Modelling molecular diffusion in the nucleus and its role in gene regulation

Karen Amaral de Oliveira

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

UM41 - UMR7104 - UMRS1258 - Institut de génétique et de biologie moléculaire et cellulaire

THÈSE présentée par :

Karen AMARAL DE OLIVEIRA

soutenue le : **17 mars 2023**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Biologie des Systèmes

Modelling Molecular Diffusion in the Nucleus and its role in gene regulation

THÈSE dirigée par :

M. MOLINA Nacho	PhD, Université de Strasbourg

RAPPORTEURS :

M. NICODEMI Mario	PhD, Instituto Nazionale di Fisica Nucleare
M. SNEPPEN Kim	PhD, Niels Bohr Institute

AUTRES MEMBRES DU JURY :

Mme. DEJAEGERE Annick	PhD, Université de Strasbourg
M. BÉNICHOU Olivier	PhD, Sorbonne Université

Karen Amaral de Oliveira

Modelling Molecular Diffusion in the Nucleus and its role in gene regulation

Doctoral Thesis presented to the Graduate Program from the École Doctorale des Sciences de la Vie et de la Santé of the Université de Strasbourg as a requirement to obtain the degree of Doctor in Biophysics

Université de Strasbourg

École Doctorale des Sciences de la Vie et de la Santé

Nacho Molina

Université de Strasbourg

2023

Statement of Authorship

I hereby declare that the thesis submitted is my own work. All direct or indirect sources used are acknowledged as references. I further declare that I have not submitted this thesis at any other institution in order to obtain a degree.

Vocês sabem que é para vocês. ***Sempre*** *é, e será, para e por vocês.*

Acknowledgements

There is a gigantic and non-exhaustive list of people I should thank for being with me on this ride, and I'm scared of forgetting someone. I emphasize: if I forgot your name and you know I thanked you during my PhD be assured that I meant to thank you but was just forgetful, as I can be from time to time.

First and foremost, Nacho Molina, my PhD supervisor was open to taking me in this more than three years saga - with ups and downs and a pandemic in the middle! He is an amazing boss and PI (scientifically outstanding, really), but more than that: Nacho is a cool guy and a person I could talk to without being afraid of something or another. Which during a pandemic was what kept me a bit less insane than I could have turned into. Thank you, Nacho! It was a pleasure being your student.

Now, I would like to thank all the members of Nacho Molina's lab (past and present) for all the good laughs, scientific discussions and friendships we built together: my cutie pies (Abdul, Delphine, Hussain and Lasha), my princess (Sara), my culinary test subject (Olivier), Anaïs (and her natural excellence to reign us in) and the boys that I surrounded myself from the beginning to the end of this PhD (Sergio, Andrea, Atilla, Gui, Samuel, David, Maulik and Timothé). Computer power, sweeties! Another thank you to my francophone friends that spent lots of time trying to make sense out of my subpar french skills: Delphine, Olivier and Timothé, you are stars!

Outside my little theoretical bubble in Strasbourg, I have many people to be thankful for letting me in! To Juliana (who's always going to be my little jujuba), Paulo, Isabel and Iris for allowing me to speak Portuguese in the middle of France. To Marina for being there with life advice, food, and friendship - and all the judgement I deserve sometimes.

In a more theoretical and less strasbourgeoise vein: I need to thank Lucas for being the most amazing human being and being my sounding board and first reader and Carneiro for bringing sane insanity from duck memes. Thank you, guys!

The cool parts of my family should also be thanked, as they were - well - cool and respectful of my very confusing career choice: is it mathematics, biology or computer science? Well, it is everything at the same time and - at times - I felt it was nothing too, so there's that. By that, I mean my parents and little brother and a couple of others. And, yes, Fernanda, you are my family too. Obviously.

Work-related, I thank all my fellow pep-net friends, I'm so glad I met this heterogeneous system a bit chaotic, but so much fun and helpful! It wasn't a race with you it

was a crazy hike with the coolest view ever - also a bit deadly given, well, COVID-19. I also thank Davide Mazza and Tom Fillot from the network for hosting me at San Raffaele and teaching me experimental work: I fell in love with it and I really hope in doing more wet labs in the future!

From my Italian front at OSR, I also need to thank everyone at OSR: in special, Alessia, Dani, Fulvio and Ema: thank you for your patience with this little mathematician full of questions about how a wet lab works. A special thank you to Samuel Zambrano from OSR for taking me in too! I had the luck of meeting great people over there and I am so glad for the opportunity of meeting them!

I also thank my committee for reading and evaluating this thesis: I am proud of the work I did and I hope it is an enjoyable read for you; trust me, I enjoyed writing it.

Last, I thank you, my reader, that didn't stop reading once you saw this is my acknowledgements page. As a thank you for reading this, some advice: please, straighten your back, unclench your jaw and maybe drink some water. You deserve it.

"I can handle whatever I stumble upon

...

Most of the time"

Bob Dylan - Most of the Time

Karen AMARAL DE OLIVEIRA

Modelling Molecular Diffusion in the Nucleus and its role in gene regulation

Résumé

La diffusion des facteurs de transcription (FTs) dans le noyau joue un rôle crucial dans la régulation transcriptionnelle. La recherche par les TF d'une séquence d'ADN spécifique est l'un des principaux facteurs de l'expression des gènes. Ainsi, les interactions entre deux FTs dues à de faibles interactions protéine-protéine (IPPs) forment des clusters de FTs, influençant leur occupation à un site cible particulier. Pour comprendre comment la structure 3D de la chromatine affecte l'agglutination des FTs, nous avons proposé un modèle pour traduire la présence des IPPs dans le noyau et vérifié comment l'agglutination affecte l'allocation des FTs, en considérant la diffusion 3D/1D comme notre mécanisme de recherche. Ensuite, une fois qu'un FT est lié à une région, il recrute l'ARN polymérase (ARNP). En outre, les FTs inductibles restent dans le cytoplasme et translocation dans le noyau par le biais du complexe du pore nucléaire (CPN) après une signalisation appropriée. Afin d'intégrer ces mécanismes, nous avons proposé un autre modèle pour comprendre la dynamique de recherche des TF et de recrutement de l'ARNP. Nous avons obtenu des solutions déterministes et stochastiques vérifiant comment la transcription est renforcée à la périphérie du CPN et confirmée par l'analyse d'imagerie de gènes spécifiques. Enfin, nous avons incorporé le processus d'exportation de l'ARNm pour vérifier les différentes concentrations de transcrits cytoplasmiques, prouvant ainsi que le volume d'ARNm disponible dépend également du CPN. Par conséquent, notre travail montre des liens pertinents entre la structure de la chromatine, l'allocation des ressources transcriptionnelles et la stochasticité de la régulation des gènes.

Recherche par les FTs, Interactions Protéine-Protéine, Recrutement de l'ARNP, exportation d'ARNm, modèle mathématique, structure de la chromatine

Résumé en anglais

The diffusion of transcription factors (TFs) within the nucleus plays a crucial role in transcriptional regulation. The TF search for a specific DNA sequence is one of the main factors in gene expression. Thus, the interactions between two TFs due to weak protein-protein interactions (PPIs) form TF clusters, influencing their occupancy at a particular target site. To understand how the 3D structure of the chromatin affects the TF agglutination, we proposed a model to convey the presence of PPIs in the nucleus and verified how the clustering affects the TF allocation, considering the 3D/1D diffusion as our search mechanism. Then, once a TF is bound to a region, it recruits RNA Polymerase (RNAP). Besides, inducible TFs remain in the cytoplasm and translocate into the nucleus through the nuclear pore complex (NPC) upon proper signalling. To incorporate these mechanisms, we proposed another model to understand the TF search and RNAP recruitment dynamics. We obtained deterministic and stochastic solutions verifying how transcription is enhanced at the NPC periphery and confirmed through imaging analysis of specific genes. Finally, we incorporated the mRNA export process to verify the different cytoplasmic transcripts concentrations proving how the volume of available mRNA is also NPC-dependent. Therefore, our work shows relevant connections between the chromatin structure, the allocation of transcriptional resources and the stochasticity in gene regulation.

TF search, Protein-protein Interactions, RNAP recruitment, mRNA export, mathematical model, chromatin structure

Contents

List of Figures

List of Tables

0 Résumé Étendu de la Thèse

Introduction

Pour maintenir sa stabilité, la cellule doit optimiser ses systèmes. L'efficacité et la précision sont donc extrêmement importantes. Par exemple, deux cellules différenciées ayant le même ADN pourraient ne pas avoir la même synthèse protéique, car leurs besoins sont différents (Roeder, 2019; Alberts, 2004). Le processus de production d'un modèle à partir d'une séquence d'ADN (c'est-à-dire un gène) vers un ARN immature est appelé transcription, qui correspond à l'un de nos principaux points d'intérêts dans cette thèse.

Chez les eucaryotes, l'ADN est très complexe et s'enroule autour des protéines appelées histones et forme le premier niveau de compaction de l'ADN dans le noyau. En raison de cette compacité de l'ADN, nous avons deux considérations importantes: (i) la plupart des gènes ne sont pas accessibles à la transcription et (ii) le noyau est un environnement encombré (Hancock, 2014). Ainsi, la modélisation de la transcription n'est pas simple comme le laisse entendre le dogme central de la biologie moléculaire.

Pour initier la transcription, un gène doit être actif, c'est-à-dire que le brin d'ADN à cette séquence doit être déroulé et ouvert pour la machinerie transcriptionnelle et est défini comme un site cible. De manière simplifiée, un type spécial de protéine appelé facteur de transcription (FT) doit trouver le site cible pour se lier et recruter l'ARN polymérase (ARNP) pour initier la transcription (Alberts et al., 2002; Turner, 2002). Par conséquent pour modéliser le processus de transcription il est nécessaire de considérer le mouvement des molécules à l'intérieur du noyau.

Étant donné que le noyau est un environnement encombré, avec la chromatine et d'autres éléments de régulation des gènes à l'intérieur, les interactions entre les molécules influencent leur processus de recherche et de recrutement (Hancock, 2014; Woringer; Darzacq, 2018). Les données expérimentales ont montré que le processus de recherche du FT n'est pas diffusif. En effet, le mécanisme le plus utilisé pour le mouvement du FT est le processus 3D/1D, également connu sous le nom de mécanisme de facilitated diffusion. Il se base sur le fait qu'une molécule diffuse en 3D si elle est libre, mais si elle est liée, glisse le long de la chromatine (Bénichou et al., 2011; Mirny et al., 2009; Izeddin et al., 2014; Avcu; Molina, 2016). De plus, les interactions entre les éléments du noyau ne se limitent pas aux interactions FT-chromatine mais les FTs interagissent aussi faiblement entre elles, formant des clusters qui influencent l'activation de la transcription (Meeussen et al., 2022; Zhang et al., 2019).

Pour approfondir nos connaissances sur les processus entourant la régulation des

gènes, nous avons proposé deux extensions différentes du modèle mathématique présenté dans (Avcu; Molina, 2016): l'une pour comprendre la formation de clusters dans les régions chromatiniennes prolifiques et l'autre pour décrire le processus de recherche du FT et de recrutement de l'ARNP jusqu'à la transcription. Nous avons également proposé un modèle de vitesse de l'ARN (Manno et al., 2018) pour analyser le séquençage de l'ARN par fractionnement (Frac-Seq).

Résultats

La transcription dépend de la vitesse à laquelle un TF trouve un site cible ; sachant cela, comprendre comment le FT occupe les régions de la chromatine est un moyen de mieux comprendre l'expression des gènes. Notre objectif est de comprendre la transcription et comment d'autres éléments du noyau influencent la régulation des gènes.

Si l'on considère le mécanisme de facilitated diffusion, la structure de la chromatine joue un rôle dans les schémas d'allocation des FTs, car les régions chromatiniennes hautement connectées sont plus susceptibles d'être atteintes. En outre, l'accessibilité des séquences d'ADN a également un impact sur l'accès du FT au site cible, car le noyau est un environnement surpeuplé. Par conséquent, nous devons tenir compte à la fois de la connectivité et de l'accessibilité pour modéliser l'allocation des FTs.

En effet, les FTs peuvent s'agglutiner dans des régions particulières, se liant faiblement et formant des agrégats qui, à leur tour, influencent l'activité transcriptionnelle. Ainsi nous avons proposé un modèle pour comprendre la formation d'agrégats et, pour interdire aux FTs de s'agglutiner dans une seule région (ce qui n'est pas biologiquement faisable). Nous avons considéré la présence d'une exclusion de volume, également appelée capacité de charge - c'est-à-dire que nous avons supposé une quantité limite de FTs qui s'adaptent dans les régions de chromatine. Notre modèle est un système d'équation différentielles ordinaires pour toutes les régions chromatiniennes de notre réseau, avec I interactions.

Nous avons implémenté des solutions numériques en utilisant ode15s de Matlab pour obtenir des solutions considérant différents taux d'association K_a, capacités de support C et concentrations nucléaires de FT. Suite à cela, nos résultats ont montré qu'en augmentant les valeurs des paramètres, on augmente le groupement autour des régions plus prolifiques.

Nous avons découvert comment le processus de facilitated diffusion modélise les recherches de FT et comment, en limitant la concentration maximale autorisée dans une région, nous influençons les modèles d'occupation des FTs. De plus, en supposant que les interactions protéine-protéine (IPP) sont une condition inhérente à la formation de nœuds de transcription et à l'influence de la transcription par les FTs, nous avons découvert que si des valeurs plus élevées de Ka semblent influencer le regroupement dans les nœuds

prolifiques, les régions avec moins de FT attendus lorsque $K_a \equiv 0$ bénéficient davantage de K_a.

En outre, en diminuant la concentration maximale de FT autorisé, nous augmentons également l'activité dans d'autres régions moins prolifiques. De faibles concentrations de FT nucléaires rendent également difficile l'agglutination, car la probabilité que deux FTs se chevauchent est réduite. Cependant, ce modèle ne décrit que la formation de l'amas de FT et, pour que la transcription ait lieu, d'autres éléments doivent être recrutés sur le site cible, comme l'ARNP par exemple.

Puisque ce ne sont pas seulement les FTs qui s'agglutinent et que les résultats expérimentaux ont montré que les ARNPs se regroupent également autour du site de début de transcription pour faciliter la transcription. De cette manière, nous avons également modélisé la recherche par les ARNP d'un FT lié et son recrutement pour la transcription. Notre système permet deux états pour les FTs: libre et diffusant en 3D dans la chromatine ou lié à une région. Nous avons considéré trois états de ARNP: libre et essayant de trouver un FT lié; lié et attendant de commencer la transcription ou en cours de transcription. L'état d'équilibre pour l'ARNP lié et en cours de transcription a montré la plus forte influence de la connectivité sur l'occupation.

Dans le modèle précédent, nous avons montré comment la concentration de FT dans le noyau influence l'occupation. De plus, certaines FTs sont endogènes dans le cytoplasme et nécessitent une activation pour entrer dans le noyau et commencer la transcription. Elles s'accumulent ensuite dans le cytoplasme, ce qui peut être compris comme un autre mécanisme d'expression génétique (Zambrano et al., 2020). Le mécanisme de translocation a été incorporé en sélectionnant au hasard certains nœuds actifs de notre réseau qui sont connectés au complexe du pore nucléaire (CPN).

Nous avons implémenté des solutions en utilisant ode15s de Matlab mais, comme l'expression des gènes d'une cellule est stochastique, nous avons intégré une version de l'algorithme stochastique de Gillespie. De ces résultats, nous avons conclu que la proximité du CPN augmente l'activité transcriptionnelle, de plus, aucune région n'a démontré une activation retardée ou une réactivation après avoir été désactivée. Des expériences de microscopie (single molecule RNA FISH et single molecule tracking) pour un TF inductible ont corroboré l'hypothèse selon laquelle la proximité du CPN améliore la transcription.

Cependant, notre modèle ne prédit que l'allocation pour la transcription du ARNP et non la concentration de l'ARNm produit, nous avons donc étendu notre modèle pour prendre en compte l'ARNm nucléaire et cytoplasmique et finalement déterminer les changements de concentration d'ARNm pour chaque gène. À partir de ces solutions, nous avons conclu que les régions connectées aux pores présentaient des concentrations d'ARNm cytoplasmiques plus élevées que les autres régions et cela en raison du processus d'exportation de l'ARNm par le noyau.

Enfin, pour modéliser l'exportation d'ARNm par gène pour les eucaryotes, nous avons proposé une extension de la vitesse de l'ARN pour considérer trois états différents de l'ARNm : l'ARNm non épissé qui est un ARN immature avant l'épissage et qui reste dans le noyau, l'ARNm épissé nucléaire et l'ARNm épissé cytoplasmique, qui est l'ARNm réellement traduit. Nous avons utilisé ce modèle pour comprendre les ensembles de données sur les gènes de séquençage et comment le traitement à l'auxine affecte la transcription.

Nous avons utilisé l'inférence bayésienne pour obtenir les paramètres de notre modèle et les prédictions de ce modèle ont prouvé que la plupart des gènes ne sont pas affectés par le traitement à l'auxine. Ensuite, nous avons sous-groupés les gènes en considérant une forte corrélation entre les lectures expérimentales et les valeurs analytiques proposées par notre modèle. Les gènes affectés sont principalement régulés à la baisse dans les cellules traitées à l'auxine et, de manière intéressante, la longueur du gène n'était pas une caractéristique déterminante pour les paramètres.

Même si les techniques expérimentales actuelles ne permettent pas encore de découvrir tous les détails de la transcription, la fusion de la recherche expérimentale et théorique est la meilleure voie pour comprendre les mécanismes de la régulation des gènes. En conclusion, la régulation des gènes n'est pas seulement un processus biologique mais peut être comprise comme un phénomène mécanistique.

En ce qui concerne la modélisation, nous sommes convaincus que nos recherches ont mis en lumière les facteurs clés de la transcription, même si nous n'avons pas pu les mettre en œuvre à l'échelle du génome ou même à haute résolution pour nos solutions numériques. Nous avons également confirmé que les modèles in silico peuvent être de bons outils d'exploration pour les systèmes complexes, comme le montre souvent la biologie.

Publications et Présentations

Publications en préparation

- Oliveira, K.A., Mazza, D., Zambrano, S., Molina, N., "Modelling Transcription Factors Search and Polymerase Recruitment Dynamics within a complex chromatin structure";
- Molina, N., Oliveira, K.A., "Protein-protein interactions effects on cluster formation".

Présentations orales

- EMBL: Chromatin and Epigenetics May 17 - 20, 2021
- EMBO: Physics of living systems: From molecules to tissues Jun 07 - 11, 2021

- 13th European Biophysics Conference Jul 24 - 28, 2021
- EMBL: Transcription and Chromatin Aug, 27-30, 2022
- 12th European Conference on Mathematical and Theoretical Biology Sep 19-23, 2022

1 Introduction

In some sense, one can say that with only four letters life is created. And not only created but maintained, evolved to fit the environmental changes, and passed to the next generations. How living organisms continue to survive is a question which can be answered on different scales and in different scientific fields and infinite perspectives. In our case, the scale is microscopic and the scientific field is mathematics.

The technology to understand and analyse molecular biology evolves every year, presenting better ways to quantify and interpret data: for example, between (Watson; Crick, 1953) and (Marini et al., 2015), more than 60 years passed and the quality of reproducible data increased exponentially. However, as technology becomes more accessible facilitating reproducibility and innovations in the methods, the use of theoretical/computational tools becomes indispensable.

Here, we aimed to integrate experimental results with theoretical models and *in silico* experiments to explain, for example, how a transcription factor (TF) searches its target site, binds the chromatin regions and recruits RNA polymerase (RNAP) to start transcription, how the chromatin influences the occupancy patterns or which mechanisms regulate mRNA synthesis. The theoretical setup used is by proposing mathematical models with ordinary differential equations systems, a tool greatly used in theoretical biology (Edelstein-Keshet, 2005; Murray, 2007).

This thesis deals with biology, more specifically gene regulation, by applying mathematical models, which is not a new theme in biology and presents a broad range of different techniques (Chen; He; Church, 1998; Chiu et al., 2012; Ay; Arnosti, 2011). Furthermore, this thesis is organized as follows: here, we present key concepts and the reasoning behind our modelling. Then, in Chapter 2, we present a model for TF searches and how by allowing protein-protein interactions (PPI) we affect the TF allocation in the chromatin and influence the clustering.

Once a TF binds to a target site, it recruits polymerase (RNAP) to start transcription. More than that, since some TFs remain in the cytoplasm and enter the nucleus upon activation, we also used this behaviour in our model as a gene expression control in Chapter 3. In eukaryotes, the result of transcription, the messenger RNA (mRNA), must be translated into a protein in the cytoplasm. Thus, we modelled the mRNA export in Chapter 4. To verify *in silico* predictions from Chapter 3, we did image analysis for two different microscopy experiments in Chapter 5.

Last, we present our conclusions and discussions in Chapter 6 for all models from this thesis. We also present some insights about other possibilities of our model.

1.1 Gene Regulation

From the central dogma of molecular biology, we have a simplified (and unidirectional) pathway in which from the DNA we obtain RNA (in a process called **transcription**) and from the RNA we synthesize any protein (which it is called **translation**) (Crick, 1958; Cobb, 2017). Of course, as far as a dogma goes the order DNA → RNA → protein is not true for every living organism: retroviruses carry an enzyme called *reverse transcriptase* which reverses the central dogma (hence the name) to RNA → DNA → RNA → protein (Alberts et al., 2002; Turner, 2002). We proposed a diagram for the central dogma of molecular biology in Fig. 1.1. This diagram represents transcription/translation in eukaryotes since we considered the existence of the nucleus. Our focus in this thesis is to model the first three steps of this dogma.

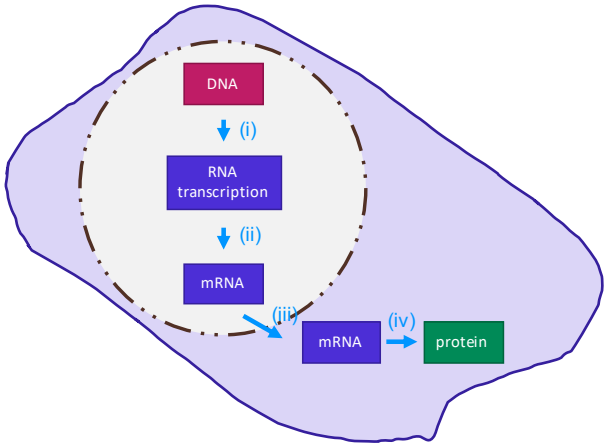

Figure 1.1 – **Diagram for the Central Dogma of Molecular Biology for Eukaryotes.** Here, we present the steps between the DNA and the protein, with all the possible controls: **(i) Initiation Control:** where the cell controls which gene is produced and its frequency of synthesis; **(ii) Transcription Control:** controls the speed of transcription, the splicing and processing; **(iii) RNA Transport Control:** controls the messenger RNA (mRNA) export to the cytoplasm; **(iv) Translation Control:** controls the translation by the ribosomes and, later, the protein activation/inaction. This diagram was adapted from (Alberts et al., 2002), p.270.

A cell can control its protein synthesis by switching a gene *on* and *off*, which is a common mechanism for almost all cellular organisms. Some genes, also known as regulatory DNA sequences, are responsible for this switch. Then, those genes are recognized by gene regulatory proteins (commonly known as transcription factors, TFs) that bind to the DNA, allowing or prohibiting the transcription in this gene (Struhl, 1999).

Different from prokaryotes, eukaryotic cells package their DNA into chromosomes to conceal its size. More than that, the compact structure of DNA with proteins (mostly histones, i.e., proteins whose main function is to act as a reel for the DNA, creating the nucleosomes) is called **chromatin**, and it prevents, for example, the entanglement of DNA

strands (Cherstvy; Teif, 2013; Maeshima; Ide; Babokhov, 2019; Li; Carey; Workman, 2007). Therefore, one can conclude that the DNA organization for eukaryotes acts as a control mechanism for gene expression.

As cells turn into more complex systems, better control over gene expression is required. For example, in multicellular cells, the differentiation process is a consequence of this genetic switch. On top of that, it is a mechanism to control the genes each cell produces - e.g., α and β cells are localized in the pancreas but while one is responsible for the glucagon (a hormone that raises the glucose in the body) and insulin (a hormone that promotes the absorption of glucose) (Peterson et al., 2020; Scharer et al., 2018). A consequence of the differentiation is that different cells have different functions and will produce different proteins even if any cell has the complete information from the DNA. Besides, an important point is that given environmental changes, a cell can alter its gene expression (Hunter, 2005; Findley et al., 2021).

Given the DNA size, for a mature cell (i.e., differentiated), most of the coding genes are inactive; and, inside the eukaryotes DNA, most of the sequences are non-coding (Maston; Evans; Green, 2006; Piovesan et al., 2019). However, even with this optimal condition for gene expression, the cell must be able to control and fine-tune each gene produced.

The regulation of the gene expression (or **gene regulation**) is an important mechanism of cell maintenance. The most fundamental step of gene regulation is the transcription initiation, as we presented in Fig. 1.1. Next, we present an in-depth description of the transcription process and its consequences, as this is fundamental for the understanding of this thesis.

