binding and dissociation (blue line)

From figure 4.19, we can observe that when the repressor value goes to 0 (the repressor is bound to the DNA template) the RNA numbers eventually go down, since the RNAps are blocked by the repressor. When the value goes to 1 (and this is because we only used 1 repressor), the repressor dissociates and allows to go back to a normal state of RNA production.

.

Chapter 5. Conclusions and future work

5.1. Conclusions

We studied the dynamics of expression of pairs of genes driven by closely spaced promoters within realistic intervals of parameter values for *E. coli*. For that, we proposed a new model for mimicking the process of transcription initiation, one RNA polymerase and one nucleotide at a time, whose dynamics is driven by the delayed stochastic simulation algorithm.

This model includes the binding of RNAp to the DNA template, the promoter search, the steps leading to open complex formation, abortive initiation and the transcription elongation. This model is the first to model the promoter region explicitly combining the traditional model of transcription initiation [1] with the nonspecific binding [12], studies of diffusion [92], abortive initiation [19] and collisions between RNAps at different stages [109]. We can use this model to study the dynamics of RNA production as a function of the arrangement, binding region, distance between both TSSs, chemical constants and repressor dynamics as these properties allow the promoters to have diverse kinetics of RNA production, not easily achievable by individual genes, or by genes that interact via transcription factors.

From the simulations of the dynamics RNA production as a function of the binding affinity we found that the value estimated by Singer and colleagues [12] is close to the saturating rate of RNA production. We also found that all the arrangements tested were less noisy then a pure Poisson process (FF < 1) and the values with the lowest Fano Factor are the ones closed to the value estimated by Singer and colleagues [12].

From the simulations with the abortive ratio we also observed that the lowest values of Fano Factor corresponded to the middle values of the abortive ratio that are able to reproduce the most observed values of Abortive to Productive Ratio in experimental studies [61].

We then studied the effects of changing asymmetrically the chemical constants of some of the steps in transcription initiation, namely, the binding of the RNAp and the isomerization. Using different arrangements we showed how highly diverse kinetics of RNA production have different degree of fluctuations in RNA numbers. We also showed how this diversity in RNA numbers may range from sub- to supra-Poissonian as a function of the dynamics of production, or even have the same noise as a pure Poisson process.

We then studied how the binding affinity and distance between TSSs affected the probabilities of binding of the RNAp to the DNA template. We found that using the standard value the binding probability was not uniform, however with a lower binding the distribution of probabilities turns uniform. With varying the distance between TSS we concluded that with a standard binding rate, there would be two cases where the distribution would also be uniform, the first is for a distance between TSSs so small to have more than one RNAp bound to it at any moment (divergent overlapping promoters). The second case would be for a distance between TSSs so large that the number of free RNAps in the cell would not be sufficient to fully occupy it.

Also from the same study we observed a discrete nature of the probabilities of binding of the RNAp to the DNA template. This discrete nature is a result of the length of DNA occupied by a RNAp being only one order of magnitude smaller than the distances between TSSs, and suggests that the placing of binding sites for repressors and activators relative to the TSSs is likely to be far from random, as they can alter in a different nature the overall probabilities of freely diffusing RNAps to reach either TSS.

We observed a great range of variability in the distributions of intervals between the productions of consecutive RNA molecules as a function of the configuration of the promoter. Different arrangements have different inherent interferences, as for example in the convergent promotes, colliding RNAps can be removed from the template, or in the divergent overlap one RNAp at the TSS during the execution of the rate limiting steps (isomerization and open complex formation) can block the other TSS from the another RNAp to start a transcription event, this interference affects mainly the tail of the time interval distribution as the mode of the distribution is related to the rate limiting steps (as observed that all the models have almost the same type of time distribution in the 0 to 60 s interval, except the model where the rate limiting steps were removed).

In terms of the correlation studies, this interference can lead to a positive correlation between consecutive choices of which of the two RNAs is transcribed next, as observed in the convergent and divergent overlapped cases. In the divergent case, it was observed a degree of anti-correlation that is accounted to the rate limiting steps that take place at both TSSs, as when this were removed, the correlation moved to a null value.

We also observed that the distance between TSSs can also have an effect both in the correlation between consecutive choices and mean RNA levels. In the divergent case, with a distance bellow 110 nucleotides there is a strong positive correlation between consecutive choices. Above this point, as the distance is increased, there is an abrupt change, and the choices become anti-correlated, due to the same reasons as before. In the convergent case, the correlation is always positive and has the tendency to grow as the distance is increased. The mean RNA levels have an opposite nature, as in the convergent case they decrease with the increase of the distance (due to the interference), and in the divergent case they increase with the increase of the distance until they saturate at around 300 nucleotides.

The study of the effects of repressors showed that these may be a means to achieve more complex patterns of behaviors, not possible otherwise. First we studied how the size could affect the dynamics of RNA production and we observed that with the standard binding affinity of the RNAp a higher size of the repressor than the RNAP (55 nucleotides) does not reduce effectively the mean RNA levels, because the binding of the repressor competes with the binding of the

RNAp. But with a lower dissociation rate of the repressor, then a repressor with a higher size is able to stay in the template long enough to compete with the RNAp.

First we used a unidirectional promoter to study the repression mechanisms at different steps of transcription. These mechanisms can either lead to similar as well as distinct kinetics of RNA numbers, depending on various factors including the number of repressor molecules in the cell. In general, for the same number of repressor molecules and same binding affinity to the TFBS, the effects on RNA production differ with the step of transcription that is repressed by the mechanism of repression.

These results are in good agreement with experimental observations that studied this type of repressions mechanisms. Schlax and colleagues [115] concluded that the most probable repression mechanism is the inhibition of closed complex formation, however another study using the lacUV5 promoter stated that repression of that promoter is likely to be achieved by the repression of the open complex formation [87]. Using an *E. coli* database, Garcia and colleagues [88] showed a broad distribution of binding regions of the repressors (but note that the majority of them are very close to respective TSSs). This observation along with the diversity of gene regulation mechanisms suggest that in different genes, repression can occur at different stages, including the promoter escape.

We then studied how the repression mechanisms affect bidirectional promoters, which have a more complex RNA production dynamics. We observed that repression in this type of promoters can lead to a bias on the mean RNA levels of both TSSs, but for example a symmetric location of the TFBS can also lead again to an unbiased production. We also observed that depending on various variables (such as the possibility of a double TFBS), repression by occlusion can either correlate or anti correlate the time series of RNA production.

Finally, we hypothesize that the multitude of regulatory steps of the dynamics of RNA production not only explains in part the observed diversity of kinetic behavior of genes in *E. coli*, but also suggests that multiple structures may give rise to similar kinetics of RNA production, thereby providing the means for the emergence of neutral evolutionary pathways.

5.2. Future work

As it was mentioned before, this model has a great ability to study the dynamics of transcription of promoters who are closely spaced. The next objective is to use this model to simulate an existing bidirectional system: the L-arabinose operon.

For this we might need to join this model with a previous model that couples transcription and translation [116] or using a simpler model of protein production like the one used in one of the groups previous models [9], since this system is positively and negatively regulated depending on the amount of arabinose present in the organism. But the system itself has a mechanism of regulating the level of arabinose using the araBAD and araC promoters and their regulators [77, 78]. For this model we will also add an activation system combined with the repressor systems already implemented. This next study will be supported by experimental single-molecule studies done by the LBD group in Tampere University.

Single-molecule studies [14] can give a better insight in how transcription works at the various levels. New studies can lead to a greater knowledge of this process so this model can be modified as our knowledge of this process grows.

Other types of bacterial operons can also be studied in the future using this model, like the bidirectional promoters: crp and yhfA [117]. This is an important system, because the crp gene encodes the cyclic AMP receptor protein, which is an important regulator of transcription initiation at various promoters in *E. coli* [118].

Chapter 6. References

[1] McClure WR (1985) Mechanism and control of transcription initiation in prokaryotes. Ann. Rev. Biochem 54: 171-204.

[2] Süel GM, Garcia-Ojalvo J, Liberman LM and Elowitz MB (2006) An excitable gene regulatory circuit induces transient cellular differentiation. Nature 440 (7083): 545-50.

[3] McAdams HH and Arkin A (1997) Stochastic mechanisms in gene expression. Proc Natl Acad Sci U S A 94 (3): 814-9.

[4] Ozbudak EM, Thattai M, Kurtser I, Grossman AD and van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. Nat Genet 31 (1): 69-73.