1.1.1 Processes around Transcription

As introduced earlier, the transcription of a gene depends on the gene regulatory proteins. Transcription is defined as the process of copying a portion of the DNA sequence into an RNA sequence, and the enzyme that reads the DNA sequence is the RNA polymerase (RNAP), which we will discuss next.

We note that throughout the scope of this text when we refer to polymerase, we mean RNA polymerase (RNAP). We are aware of the existence of DNA polymerase (DNAP), but as we are not dealing with DNA replication in this thesis, there will be no confusion with this other enzyme.

1.1.1.1 RNA Polymerase and Transcription

RNA Polymerase has a key role in transcription as it is responsible for recognizing and binding the DNA sequence, with the help of the transcription factors; breaking the bonds between the complementary nucleotides to dissociate the DNA strands; and starting

the formation of the RNA strand (also called transcript), which is the transcription process per se.

For prokaryotes, there is just one type of polymerase but for eukaryotes, until now, experiments have found five different types of RNAPs, two of them (RNAP IV and V) are plant-specific (McKinlay et al., 2017). The other three RNAPs synthesize: (i) RNAP I (or PolI) most of the ribosomal RNA (rRNA) genes; (ii) RNAP II (PolII) all protein-coding genes, microRNAs (miRNAs) and small RNAs (from spliceosomes, for example); (iii) RNAP III (PolIII) transfer RNAs (tRNAs), 5S rRNA genes and small RNAs in general (Moss; Stefanovsky, 2002; Lee et al., 2004; Guiro; Murphy, 2017; Willis, 1993). From this description, one concludes the importance of RNAP II (Pol II) for gene regulation, as it is the polymerase responsible for the coding genes while the others are polymerases for the structure and organization of genes. Indeed, it is important to remind that in eukaryotes the DNA is packaged into chromatin, which explains the specificity of eukaryotic polymerases.

RNAP II genes depend on the presence of transcription factors to bind to a DNA sequence, as the TFs orient the RNAP II at the promoter, which are DNA sequences that remain a few hundred base pairs near the TATA box (a non-coding DNA sequence that contains a consensus of T and A base pairs). Most of the sequences are promoter proximal being where the preinitiation complex (PIC) is formed (Wang; Stumph, 1995; Tsai, 2000). In simple terms, the PIC is the protein complex responsible for orienting the RNAP II at the transcription start sites (TSS) and prepares DNA for transcription (Kornberg, 2007; Luse, 2013), which we exemplified in Fig. 1.2.

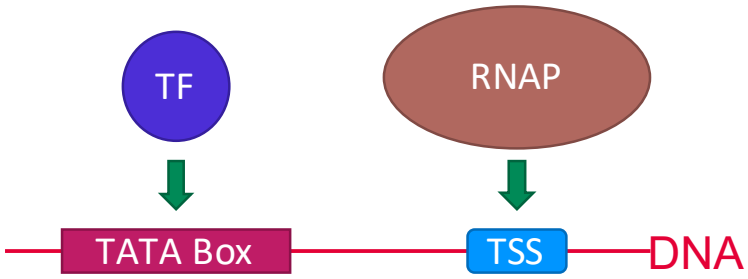

Figure 1.2 – **Simplified Schematics for the PIC Formation.**

The PIC formation is the first step towards transcription initiation and, since its assembly is a complex process, it is one of the most susceptible steps to regulation (Petrenko et al., 2019; Soutoglou; Talianidis, 2002). Another important fact is that the PIC remains attached to the DNA once the RNAP starts transcription, facilitating the recruitment of another available RNAP to transcribe at that TSS (He et al., 2013).

Once the PIC is assembled, the gene is switched *on* and the RNAP is released to start the transcription, in a process called elongation (Veloso et al., 2014; Pokholok; Hannett;

Young, 2002). Even if the main goal of the polymerase is to finish transcription, sometimes the RNAP can be stalled waiting for the cell signals to enter the gene body, in a process called polymerase pausing (Wade; Struhl, 2008; Core; Adelman, 2019). Environmental changes can cause the polymerase pausing, for example, the decreasing oxygen availability (also known as hypoxia) or the presence of proinflammatory signals cause polymerase pausing (Liu; Kraus; Bai, 2015; Yang et al., 2022; Barboric et al., 2001).

From the RNAP front, we stated that gene regulation can occur early on in the PIC formation, for example. However, mechanisms such as polymerase pausing can only occur after the PIC assembly and are usually necessary to redefine chromatin accessibility to transcription. Since the reorganization of chromatin is an important step, we need to comprehend the mechanisms behind chromatin reorganization in the cell. One of those mechanisms is the Enhancer-Promoter Interactions which we present next.

1.1.1.2 Enhancer-Promoter Interactions

In prokaryotes, gene regulation is a straightforward process in which the cell has activators and repressors as the proteins that turn their genes *on* and *off* (Ishihama, 2012). As we stated previously, gene regulation is far more complex in eukaryotes but all the eukaryotic promoters need some sort of activators to start the assembly of the transcriptional machinery (Ma, 2011).

The DNA sequences where the activators bind are called enhancers as their presence increases transcription rate; and, differently from the promoters that need to be upstream to the gene (i.e., considering transcription occurs from the 5′-end to the 3′-end of the DNA, upstream is closer to the 5′-end and downstream is closer to the 3′-end), the enhancers do not require directionality, being found even in intronic parts of genes (Arensbergen; Steensel; Bussemaker, 2014; Pennacchio et al., 2013). Another difference is that the distance between an enhancer and a promoter varies: from 50-100 kb in *Drosophila* to 1Mb in the Sonic hedgehog protein (SHH) (Kyrchanova; Georgiev, 2021; Lettice, 2003). In fact, given there are plenty of enhancers in the DNA and their distance from the promoter is not fixed, the pairing of an enhancer-promoter must not consider the linear distance between them (Schoenfelder; Fraser, 2019).

To solve the long-distance problem, the DNA must have a mechanism to approximate the enhancer and the promoter, leading to their interaction. One accepted mechanism for the chromatin to decrease this distance is by looping itself (Whalen; Truty; Pollard, 2016). We present a simplified version of the chromatin looping in Fig. 1.3 and show how it facilitates Enhancer-Promoter Interactions.

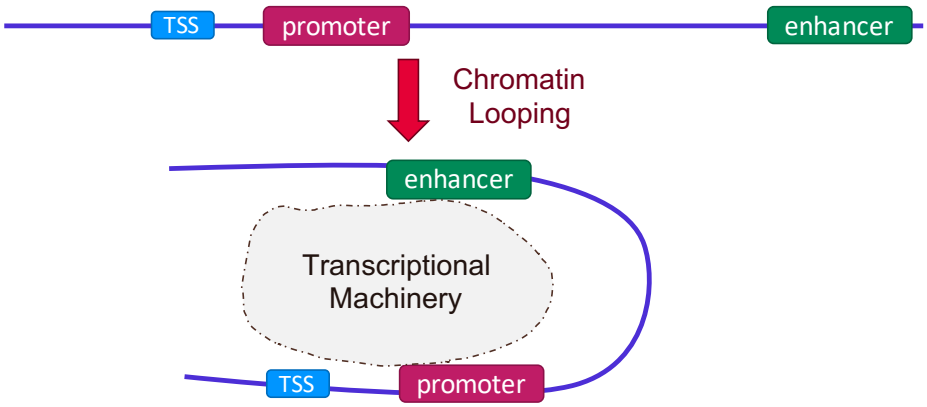

Figure 1.3 – **Simplified Schematics for the Chromatin Looping and Enhancer-Promoter Interactions.** Here, we can see how fairly distant enhancers and promoters interact by looping chromatin, which facilitates their interaction. By **Transcriptional Machinery**, we meant all the molecules involved in transcription such as TFs (general and specific), RNAPs and mediators.

In broader terms, chromatin looping is a process that occurs when the enhancer and promoter of a DNA sequence are in the same chromosome (i.e., in the cis configuration. The trans configuration occurs when the elements are found in both chromosomal alleles equally) and closer in proximity than its silencers (Reuveni et al., 2018; Kadauke; Blobel, 2009). Of course, different molecules play a role in chromatin looping even if not all the key players are currently known. For example, evidence suggests CTCF (a TF involved in insulator activity) and cohesin facilitates enhancer–promoter interactions (Ren et al., 2017). Yet, it is clear that once all the responsible elements are identified, our comprehension of gene regulation will increase exponentially.

However, for this thesis, the knowledge that enhancer-promoter interactions shape the chromatin and thus gene regulation is enough. Several techniques to understand the chromatin configuration were developed over the years, most of them based on the chromosome conformation capture (3C), a technique that uses the contact frequencies between two DNA sequences from a cell (Wit; Laat, 2012).

1.1.1.2.1 Hi-C and Chromatin Structure

The Hi-C technique (Lieberman-Aiden et al., 2009) is a combination of the 3C and next-generation sequencing (NGS) - a fast parallel sequencing technology to obtain sequencing data (Wit; Laat, 2012; Behjati; Tarpey, 2013). As stated previously, the 3C technique measures the frequency of two DNA sequences to physically associate in 3D. The method is based on the likelihood of those two sequences to have a formaldehyde-crosslinking - which is a tool to detect and quantify, for example, interactions between protein-DNA or between two chromatin fibres (Hoffman et al., 2015; Miele; Dekker, 2008) - and, once we obtain the crosslink, the chromatin is solubilized and fragmented.

However different from the 3C - that supports just a set of *loci* to quantify the crosslinking, a technique interesting to verify enhancer-promoter interactions, for example, the Hi-C technique accepts the sequencing of all fragmented chromatin (Belton et al., 2012). Furthermore, as the technology improves, new uses/improvements for an established technique shall occur (Lafontaine et al., 2021; Díaz et al., 2018; Niu et al., 2019).

With the Hi-C data, one can infer the structure of the chromatin - as previously predicted by fractal globule or equilibrium globule experiments (Mirny, 2011; Sanborn et al., 2015; Pal; Forcato; Ferrari, 2018). To reconstruct the structure, we use the sequencing data to create contact maps from all the chromosomes (Pal; Forcato; Ferrari, 2018; Oluwadare; Highsmith; Cheng, 2019; Galitsyna; Gelfand, 2021). In this thesis, we used the information from Hi-C data available from (Rao et al., 2014) to construct our chromatin network.

From the construction of a chromatin network, we still need to understand how the molecules explore the nucleus, as it is a crowded environment with different compartments and proteins interacting inside, attracting and repulsing each other. Next, we present the diffusion inside and later outside the nucleus.

1.2 Diffusion in the Nucleus

The nucleus is an organelle unique to eukaryotes, and the actual mechanism behind its emergence is still unknown. However, without it, most of the multicellular organisms (for multicellular prokaryotes, (Mizuno et al., 2022)) would not exist (Devos; Gräf; Field, 2014).

One can understand the nucleus as a colloidal system, with different molecules of different sizes and functions crowding it. Therefore, the spatial nuclear organization allows the formation of compartments with specific functions: for example, the nucleolus is where rRNA and tRNA are transcribed, i.e., the clustering of those genes facilitates the access of their regulatory elements (Hancock, 2014; Meldi; Brickner, 2011). It should be noted that the nuclear compartments are dynamical elements inside the nucleus, which is also changing - for example, cell cycle changes (Zidovska, 2020).

Therefore, analysing the mechanisms behind the diffusion of molecules in the nucleus is a way to understand gene regulation. We define **diffusion** as the movement of a molecule from a highly occupied region to a lowly or unoccupied one without outside forces. By Fick's law, we know the flux of diffusion is proportional to this concentration gradient (Pollak; Siegmund, 1985; Timney et al., 2016; Leijnse et al., 2012).

The diffusion can be understood as a random walk of molecules and, in the nuclear case, macromolecules are frequently colliding with other elements inside the nucleus (e.g., chromatin) and slow them down, which is called an anomalous diffusion (i.e., a

diffusive process with a non-linear in time mean squared displacement (MSD), being subdiffusive - below diffusion - or superdiffusive - above diffusion) (Weiss, 2014; Hancock, 2014; Woringer; Darzacq, 2018). Yet, the subdiffusion is at times an optimal feature for nuclear molecules, protein complex formation or signal propagation enhanced in a subdiffusive condition (Guigas; Weiss, 2008; Banks; Fradin, 2005).

There are plenty of different molecules in the nucleus, with distinct movement patterns. As one of the first steps in the PIC assembly is a TF binding to a promoter, and understanding the TF search process is also a way to understand gene expression, we focus on TF diffusion.

1.2.1 TF diffusion

Both gene expression and regulation depend on how the TF finds a target site, as the first step for transcription is a TF binding a promoter. Since the distance between the transcriptional elements is not constant and the numerous non-coding sequences in DNA for eukaryotes, the mechanisms behind the TF finding its target site must be efficient in order to maintain cell stability (Hager; McNally; Misteli, 2009).

Here, efficiency means not only producing the necessary gene but also fast-starting transcription with the smallest use of energy possible. The first step is to understand how the TF diffuses in the nucleus which is not straightforward: interactions with TF-chromatin or TF other molecules influence the diffusion and the nuclear size and shape as well, in which, in turn, affect the gene expression (Gorski; Dundr; Misteli, 2006; Zon et al., 2006).

Again, the nucleus is not an empty environment and all the structures inside of it influence the movement: surprisingly, crowding or DNA looping for example accelerate the TF (Li; Berg; Elf, 2009; Banks; Fradin, 2005). TF-chromatin interactions are more diverse: TFs spend longer periods in more compact and redundant regions of the chromatin (heterochromatin, responsible for gene silencing, for example) and it is faster in open regions of the chromatin (euchromatin, responsible for active transcription), meaning a TF can identify easily its target site in euchromatin than in heterochromatin (Bancaud et al., 2012; Morrison; Thakur, 2021).

Therefore, the chromatin structure *and* its accessibility influence the TF search process, meaning a completely diffusive process - either diffusion or anomalous diffusion - does not represent the TF movement. To incorporate the structure regulatory potential, researchers have proposed the facilitated diffusion as a mechanism to describe TF movement (Mirny et al., 2009; Bauer; Metzler, 2013).

1.2.1.1 Facilitated Diffusion

The **facilitated diffusion** is a mechanism first proposed by Riggs in 1970 after his experiments *in vitro* of the lac repressor association was around two orders of magnitude faster than the theoretical values obtained considering a diffusion-only process. The basic idea is that a 3D diffusion must be paired with another diffusive process on a reduced dimensionality, explaining the other name for facilitated diffusion: 3D/1D process (Woringer; Darzacq, 2018; Mirny et al., 2009).

The basic concept is that a molecule, in its search for a target site, binds and unbinds the DNA and, if bound to a non-specific sequence, slides along the strand, which is a 1D diffusion/Brownian motion process. Once unbound, this molecule is free to 3D diffuse again (Mirny et al., 2009; Kampmann, 2005). We proposed a schematic for this process in Fig. 1.4.

However, while as a concept is easy to understand, the understanding behind the 3D/1D switch is non-trivial and neither is how to theoretically model the diffusion of a chromatin-bound protein (Bénichou et al., 2011). More than that, chromatin-TF interactions are key factors for TFs to either slide (1D) or jump (3D) and the structure of chromatin is a complex system by itself (Woringer; Darzacq; Izeddin, 2014; Almassalha et al., 2017; Bancaud et al., 2012).

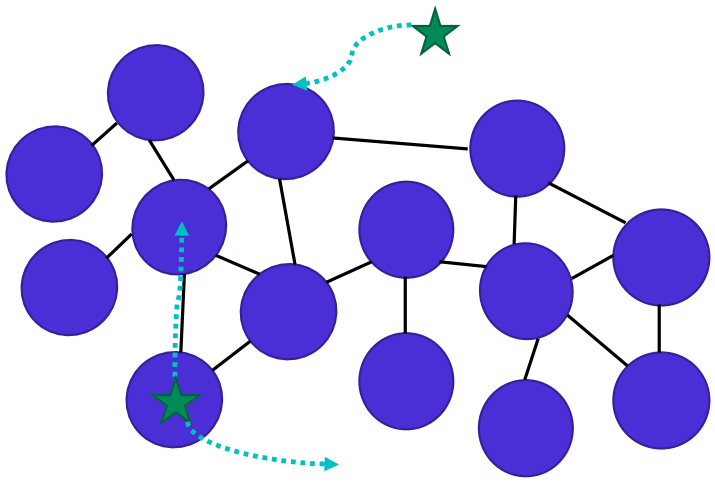

Figure 1.4 – **Facilitated Diffusion Schematics.** Where a molecule (green star) once interacting with a region (blue node) can either leave to 3D diffuse or slide to a neighbour (which is represented by the presence of an edge between two regions). The movement choices are represented by dotted lines.

Still, experiments have shown that the facilitated diffusion models are good descriptions of the reality of a molecule diffusing in the nucleus. One robust experiment to identify this facilitated diffusion behaviour is the Single-Molecule Tracking (Izeddin et al., 2014; Jonge et al., 2022).

1.2.1.2 Single-Molecule Tracking

Different experiments have helped scientists to obtain snapshots of what happens inside the cell during its life cycle: from mitosis to cell differentiation to apoptosis. Still, to perform some experiments, the cell must be fixed - i.e., the ability to have a time evolution of the cellular processes is lost due to the techniques killing the cell.

The advent of super-resolution microscopy (SMR) allowed scientists to obtain live imaging from their cell lines - which facilitates the achievement of spatial and temporal data acquisition (Hoboth; Šebesta; Hozák, 2021; Dange; Joseph; Grünwald, 2010). Usually, experiments overlook cell stochasticity by averaging the results and assuming all cells in the same conditions behave the same way. Of course, this is a huge assumption that does not represent reality as gene expression is a fairly stochastic process (Elowitz, 2002; Anink-Groenen et al., 2014).

Single-Molecule Tracking (SMT, also known as Single-Particle Tracking, SPT) is a technique to track, as the name indicates, live-image single-molecules fused fluorescent probes over some time, reconstructing their movement. SMT offers high precision, as each fluorescent molecule can be tracked and is quite useful to understand the dynamics of chromatin-associated molecules (e.g., TFs and RNAPs). However, every technique has its drawbacks and for SMT the trajectory length, spatial precision, and the fluorescent probe ensure we can only obtain a few seconds at a time due to the photobleaching, and the constructed trajectory is only in two-dimensional space instead of 3D (Manzo; Garcia-Parajo, 2015; Kuchler et al., 2022; Liu et al., 2016).

Therefore, SMT is a potent technique to verify the facilitated diffusion mechanism from a TF searching for a target site, even if the search and the SMT times are in two different time scales, i.e., it is not possible to track the complete TF search time in the nucleus (Swift; Coruzzi, 2017; Mazzocca et al., 2021). In fact, SMT results have shown that TFs can have two explorer behaviours: (i) **global**, in which the TF explores more and far away regions, and (ii) **local**, in which the TF remains in one predetermined area, proving the TF search is not a pure diffusive process in the random walk sense (Izeddin et al., 2014; Liu et al., 2021; Jonge et al., 2022; Li; Varala; Coruzzi, 2015).

Eukaryotes developed the nucleus and separated the transcription and translation process, which is one mechanism for gene regulation since mRNAs do not require the same stability as the DNA - as the mRNA has a shorter lifespan and a fast mRNA degradation facilitates the reuse of its ribonucleotides (Turner, 2002; Nouaille et al., 2017). Next, we present some key concepts of the eukaryotic mRNA, as its export is the key point in Chapter 4.

1.3 mRNA export

Most of the eukaryotic genes have two different categories of sequences inside themselves: the **introns**, intervening sequences, which are non-codding and the **exons**, expressed sequences, which are coding sequences. Some genes, however, do not have introns, for example, histones (the proteins which the DNA coils into to form chromatin) are exonic genes. Of course, those non-codding sequences are removed from the mRNA before translation occurs in a process called splicing, which occurs inside the nucleus (Alberts, 2004; Tilgner et al., 2012). At a first glance, the presence of introns seems a waste of transcriptional resources, but researchers have suggested that the presence of introns influences the alternative splicing process (in which isoforms of the mRNAs are generated from the same gene), enhances transcription for some genes, and even helps with mRNA export (Roy; Gilbert, 2006; Jo; Choi, 2015).

Not all mRNAs transcribed are viable for transcription, some of them are downright dangerous to the cell if translated; thus, the cell needs a mechanism to select mature mRNA from the debris around in the nucleus. The solution is the presence of the nuclear pore complex (NPC), which recognizes and transports processed (mature) mRNA for translation, with the transport being only from the nucleus to cytoplasm in this case (Magistris, 2021; Xie; Ren, 2019; Misteli, 2007). The presence of the nucleus is another mechanism for gene regulation, as it leads to the separation between transcription and translation in eukaryotes, as discussed previously and represented in Fig. 1.1.

However, the nucleus is a crowded environment and the transport to the NPC is an important mechanism, given that time and efficiency are important factors. Naked mature mRNAs (i.e., only the mRNA molecule) are not found in the cells, and as soon as the processing (splicing for example) starts, the assembling process of the pre-mRNA-protein complex (pre-mRNPs) also begins by the association of the pre-mRNA with RNA binding proteins (RBPs), that later mature to ribonucleoprotein complexes (mRNPs) (Björk; Wieslander, 2014; Magistris, 2021). Note the functions of RBP are plenty: from alternative splicing to RNA modification and translation (Glisovic et al., 2008).

The mRNP complex is the molecule actively transported through the NPC, and the complete mRNA export process varies from 5 to 40 minutes (Mor et al., 2010). Another important function of the RBP is responsible for mRNA export: as soon as the mRNA processing ends the export must start. We can understand the mRNA export as a three-step process for the RBPs (Björk; Wieslander, 2014; Magistris, 2021):

1. **Assembling of the mRNP Complex**: as the mRNA processing starts, the mRNP complex gets concomitantly assembled.

2. **Translocation through Nucleus**: once the mRNP complex is released, the mRNPs

diffuse in the interchromatin channels (i.e., a region with low average DNA concentration and near the NPC (Cremer et al., 2020)). Some experimental results showed this complex diffusion coefficient in humans is around $0.004 - 0.006\ \mu m^2/s$ (Mor et al., 2010). Then, the mRNP acquires the export receptors.

3. **Docking into the NPC**: the mRNP docks into the NPC, rearranges itself and is effectively exported (Becskei; Mattaj, 2005).

Thus, the mRNA export process is more complex than a molecule freely diffusing in the nucleus. Again, there is a lot of information to uncover about the molecular interactions and mRNA export, being either not detected mechanisms or new RBP molecules to identify.

Yet, with this knowledge of how complex the mRNA export is, the process of developing a model to represent all the intricacies of any cellular mechanism is not currently viable. Next, we present models proposed in previous literature to represent the dynamics of a TF searching from a target site.

1.4 Modelling TF

When we consider the size of the datasets from gene expression/regulation, the ability to theoretically predict some behaviours and patterns with the corroboration of experiments is advantageous, and one especially useful technique is modelling. A model is a representation simplified and scaled of a more complex system. In a system as broad and complex as the TF, there are infinite ways to model it, depending always on the point of view and/or the technique used (Ay; Arnosti, 2011).

For example, one can model biological systems as a stochastic process: a gene expression model in which the TF role is implicitly defined, as the transcription equation is a three-state model for DNA, RNA and protein (Robert, 2019) or the use of gaussian processes to determine TF activity (Gao et al., 2008). Or even computational modelling sequencing to determine TF-DNA binding sites with deep-learning (Zhao et al., 2016) or using machine learning to predict pairwise the TF-TF complex binding the DNA sequence (Antikainen; Heinonen; Lähdesmäki, 2022).

In this thesis, we opted for working with ordinary differential equations, a frequent and common tool for modelling biological systems. Once we consider TF mathematical modelling with differential equations, the field is still broad and yet to be fully explored: for TF modelling, one can model the interaction between a gene induction that can turn off a TF (e.g., Brn2 and Nanog (Sokolik et al., 2015)), two TFs interacting regulates the third one (e.g., Oct4-Sox2 complexes and the positive regulation of Nanog (Glauche; Herberg; Roeder, 2010)) or modelling the liquid-liquid phase separation of TF-droplets increasing gene expression (Schneider et al., 2021).

Gene regulatory network models can be used to understand TF binding and the TFs competition, using a master equation to incorporate the cell stochasticity (Hettich; Gebhardt, 2018). In our model, in addition to considering interactions between TFs, we also consider how the chromatin structure influences the search process for a binding site, as we extended the model in (Avcu; Molina, 2016) to obtain our different gene expression models. We incorporated the stochasticity expected from the cells by applying the Gillespie Stochastic Algorithm to our ODE system from Chapter 3.

Thus, one can conclude that the possibilities of modelling TF interactions are as diverse and complex as the TF itself. Still, there are different and interesting methods to fully comprehend the role of TFs in gene expression - some of them are yet to be uncovered.

2 Transcription Factor Searches and TF Clustering

Disclaimer: Some parts of this chapter can be found in Molina's team paper *"Chromatin structure shapes the search process of transcription factors"*

Cellular processes depend on producing the suitable protein at an optimal rate; thus, the recognition of a specific motif in the DNA is extremely important for cell maintenance and, consequently, its permanence. From the Central Dogma, the cell transcribes DNA sequences into RNA to start the protein synthesis, using transcription factors (TFs) (Roeder, 2019). Given the size of the DNA, both DNA and TFs develop mechanisms to facilitate the process of finding the right region for the right protein, like TFs condensates (Shrinivas et al., 2018; Liu et al., 2014) or transcriptional activation or repression (Pugh; Tjian, 1990; Mitarai; Semsey; Sneppen, 2015).