[5] Elowitz MB, Levine AJ, Siggia ED and Swain PS (2002) Stochastic gene expression in a single cell. Science 297 (5584): 1183-6.

[6] Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. 81 (25): 2340-2361.

[7] Arkin A, Ross J and McAdams HH (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. Genetics 149 (4): 1633-48.

[8] Ribeiro AS, Smolander OP, Rajala T, Häkkinen A and Yli-Harja O (2009) Delayed stochastic model of transcription at the single nucleotide level. J Comput Biol. 16 (4): 539-53.

[9] Ribeiro AS, Zhu R and Kauffman SA (2006) A general modeling strategy for gene regulatory networks with stochastic dynamics. J Comput Biol 13: 1630-1639.

[10] McClure WR (1980) Rate-limiting steps in RNA chain initiation. Proc Natl Acad Sci U S A 77 (10): 5634–5638.

[11] Brunner M and Bujard H (1987) Promoter recognition and promoter strength in the Escherichia coli system. EMBO J 6 (10): 3139-44.

[12] Singer P and Wu C-H (1987) Promoter Search by Escherichia coli RNA polymerase on a Circular DNA Template. The Journal of Biological Chemistry 262 (29): 14178-14189

[13] Mishra RK and Chatterji D (1993) Promoter search and strength of a promoter: two important means for regulation of gene expression in Escherichia coli. J. Biosci. 18 (1): 1-11

[14] Bai L, Santangelo TJ and Wang MD (2006) Single-molecule analysis of RNA polymerase transcription. Annu. Rev. Biophys. Biomol. Struct. 35: 343–60.

[15] Buc H and McClure WR (1985) Kinetics of Open Complex Formation between Escherichia coli R N A Polymerase and the lac UV5 Promoter. Evidence for a Sequential Mechanism Involving Three Steps. Biochemistry 24 (11): 2712-2723.

[16] Kontur WS, Saecker RM, Davis CA, Capp MW and Record MT Jr (2006) Solute probes of conformational changes in open complex (RPo) formation by Escherichia coli RNA polymerase at

the lambdaPR promoter: evidence for unmasking of the active site in the isomerization step and for large-scale coupled folding in the subsequent conversion to RPo. Biochemistry 45 (7): 2161-77.

[17] Djordjevic M and Bundschuh R (2008) Formation of the open complex by bacterial RNA polymerase--a quantitative model. Biophys J. 94 (11): 4233-48.

[18] Gralla JD, Carpousis AJ, and Stefano JE (1980) Productive and Abortive Initiation of Transcription in Vitro at the lac UV5 Promoter. Biochemistry 19: 5869-5873.

[19] Hsu LM (2002) Promoter clearance and escape in prokaryotes. Biochimica et Biophysica Acta 1577: 191– 207.

[20] Margeat E, Kapanidis AN, Tinnefeld P, Wang Y, Mukhopadhyay J, et al. (2006) Direct observation of abortive initiation and promoter escape within single immobilized transcription complexes. Biophys J. 90 (4): 1419-31.

[21] Kapanidis AN, Margeat E, Ho SO, Kortkhonjia E, Weiss S, et al. (2006) Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. Science 314(5802): 1144-7.

[22] Revyakin A, Liu C, Ebright RH and Strick TR (2006) Abortive Initiation and Productive Initiation by RNA polymerase Involve DNA Scrunching. Science 17 314(5802): 1139-1143.

[23] Beck CF and Warren RA (1988) Divergent promoters, a common form of gene organization. Microbiol. Rev 52(3): 318–326.

[24] Tommasi S and Pfeifer GP (1999) In vivo structure of two divergent promoters at the human PCNA locus: Synthesis of antisense RNA and S phase-dependent binding of E2F complexes in intron 1. J Biol Chem 274 (39): 27829-38.

[25] Adachi N and Lieber MRL (2002) Bidirectional gene organization: a common architectural feature of the human genome. Cell 109 (7): 807-9.

[26] Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, et al. (2004) An Abundance of Bidirectional Promoters in the Human Genome. Genome Res. 14 (1): 62-6.

[27] Collado-Vides J, Magasanik B and Gralla JD (1991) Control site location and transcriptional regulation in Escherichia coli. Microbiol Rev. 55 (3): 371-94.

[28] Rojo F (1999) Repression of transcription initiation in bacteria. J Bacteriol. 181 (10): 2987-91.