Experimental evidence suggests that the 3D structure influences the TFs' search processes. However, not only the structure and accessibility of chromatin influence the binding of a TF to a region, as weak protein-protein interactions (PPI) may cause the clustering formation of TFs, which can be defined as a high local concentration of TFs on some specific nuclear regions (Cherstvy; Teif, 2013; Zhang et al., 2019; Gibcus et al., 2018; Fudenberg et al., 2018; Nagamine; Kawada; Sakakibara, 2005). These transient TF condensates help to stabilize DNA binding, recruit RNA polymerase (RNAP) and activate transcription (Hnisz et al., 2017; Brodsky et al., 2020; Bompadre; Andrey, 2019; Quintero-Cadena; Lenstra; Sternberg, 2020), which we will discuss in depth in Chapter 3.

In this chapter, we present the model previously developed by Molina's team to understand TF searches in chromatin (Avcu; Molina, 2016) and later we try to understand how the structure creates a rich environment for cluster formation.

2.1 Transcription Factor Searches and Importance of the Structure and Accessibility for its Occupation

As we discussed previously, the TF movement does not obey a pure diffusive process, but rather it is a facilitated diffusion mechanism (or 1D/3D process where the TF either slides the DNA (1D) or moves freely in solution (3D) (Mirny et al., 2009; Avcu; Molina, 2016; Bauer; Metzler, 2012)). Such approximation to represent TF movement is widely used in literature, (Bauer et al., 2015; Hettich; Gebhardt, 2018; Zabet; Adryan,

2013; Slutsky; Mirny, 2004). Experiments such as *in vivo* Single-Molecule Tracking (SMT) experiments (Kuhn et al., 2021; Xiao; Hafner; Boettiger, 2020), have shown that structure is also an important component for TF occupancy in chromatin regions.

To incorporate the chromatin structure as a component for TF diffusion, we admitted the existence of a network in which if there is a connection between two nodes (i.e., regions), we have an edge. Thus, to reach a specific region i from region j, a path between those regions must exist. To build our network mathematically, let $\mathbf{N}$ be the set of nodes (regions) of our network. Then, we assumed our network as a symmetric matrix, A, as for each pair of nodes of our network, i and j, we have:

$$A(i,j) = A(j,i) = \begin{cases} 1, \text{if exists a connection between region } i \text{ and } j; \\ 0, \text{otherwise.} \end{cases}$$

We assumed $\forall i, A(i,i) \equiv 1$, i.e., we admitted the existence of *self-loops* for our network, meaning each region has a connection with itself. With our mathematically proposed network, we defined the number of connections of the region i as the sum of all the nodes connected to region i, i.e., let L be the size of our network (the number of nodes, $|\mathbf{N}| = L$), we have:

$$d_i = \sum_{j}^{L} A_{ij} \ . \tag{2.1}$$

The TF movement from region to region is represented in Fig. 2.1. For example, a TF localized in the light blue node can move to its neighbours' nodes (magenta); in this case, a TF has 6 different choices to move to the next nodes and another one to *stay* in this node. Therefore, the number of connections each node has is important to either leave the node or be reached.

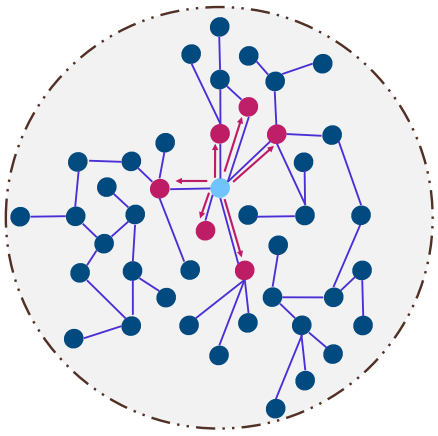

Figure 2.1 – **Representation for the TF movement in a network inside the nucleus.** Here, a TF localized in the light blue node can move in the next step to the nodes in magenta or remain in its node, we assumed each node is equally probable.

To describe the TF movement through our network, we need to define the probability to move from region j to region i, $\forall i, j \in \mathbf{N}$. First, we supposed there is no bias toward specific nodes in the network, i.e., all connected nodes have the same probability. Since each region might a different number of connections, we assumed the probability for region i, is the inverse of Eq. (2.1), and we defined this probability of moving from region j to region i as:

$$M_{i \leftarrow j} = \frac{A_{ij}}{d_j} \ . \tag{2.2}$$

This part of the model represents the 3D diffusion of TFs through chromatin. However, the more accepted models for TF movements assume the mechanism of facilitated diffusion (3D/1D process). To incorporate the sliding process (1D diffusion), we considered the residence times, which represent how long a TF spends in a particular region before sliding off, (Avcu; Molina, 2016). Of course, if a TF remains for longer periods in a region, this region is more likely to recruit the transcriptional machinery (Azpeitia; Wagner, 2020; Popp; Hettich; Gebhardt, 2020).

Thus, for our model, the waiting time should be region-specific with the length of our network size, L. Of course, as chromatin is a complex of proteins (mainly histones) and DNA (Alberts, 2004), the DNA sequence is region-specific. Also, the TFs are DNA-specific, i.e., it looks for their binding sequence. Besides, chromatin is also packed and we need open chromatin (euchromatin) for the TF to bind a region, which is more responsive to active histone marks. To evaluate the residence times, we used the mean first-passage time of a random walk and single molecule microscopy data to create a bimodal distribution, for active and inactive times (Avcu; Molina, 2016). Therefore, the probability of finding a TF in region i depends on the master equation in Eq. (2.3).

$$\frac{dp_i(t)}{dt} = -k_i p_i(t) + \sum_j^L M_{i \leftarrow j} k_j p_j(t) \ . \tag{2.3}$$

In which k_i is the region-specific exiting rate, i.e., the inverse of the residence time for region i. Here, the time-changing probability depends on the TF that leaves the region i and the sum of all the possibilities of the region i to be reached, given the concentrations in those other regions. We transformed our model into a matrix form, i.e., let $\mathbf{K}$ the diagonal matrix of $(k_i)_{i=1}^L$ and $\mathbf{M}$ the matrix of movement in the network,

$$\frac{d\mathbf{p}(t)}{dt} = -(\mathbf{K} - \mathbf{MK})\mathbf{p}(t) = -\Omega \mathbf{p}(t),$$

where Ω is the transition rate matrix. This ordinary differential equation (ODE) is easily solvable with calculus techniques, and

$$\mathbf{p}(t) = \mathbf{p}_0 e^{-\Omega t} \ . \tag{2.4}$$

We used the Hi-C data for Chromosome 19 for CH12 Mouse Lymphoblasts in $5kb$ resolution to generate our network (Rao et al., 2014), we threshold the values from the

contact frequencies to create our fixed network; and, in Fig. 2.2, we present the correlation matrix for this network. From Fig. 2.2, we can see the regions are connected in blocks, creating sub-topologies inside the Topologically Associated Domains (TAD). Also, since we obtained our structure from experimental data, there are some regions in which the Hi-C could not get any reads, creating blocks of non-connected regions in our matrix.

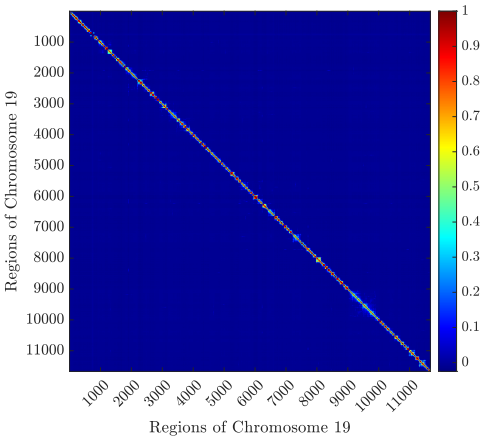

Figure 2.2 – **Correlation Matrix for Chromosome 19 for CH12 Mouse lymphoblasts from (Rao et al., 2014)**. Here, we can see the formation of different topologies in smaller intervals inside the network.

To understand the parameters for the model in Eq. (2.3), we present Fig. 2.3 the probability density for **A** the number of connections and **B** $\log_2$ of the residence times. The number of connections shows two different groups: the connected regions, a bigger part of the network, and barely-connected regions, which are artefacts from the Hi-C measurement. The residence times we can see the bi-modality of inactive and active regions is an expected result in **B**, and most of the regions remain inactive.

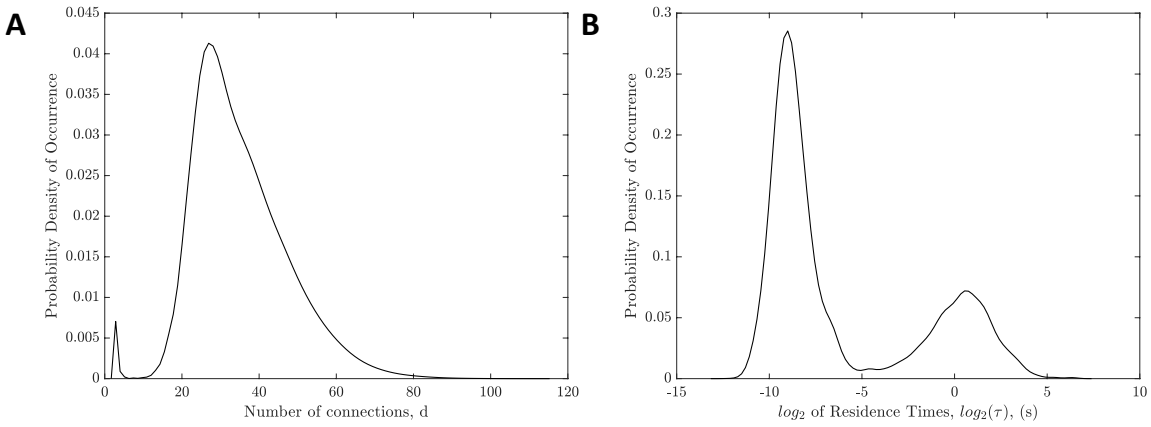

Figure 2.3 – **Probability density of occurrence for the parameters in Eq. (2.3). A** Number of connection, d and **B** $\log_2$ of residence times, $\log_2(\tau)$.

We studied the steady-states for model Eq. (2.3), which is a way to obtain the TF occupancy pattern. To do so, we admitted no variation in our system ($\mathbf{p}'(t) = 0$) and

the Laplacian matrix (Strogatz, 2015; Avcu; Molina, 2016). Thus, for any region i in the network, the steady state in this region is:

$$p_i = [T]\frac{\tau_i d_i}{\sum_j \tau_j d_j} \ , \tag{2.5}$$

in which τ_i is the residence time and d_i the number of connections for region i. The normalization for the steady state is given by $\sum_j \tau_j d_j$, and $[T]$ is the nuclear TF concentration. Given Eq. (2.5), we solved our system for RelA ($p65$) TF binding sites and the chromatin network from (Rao et al., 2014).

2.1.1 Simulating the steady-states

From Eq. (2.5), the number of connections and the residence times, we evaluated the TF occupancy for Eq. (2.3), considering 2000 TFs inside the nucleus ($[T] = 2000$). In Fig. 2.4 **A**, we proposed the solution for all the regions, where the TF concentration is less than 20% of the number of regions. Since accessibility and connectivity are important factors for occupancy, we verified how some regions retain TFs and others are expected to remain unoccupied.

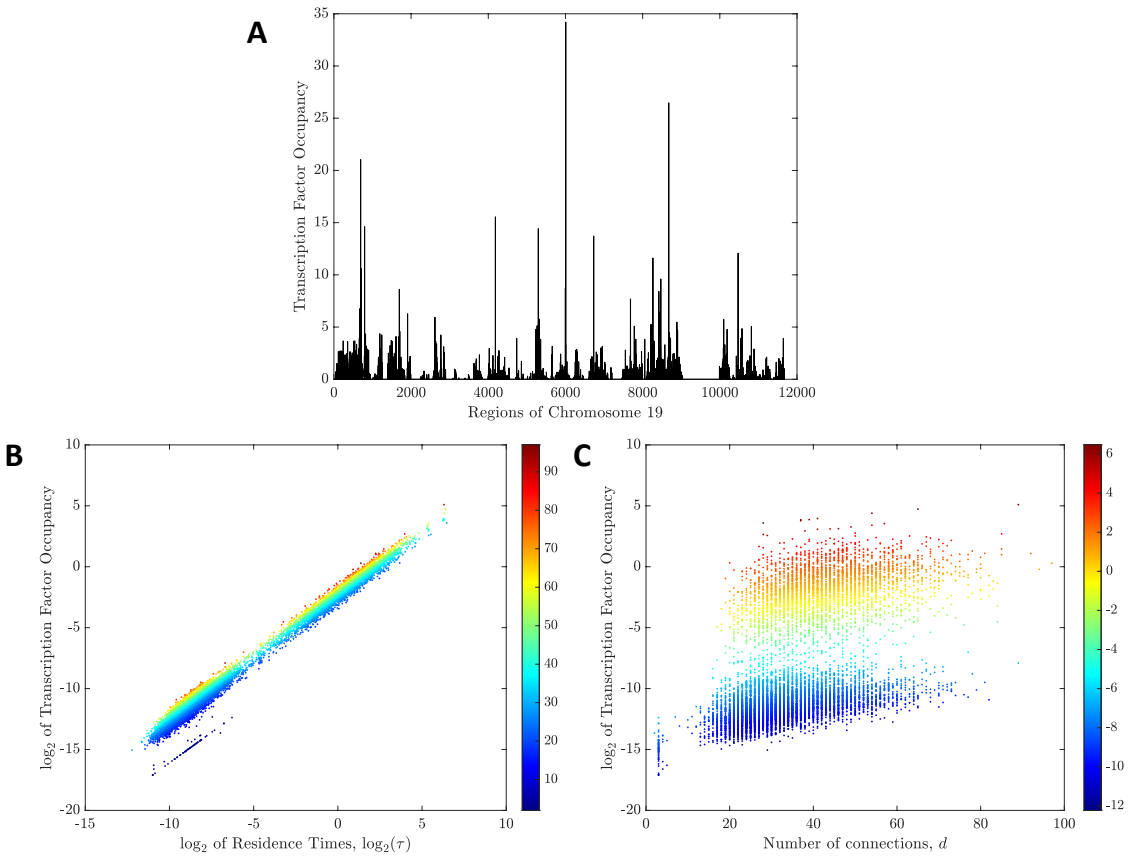

Figure 2.4 – **Solutions for Eq. (2.5) given our network and residence times. A** Transcription Factor Occupancy. **B - C** $\log_2$ of TF Occupancy, with **B**, $\log_2$ of the residence times (coded by the number of connections) and **C** number of connection (coded by the $\log_2$ of the residence times).

In Figs. 2.4 **B** and **C**, we analyzed the solution in Fig. 2.4 **A** and compared with the values of the residence times, **B**, and the number of connections, **C**. Fig. 2.4 **B** showed how the $\log_2$ of the residence times in which each value has the number of connections. The low connected regions (present in Fig. 2.3 **A**) create an outlier from the occupancy pattern. The number of connections increases the chances of being occupied as expected from Eq. (2.5).

To verify how the regions' connectivity affects the occupancy, we proposed Fig. 2.4 **C**, where the outliers in Fig. 2.4 **B** are verified in the bottom left of Fig. 2.4 **C**. Since our network is binary (we assumed whether there is a connection between the regions or not), we can see the discrete values of d. Another interesting behaviour is the two clouds of the concentration dependence of the $\log_2(\tau)$ by the activity (and inactivity). Both structure and accessibility influence the TF occupancy as we presented in Eq. (2.5).

2.1.2 Specific Influences from Structure and Accessibility on the Occupancy

From Eq. (2.5), both the connectivity (d_i) and the accessibility/sequence (τ_i) affect the TF occupancy patterns. To showcase the specific effects of the structure and the residence times on the TF occupancy, we proposed two different conditions: (i) a fully-connected network with the same network size L and (ii) an averaged residence time equal in each region, i.e., we have a 3D diffusive process, instead of a facilitated diffusion one.

For the fully-connected network, we assumed all the regions are connected and the only region-specific parameter is the residence times, creating a network in which all the nodes are reached with only one step, facilitating, even more, the TF diffusion. This means the structure does not influence the occupancy since, $\forall i \in \mathbf{N}, d_i = L$.

We present the steady-states considering the fully-connected network in Fig.2.5 **A**, in which we can see both increases and decreases in TF concentration when we compared with Fig. 2.4 **A**. The increase in occupancy in some regions is explained by the connectivity and the decrease is the effects of the TF nuclear concentration and the residence times. The unoccupied regions around the $9000 - 10000$ interval remain unoccupied considering this network, meaning the lack of TF occupancy is due to inaccessibility/ residence times values.

Since all the regions have the same number of connections, the only region-specific parameter is τ. Thus, we verified the linear influence of τ in our steady-state for the fully-connected network in Fig. 2.5 **B**. This means the occupancy is only affected by the residence times, i.e., the sequence specificity and accessibility of the region.

Still, we need to verify how the occupancy was affected by the structural change in the network and we verified the differences in concentration in Fig. 2.5 **C**. First, the outliers with low concentrations had a slight increase in value due to the expansion of connectivity

everywhere. The improvement pattern in concentration for regions with smaller residence times in comparison with the results from Fig. 2.4. Since we have limited transcriptional resources, the occupancy in more active regions is decreased in part of the regions showing active transcription sites and it is a consequence of all regions being reached in one step. This result guarantees that changes in connectivity affect how the molecules occupy the space, which was expected.

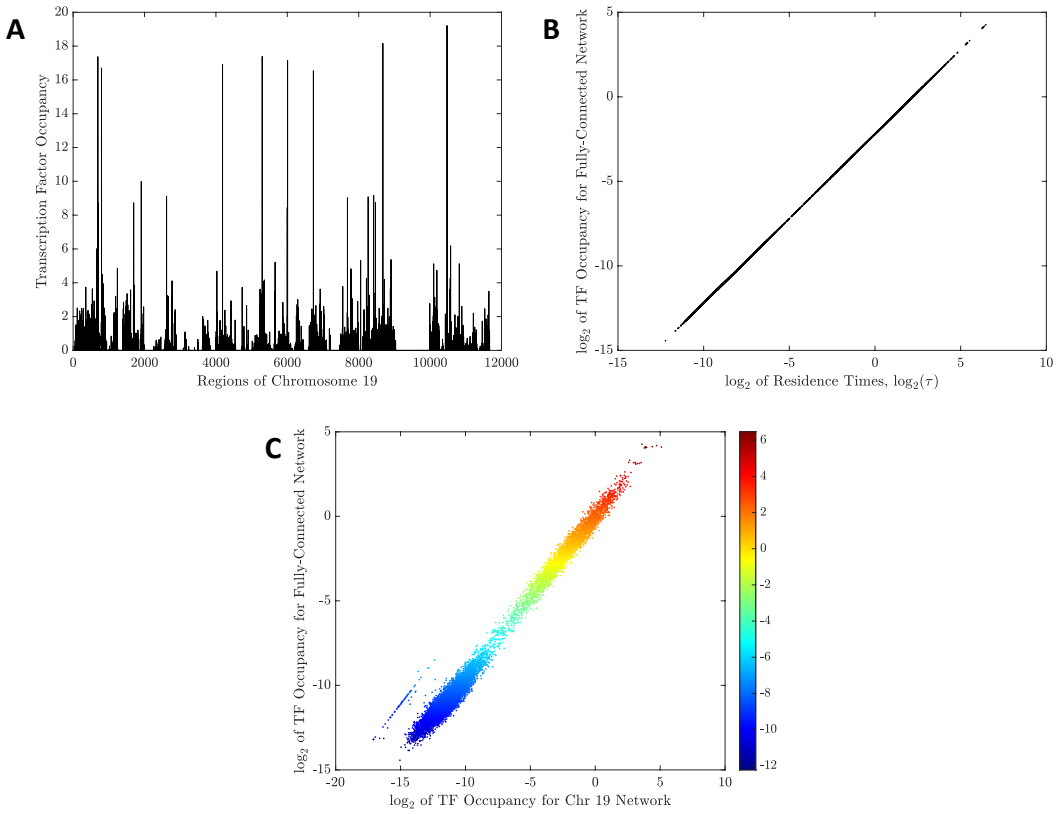

Figure 2.5 – **Steady-state solutions considering a fully-connected network. A** Steady-state occupancy per region; **B-C** $\log_2$ of TF steady-state in **A** where **B** $\log_2$ of residence times and **C** $\log_2$ of TF steady-state in Fig. 2.4 **A**, with $\log_2 \tau$ values.

To improve our knowledge of the residence times and their effects on TF occupancy, we proposed each region has the same residence time, $\langle \tau \rangle = 0.7881$ seconds, which is the average of all the residence time values for our solution in Fig. 2.4. In this case, we can see how the TF occupancy might occur in a system with only 3D diffusion, and we present the occupancy pattern for this condition in Fig. 2.6 **A**.

Since we admitted $[T] = 2000$, we do not have TF concentration higher than 1 in any region, i.e., $\forall i \in \mathbf{N} : [T_i] < 1$, with an average occupancy of $\langle [T_i] \rangle = 0.1714$ TF's molecules. From Fig. 2.6, we can see how the number of connections has an impact on the TF distribution, and how a diffusive process creates a pattern of reduced TF concentration everywhere, which is not biologically feasible since TFs are region-specific parameters.

Thus, this steady-state also proves why a 3D diffusive process only is not an optimal way to understand the TF search process.

As expected, occupancy is linearly dependent on the number of connections, as we show in Fig. 2.6 **B**. However, the effects of d in the occupancy are less effective in increasing TF occupancy per region than τ values, creating a more spread-like occupancy pattern, with the absence of unoccupied regions.

Once we compared $\log_2$ of Fig. 2.4 **A** with $\log_2$ of Fig. 2.6 **B**, we can see two separate clouds of TF occupancy. The two clouds represent the separation between active and inactive regions, as we have shown in Fig. 2.3 **B** and how the connectivity improves the occupancy in those two states. Another interesting result is the set of completely off outliers with low concentration in the bottom left corner of Fig. 2.6 **C**: those regions were the less occupied regions and they correlate with regions almost disconnected from our network, represented in Fig. 2.3 **A**. Therefore, we can see how incorporating a facilitated diffusion mechanism into TF search improves our abilities to predict TF occupancy.

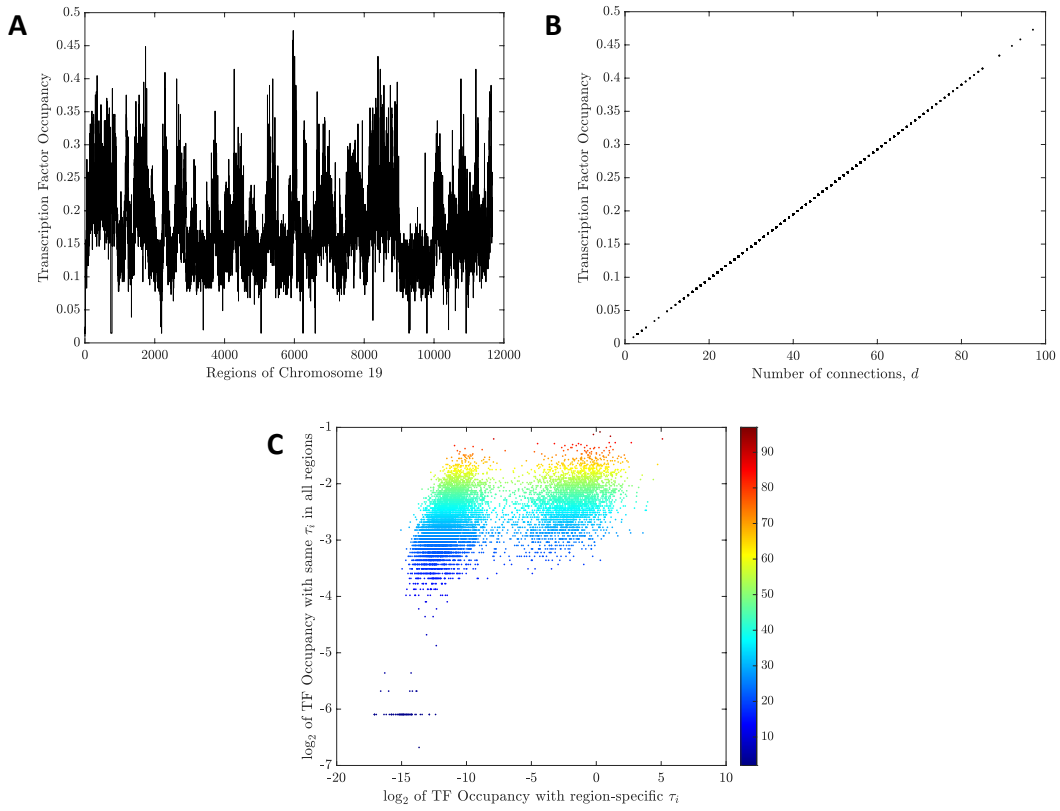

Figure 2.6 – **Steady-state solutions considering only a diffusive process, i.e., all the regions have the same residence times, $\langle\tau\rangle$. A** Steady-state occupancy per region; **B-C** $\log_2$ of TF steady-state in **A** where **B** number of connections per region, d, and **C** $\log_2$ of TF steady-state in Fig. 2.4 **A**, with d values.

The past examples were extreme and artificial conditions we created to verify how the TF occupancy pattern can emerge from a network fully connected and without region-

specificity for binding. To take our analysis one step further, we proposed randomizing the parameters to maintain the same parameter set conditions from Fig. 2.3 but in different (randomized) regions.

2.1.3 Randomizing the Network and Residence Times

As we wanted to understand how parameters changes affect the occupancy in a less extreme way as described in the previous subsection, we defined two different networks with the same parameters values from Figs. 2.3 **A** and **B**. We opted to randomize the conditions in two different ways: (i) the connectivity, d, i.e., we fixed the residence time vector τ but with different connections between regions (since connections often change due to the chromatin movement inside the nucleus), and (ii) the residence times with the original network, i.e., since TFs are region-specific, we consider that our system represents a different TF.

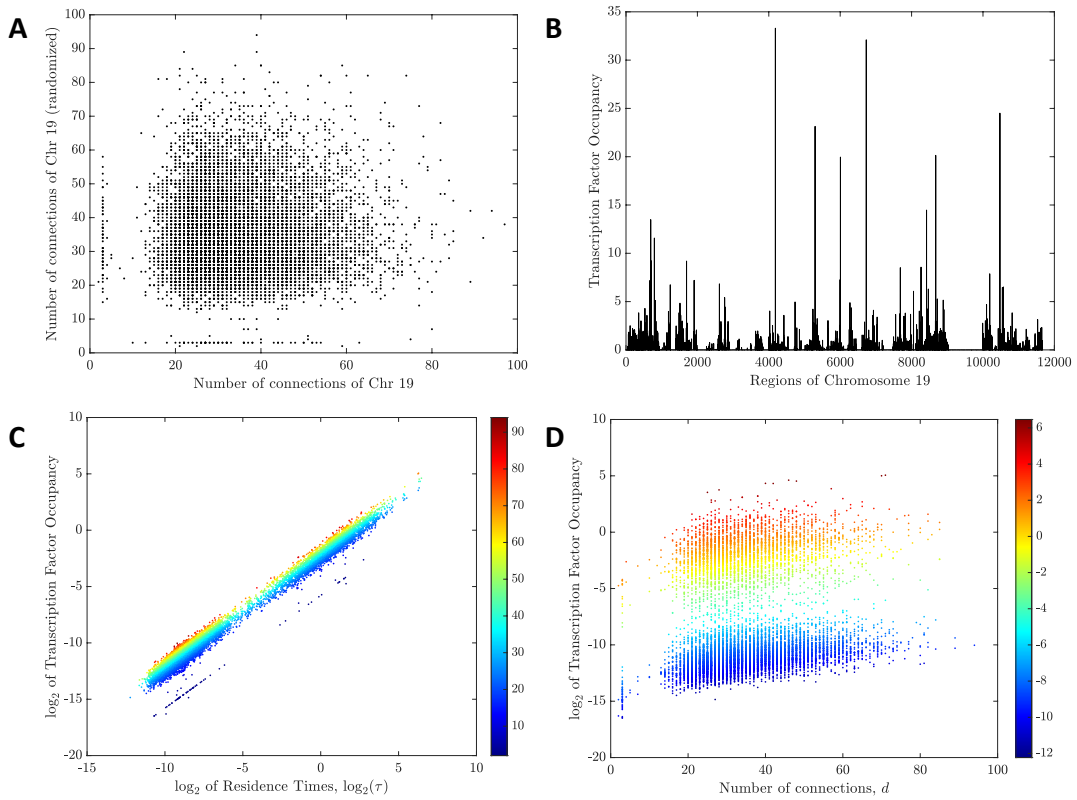

Figure 2.7 – **Steady-state solutions considering a new randomized version of our network connectivity. A** The connectivity changes between our initial network and the new network **B** Steady-state occupancy per region; **C-D** $\log_2$ of TF steady-state in **A** where **C** $\log_2$ of the residence times, and **D** number of connections per region, d.

In Fig. 2.7, we present the results after rearranging the number of connections d_i but maintaining the same active/inactive regions. In Fig. 2.7 **A**, we showed the difference

between our initial network and the randomized version of it, to prove the changes are big enough between both networks.

Given our new network, we showed its steady-state (Eq. (2.5)) in Fig. 2.7 **B**, in which the occupancy pattern does not deviate as much from Fig.2.4 **A** because of the same τ. However, as Eq. (2.5) defined, both d_i and τ_i determine the expected TF occupancy in the region i, thus the new network created a different occupancy pattern.

We checked how the new network changed our occupancy in Figs. 2.7 **C** and **D**: in **C** we verified the occupancy pattern over the log_2 of the residence times and in **D** the number of connections. Both figures showed a similar pattern to the ones in Figs. 2.4 **B** and **C**.

If compared Fig. 2.7 **C** with Fig. 2.4 **B**, we obtained regions with low TF occupancy and active regions, i.e., such regions are less accessible even if they are attractive. This result does not mean those regions are not reached over time: TFs explore chromatin very fast, but this subset of active regions but with a low number of connections, namely the ones present in the left part of Fig. 2.3 **A**, which affects the probability of finding a TF in those regions.

The cloud that separates the regions between active and inactive subsets is once again present in Fig. 2.7 **D**, similar to the ones present in Fig. 2.4 **C**. The outlier from Fig. 2.7 **C** is split between the two clouds of activity.

It is clear a change in the network affects the TF occupancy - as defined by Eq. (2.5); thus, in Fig. 2.8, we proposed a comparison between the two networks. In Fig. 2.8 **A**, we verified how the new network is different from the expected. Another interesting result is the three outliers: one with increased TF concentration and the other two with reduced TF concentrations for the network. The effects of τ are maintained in both networks since they are the same, as we showed in Fig. 2.8 **B**.

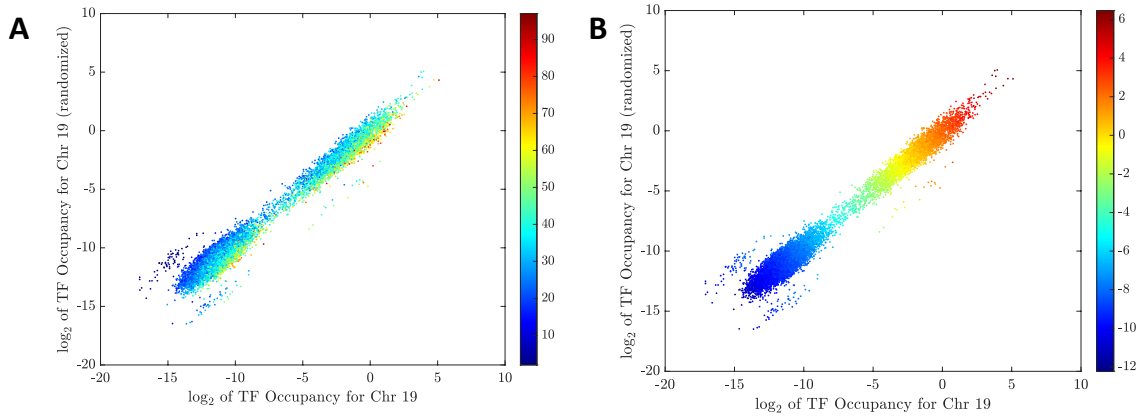

Figure 2.8 – log_2 **of TF occupancy for the randomized network over the** $\log_2$ **for the TF occupancy from Fig. 2.4 A.** With the values from **A** number of connections d and **B** residence times, τ.

Our result for a different network showed how the differences in connectivity influence the occupation pattern, which is an interesting result since cells are stochastic and chromatin is free inside the nucleus, creating and destroying connections between regions. Next, we randomized the residence time values, τ, which can be understood as a different TF exploring the same network, and presented the results in Fig. 2.9.

The randomized values of τ were present in Fig. 2.9 **A** to prove how different one vector is from the other. Since those values are in $\log_2$ space, we have 4 separate clouds of residence times. We also fixed the network from Fig. 2.4, as we aimed to create a different pattern for a different TF. The occupancy pattern per region is present in Fig. 2.9 **B** and it shows the preference to bind different regions, as this model can be understood as a representation of another transcription factor active in the nucleus. From the steady states, it is clear this new TF binding to our network presents a decrease in occupancy, proving the connectivity's importance in TF occupation.

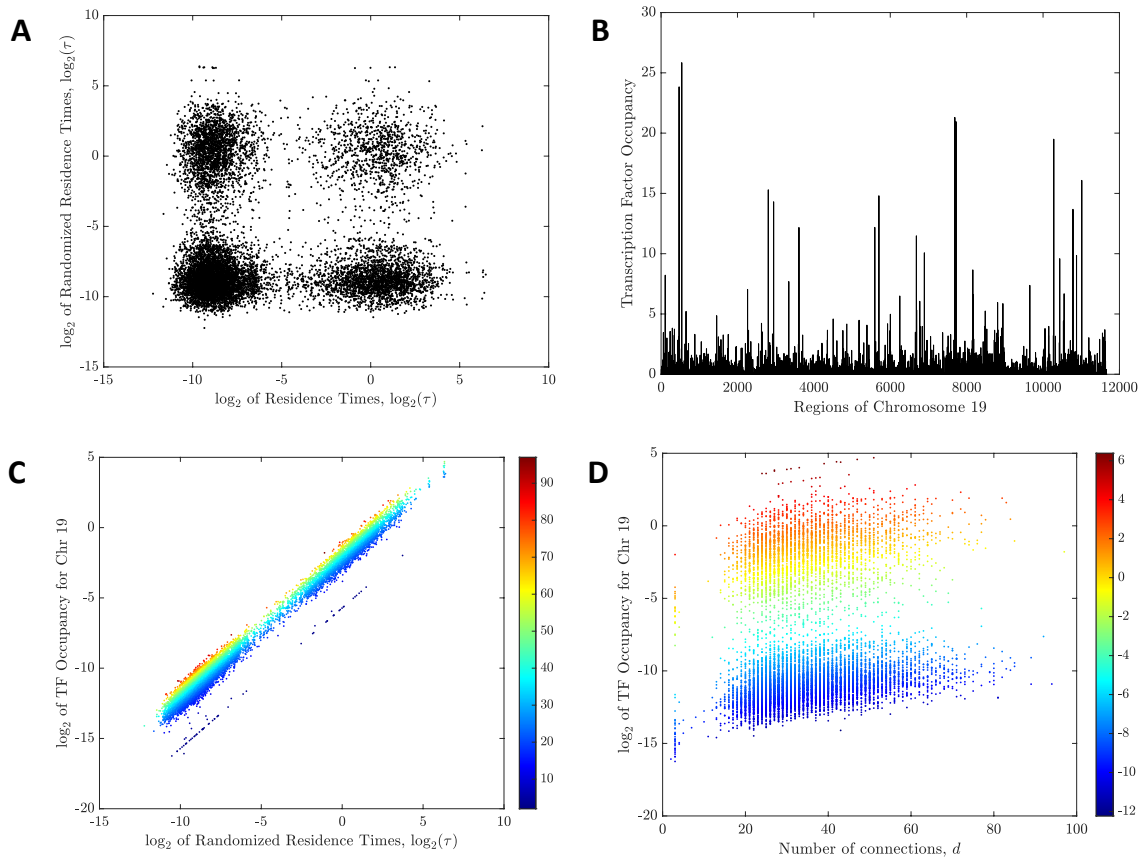

Figure 2.9 – **Steady-state solutions considering a new randomized version of the residence times. A** The differences between $\log_2 \tau$ and $\log_2 \tau^R$, the randomized τ **B** Steady-state occupancy per region; **C-D** $\log_2$ of TF steady-state in **A** in which **C** $\log_2$ of the residence times, and **D** number of connections per region, d.

Similar to the randomized network we previously generated, Figs. 2.8 **C** and **D**, we proposed Figs. 2.9 **C** and **D** to compare d and τ effects on the occupancy likelihood.

Fig. 2.9 **C** shows the $\log_2$ of the steady-states in Fig. 2.9 **B** over the $\log_2(\tau^R)$, i.e., the randomized values of τ. The combination of $\tau_i d_i$ shows a subset outlier of low-occupied regions for higher expected residence times (i.e., active regions). However, differently from our previous result, we had a decrease in the maximum concentration as a consequence of changing the TF for that fixed network, but with an increase in more highly-occupied regions (more than 10 TFs in the region).

To compare how changes in accessible regions impact TF occupancy, we present in Fig. 2.10 the $\log_2$ of Fig. 2.9 **B** over Fig. 2.4 **A**. Different from the randomized network, the concentration for the system with τ^R behaves similarly as Fig. 2.9 **A**. In Fig. 2.10 **A**, we showed how the low connectivity for a region does not imply low TF concentration and how the number of connections has a weaker influence in the steady state. When we consider the values of $\log_2 \tau$, Fig. 2.10 **B**, we verified how our system depends on the time a TF spends bound to a region.

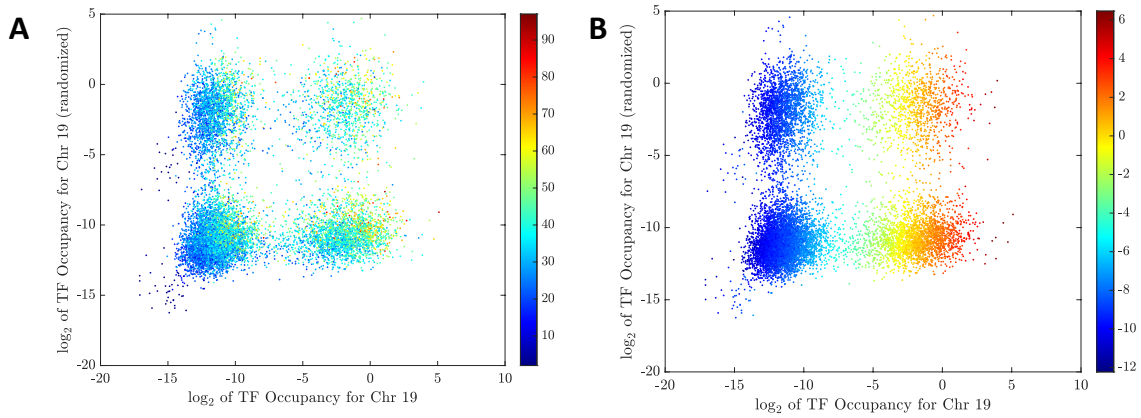

Figure 2.10 – ***log*$_2$ of TF occupancy for the randomized residence time over the $\log_2$ for the TF occupancy from Fig. 2.4 A.** Each point also presents **A** number of connections d and **B** residence times, τ.

All our analyses showed how both the number of connections and the residence times impact the expected behaviour of a TF. The network changes proved how structural changes facilitate or hinder access to particular regions, as chromatin is not a stable network as we supposed (Dion; Gasser, 2013; Soutoglou; Misteli, 2007). The activity (or inactivity) of a region to transcription is also a mechanism of gene regulation, thus changes in accessibility influence how a TF binds chromatin.

From all those studies, we saw some regions were so attractive to the TFs that we had the emergence hubs of transcription factors. Those TF hubs can also interact between themselves, i.e., those TFs cluster together to maintain the transcription in a particular region (Wollman et al., 2017; Liu et al., 2014; Johnnidis et al., 2005). As a tool to comprehend this mechanism, we expanded the model in Eq. (2.3) to incorporate the TFs protein-protein interactions in the next section.

2.2 TF Cluster Formation and the need for Volume Exclusion

In the previous section, we discussed how the structure affects the search and in consequence the TF occupancy patterns and also how the accessibility influences the TF binding. Both characteristics generated highly-occupied regions for TFs which is due to how attractive those regions are to the TFs, which indicates other mechanisms are at play than just the 3D/1D process. One answer for this clustering of TFs is the presence of weak protein-protein interactions between the TFs as a mechanism for gene expression (Cherstvy; Teif, 2013; Zhang et al., 2019; Gibcus et al., 2018; Fudenberg et al., 2018; Nagamine; Kawada; Sakakibara, 2005). This difference in TF concentration through the nucleus is a well-known feature in cells (Meeussen et al., 2022).

To model the genome-wide effects of protein-protein interactions on TF occupancy and, eventually, transcription, we extended Eq. (2.3) from the previous section and (Avcu; Molina, 2016). To include protein-protein interactions into the model, we assumed that each TF molecule can interact with up to I other molecules (as the molecules' geometry influences the interactions, I is a small number compared with the size of the network, L - i.e., $I << L$), and these interactions are formed and broken with the constant association and dissociation rates K_a and K_d.

A cartoon representation of this system is present in Fig. 2.11. There, we assumed four different scenarios for protein-protein interaction: (1) Zero or No Interactions (blue star, no dashed lines); (2) One Interaction (pink star and one dashed line); (3) Two Interactions (green star and two dashed lines); and (4) Three Interactions (yellow star with three dashed lines). Therefore, to model this system, we first assumed that only molecules with zero interactions are free to jump from a given region to a new one provided that both are connected since TF-TF interactions restrain the movement to another region.

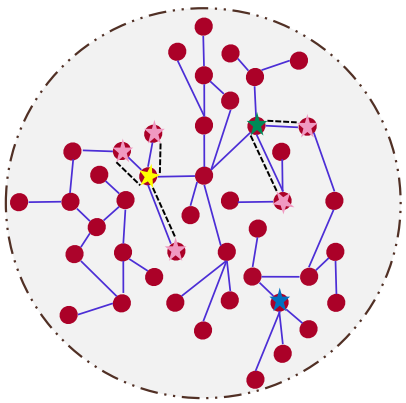

Figure 2.11 – **Cartoon representation of Protein-Protein Interactions.** In which the blue star without a dashed line represents a region without protein-protein interactions; the pink star with one dashed line one protein-protein interaction; the green star with two dashed lines connecting it to two protein-protein interactions; and the yellow star connected to three dashed lines represents three protein-protein interactions.

From the results of the previous section, we found some regions had high TF concentration (in those cases, $[T_i] > 10$, even if $[T] < L$). To avoid aggregation of all TF molecules in one single region, we imposed that all regions have an equally limited space to fill with TFs, which is feasible since each region has the same size, even if by the residence times, τ, some regions can be more or less open to TF binding.

Thus, we introduced a constant carrying capacity C that represents the maximum number of TFs that a region can support. The presence of a constant to limit the occupancy in a region is also a necessity to make our model more realistic, as there are no regions with infinite space. With our assumptions and considering Eq. (2.3) as a good approximation for the TF search process in chromatin, we present our model in Eq. (2.6).

$$\begin{cases} \dfrac{dp_i^0}{dt} = -k_i p_i^0 + \displaystyle\sum_j \frac{(C - p_i^*)A_{ij}k_j p_j^0}{\sum_q (C - p_q^*)A_{jq}} - K_a f_i p_i^0 + K_d p_i^1 \ ; \\ \dfrac{dp_i^\alpha}{dt} = K_a f_i p_i^{\alpha-1} - K_d p_i^\alpha - K_a f_i p_i^\alpha + K_d p_i^{\alpha+1} \ ; \quad 0 < \alpha < I \\ \dfrac{dp_i^I}{dt} = K_a f_i p_i^{I-1} - K_d p_i^I \ ; \end{cases} \tag{2.6}$$

where p_i^α is the number of TF molecules in region i that have α interactions, $(0 < \alpha < I)$. The two first terms of the first equation ($\alpha = 0$) describe the search for a binding site is similar to the one proposed in Eq. (2.3); however, here, we considered that the probability of jumping into a region is proportional to the free space available there. The formation of protein-protein interactions is modelled as a second-order reaction governed by the association rate K_a, which is not region-specific, i.e., all regions have the same capability to form clusters. We also assumed the dissociation rate, K_d, is not region-specific and we supposed $K_d = 1s^{-1}$.

A new protein-protein interaction depends on the number of available TF molecules in the neighbourhood can establish a new protein-protein interaction. This quantity of TF's availability is expressed as $f_i = \sum_j \sum_{\beta=0}^{I-1} p_j^\beta A_{ij}$, i.e., a TF in the region i depends on TF concentration with less than I interactions in all regions connected to itself. Finally, we defined $p_i^* = \sum_{\beta=0}^{I} p_i^\beta$ as the total concentration of TFs in region i, since this concentration should be smaller than the carrying capacity C. Given there is a limited amount of TFs inside the nucleus, we also define the total TF concentration of our system as $[T] = \sum_j p_j^*$.

Different from Eq. (2.3), Eq. (2.6) is not easily solvable. Thus, the understanding of its stability is the main tool to understand the patterns that can emerge from our model, which we present in the next subsection.

2.2.1 Parameters and Stability Analysis

The first point to consider about our model in Eq. (2.6) is that the system has $(I + 1) \times L$ equations. Of course, this means our model is more complex to solve and

understand than the previous one, but we still can obtain its behaviour and how stable are the solutions.

Considering the steady state for our system (i.e., when we have no variation over time (Strogatz, 2015; Murray, 2007)), we obtained the following solution for any $\beta \geq 0$ depending on p_i^0 in Eq.(2.7).

$$p_i^\beta = (K_a f_i)^\beta p_i^0 \ . \tag{2.7}$$

Note that this critical point is *uniquely* defined but not unique. More than that, f_i is still dependent on all p_i^γ values except for $\gamma = I$. However, we still can write the set of all critical points, $\mathcal{C}$, as one in which the solutions obey Eq.(2.7).

Considering proteins are very complex and long compounds, the number of possible interactions in our system must remain low enough to be biologically feasible. Thus, we supposed the number maximum of interactions $I = 3$ and we defined $\mathbf{p}^0 = (p_1^0, ..., p_L^0) \in \mathbb{R}_+^L$ (as our system is counted as molecular concentration, we supposed only non-negative real numbers) and $\mathbf{f} = (f_1, ..., f_L) \in \mathbb{R}_+^L$ (as a direct consequence of $\mathbf{p}^0$). The first vector contains all the no-interacting TF concentrations for $i \in [1, L]$ and the second represents the chances for each region to *associate*. We also note $K_a \in \mathbb{R}_+$.

$$(\mathcal{C}_0, \mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3) = (\mathbf{p}^0, K_A \mathbf{f}\mathbf{p}^0, (K_A \mathbf{f})^2 \mathbf{p}^0, (K_A \mathbf{f})^3 \mathbf{p}^0), \tag{2.8}$$

Note that $(\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0})$ is a feasible critical point in Eq. (2.8). However, we analyzed the behaviour in respect of K_a for the case where $\mathbf{p}^0 \neq \mathbf{0}$, i.e., the case in which the solutions obey Eq. (2.7). Since we assumed all the parameters are non-negative, $\mathbf{p}^0, \mathbf{p}^1, \mathbf{p}^2, \mathbf{p}^3 \geq 0$.

- If $0 < K_a < 1$, then $\forall \beta$, $\mathbf{p}^{\beta-1}\mathbf{f} > \mathbf{p}^\beta$. So, the effect of protein-protein interactions is present but only slightly, i.e., the TFs will occupy more places with a smaller number of interactions between regions.
- If $K_a > 1$, then $\mathbf{p}^{\beta-1}\mathbf{f} < \mathbf{p}^\beta$, $\forall \beta$. Then, we expect stronger effects of cluster formations in higher levels of interactions.
- If $K_a = 1$, then we expect the TFs to be distributed over states as the ones present in Eq. (2.8).
- If $K_a \longrightarrow 0$, then the effect of protein-protein interactions is negligible and thus not altering the obtained on $\mathbf{p}^0$.

In Eq. (2.8), we proposed the steady-states for our system; however, the stability of this critical point is not known. Since our model complexity creates a high-order polynomial as the characteristic polynomial, we opted against furthering the analytical approach to understand this model. Thus, we defined the matrix version of our model, i.e., let κ be

the vector of exiting rates, with $\kappa \in \mathbb{R}^L$, $\mathbf{A} \in \mathbb{R}^L \times \mathbb{R}^L$, which we used in our numerical solutions.

$$\begin{cases} T^0 = \dfrac{d\mathbf{p}^0}{dt} = -\kappa\mathbf{p}^0 + \dfrac{\kappa(C-\mathbf{p}^*)\mathbf{A}\mathbf{p}^0}{(L-[T])\mathbf{A}} - K_a\mathbf{f}\mathbf{A}\mathbf{p}^0 + \mathbf{p}^1 \ ; \\ T^1 = \dfrac{d\mathbf{p}^1}{dt} = K_a\mathbf{f}\mathbf{A}\mathbf{p}^0 - \mathbf{p}^1 - K_a\mathbf{f}\mathbf{A}\mathbf{p}^1 + \mathbf{p}^2 \ ; \\ T^2 = \dfrac{d\mathbf{p}^2}{dt} = K_a\mathbf{f}\mathbf{A}\mathbf{p}^1 - \mathbf{p}^2 - K_a\mathbf{f}\mathbf{A}\mathbf{p}^2 + \mathbf{p}^3 \ ; \\ T^3 = \dfrac{d\mathbf{p}^3}{dt} = K_a\mathbf{f}\mathbf{A}\mathbf{p}^2 - \mathbf{p}^3 \ . \end{cases}$$

Again, we remind that $\mathbf{p}^* = \mathbf{p}^0 + \mathbf{p}^1 + \mathbf{p}^2 + \mathbf{p}^3$ and $\mathbf{f} = \mathbf{p}^0 + \mathbf{p}^1 + \mathbf{p}^2$. Thus, it is clear that evaluating our model analytically is not as simple as our previous model. However, a tool to understand our model is to obtain its numerical solutions.