[29] Shearwin KE, Callen BP and Egan JB (2005) Transcriptional interference – a crash course. TRENDS in Genetics 21 (6): 339-45.

[30] Gillespie DT (2007) Stochastic simulation of chemical kinetics. Annu. Rev. Phys. Chem. 58: 35-55.

[31] Gillespie DT (1992) A rigorous derivation of the chemical master equation. Physica A 188: 404-425.

[32] Wilkinson DJ (2006) Stochastic Modelling for Systems Biology. Chapman & Hall/CRC Mathematical & Computational Biology Series.

[33] Lloyd-Price J (2011) Simulating stochastic chemical kinetics with dynamic compartmentalization at runtime. Master of Science Thesis. Tampere University of Technology.

[34] Gillespie DT (1976) A general method for numerically sampling the stochastic time evolution of coupled chemical reactions. J. Comput. Phys. 22: 403-434.

[35] Gibson MA and Bruck J (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. J. Phys. Chem. A 104: 1876-1889.

[36] Li H and Petzold L (2006) Logarithmic direct method for discrete stochastic simulation of chemically reacting systems. Technical report. Department of Computer Science, University of California: Santa Barbara.

[37] Zhu R, Ribeiro AS, Salahub D and Kauffman SA (2007) Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models. J Theor Biol 246: 725-745.

[38] Golding I, Paulsson J, Zawilski SM and Cox EC (2005) Real-Time Kinetics of Gene Activity in Individual Bacteria. Cell 123: 1025–1036.

[39] Yu J, Xiao J, Ren X, Lao K and Xie XS (2006) Probing gene expression in live cells, one protein molecule at a time. Science 311: 1600-1603.

[40] Zhang G and Darst SA (1998) Structure of the Escherichia coli RNA polymerase alpha subunit amino-terminal domain. Science 281 (5374): 262-6.

[41] Minakhin L, Bhagat S, Brunning A, Campbell EA, Darst SA, et al. (2001) Bacterial RNA polymerase subunit ω and eukaryotic RNA polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly. Proc Natl Acad Sci U S A. 98 (3): 892-7.

[42] Darst SA, Opalka N, Chacon P, Polyakov A, Richter C, et al. (2002) Conformational flexibility of bacterial RNA polymerase. Proc Natl Acad Sci U S A 99 (7): 4296-301.

[43] Murakami KS, Masuda S and Darst SA (2002) Structural Basis of Transcription Initiation: RNA Polymerase Holoenzyme at 4 Å Resolution. Science 296 (5571): 1280-1284

[44] Campbell EA, Korzheva N, Mustaev A, Murakami K, Nair S, et al. (2001) Structural mechanism for rifampicin inhibition of bacterial rna polymerase. Cell 104 (6): 901-12.

[45] Vassylyev DG, Sekine S, Laptenko O, Lee J, Vassylyeva MN, et al. (2002) Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A˚ resolution. Nature 417 (6890): 712-9.

[46] Sousa R, Chung YJ, Rose JP and Wang BC (1993) Crystal structure of bacteriophage T7 RNA polymerase at 3.3 A resolution. Nature 364 (6438): 593-9.

[47] Liu C and Martin CT (2002) Promoter clearance by T7 RNA polymerase. Initial bubble collapse and transcript dissociation monitored by base analog fluorescence. J Biol Chem. 277 (4): 2725-31.

[48] Tunitskaya VL and Kochetkov SN (2002) Structural-functional analysis of bacteriophage T7 RNA polymerase. Biochemistry (Mosc) 67 (10): 1124-35.

[49] Steitz TA (2004) The structural basis of the transition from initiation to elongation phases of transcription, as well as translocation and strand separation, by T7 RNA polymerase. Curr Opin Struct Biol 14 (1): 4-9.

[50] Gruber TM and Gross CA (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. Annu Rev Microbiol 57: 441-66.

[51] Place C, Oddos J, Buc H, McAllister WT and Buckle M (1999) Studies of contacts between T7 RNA polymerase and its promoter reveal features in common with multisubunit RNA polymerases. Biochemistry 38 (16): 4948-57.

[52] Ross W, Gosink KK, Salomon J, Igarashi K, Zou C, et al. (1993) A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. Science 262 (5138): 1407-1413.