2.2.2 Numerical Solutions and Parameter Set Exploration

To solve our protein-protein interaction model and explore how the association rate, K_a, and the carrying capacity, C influence the TF occupancy. First, we need to define our network and exiting rates for all our numerical solutions and, since the system has $(I+1) \times L$ equations, we opted to reduce the network in Fig. 2.3 **A** by averaging in blocks of 7, and removing its outliers; this result is present Fig. 2.12 **A**. We also averaged Fig. 2.3 **B** in blocks of 7 and removed the values from the outliers, and normalized the values of τ such that $\langle\tau\rangle = 1s$, Fig. 2.12 **B**

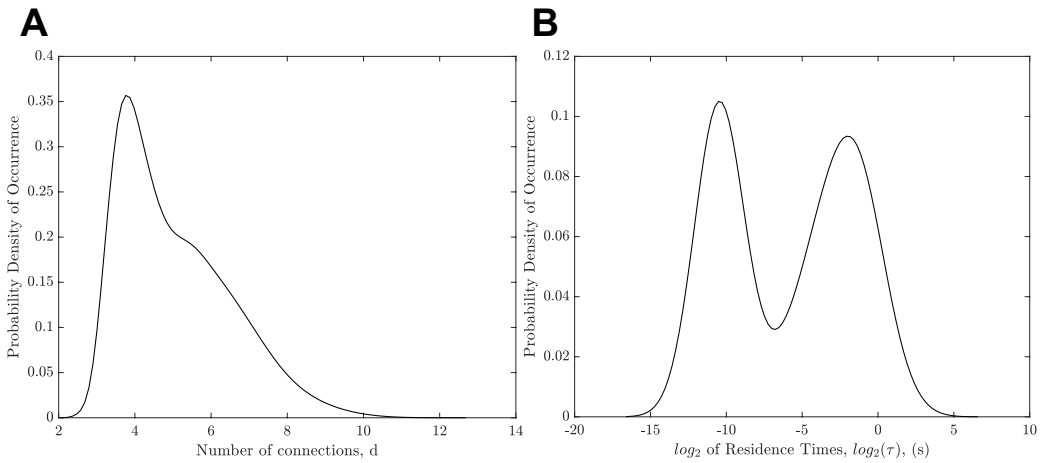

Figure 2.12 – **Probability density of occurrence for the reduced network from Fig. 2.3. A** Number of connections, d and **B** $\log_2$ of residence times, $\log_2(\tau)$.

As we aim to explore cluster formation, the two parameters in Eq. (2.6) are essential to comprehend the clustering around a specific chromatin region: the association rate (K_a) and the carrying capacity (C), since the association rate rule, the strength of clustering the system can achieve and the carrying capacity works as a limiting the full occupancy to prevent the aggregation in only a few attractive regions. With that, we supposed our

model has a fixed total concentration of 1000 TFs for our scaled network and residence times, i.e., $[T] = 1000$ TFs.

Therefore, we solved our model using the deterministic approach from Matlab's ode15s (Gupta; Wallace, 1975; Shampine; Reichelt, 1997), considering different values for both K_a and C while fixing the network, k_i, and $[T]$. We implemented our solution for 400 seconds, as since the movement of the TFs, and supposed the following values for $K_a = \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 1, 1.5, 2, 2.5\}$ and $C = \{5, 10, 15, 20, 25, 30, 35, 40\}$. We emphasize our simulations are not representatives of the steady-state but rather a study of the expected behaviour after 400 seconds. Examples of our results for the extreme values of K_a and C are present in Fig. 2.13.

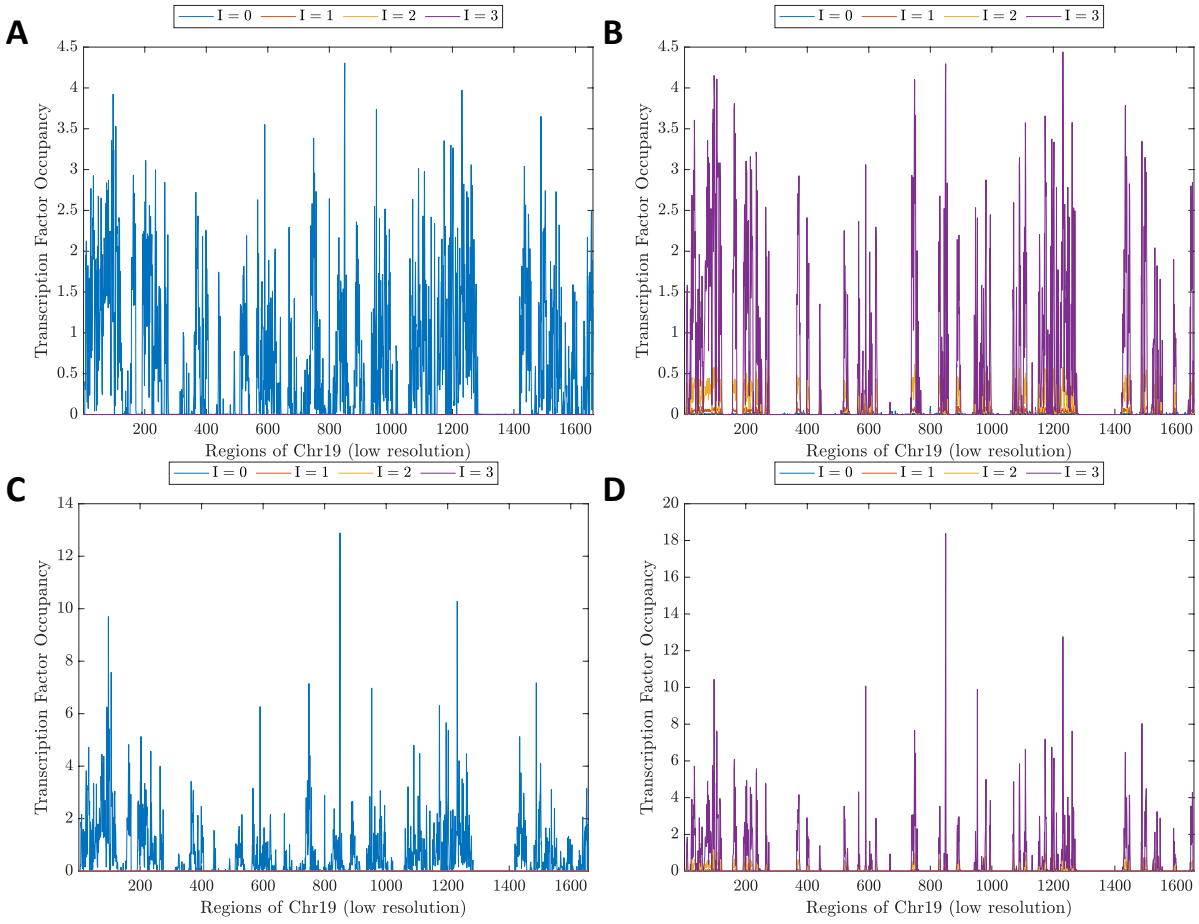

Figure 2.13 – **Numerical Solutions for Eq. (2.6) with different values and K_a and C. A-B** $C = 5$ with **A** $K_a = 0$ and **B** $K_a = 2.5$. **C-D** $C = 40$ in which **A** $K_a = 0$ and **B** $K_a = 2.5$. Here, we used the result from $t = 400$ seconds.

From Fig. 2.13, we can see how the values K_a and C influence the occupancy: first, for all the figures an inaccessible interval appears between regions 1200 and 1500. Such inaccessibility being sustained in Fig. 2.13 **A** and **B** is a sign of the difficulty of finding a TF over that region interval. In Fig. 2.13 **A**, the carrying capacity limits the TF concentration, thus stimulating the occupancy in other less attractive regions. Between

Fig. 2.13 **A** and **B**, the aggregation due to K_a increases the occupancy in some regions while decreases in the rest, as both figures have the same low carrying capacity, $C = 5$ - i.e., each region can only admit 5 TF molecules maximum. This clustering reproduces a transcription factory for genes in those specific regions, which facilitates transcription there and debilitates in other parts of our network.

Once we compare Figs. 2.12 **A** and **C**, we verified how allowing more TF TFs to agglutinate in C increases the occupancy in prolific regions. Thus, granting the maximum TF concentration in our model benefits the transcription in the regions with better connectivity and residence times, given those regions are more attractive for the TFs. Again, the presence of a non-negative association rate increases the occupancy in prolific regions, a result that corroborates the assumption of cluster formation improving transcription.

Fig. 2.13 showed how sensitive is the model to parameter changes. Next, we analyze the effects of (i) the association rate and (ii) the carrying capacity given fixed values of C and K_a, respectively.

2.2.2.1 Association rate to tune gene expression

The association rate influence on our model was previously verified in our stability analysis and Fig. 2.13. To improve the understanding of K_a, we fixed C (i.e., we assumed $C = 15$), and implemented the numerical solutions for different values of K_a. With a fixed value for the carrying capacity and nuclear TF concentration, we validated how our model behaves with distinct values of K_a, some solutions are present in Fig. 2.14.

Fig. 2.14 **A** presents a concentration of TFs lower than the carrying capacity since the value of C considers all the protein-protein interaction levels. Since this figure is a consequence of the low association rate, we have that numerous of the TFs remain without protein-protein interactions. Thus, while such K_a creates clusters, the distribution of protein-protein interactions is more even, which is not the case for the other figures in Fig. 2.14. Another interesting result is the maximum TF concentration is less than 1/3 of the carrying capacity after 400 seconds.

However, in Figs. 2.14 **B**, **C** and **D**, we can see how the association rates directly influence the TF concentration in $I = 3$ and, in consequence, decrease previously occupied regions. Thus, the TFs with a higher propensity of clustering will occupy their target regions in transcriptional hubs, as predicted by Eq. (2.8).

More in-depth, in Fig. 2.14 **B**, the regions with small TF concentrations are mostly in the zero protein-protein interactions state, $I = 0$. The maximum TF concentration found in the system doubled between **A** and **B** but the values for K_a in **B** are five times bigger than the K_a in **A**. The concentration rises for $I \geq 2$ states is found in Fig. 2.14 **C** and **D** as well, but the maximum value difference between them is less abrupt than the

one present in Figs. 2.14 **A** and **B**.

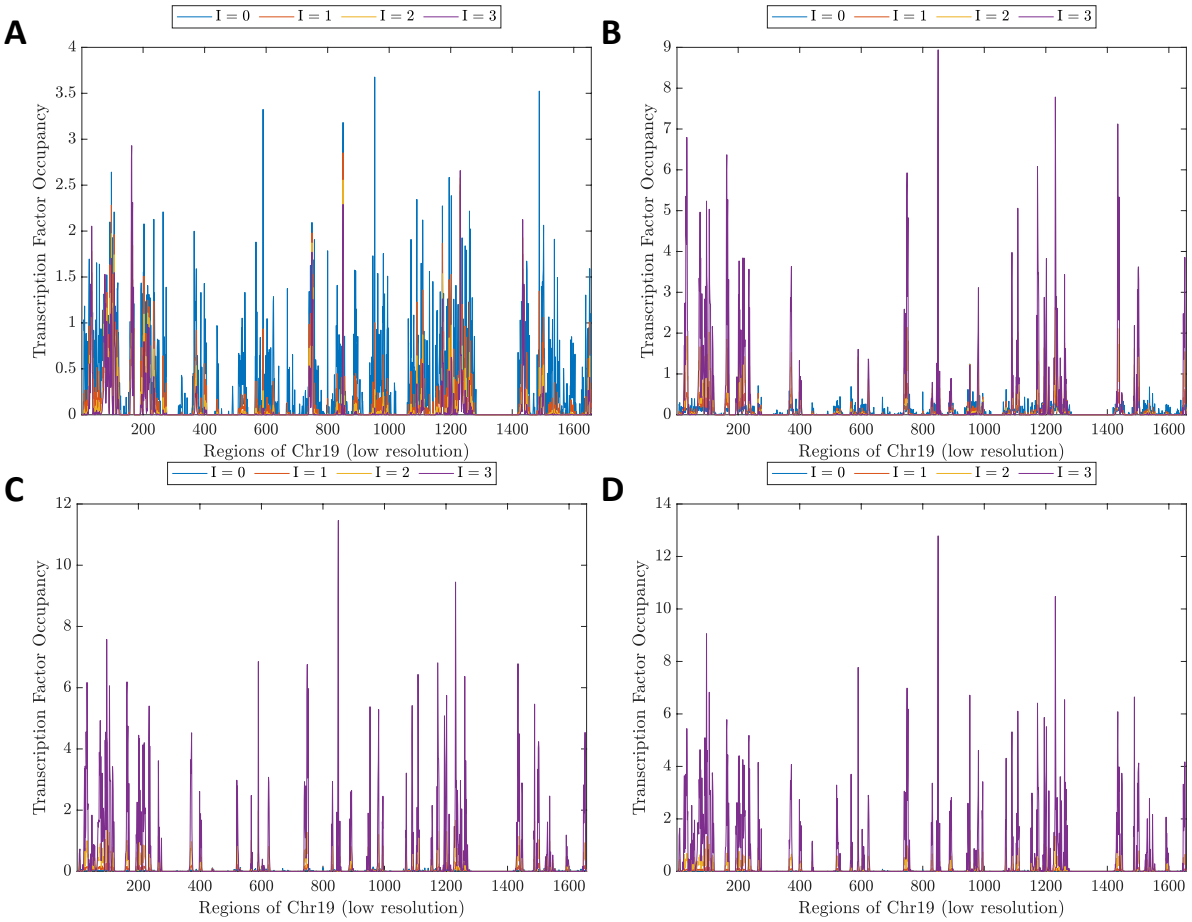

Figure 2.14 – **Numerical Solutions for our model in Eq. (2.6) for different values of K_a and $C = 15$.** Where, **A** $K_a = 0.05$; **B** $K_a = 0.25$; **C** $K_a = 1$ and **D** $K_a = 2$.

It is clear that concentration and PPI are affected by the values of K_a, as it increases the chances of two or more TFs to weakly interact. The carrying capacity C has an indirect impact on the cluster formation since the occupancy in states with PPI depends on the occupancy in $I = 0$ which is affected by C, Eq. (2.6). However, if our system allows a higher TF concentration in its nodes, the occupancy pattern changes its distribution opting for preferential regions, as we showed in Fig. 2.13.

As discussed earlier, the aggregation in specific (more prolific) regions forces the eviction in less attractive regions of our network, since transcriptional resources are limiting factors for gene expression. We present the analysis for specific regions and fixed carrying capacities in Fig. 2.15.

In Fig. 2.15 **A**, we verified two characteristics for this specific region: first, we demonstrated a fast increase in PPI once K_a in non-negative for both values of C, which is a consequence of a region being well-connected in the network. Then, as a consequence of the region being active but not preferential sequence-wise, while the TFs are more

likely to occupy this region than most of the network, once the values of K_a rise, its TF concentration drops as more attractive region recruit those TFs, clustering TFs in more attractive than the region present in Fig. 2.15 **A**. The second characteristic is more prominent for higher values of C, as it allows more Tfs per region.

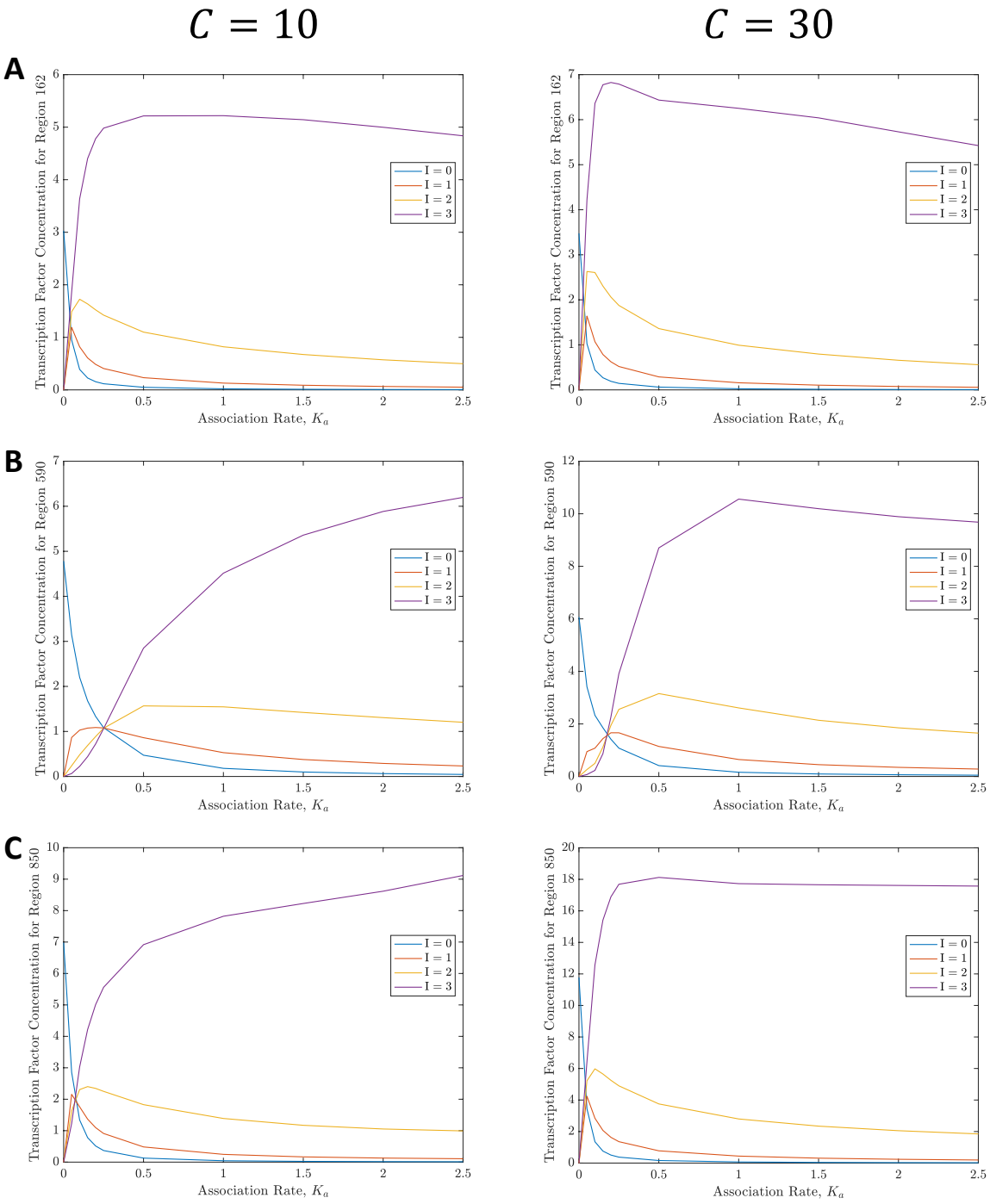

Figure 2.15 – **Association rate over TF occupancy for different regions and with different values of C, $C = 10$ and $C = 30$. A** Region 162. **B** Region 590. **C** Region 850. The τ and d values for the regions are present in Table 1.

In a sense, the region from Fig. 2.15 **B** is the opposite of the one from Fig. 2.15 **A**.

First, it shows a delay in higher states of PPI, a consequence of being less connected than its predecessor and since it entraps TFs more efficiently and its cluster increases with more effective values of K_a. Yet, similar to the previous region the allowance of more TFs in the regions creates a higher tendency to bind those prolific regions (i.e., highly connected and with better and accessible binding motifs), impacting the other regions' allocation.

Fig. 2.15 **C** is a very prolific region and we verified both the overshoot of aggregation for being highly connected and the sustained occupancy with the values of K_a. Another important result is that for given low values of C the TF concentration in that particular region is approximately C, which is not the case for higher values of C.

To verify how K_a affects our model on a global scale, we fixed both $[T]$ and C and analyzed the occupancy pattern at $t = 400s$ for all our values of association rate in Fig. 2.16 **A**, and we verified how by increasing the efficacy in clustering, bigger clusters form around specific regions, reducing the TF occupancy in other regions, as we verified also in Fig. 2.15. More so, we compared the occupancy changes between efficient clustering ($K_a = 2.5$) and the absence of clustering ($K_a = 0$) by evaluating the log of the ratio between their allocation patterns $t = 400s$ in Fig. 2.16 **B**, which we explicitly show how most of the regions in our network are negatively impacted by the presence of clustering formation, since few regions hoard more TFs inside, arresting the TF search.

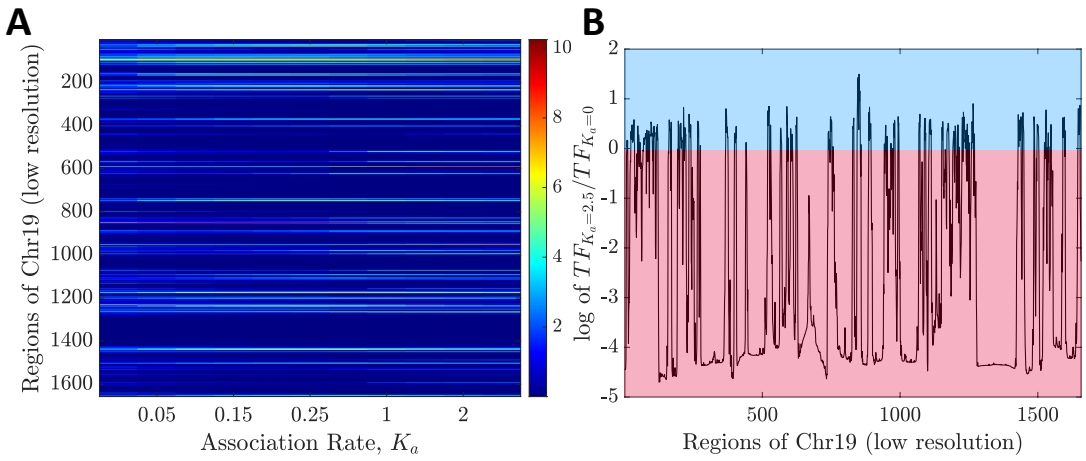

Figure 2.16 – **Impact of the association rate over TF occupancy for fixed $[T]$ and C (1000 TFs and 10 TFs, respectively) at $t = 400s$ A** Heatmap for all values of K_a in which we can see how the effectiveness of cluster formation influences the occupancy pattern **B** log of the ratio between the maximum K_a considered in our simulations ($K_a = 2.5$) and the absence of clustering ($K_a = 0$), showing the steep decrease in concentration in the less attractive regions to favour other regions.

Thus, the results from Figs. 2.15 and Fig. 2.16 justify an in-depth analysis of the carrying capacity effects for our model and in gene regulation.

2.2.2.2 Carrying Capacity as a mechanism to control the transcription levels

By limiting the maximum concentration of a TF in any chromatin region, we clearly affect the TF search process. In the previous subsection, we verified how the association rate impacts the cluster formation, hoarding available transcription factors in specific regions with higher PPI states.

To understand the global behaviour that emerges from varying the values of C, we implemented the numerical solutions for the same K_a and $[T]$ but with different values of C. This result is in Fig. 2.17, in which we admit $K_a = 1$ and $[T] = 1000$.

As discussed earlier, allowing more TFs to allocate in the regions create an impoverishment of transcriptional resources in less prolific regions and such a pattern is once again verified when we compared all the subfigures in Fig. 2.17. More than that, as expected, peaks of TF concentration also increase with C. It is important to notice that with $K_a = 1$, the cluster formation is strong but not all the regions have a maximum concentration in $I = 3$, as Fig. 2.17 **A** shows. Since the total nuclear concentration is fixed in all the numerical solutions, regions with low TF concentrations further reduce their concentration which is transferred to more attractive regions.

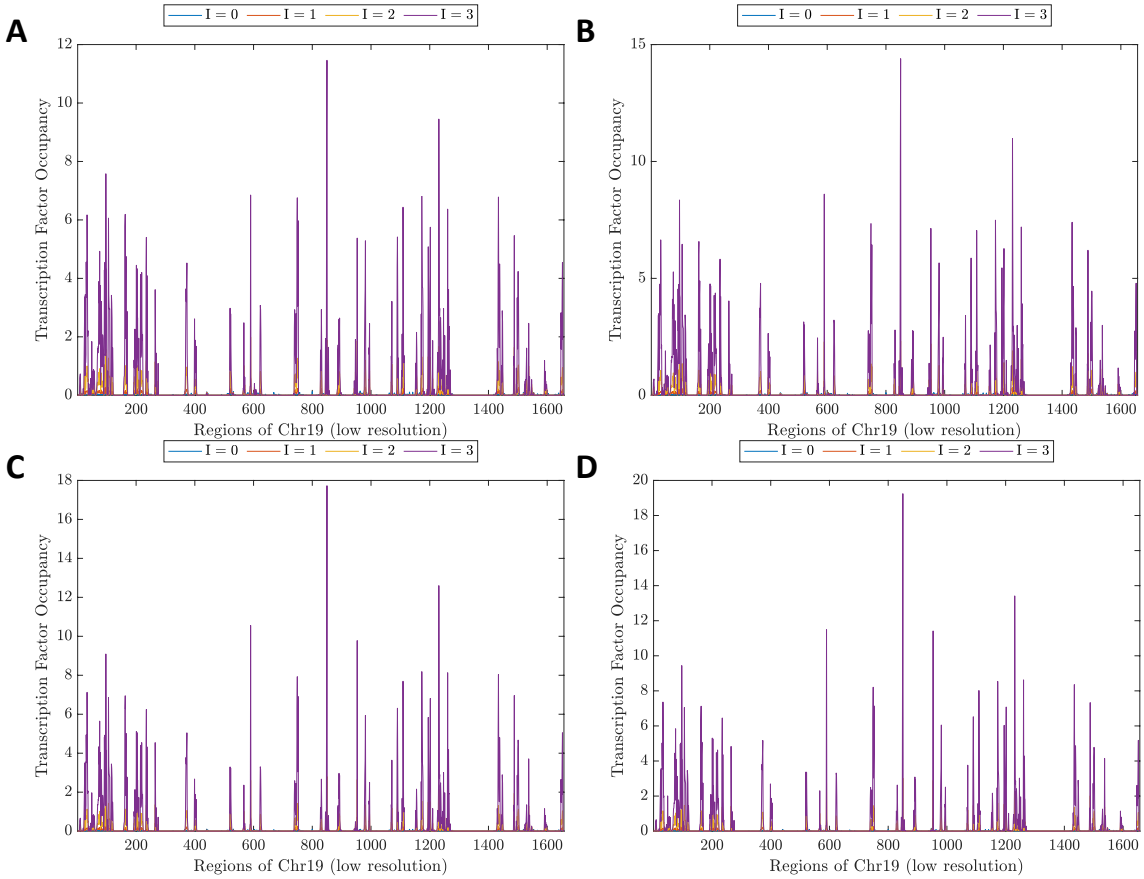

Figure 2.17 – **Numerical Solutions for our model in Eq. (2.6) for different values of C and $K_a = 1$.** In which, **A** $C = 15$; **B** $C = 20$; **C** $C = 30$ and **D** $C = 40$.

Another interesting result is that while our numerical solutions admitted up to 40 TFs/region, even the most occupied region does not reach C - as shown in Figs. 2.17 **C** and **D**, for example. Thus, we concluded that while C helps to control the TF's agglutination in a particular set of regions (or even just one), the search mechanism (Eq. (2.3)) used by the TF also works as an auto-regulating tool for TF allocation. The evolution of the C values also shows how groups of chromatin regions with higher concentrations in the highest PPI state ($I = 3$) were also reduced by allowing more TFs per region, i.e., increasing the clustering.

In the previous subsection, we verified how different values of K_a influence TF behaviour and cluster formation. Thus, we demonstrated how a lower than 1 association rate and low C influence the sum over all PPI states, cluster formation and, as a consequence of it, transcription in Fig. 2.18.

An important feature of this numerical solution is that since we consider a low value of C, the regions are more spread through less active/connected regions than the results in Fig. 2.17. Since we considered $K_a < 1$, this result helps us to understand how C influences the occupancy pattern and how regions with less expected activity are active if there is any limitation on the number of molecules allowed in each chromatin region.

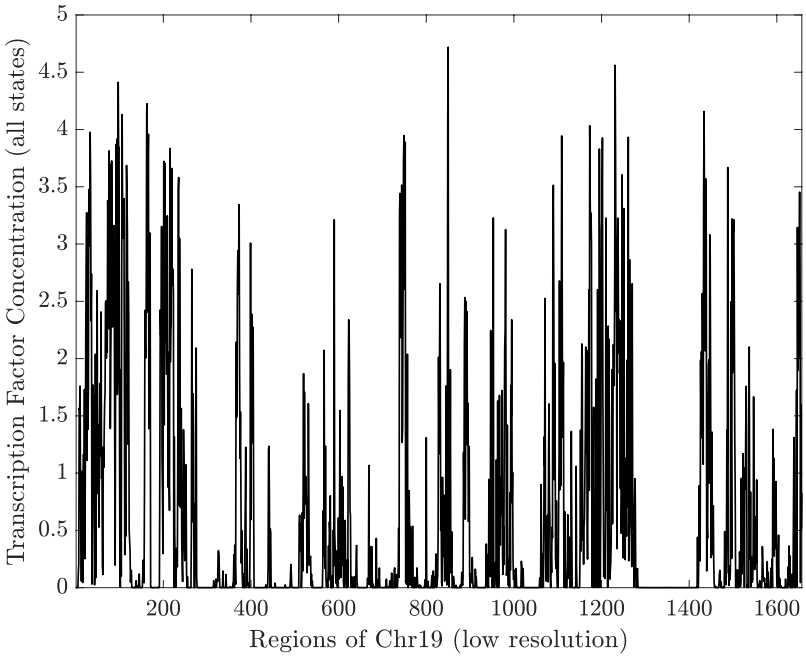

Figure 2.18 – **Sum over all PPI states of the numerical solution for our model in Eq. (2.6) for $C = 5$ and $K_a = 0.25$.**

This result implies C can be used as a mechanism to regulate gene expression and as a way to understand the TF placement considering regions with small binding affinity. Even in a case where the occupancy in a node is limited, some regions maintain their

unoccupied features, as the regions inside the interval $1200 - 1400$ show; i.e., some regions are so unlikely to be open to transcription in our model that even in a restricted system, we did not find TF occupancy.

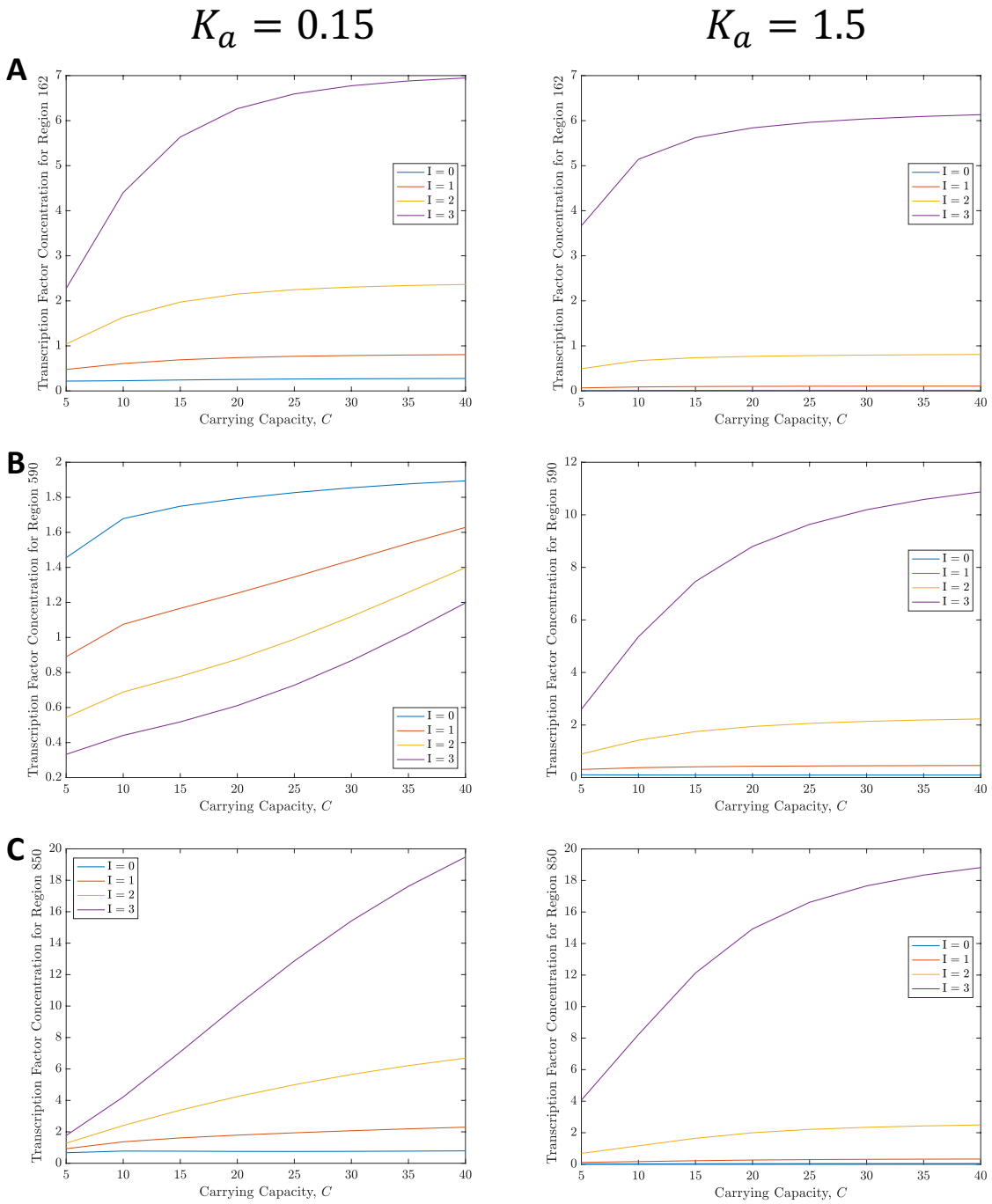

Figure 2.19 – **Carrying Capacity over TF occupancy for different regions and with different values of K_a, $K_a = 0.15$ and $K_a = 1.5$. A** Region 162. **B** Region 590. **C** Region 850.

However, the K_a value considered does not increase the cluster formation only in the highly active regions and the total nuclear TF concentration is low enough to force the TF occupancy everywhere. And it is clear that by increasing $[T]$, one also affects TF

occupancy.

In Fig. 2.17, we showed how different values of C affect the TF allocation. To verify how specific regions are affected by parameter changes, we propose Fig. 2.19, using the same regions from Fig. 2.15, but for two fixed values of K_a : (i) $K_a = 0.15$ and (ii) $K_a = 1.5$. We emphasize that the τ and d values for the regions are present in Table 1.

In Fig. 2.19 **A**, which has a well-connected region with a small residence time, we have the majority of the TFs in this region in the maximum PPI as possible ($I = 3$). The other states of PPI show higher concentration for the smaller value of K_a ($K_a = 0.15$), which is expected since K_a is responsible for the clustering in our model. For $K_a = 1.5$ the clustering in this region is high thus for $\beta \leq 2$, $\mathbf{p}^\beta \to 0$; however, when we compare both $\mathbf{p}^3$, the one with smaller association rate is bigger than the other, i.e., the clustering in the first condition is less effective everywhere, but this region benefits from it.

The effectiveness to form clusters is a behaviour we obtained in Fig. 2.19 **B**, in a low-connected region with small K_a values, we obtained that $\mathbf{p}^{\beta-1} > \mathbf{p}^\beta$, $\forall\beta$. The same result was shown in Fig. 2.15 **B**. Yet, with higher values of K_a the pattern has the same behaviour as Fig. 2.19 **A**.

Last, in Fig. 2.19 **C** we have a similar pattern from Fig. 2.19 **A**, but with more than twice the concentration for $\mathbf{p}^3$ in **C** and an almost linear pattern for $K_a = 0.15$ for the maximum PPI, i.e., $\mathbf{p}^3 \propto C$. For $K_a = 1.5$ in **C**, the concentration in the $\mathbf{p}^3$ state remains bigger since $C = 5$ TFs, but at $C = 40$ TFs, the concentration is slightly less than the one for $K_a = 0.15$.

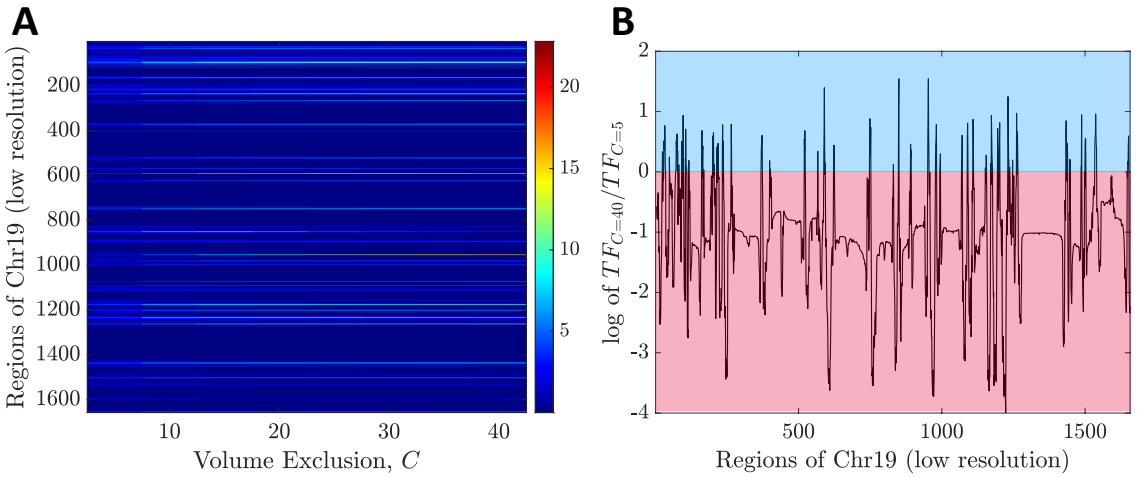

Figure 2.20 – **Carrying Capacity over TF occupancy for all the regions with $[T] = 1000$ TFs and $K_a = 1.0$ at $t = 400s$. A** Heatmap to analyze how the occupancy landscape changes by allowing more TFs into the regions, resulting in bigger clusters. **B** log of the occupancy from $C = 40$ TFs over the obtained in $C = 5$ TFs.

To understand the global effects of C on TF occupancy, we propose in Fig. 2.20

a similar analysis to the one in Fig. 2.16. Thus, we fixed the values for $[T]$ and K_a and analyzed how by increasing more TFs to occupy a region, we deplete other regions of TFs, i.e., by increasing C, the even TF spread showed in Fig. 2.19 is less likely.

In Fig. 2.20 **A**, we can see this increase in occupancy in specific regions as we allow more TFs to occupy the regions at the same time. Then, we compared simulations with the highest ($C = 40$ TFs) and the lowest ($C = 5$ TFs) values of C by dividing the occupancy profile at $t = 400s$ of $C = 40$ by the one from $C = 5$ and applying the natural logarithm, Fig. 2.20 **B**. Fig. 2.20 **B** again showed that few regions are favoured and, as a consequence, most of the regions decrease their TF numbers.

From the results in Fig. 2.20, one can conclude that the limited space inside active chromatin regions can be understood as a mechanism to control gene expression and force the TF to remain in the search process for a target site, which corroborates the fact gene activity is not increased given clustered TFs. In other words, one can understand the carrying capacity as a mechanism to control the TF allocation and, as a direct consequence of this allocation, polymerase (RNAP) recruitment to start transcription. Besides, since the structure and accessibility also regulate the occupancy, one can inquire about the consequences of increasing the total TF concentration in our numerical solutions, a result we present next.

2.2.2.3 Nuclear TF Concentration is an internal control for TF occupancy

By limiting transcriptional resources available, the cell can regulate transcription levels by itself, as we showed the TF preference for clustering around prolific regions. To analyze the impact of $[T]$ in our model, we implemented numerical solutions considering different nuclear concentrations, i.e., $[T] = \{10, 30, 50\} \bigcup \{100 : 100 : 500\} \bigcup \{700, 900\} \bigcup \{1000 : 50 : 1600\}$ for 400 seconds.

To demonstrate how the concentration influences our numerical solutions, we proposed Fig. 2.21 with fixed values of K_a and C and changing the TF nuclear concentration. In our model, a TF must occupy a region, any increase of TFs in the system increases the TF allocation in the regions, which is shown in Fig. 2.21. As expected, the concentration effect also affects TF clustering, and the lack of available TFs reduces the chances of two TFs to interact, Fig. 2.21 **A**.

By increasing the values of $[T]$, we also favour clustering, meaning higher concentrations of nuclear TFs improve the chances of two (or more) TFs clustering. Besides, the direct impact of a TF concentration in the nucleus on the gene expression is a well-known feature and we showed how regions with low TF concentrations in Figs. 2.21 **A** and **B** for example had a boost in higher $[T]$ environments - Figs. 2.21 **C** and **D** - proving the lack of resources can be used as a gene regulatory tool even if the fast increase in concentration may not be optimal for the cell (Koşar; Erbaş, 2022).

Thus, we verify the global effects of changing the TF concentration in Fig. 2.22 by presenting the total TF occupancy per region at $t = 400$ seconds in the function of the total TF, $[T]$ for fixed values of K_a and C. Fig. 2.22 shows how changes in the volume of TFs available influence the occupancy and how the efficiency of clustering, K_a, and the maximum TF allowed in a region, C, also play roles in the TF organization in the chromatin.

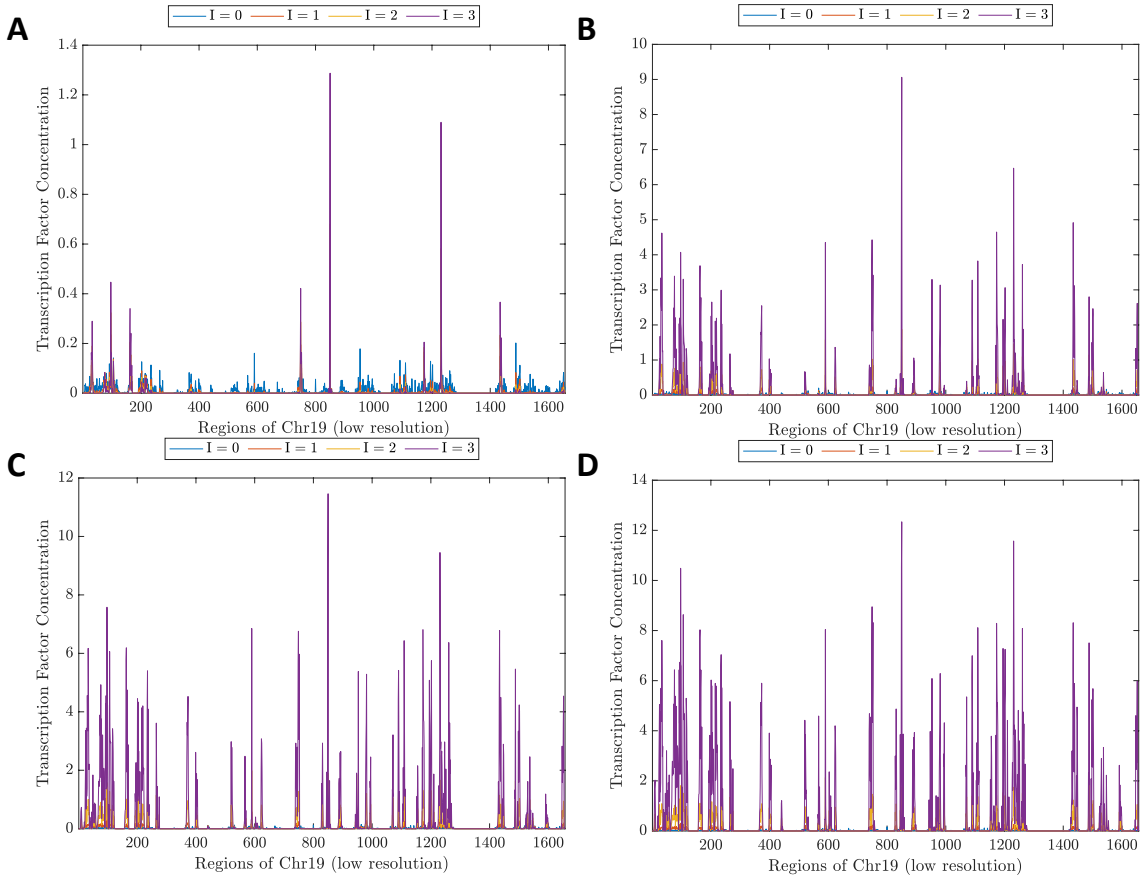

Figure 2.21 – **Numerical Solutions for our model in Eq. (2.6) for different values of $[T]$ and fixed $K_a = 1$ and $C = 15$.** In which, **A** $[T] = 50$; **B** $[T] = 500$; **C** $[T] = 1000$ and **D** $[T] = 1500$.

In Figs. 2.22 **A** and **B**, we assumed the same carrying capacity for different values of K_a and we demonstrated how the association rate affects the TF occupancy, stimulating the clustering in some regions and decreasing the TF concentration in other occupied regions. We previously showed how a smaller value of C increases the activity in less attractive regions which decreases for higher values of C as more TFs are allowed in each region - Figs. 2.22 **C** and **D** - and we verified this result for different $[T]$ values. Therefore, we proved how lowering the global TF availability decreases the chances of clustering and impacts the transcription.

In Fig. 2.22, we showed how each region is occupied once we increase the TFs available, showing how most of the regions present low TF concentrations as the clustering

around specific regions increases. To check the average patterns from the regions, we propose the cluster analysis in Fig. 2.23 for a low value of C to force a higher occupancy in less preferential regions, given two different K_a values, $K_a = 0.15$ in Fig. 2.23 **A** and $K_a = 1.5$ in Fig. 2.23 **B**. Notice the existence of an inactive cluster in both Fig. 2.23 **A** and **B**, which represents most of the regions (since in Figs. 2.22 **A** and **B**, we verified a consistent low to no TF occupancy in most of the regions).

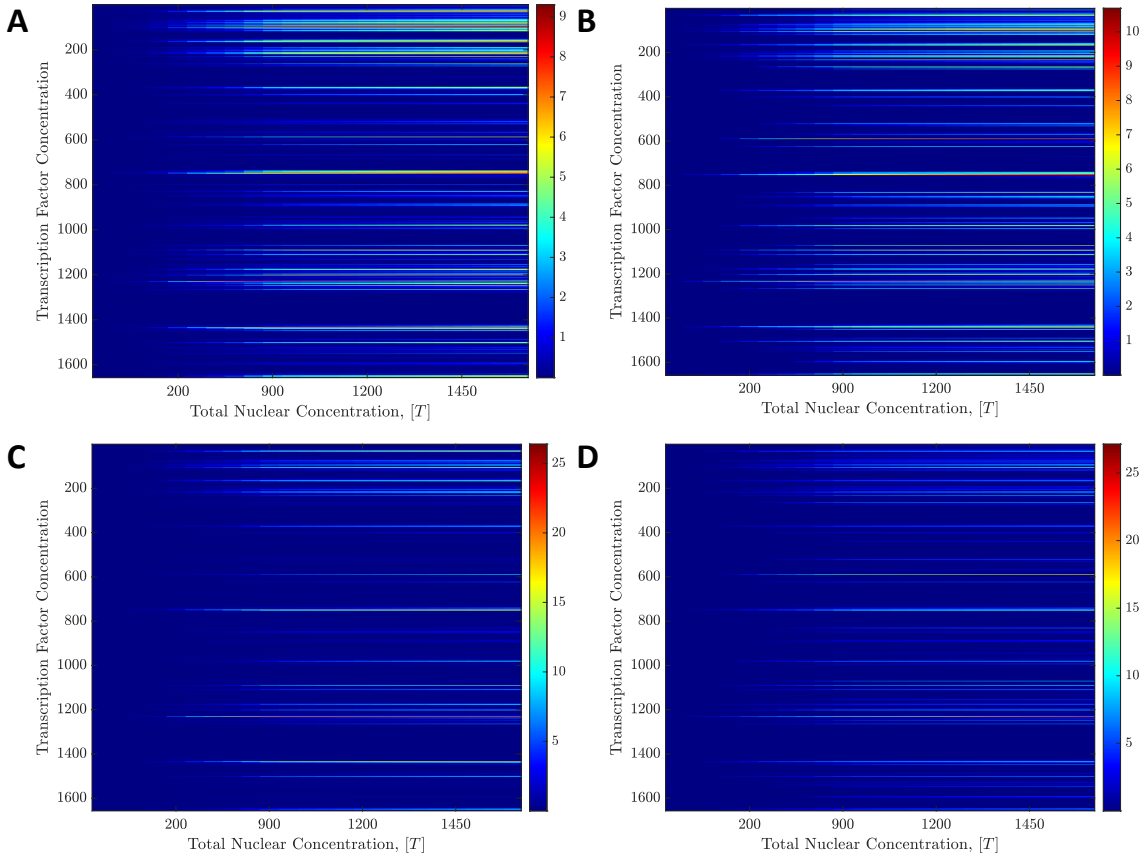

Figure 2.22 – **Heatmap of total TF concentration over the different values of** $[T]$ **and fixed values for** K_a **and** C**. A** $K_a = 0.15$ and $C = 10$; **B** $K_a = 1.5$ and $C = 10$; **C** $K_a = 0.15$ and $C = 30$; and **D** $K_a = 1.5$ and $C = 30$.