[53] Gourse RL, Ross W and Gaal T (2000) UPs and downs in bacterial transcription initiation: the role of the alpha subunit of RNA polymerase in promoter recognition. Mol Microbiol. 37 (4): 687-95.

[54] deHaseth PL, Zupancic ML and Record MT Jr (1998) RNA polymerase-promoter interactions: the comings and goings of RNA polymerase. J Bacteriol 180 (12): 3019-25.

[55] Bucher P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J Mol Biol 212 (4): 563-78.

[56] Nikolov DB and Burley SK (1997) RNA polymerase II transcription initiation: a structural view. Proc Natl Acad Sci U S A 94 (1): 15-22.

[57] Browning DF and Busby SJW (2004) The regulation of bacterial transcription initiation. Nature Reviews Microbiology 2: 57-65.

[58] Saecker RM, Record MT Jr and deHaseth PL (2011) Mechanism of Bacterial Transcription Initiation: RNA Polymerase - Promoter Binding, Isomerization to Initiation-Competent Open Complexes, and Initiation of RNA Synthesis. J Mol Biol 412 (5): 754-71.

[59] Sclavi B, Zaychikov E, Rogozina A, Walther F, Buckle M, et al. (2005) Real-time characterization of intermediates in the pathway to open complex formation by Escherichia coli RNA polymerase at the T7A1 promoter. Proc Natl Acad Sci U S A 102 (13): 4706-11.

[60] Vo NV, Hsu LM, Kane CM and Chamberlin MJ (2003) In vitro studies of transcript initiation by Escherichia coli RNA polymerase. 3. Influences of individual DNA elements within the promoter recognition region on abortive initiation and promoter escape. Biochemistry 42 (13): 3798-811.

[61] Hsu LM, Cobb IM, Ozmore JR, Khoo M, Nahm G, et al. (2006) Initial transcribed sequence mutations specifically affect promoter escape properties. Biochemistry 45 (29): 8841-54.

[62] Goldman SR, Ebright RH and Nickels BE (2009) Direct detection of abortive RNA transcripts in vivo. Science 324 (5929): 927-8.

[63] Hansen UM and McClure WR (1980) Role of the sigma subunit of Escherichia coli RNA polymerase in initiation. II. Release of sigma from ternary complexes. The Journal of Biological Chemistry 255 (20): 9564-70.

[64] Raffaelle M, Kanin EI, Vogt J, Burgess RR and Ansari AZ (2005) Holoenzyme switching and stochastic release of sigma factors from RNA polymerase in vivo.Mol Cell 20 (3): 357-66.

[65] Mooney RA, Darst SA and Landick R (2005) Sigma and RNA polymerase: an on-again, off-again relationship? Mol Cell 20 (3): 335-45.

[66] Lutz R, Lozinski T, Ellinger T and Bujard H (2001) Dissecting the functional program of Escherichia coli promoters: the combined mode of action of Lac repressor and AraC activator. Nuc Ac Res 29: 3873–3881.

[67] Singh SS, Typas A, Hengge R, Grainger DC (2011) Escherichia coli σ70 senses sequence and conformation of the promoter spacer region. Nucleic Acids Res. 39(12):5109-18.

[68] Singer PT and Wu CW (1988) Kinetics of promoter search by Escherichia coli RNA polymerase. Effects of monovalent and divalent cations and temperature. J Biol Chem 263 (9): 4208-14.

[69] Suh WC, Leirmo S and Record MT Jr (1992) Roles of Mg2+ in the mechanism of formation and dissociation of open complexes between Escherichia coli RNA polymerase and the lambda PR promoter: kinetic evidence for a second open complex requiring Mg2+. Biochemistry 31 (34): 7815-25.

[70] Cooper GM (2000) The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates.

[71] Hawley DK and McClure WR (1983) Compilation and analysis of Escherichia coli promoter DNA sequences. Nucleic Acids Res 11 (8): 2237–2255.

[72] Harley CB and Reynolds RP (1987) Analysis of E.coli promoter sequences. Nucleic Acids Res 15: 2343–2361

[73] Taylor K, Hradecna Z, Szybalski W (1967) Asymmetric distribution of the transcribing regions on the complementary strands of coliphage lambda DNA. Proc Natl Acad Sci U S A 57 (6): 1618-25.