The other clusters show different occupancy behaviours depending on K_a. The highly concentrated clusters in Fig. 2.23 **A** show a sigmoidal behaviour and the less concentrated clusters present a slower increase in TFs as the concentration of TFs increases, i.e., while all the regions are favoured by the more TFs available in the sense of occupation, attractive regions have a preference for the *extra* TFs. Fig. 2.23 **B** has a more stringent increase for the higher occupied clusters, with some averages approaching the maximum number of TFs allowed, and the increase in occupancy is verified in different levels for all the clusters *except* the low TF concentration cluster, which decreases. Figs. 2.23 **A** and **B** showed how the increase K_a depletes the TFs from less prolific regions to increase the clustering around more attractive regions.

To understand how each value from Figs. 2.22 **A** and **B** deviates from the mean, we calculate their z-score and then clustered the results - Figs. 2.23 **C** and **D**. Fig. 2.23 **C** showed an increasing tendency for each cluster, i.e., for this K_a value all the regions benefit from the increase of $[T]$, even if not linearly. However, for Fig. 2.23 **D** we uncovered that for some regions the increase of $[T]$ is not beneficial for its TF allocation, demonstrating how the increase in K_a facilitates the TF clustering and, in turn, decreases the TF concentration in less attractive regions since we have a limited amount for $[T]$. More so, the regions with low TF concentration showed a linear behaviour for both z-scores.

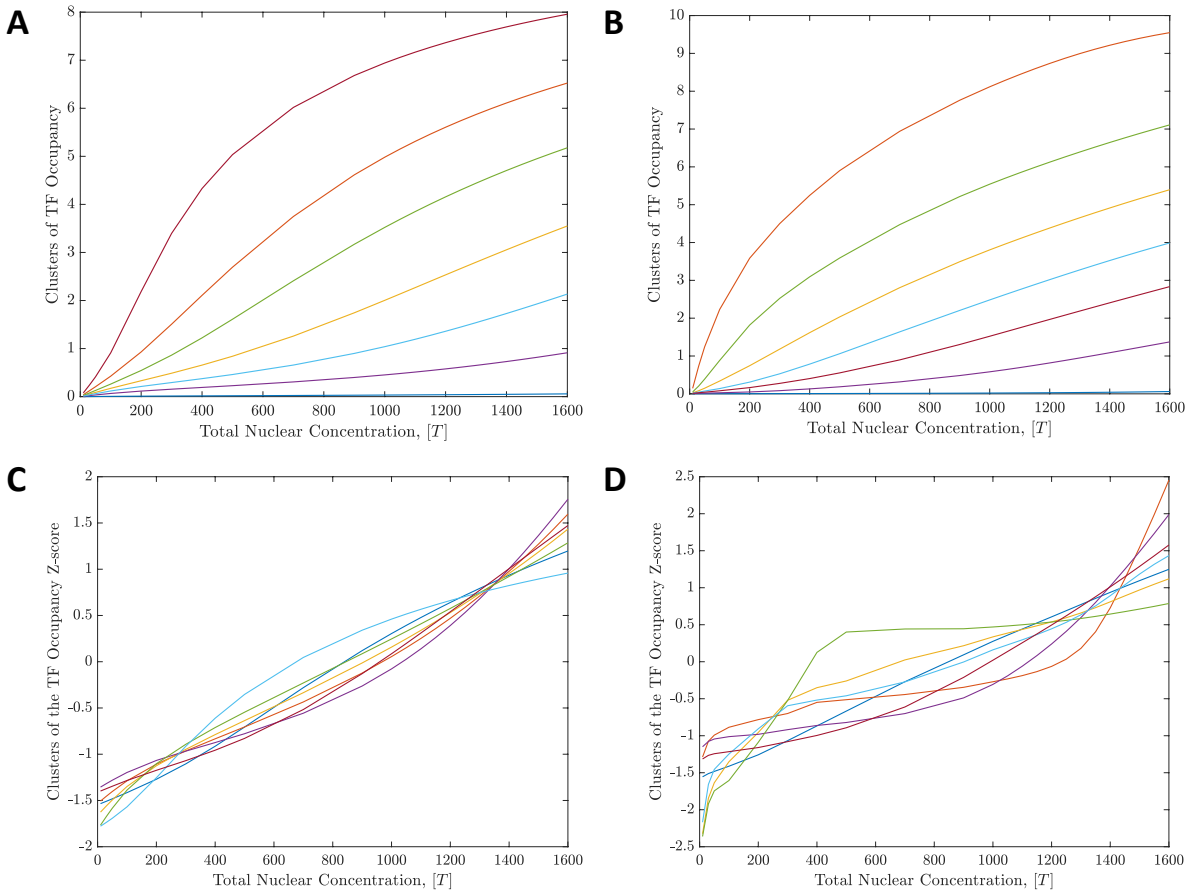

Figure 2.23 – **Cluster analysis of the TF occupancy over the nuclear TF concentration and fixed $C = 10$ and different values of K_a. A-B** the clustering of averaged TF occupancy and **C-D** the clustered averages of z-scores. With the following values for **A-C** $K_a = 0.15$ and **B-D** $K_a = 1.5$.

Similar to what we proposed in Figs. 2.15 and 2.19, we select specific regions to see how increasing $[T]$ affects the TF allocation in those regions for two different carrying capacities. Here, we omitted the difference in PPI levels to verify how the values of K_a compare per region. Those results are present in Fig. 2.24, with the values of τ and d available in Table 1.

Fig. 2.24 **A** shows the linear behaviour as $[T]$ increases for $K_a = 0$ - i.e., no presence of protein-protein interactions and the increase in TF allocation between the two subfigures

of **A** is due to the increase in C. The presence of K_a increases the total TF occupancy at this region but with different behaviours: for smaller values of $[T]$, $K_a = 0.15$ behaves as $K_a = 0$, which is explained by the low number of TFs available and low efficiency of our system to form clusters, which is not the case for $K_a = 1.5$. Interestingly, the TF concentration at this region is higher for the $K_a = 0.15$ once we have $[T] = 200$ TF molecules. This means low efficiency in form clusters is advantageous for this region, as it entraps more TFs inside.

Higher values of K_a increases the occupancy for Fig. 2.24 **B** and lower values of K_a present a TF occupancy lower than without the presence of association rates ($K_a = 0$), except for higher values of $[T]$ and allowing more TF molecules per region. Besides, for lower values of $[T]$ all values of K_a present the same concentration of TFs. Thus, this region requires higher efficiency in the formation of TF-TF interactions to benefit from the clustering, a different result from Fig. 2.24 **A**.

Last, Fig. 2.24 **C** showed both benefiting from both types of values for K_a: while for $C = 10$ $K_a = 1.5$ increased the TF concentration up to the maximum allowed for higher values of $[T]$, for $C = 30$, $K_a = 0.15$ presented higher TF allocation for the interval $200 \leq [T] \leq 1400$. Again, for smaller values of $[T]$, $K_a = 0.15$ presented the same occupancy as $K_a = 0$.

The results in Fig. 2.24 demonstrated how higher association rates increase the clustering around the more prolific regions, other regions are impaired by the lack of transcriptional resources available. Of course, by increasing the maximum of TFs inside a region, we also favour the higher propensity of clustering of those regions. More so, by changing the values of nuclear TFs available in the system, we corroborated the $[T]$ effects on the TF allocation, proving how this value can be also understood as a mechanism to control gene expression and how $[T]$ impacts to cluster formation.

In this chapter, we understood how the presence of volume exclusion leads to different TF occupancies and how cluster formation impacts gene expression. We found that while C is a structural parameter that directly affects the $\mathbf{p}^0$ its presence force the TF to allocate in less prolific regions. More than that, in cases with small C, the association rate increases concentration around those prolific regions.

The number of TFs available also influences clustering, as lower TF concentrations block PPIs from occurring and higher TF concentrations allow less attractive regions to be occupied by TFs; both affect transcription in the long run. Thus, the cell can regulate its expression by regulating the number of TFs for transcription.

Therefore, we proved how the cluster of TFs can be used in transcription as a mechanism for gene regulation if combined with limiting the TF occupancy in the regions to avoid over-clustering in just a few regions. In Chapter 6, we present a more in-depth

discussion of our results.

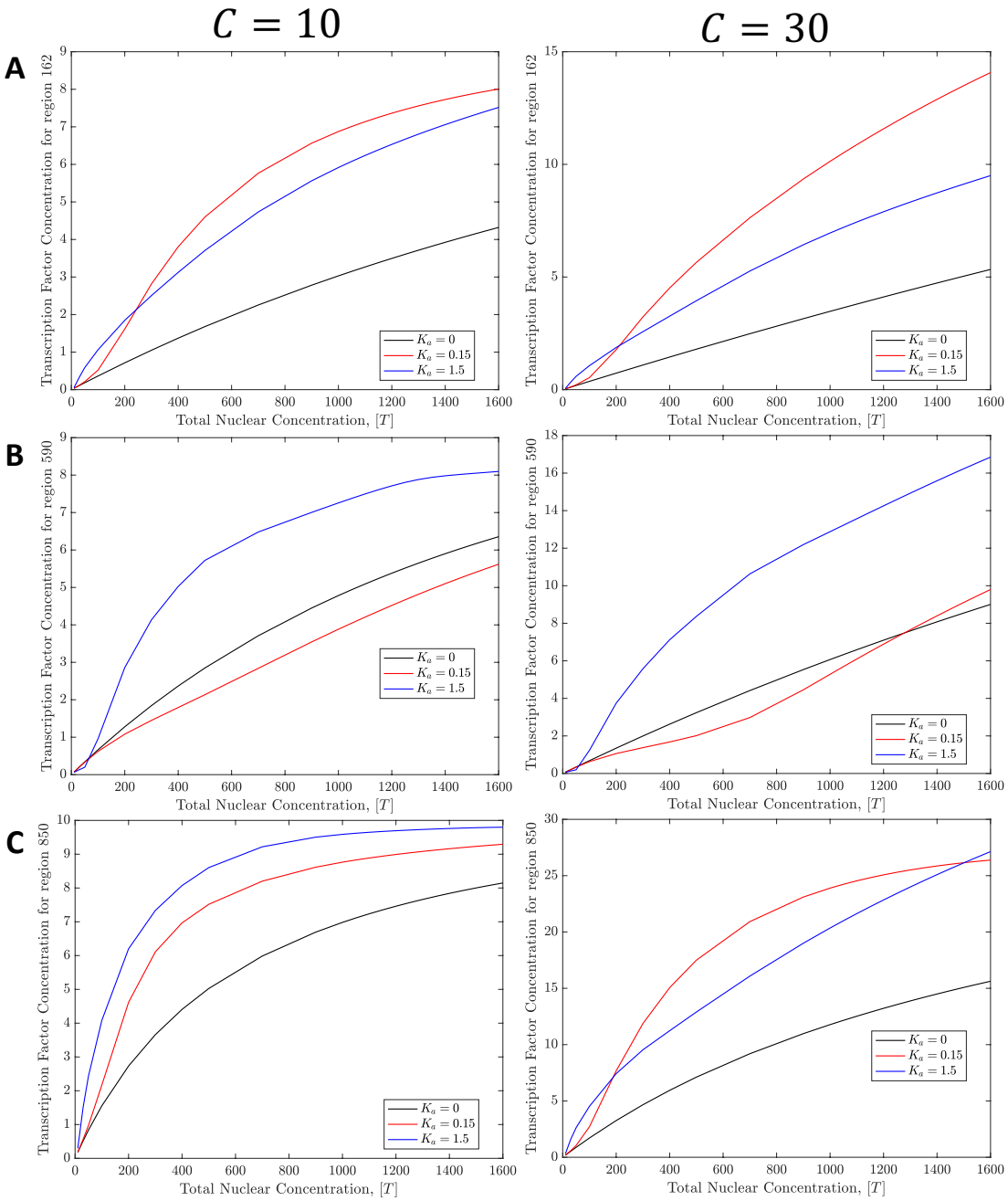

Figure 2.24 – **Total TF concentration over TF occupancy for different regions and with different values of C, $C = 10$ and $C = 30$. A** Region 162. **B** Region 590. **C** Region 850.

From both models - Eqs. (2.3) and (2.6) - we understood how the TF searches and occupies chromatin, which is the first step in transcription. The next biological step is to consider the polymerase recruitment and actual transcription. Thus, we incorporated those two steps in Chapter 3 and Chapter 4, respectively to our model of TF search. We aim to understand gene regulation, in which we opted to adapt the search dynamics and limited the TF occupancy by lowering the nuclear TF concentration.

3 Modelling Transcription Factor searches and RNA Polymerase recruitment

Disclaimer: Some parts of this chapter can be found in my paper *"Modelling Transcription Factors Search and Polymerase Recruitment Dynamics within a complex chromatin structure"*

One of the main objectives of this thesis is to present a mathematical model that explains some of the intricacies of gene expression using simple techniques that still make biological sense. To achieve that, we split the two fundamental steps of protein production (i.e., Transcription - where the cells copy the DNA into RNA and Translation - the cells use the RNA as a template for protein production) as we described in Chapter 1. In this chapter, we present a model for the transcription mechanism in which one key point for controlling gene expression is the presence of RNA polymerase II (RNA Pol-II) since it is a fundamental protein complex in messenger RNA (mRNA) production (Alberts et al., 2002; Coulon et al., 2013; Hager; McNally; Misteli, 2009).

We can think of the eukaryotic transcription process as the following five steps: (i) preinitiation complex (PIC) formation, in which the PIC is a complex of Pol-II and TFs, Fig. 1.2; (ii) PIC activation; (iii) transcription initiation; (iv) Promoter liberation; (v) elongation (Roeder, 1996; Nikolov; Burley, 1997; Roeder, 2019; Petrenko; Struhl, 2021). With this complex mechanism in mind, we incorporated our TF search model into the PIC formation process, since the accessibility of a region is essential for transcription (Avcu; Molina, 2016) and Chapter 2.

The effectiveness of a TF in finding a target site is a turning point for understanding gene expression and facilitating gene regulation, the PIC formation is an interesting mechanism to model per se (Wunderlich; Mirny, 2009; Bruneau, 2010). However, since this model proposes a mechanistic way to understand transcription, we also incorporated steps (ii), (iii) and (iv). We ignored step (iv) because we decided to not explicitly work with enhancer-promoter interactions for this model as they are intrinsic to our network structure (Schoenfelder; Fraser, 2019).

Given what we already discussed in Chapters 1 and 2, *in vivo* single-molecule tracking (SMT) experiments proved diffusion is not a good approximation to describe the TF search process (Kuhn et al., 2021; Xiao; Hafner; Boettiger, 2020) and the most common method to model this process is by assuming the facilitated diffusion mechanism (also called 1D/3D process), i.e., a TF randomly searches the chromatin structure (3D diffusion) but also slides towards the region's nearest neighbours (1D) plenty of examples of this method

being used for TF search mechanisms models (Mirny et al., 2009; Avcu; Molina, 2016; Bauer; Metzler, 2012; Bauer et al., 2015; Hettich; Gebhardt, 2018). However, no other model considered RNAP recruitment as an integral part of TF search mechanism/transcription and how RNAP fine-tunes gene expression (Dergai; Hernandez, 2019).

In layman's terms, once a searching TF (i.e., a TF looking for an accessible DNA sequence, an *on* gene) finds such a target site, this TF attracts the RNA Polymerase (RNAP), forming the preinitiation complex, and eventually leading to transcription. In Fig. 3.1, we present a simplified representation for TF search (represented by the cherry) in chromatin (in which each node is a chromatin region and each edge represents the connections between two regions similar to the presented in Fig. 2.1), trying to find an active region (here, represented by the light blue node) and an RNAP (the green PAC-man type) explores the network searching the active TF. We emphasize that even if a protein such as TF or RNAPs is considered inside a closed region, we meant they are nearby not effectively bound.

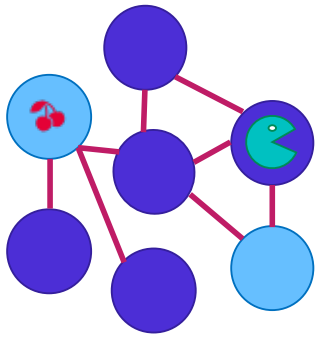

Figure 3.1 – **Cartoon representation for RNAP recruitment**. Here, we present the chromatin structure in two different states: (i) Open or "active"(light blue) and (ii) Closed or "inactive"(dark blue) the interactions between chromatin regions are represented by the edges. Considering the PIC formation, we present the TF as the red cherry in an open region and recruit an available RNAP (the green PAC-man type).

Since both the TF and the RNAP explore the chromatin, both proteins need a structural component in their searching/recruiting processes; thus, we used the same mechanism described in Chapter 2, coupling TF/RNAP dynamics in only one model. Once again, we integrated high-resolution information of the 3D structure of chromatin with DNA-protein interactions from Hi-C data (Lieberman-Aiden et al., 2009; Berkum et al., 2010).

Similar to any other biological process, the volume of available resources impacts the final volume of the product. Thus, limiting transcriptional resources surely affects transcriptional activity. More so, while some transcription factors remain inside the nucleus at all times, for example, the basal transcription factors or the Sp1 transcription factor which is involved in cell differentiation and growth, and, in this case, we can assume the

total concentration of this TF over time remains constant; other transcription factors need activation to function in two different ways: (i) developmental (e.g., GATA) or (ii) signal-dependent, which requires an external signal to get activated (e.g., p53) (Brivanlou; Darnell, 2002).

The signal-dependent transcription factors have a subgroup of TFs that remain outside the nucleus and, upon activation, they start a translocation process and the transcription of their target genes to later re-accumulate in the cytoplasm. One TF family that presents translocation behaviour is the nuclear factor κb (NF-κb). NF-κb is a rapid-acting TF and was originally identified in the immunoglobulin regulation in κ-light chain expression in B lymphocytes but now is a known TF for inflammatory and immunity mechanisms, being also a key TF for human cancer studies.

NF-κb can induce and maintain a chronic inflammation leading to tumour initiation by stimulating cell proliferation and preventing apoptosis, for example (Xia; Shen; Verma, 2014; Liu et al., 2017). After NF-κb receives any activation stimuli (cytokines, DNA damage, UV radiation, etc), it translocates into the nucleus, starts transcription and re-accumulates in the cytoplasm in around 60 minutes. Since NF-κb can be activated by proinflammatory cytokines (i.e, small proteins used in cell communication used for upregulation of inflammatory processes), it is an excellent example of extracellular stimuli affecting transcription activation (Trask, 2012; Zhang; An, 2007; Noursadeghi et al., 2008; Zambrano et al., 2020; Xia; Shen; Verma, 2014; Liu et al., 2017).

In this chapter, we proposed a model considering two TF flux dynamics: one with the TF import and the other with a translocation/re-accumulation process or an import/export mechanism. We solved our model using deterministic and stochastic techniques.

3.1 Mathematical model for TF diffusion and RNAP recruitment

Literature has proposed a myriad of gene regulation mechanisms for eukaryotes. For example, the existence of the nuclear membrane separating cytoplasm and nucleus also separates transcription from translation, which means a huge number of mRNAs are degraded before translation, creating a fail-safe protocol for the cell.

One extremely important mechanism is chromatin, which condenses DNA and inhibits transcription unless the cells demand it and the remodelling process occurs. Besides that, the interactions between chromatin regions create a complex 3D structure that we can reconstruct from Hi-C experiments, as presented in the previous chapter (Pal; Forcato; Ferrari, 2018; Lieberman-Aiden et al., 2009; Johnstone et al., 2020; Berkum et al., 2010).

Given the model proposed on (Avcu; Molina, 2016) and Chapter 2, we present a model on how the TF explores the chromatin structure and recruits the RNAP to

initiate transcription. Again, we defined the size of our network, which correlates with the resolution of our model as L, meaning our network has L different chromatin regions. We incorporated the structure by using the adjacency matrix $\mathbf{A}$ and that for any two regions i, j they either have a connection or not, i.e., let A_{ij} be the expression to evaluate the link between i and j is $a_{ij} \neq 0$ if there is a connection between i and j and 0 otherwise.

It should be noted $\mathbf{A}$ is a symmetric matrix (i.e., $a_{ij} = a_{ji}$) since we consider that the connections between regions are non-oriented. Since the chromatin moves through the nucleus, the connections between non-neighbouring regions (regions with linear distances between them bigger than 1) can change, but we fixed our network, which can be accepted in small time scales (smaller than the cell cycle).

Each region has a different number of connections with the other regions as not all regions are connected. Considering the contact degree of a node, the probability of a protein jumping from region j to region i depends on their connection and the number of associations of the region j, d_j, i.e., a highly connected region has higher chances of being reached, as exemplified in Fig. 3.2 (Avcu; Molina, 2016) and Chapter 2,. Once more, we express this probability as:

$$M_{i \leftarrow j} = \frac{A_{ij}}{d_j} \ .$$

We propose the TF search mechanism for this model as a Free TF (T^f) explores the chromatin network with an effective diffusion rate, k_{3D}^T independent on the structure, i.e., it is the same everywhere in our network. Upon reaching a target site, the binding process occurs and the TF leaves this free state to enter the bound state with a binding rate k_b^T. Any TF has a specific genomic sequence to bind, also called motif and the sequence of nucleotides affects the likelihood of a TF binding a region. Even if a sequence is a good motif for a TF, if there are no active histone marks in that particular region transcription can not start (Guertin; Lis, 2013; Aptekmann et al., 2022).

Therefore, one may conclude some target sites are more prolific than others - either by the region's accessibility or the sequence motifs - implying some regions have lasting effects on TFs. Given how the promiscuity of a region affects the time a TF remains at that particular region, we can use single-molecule microscopy techniques to obtain that TF binding times (Izeddin et al., 2014). Since each TF has its preferred motif, the binding times are also motif dependent and, by consequence, structure-dependent. Similarly to the proposed in Chapter 2, we used this motif-sensibility as the rate of a bound TF freeing itself, i.e., an exiting rate, k_i, where we also incorporated the chromatin region and accessibility (i.e., $k = (k_i)_{i=1}^L$ for a $L-$sized network).

To model the recruitment of RNAP binding to a region occupied by a bound TF and transcribing, we should consider an intermediate state between free and transcribing since transcription initiation is not immediate, i.e., in our model, we consider three

different states for the RNAP: (i) Free RNAP (P^f); (ii) Bound RNAP (P^b) and (iii) Transcribing RNAP (P^t). To start transcription, the RNAP needs to be recruited by a bound TF (Brouwer; Lenstra, 2019); our RNAP dynamics is the following: a free RNAP P^f diffuses with an effective diffusion rate k_{3D}^P looking for a bound TF (T^b) to bind. Once this free RNAP finds a Bound TF, it binds the region with a binding rate k_b^P; P^f then becomes a bound RNAP P^b. Then, this P^b initiates transcription at an initiation rate k_I entering the transcribing state P^t. Note the transition between Bound and Transcribing RNAP is independent of T^b presence because the bound TF role is to recruit the RNAP to a region, being allowed to unbind itself after the RNAP binds it. While transcribing, the RNAP reads the strand of DNA in that region to produce mRNA. After the production is finished, the transcribing RNAP re-enters the free state with an elongation rate, k_ε. The final product, mRNA is not considered in our model in this Chapter since our interest is in where transcription is more likely to occur but in the next chapter, we present an extension of our model considering the presence of mRNA export.

After a gene gets activated, the TF needs to find it and then bind this target site. Later, the bound TF shall recruit the RNAP, so we propose in Fig. 3.2 the interaction between regions i and j within a representation of the 3D structure in a non-oriented graph in which each node is a chromatin region and each edge is a connection between two regions, and in Fig. 3.2 inset, we present the interactions of the state for region i, in which we can see the three steps for eukaryotic transcription: the TF search and RNAP recruitment are the first step, PIC formation; the RNAP actively binding to a region is the PIC activation and once it enters the transcribing state we have the transcription initiation step and finally once the RNAP is released, we have the final transcription step, elongation.

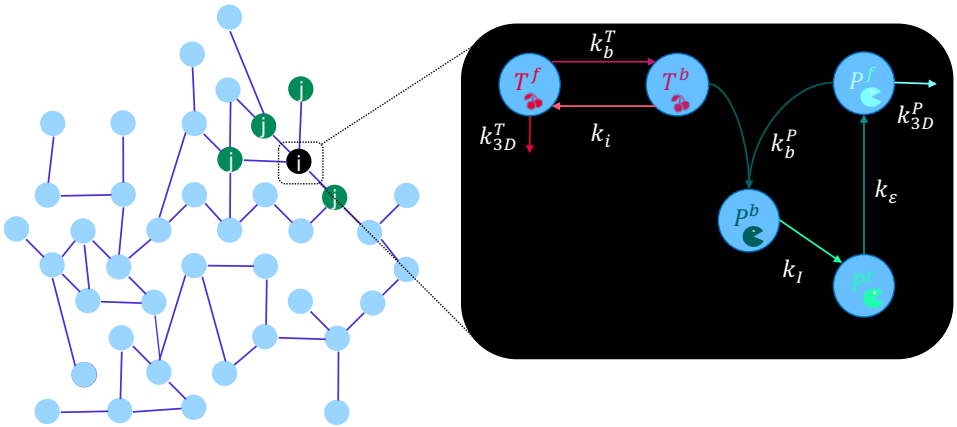

Figure 3.2 – **Our model TF/RNAP schematics.** The graph represents our network where the nodes are the regions and the edges are the connections between them. The node in black represents the region i and all nodes in green represent potential regions j from where i can be reached. In the black box (inset) we can see all the reactions between the TF (cherries) and RNAP (PAC-men) states and how each state interacts with the other.