[74] Dodd IB, Shearwin KE and Egan JB (2005) Revisited gene regulation in bacteriophage λ. Current Opinion in Genetics & Development 15 (2): 145-152

[75] Zeng L, Skinner SO, Zong C, Sippy J, Feiss M, et al. (2010) Decision Making at a Subcellular Level Determines the Outcome of Bacteriophage Infection. Cell 141(4): 682-691.

[76] Panchal CJ, Bagchee SN, and Guha A (1974) Divergent Orientation of Transcription from the Arginine Gene ECBH Cluster of Escherichia coli. J Bacteriol 117(2): 675–680.

[77] Schleif R, Hess W, Finkelstein S and Ellis D (1973) Induction Kinetics of the l-arabinose Operon of Escherichia coli. Journal of Bacteriology 115(1): 9-14.

[78] Schleif R (2000) Regulation of the L-arabinose operon of Escherichia coli. Trends Genet 16 (12): 559-65.

[79] Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, et al. (1997) The complete genome sequence of Escherichia coli K-12. Science 277 (5331): 1453-62.

[80] Warren PB and ten Wolde PR (2004) Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of Escherichia coli. J Mol Biol 342 (5): 1379-90.

[81] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291 (5507): 1304-51.

[82] Carpousis AJ and Gralla JD (1985) Interaction of RNA polymerase with lacUV5 promoter DNA during mRNA initiation and elongation. Footprinting, methylation, and rifampicin-sensitivity changes accompanying transcription initiation. J Mol Biol 183 (2): 165–177.

[83] Metzger W, Schickor P and Heumann H (1989) A cinematographic view of Escherichia coli RNA polymerase translocation. The EMBO Journal 8 (9): 2745- 2754.

[84] Schickor P, Metzger W, Werel W, Lederer H and Heumann H (1990) Topography of intermediates in transcription initiation of E.coli. EMBO J 9 (7): 2215-20.

[85] Record MT Jr, Reznikoff WS, Craig ML, McQuade KL and Schlax PJ (1996) Escherichia coli RNA Polymerase (Eσ70), Promoters, and the Kinetics of the Steps of Transcription Initiation. Second Edition of Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology 2: 792-821.

[86] Gama-Castro S, Salgado H, Peralta-Gil M, Santos-Zavaleta A, Muñiz-Rascado L, et al. (2011) RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). Nucleic Acids Res 39 (Database issue): 98-105.

[87] Sanchez A, Osborne ML, Friedman LJ, Kondev J and Gelles J (2011) Mechanism of transcriptional repression at a bacterial promoter by analysis of single molecules. EMBO J 30 (19): 3940-6.

[88] Garcia HG, Sanchez A, Kuhlman T, Kondev J and Phillips R (2010) Transcription by the numbers redux: experiments and calculations that surprise. Trends Cell Biol 20 (12): 723-33.

[89] Greive SJ and von Hippel PH (2005) Thinking quantitatively about transcriptional regulation. Nature Reviews Molecular Cell Biology 6: 221-232.

[90] Carey J, Lewis DE, Lavoie TA and Yang J (1991) How does trp repressor bind to its operator? J Biol Chem 266 (36): 24509-13.

[91] Harada Y, Funatsu T, Murakami K, Nonoyama Y, Ishihama A, et al. (1999) Single-Molecule Imaging of RNA Polymerase-DNA Interactions in Real Time. Biophysical Journal 76: 709–715.

[92] von Hippel PH and Berg OG (1989) Facilitated Target Location in Biological Systems. The Journal of Biological Chemistry 264 (2): 675-678.

[93] Bremer H, Dennis P and Ehrenberg M. (2003) Free RNA polymerase and modeling global transcription in Escherichia coli. Biochimie 85 (6): 597-609.

[94] Herbert M, Kolb A, and Buc H (1986) Overlapping promoters and their control in Escherichia coli: the gal case. Proc Natl Acad Sci U S A 83 (9): 2807–2811.

[95] Strainic MG Jr, Sullivan JJ, Collado-Vides J and deHaseth PL (2000) J Bacteriol 182 (1): 216-20. Promoter interference in a bacteriophage lambda control region: effects of a range of interpromoter distances.

[96] Dodd IB, Shearwin KE and Sneppen K. (2007) Modelling transcriptional interference and DNA looping in gene regulation. J Mol Biol 369(5):1200-13.

[97] Xue X, Liu F and Ou-Yang Z (2008) A Kinetic Model of Transcription Initiation by RNA Polymerase. J. Mol. Biol. (2008) 378, 520 – 529.

[98] Vo NV, Hsu LM and Kane CM and Chamberlin MJ (2003) In vitro studies of transcript initiation by Escherichia coli RNA polymerase. 2. Formation and characterization of two distinct classes of initial transcribing complexes. Biochemistry 42 (13): 3787-97.

[99] Hsu LM (2009) Monitoring abortive initiation. Methods 47 (1): 25-36.

[100] Phroskin S, Rachid Rahmouni A, Mironov A, and Nudler E (2010) Cooperation between translating ribosomes and RNA polymerase in transcription elongation. Science 328 (5977): 504-508.

[101] Callen BP, Shearwin KE and Egan JB (2004) Transcriptional Interference between Convergent Promoters Caused by Elongation over the Promoter. Molecular Cell, Vol. 14, 647–656

[102] Dahirel V, Paillusson F, Jardat M, Barbi M and Victor JM (2009) Nonspecific DNA-protein interaction: why proteins can diffuse along DNA. Phys Rev Lett 102 (22): 228101.

[103] Ribeiro AS and Lloyd-Price J (2007) SGN Sim, a Stochastic Genetic Networks Simulator. Bioinformatics 23 (6): 777-779.

[104] Taniguchi Y, Choi PJ, Li G-W, Chen H, Babu M, et al.(2010). Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. Science 329 (5991): 533-538.

[105] Bernstein JA, Khodursky AB, Lin P, Lin-Chao S and Cohen SN (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays 99 (15): 9697-702.

[106] Sundararaj S, Guo A, Habibi-Nazhad B, Rouani M, Stothard P, et al. (2004) "The CyberCell Database (CCDB): a comprehensive, self-updating, relational database to coordinate and facilitate in silico modeling of Escherichia coli" Nucleic Acids Res 32 (Database issue): 293-295.

[107] Wang Q, Tullius TD and Levin JR.(2007) Effects of discontinuities in the DNA template on abortive initiation and promoter escape by Escherichia coli RNA polymerase. J Biol Chem. 282 (37): 26917-27.

[108] Tang G, Roy R, Bandwar RP, Ha T and Patel SS (2009) Real-time observation of the transitionfrom transcription initiation to elongation of the RNA polymerase. PNAS 106 (52): 22175-22180

[109] Sneppen K, Dodd IB, Shearwin KE, Palmer AC, Schubert RA, et al. (2005) A Mathematical Model for Transcriptional Interference by RNA Polymerase Traffic in Escherichia coli. J. Mol. Biol. 346: 399–409.

[110] Ribeiro AS, Häkkinen A, Mannerström H, Lloyd-Price J and Yli-Harja O (2010) Effects of the promoter open complex formation on gene expression dynamics. Phys Rev E Stat Nonlin Soft Matter Phys 81(1 Pt 1): 011912.

[111] Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, et al. (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science 271(5253): 1247-54.

[112] Horton N, Lewis M and Lu P (1997) Escherichia coli lac repressor-lac operator interaction and the influence of allosteric effectors. Journal of Molecular Biology 265(1) 1-7.

[113] So LH, Ghosh A, Zong C, Sepúlveda LA, Segev R, et al. (2011) General properties of transcriptional time series in Escherichia coli. Nat Genet 43(6): 554-60.

[114] Lopez PJ, Guillerez J, Sousa R and Dreyfus M (1998) On the mechanism of inhibition of phage T7 RNA polymerase by lac repressor. J Mol Biol 276 (5): 861-75.

[115] Schlax PJ, Capp MW and Record MT Jr (1995) Inhibition of transcription initiation by lac repressor. J Mol Biol 245(4): 331-50.

[116] Mäkelä J, Lloyd-Price J, Yli-Harja O and Ribeiro AS (2011) Stochastic sequence-level model of coupled transcription and translation in prokaryotes. BMC Bioinformatics 12: 121

[117] Hanamura A and Aiba H (1991) Molecular mechanism of negative autoregulation of Escherichia coli crp gene. Nucleic Acids Res 19 (16): 4413-9.

[118] Zheng D, Constantinidou C, Hobman JL, Minchin SD (2004) Identification of the CRP regulon using in vitro and in vivo transcriptional profiling. Nucleic Acids Res 32 (19): 5874-93.