From Fig. 3.2 inset and our dynamics description, we present our mathematical model in Eq. (3.1), which is an ordinary differential equation (ODE) system for the concentration of TF/RNAP states over time in all L chromatin regions. Here, each state has L equations, which means that instead of working with equations dependent on time and space, we used only ordinary differential equations to represent our dynamics, following our aim to model a complex process as transcription as a simple mathematical model.

We decided to not change the network over time - i.e., the time scale is short enough for us to not enter in the cell-cycle chromatin changes, and we do not consider the presence of chromatin remodellation. We also defined our system as a closed one - i.e., our system does not produce or degrade TFs and RNAPs. Another feature is we do not consider loss between states, i.e., if we sum over all states, we return the description of the TF and RNAP movements through our network.

$$\begin{cases} \dfrac{dT_i^f}{dt} = -k_{3D}^T T_i^f + \sum\limits_j k_{3D}^T M_{i\leftarrow j} T_j^f - k_b^T T_i^f + k_i T_i^b \ ; \\ \dfrac{dT_i^b}{dt} = k_b^T T_i^f - k_i T_i^b \ ; \\ \dfrac{dP_i^f}{dt} = -k_{3D}^P P_i^f + \sum\limits_j k_{3D}^P M_{i\leftarrow j} P_j^f - k_b^P T_i^b P_i^f + k_\varepsilon P_i^t \ ; \\ \dfrac{dP_i^b}{dt} = k_b^P T_i^b P_i^f - k_I P_i^b \ ; \\ \dfrac{dP_i^t}{dt} = k_I P_i^b - k_\varepsilon P_i^t \ . \end{cases} \tag{3.1}$$

From experimental results, we know that the TF diffuses faster than the RNAP (Gorski; Dundr; Misteli, 2006; Klumpp, 2013), which considering TFs are smaller proteins than the RNAPs and have smaller weights this difference in speed is a consequence of classical mechanics (Maeshima et al., 2015). From this, we considered different search rates (effective diffusion rates) and also that the binding rates are affected by their size differences (Sun et al., 2020).

Our model in Eq. (3.1) has had too many parameters; so, for the TFs we assumed two things: (i) the search and binding rates for the TF are proportional, and, (ii) 70% of the TFs remain in the free state, since not all regions have a good motif and/or are accessible. With those two assumptions, we define k_b^T as a function of k_{3D}^T. However, since the RNAP's binding mechanism is dependent on the presence of a bound TF in the region to occur, it is non-linear and we defined this parameter as proportional to the product of its search rate and the effectiveness of finding a bound TF. With this in mind, we present in Eq. (3.2) expressions for the binding rates.

$$k_b^T = \alpha k_{3D}^T \ ; \ k_b^P = k_{3D}^P \left(\frac{k_b^P}{k_{3D}^P} \right) = k_{3D}^P q \ . \tag{3.2}$$

From the assumptions in Eq. (3.2) and the model in Eq. (3.1), we present the simplified model, which is still a dimensional model (i.e., we retain our dimensions in each

equation and represent the concentration in molecules/s for all the regions) with all the dimensions maintained. We present our working model in Eq. (3.3) which can be used to explain the occupancy patterns that can emerge considering the structure and accessibility of chromatin regions, as literature has described (Li; Carey; Workman, 2007; Woringer; Darzacq; Izeddin, 2014), with constant concentrations of TFs and RNAPs.

$$\begin{cases} \dfrac{dT_i^f}{dt} = -k_{3D}^T T_i^f (1+\alpha) + \sum\limits_j k_{3D}^T M_{i\leftarrow j} T_j^f + k_i T_i^b \ ; \\ \dfrac{dT_i^b}{dt} = \alpha k_{3D}^T T_i^f - k_i T_i^b \ ; \\ \dfrac{dP_i^f}{dt} = -k_{3D}^P P_i^f (1+qT_i^b) + \sum\limits_j k_{3D}^P M_{i\leftarrow j} P_j^f + k_\varepsilon P_i^t \ ; \\ \dfrac{dP_i^b}{dt} = k_{3D}^P q T_i^b P_i^f - k_I P_i^b \ ; \\ \dfrac{dP_i^t}{dt} = k_I P_i^b - k_\varepsilon P_i^t \ . \end{cases} \tag{3.3}$$

However, as we discussed previously, our interest lies in understanding how depleting a transcriptional resource as important as the available TF affects cell activity. Thus, we defined a translocation function for the transcription factor to represent this activation process, which we also numerically implemented, but first, we present our studies for the equilibrium of Eq. (3.3).

3.2 TF/RNAP occupancies are affected by chromatin structure and residence times

The first study we did with our model an ODE system is to analyze its equilibrium and stability, as we can use these results to understand global behaviours from our model, helping us to predict and interpret the patterns emerging from our system and how the network and parameters can affect transcription (Strogatz, 2015). Since our system is a closed one, we define the fixed concentration for all time t as the sum over all TF and RNAP states respectively, i.e., $[T] = \sum T_i^f + T_i^b$ and $[P] = \sum P_i^f + P_i^b + P_i^t$.

Then, we define the search time as the time a protein spends looking for its target: for the TF, this means looking for a target site (thus, $\tau_{3D}^T = (k_{3D}^T)^{-1}$) and for the RNAP, it means looking for a bound TF (similarly, $\tau_{3D}^P = (k_{3D}^P)^{-1}$). Next, we can define the time a TF stays at a specific region is called the residence time, $\tau_i = k_i^{-1}$. Since k_i is sequence-dependent, this means the time spent in a particular node of our network is also region-specific (Zabet; Adryan, 2013). As transcription is not immediate, we need to consider a waiting time for the transcription to begin after it's bound to a target site, i.e., the initiation time, $\tau_I = k_I^{-1}$ (Butler; Kadonaga, 2002; Mao et al., 1992). Finally, the elongation time is the time measure for mRNA synthesis and RNAP liberation, $\tau_\varepsilon = k_\varepsilon^{-1}$ (Tang et al., 2009; Wade; Struhl, 2008; Pokholok; Hannett; Young, 2002).

We present the steady-state expressions for all TF and RNAP states in Eq. (3.4). From our equations, we can see the dependency on the number of connections a node has, meaning any occupancy pattern will closely obey the network's connectivity pattern. The values of d_i and τ_i are the same present in Fig. 2.12 **A** and **B**, respectively to facilitate the numerical integration. As discussed earlier, τ_i is a region-specific parameter and affects the steady-state occupancy for non-free states (namely, Bound TF and RNAP and Transcribing RNAP). This means the accessibility of a chromatin region, a mechanism eukaryotic cells possess to protect the cell from misproducing a protein, affects the occupancy for protein complexes, which later affects the transcriptional activity.

Thus, the structure is an important component for gene expression in two different levels: (i) in the nucleotide level, which is derived from the motifs and (ii) in the chromatin level, i.e., the number of connections. In addition, from the non-linearity to enter the Bound RNAP state, the dependency of d_i is squared and represents a higher impact from the network for this state and the following state, Transcribing RNAP.

$$\begin{aligned} T_i^f = [T]\frac{\tau_{3D}^T d_i}{N^T} \;;\; T_i^b = [T]\frac{\alpha\tau_i d_i}{N^T} \;;\; P_i^f = [P]\frac{\tau_{3D}^P d_i N^T}{N^P} \;; \\ P_i^b = [P]\frac{([T]q\alpha\tau_i\tau_I)\, d_i^2}{N^P} \;;\; P_i^t = [P]\frac{([T]q\alpha\tau_\varepsilon\tau_i)\, d_i^2}{N^P} \end{aligned} \tag{3.4}$$

with the following normalization expressions: (i) for the TF, $N^T = \sum_k \left(\tau_{3D}^T + \alpha\tau_k\right) d_k$; and (ii) $N^P = \sum_k \left(\tau_{3D}^P + \frac{q\alpha[T]\tau_k d_k}{N^T}\left(\tau_I + \tau_\varepsilon\right)\right) d_k$, for the RNAP.

It is clear how the region-specific parameters affect the states' occupancies. Proving our linearity claim for the Free states (TFs and RNAPs), we present Figs. 3.3 **A** and **B**. We show the linear dependence of the $\log_2$ of the free states have on the $\log_2$ of the number of connections, $\log_2(d)$, labelling them with the different values of τ, proving their independence from the residence times - i.e., the time a TF(or RNAP) spends bound in a region does not influence the occupancy for the free state. Between Figs. 3.3 **A** and **B**, there is a difference of amplitude for the values which are caused by the fact $\tau_{3D}^T \neq \tau_{3D}^P$ and their different concentrations, $[T] < [P]$.

From Eqs. (3.4) and Fig. 3.2, we know the Bound states are both dependent on τ_i and d_i, we see the non-linearity in the form of the product $\tau_i d_i$ which proposes occupancy is a consequence of a combination of (a) motif/accessibility and (b) connections between regions. This relation for the bound states is found in Figs. 3.3 **C**-**D**. In Fig. 3.3 **C**, we confirm the TF's tendency to bind prolific regions, by plotting the $\log_2$ of the steady-states over $\log_2(\tau)$, labelled by the number of connections, in which we show the occupancy is affected by both parameters - this means two regions with the similar residence times are occupied depending on their connectivities. In Fig. 3.3 **D**, we present the $\log_2$ of the Bound RNAP steady-state over $\log_2(\tau)$ again labelled with the number of connections, proving the squared effect for d. Still, we need to verify the stability of this occupancy and, to achieve that, which we present next.

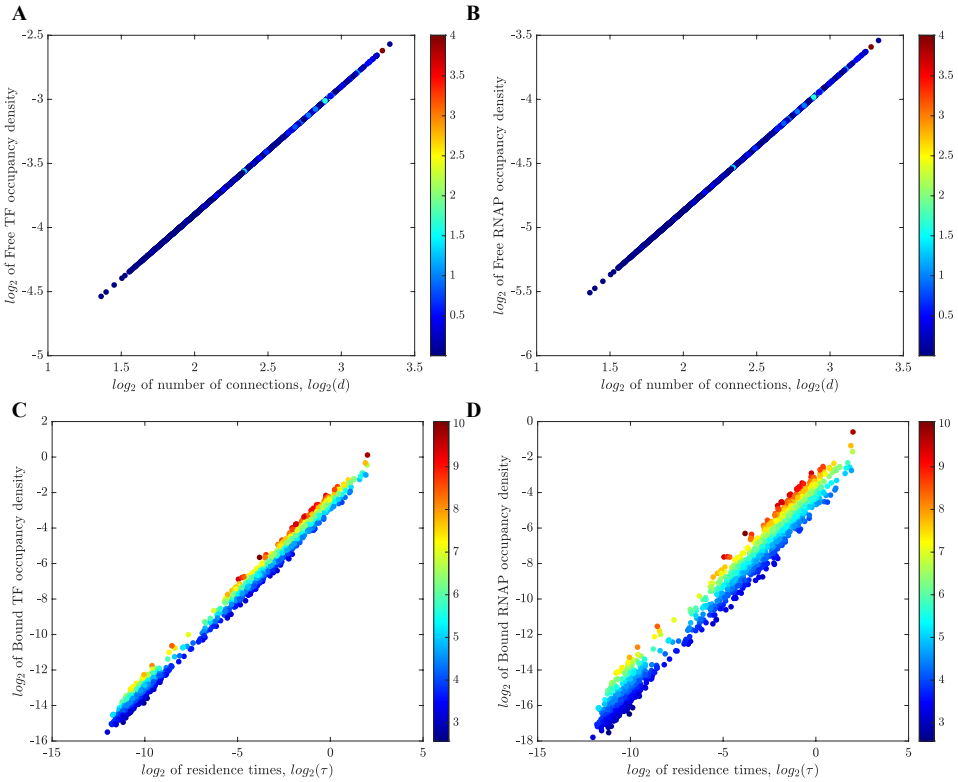

Figure 3.3 – **Our model characteristics and equilibrium. A-B** $\log_2$ of the steady-states values for Free states over the $\log_2$ number of connections labelled with the τ values, verifying its linearity, where: **A** Free TFs and **B** Free RNAPs. **C-D** Steady-states values for Bound states over the residence times labelled with d_i, proving both d_i and τ_i affect the occupancy, and the squared effect d_i have on Bound RNAP, as presented in (3.4). **C** Bound TF. **D** Bound RNAP.

3.2.1 Stability studies and the dependency on the Bound TF state

From Eq. (3.3), we calculated our model Jacobian Matrix which $t_i^* = T_i^{*'}$ and $p_i^* = P_i^{*'}$ with $*$ representing the different states (Strogatz, 2015; Edelstein-Keshet, 2005).

$$J = \begin{bmatrix} A & B & 0 & 0 & 0 \\ C & D & 0 & 0 & 0 \\ 0 & E & F & 0 & G \\ 0 & H & I & J & 0 \\ 0 & 0 & 0 & K & L \end{bmatrix} ,$$

with the following non-zero terms and δ_{ik} being the Dirac delta between i and k (i.e., $\delta_{ik} = 1$ iff $i = k$):

$$A = -k_{3D}^T \delta_{ik} + \sum_j M_{k \leftarrow j} k_{3D}^T \delta_{kj} - \alpha k_{3D}^T \delta_{ik} \ ; \qquad C = \alpha k_{3D}^T \delta_{ik} \ ;$$

$$B = k_i \delta_{ik} \ ; \qquad D = -k_i \delta_{ik} \ ;$$

$$E = -k_{3D}^P q P_i^f \delta_{ik} \ ;$$

$$F = -k_{3D}^{P}(1 + qT_i^b)\delta_{ik} + \sum_j M_{k\leftarrow j} k_{3D}^{P}\delta_{kj} \; ; \qquad J = -k_I\delta_{ik} \; ;$$
$$G = k_\varepsilon\delta_{ik} \; ; \qquad K = k_I\delta_{ik} \; ;$$
$$H = k_{3D}^{P} q P_i^f \delta_{ik} \; ;$$
$$I = k_{3D}^{P} q T_i^b \delta_{ik} \; ; \qquad L = -k_\varepsilon\delta_{ik} \; .$$

From this, we calculate the characteristic polynomial by $p(\lambda) = \det(J - \lambda Id)$, which Id is the $5L \times 5L$ identity matrix (as our network has L regions).

$$p(\lambda) = \det(J - \lambda Id) =$$
$$\left(-k_{3D}^{T}\delta_{ik} + \sum_j M_{k\leftarrow j} k_{3D}^{T}\delta_{kj} - \alpha k_{3D}^{T}\delta_{ik} - \lambda\right)\begin{vmatrix} D-\lambda & 0 & 0 & 0 \\ E & F-\lambda & 0 & G \\ H & I & J-\lambda & 0 \\ 0 & 0 & K & L-\lambda \end{vmatrix} -$$
$$- k_{3D}^{T}\alpha\delta_{ik}\begin{vmatrix} B & 0 & 0 & 0 \\ E & F-\lambda & 0 & G \\ H & I & J-\lambda & 0 \\ 0 & 0 & K & L-\lambda \end{vmatrix} .$$

Calculating the determinants from this system, we obtained the following expression for the characteristic polynomial:

$$p(\lambda) = \left(-k_{3D}^{T}\delta_{ik} + \sum_j M_{k\leftarrow j} k_{3D}^{T}\delta_{jk} - k_{3D}^{T}\alpha\delta_{ik} - \lambda\right)\left[(-k_i\delta_{ik} - \lambda)\left(-k_{3D}^{P}(1 + q\mathbf{T}_i^b)\delta_{ik} + \sum_j M_{k\leftarrow j} k_{3d}^{P}\delta_{jk} - \lambda\right)(k_I\delta_{ik} - \lambda)(-k_\varepsilon\delta_{ik} - \lambda)\right] -$$
$$-\alpha k_{3D}^{T}\delta_{ik}\left[k_i\delta_{ik}\left(-k_{3D}^{T}(1 + q\mathbf{T}_i^b)\delta_{ik} + \sum_j M_{k\leftarrow j} k_{3D}^{P}\delta_{ik} - \lambda\right)(-k_I\delta_{ik} - \lambda)(k_\varepsilon\delta_{ik} - \lambda) + k_\varepsilon\delta_{ik} k_{3D}^{P} q\mathbf{T}_i^b\delta_{ik} k_I\delta_{ik}\right] .$$

We can use our parameters to evaluate the stability conditions given the roots of this polynomial, but since it is a high-order polynomial in a complex network, we opted for omitting the roots. However, even without presenting the roots for this polynomial, we verified how the values of Bound TFs are explicitly found in the polynomial, which is a surprising side-effect of the non-linearity due to this state.

More so, we conclude the stability relies strongly on the network, and the T^b occupation pattern present in Eq. (3.4), which is transcription factor specific thus forcing any occupancy in the equilibrium to depend on the active DNA sequences. Besides, the exiting rate, k_i, is also TF-specific and, as Eq. (3.4) showed, the bound TF occupancy is

determined by its inverse, which consequently affects RNAP recruitment. Thus far, the chromatin network may change over time, but we opted to work with a fixed structure for this model and we remind our system is a closed one with no TFs or RNAPs created or degraded, because of the considered time scale that allows us to admit a fixed chromatin network.

Next, since this model does not consider the TF translocation upon activation (Pugh; Tjian, 1990; Noursadeghi et al., 2008; Allen et al., 2000; McBride, 2002) and how the change between the nuclear concentration of TFs affects the transcriptional activity. We propose two different translocation dynamics: **(1) Import Flux:** which the cytoplasmic TFs translocate into the nucleus and remain inside and **(2) Import/Export Flux:** the cytoplasmic TFs get activated, translocate into the nucleus and later translocate back to the cytoplasm.

3.2.2 Transcription activity determined by the nuclear TF concentration

The translocation process depends on the nuclear pore complex (NPC), which is responsible for transportation from the cytoplasm to the nucleus and vice-versa (Strambio-De-Castillia; Niepel; Rout, 2010; Peters, 2005). We proposed two different flux dynamics the **import-only flux**, in which the cytoplasmic TF translocates into the nucleus and remains inside the nucleus and it can be understood as an inactive TF gets activated; and the **import-export flux**, where we have two steps: the translocation and cytoplasmic reaccumulation, like the one found in STAT TFs (Meyer; Vinkemeier, 2007; McBride, 2002) and NF-κb (Zambrano et al., 2020; Noursadeghi et al., 2008), for example.

Thus, we consider a nuclear pore translocation in which some nodes from our network are connected to the NPC while many others do not: i.e., to enter the nucleus, a cytoplasmic TF has to enter through those pore-connected regions. To incorporate these dynamics in our system in (3.3), we establish that some regions of our network that are connected to these pores (i.e., the region has a first-degree contact with the nuclear pores since in a highly-connected network as the chromatin, any region is few jumps away from the nuclear pore depending on the number of pores of our system), and any TF entering the nucleus starts in the free state. In Fig. 3.4, we present the representation of the pore connectivity from a network in which we can see different steps away from the pore: the pink colour represents the first-degree connectivity and the following colours are not-connected but classified by their closeness to a pore-connected node.

The Free TF diffuses fast and our network can be interpreted as highly connected since any node needs few steps in the network to be reached; yet, we still need to verify the influence of the pore-connectivity on the transcriptional activity. Considering $f_x(t)$ the TF-specific flux of molecules from outside the nucleus with constant RNAP nuclear concentration, as RNAPs are not found outside the nucleus during the cell-growth phase.

We define the proximity to a nuclear pore as:

$$\mathcal{F}_X^i(t) = \begin{cases} f_x(t), & \text{if region } i \text{ is connected to a pore;} \\ 0, & \text{otherwise.} \end{cases}$$

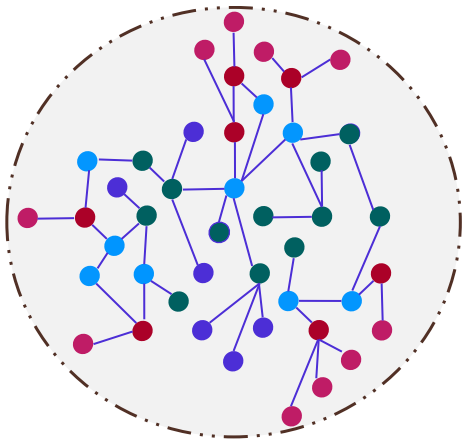

Figure 3.4 – **Diagram to represent a network and nuclear pores connectivity.** Here, we represent our network inside the nucleus (defined by the brown dashed line) and we colour-coded following its closeness to a nuclear pore, i.e., in **pink**, we represent the regions connected to pores; in **red** their immediate neighbours; in **light blue** the 2-steps away from the pore nodes; in **green** the 3-steps away nodes; and in **dark blue/purple** the 4-steps away nodes.

However, as stated previously, the import flux of TFs into the nucleus can have at least two dynamics: **(i) Import Flux** and **(ii) Import/Export Flux**, in which both consider the initial lack of nuclear TFs, but the last dynamics also state that the TF starts its exportation after reaching a maximum transcriptional resource. Since these two dynamics describe two different behaviours, $\mathcal{F}_X^i(t)$ is different in each case. So, to grasp how the different flux dynamics affect our model, we analyze these dynamics separately.

3.2.2.1 Import Flux Function

Each cell has a finite number of TFs/RNAPs molecules, so consider $[T]$ as the TF nuclear concentration, with a r number of nuclear pores ($r < L$) from where the TFs can enter the nucleus with a constant rate until $[T] = [T_{max}]$, i.e., reaching the maximum concentration of TFs for our system. After $[T] = [T_{max}]$, we recover the dynamics from Eq. (3.3). We represent this import dynamics in Eq. (3.5):

$$\frac{dT_i^f}{dt} = -k_{3D}^T T_i^f + \sum_j k_{3D}^T M_{i\leftarrow j} T_j^f - \alpha k_{3D}^T T_i^f + k_i T_i^b + \underbrace{\mathcal{F}_X^i \left([T_{max}] - \sum T_i^f + T_i^b \right)}_{\textbf{import}} . \quad (3.5)$$

In Fig. 3.5, we present a representation of our flux dynamics, showing our initial system without nuclear TFs (represented as green ellipses), and after some time, our system reaches its total concentration stopping the TF import. We also convey how the pores are a fundamental part of our model, as the TFs only enter the nucleus through a pore-connected node.

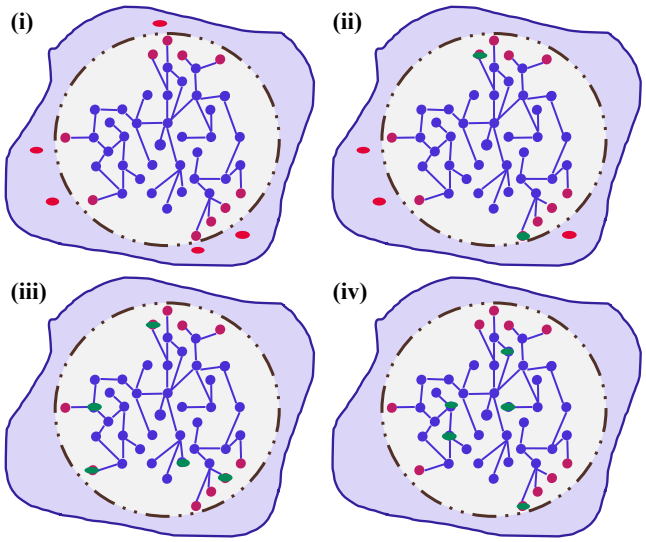

Figure 3.5 – **Cartoon Representation for Import Flux Function.** We present a cell in which we represent the cytoplasmic TF as a red ellipse and nuclear TF as a green ellipse. In our chromatin network, we represent regions connected to pores as **pink** nodes and **dark blue/purple** otherwise. The different time points represent the changes in concentration over time: **(i)** $t = 0$, where our system does not have nuclear TFs; **(ii)** our system after some time. Here, the cytoplasmic TF concentration is non-zero; **(iii)** at this time, our system only has nuclear TFs; and **(iv)** Final time, where the TFs are diffusing in our system, but not exporting.

Considering our model in Eq. (3.3) and the TF-flux described in Eq. (3.5), we have our import-only TF translocation. Our model is a non-homogeneous and non-linear ODE system; so, to facilitate the analysis, we split our model considering the two types of protein complexes, TF and RNAP.

3.2.2.1.1 Analytical Solution for the TF

Let L be the size of our network, i.e., any region $i \in \{1, ..., L\}$, and consider $\mathbf{T}$ as the TF variable like $\mathbf{T} = [T^f T^b]^T$. We define $\mathbf{1}$ as the $L \times L$ matrix with all entries 1, $\mathcal{F}_X[T] = (\mathcal{F}_X^i)_{i=1}^L \in \mathbb{R}^L$, as described in Eq. (3.5), we also define $\mathcal{J}_X = F_X^i[T_{max}]$. Let Id the $L \times L$ identity matrix, $\kappa_{3D}^T = k_{3D}^T \cdot Id$, $\kappa = k \cdot Id$, and M the probabilities matrix form

($M \in \mathbb{R}^{L\times L}$). We rewrite our system as:

$$\frac{d\mathbf{T}}{dt} = \underbrace{\begin{bmatrix} -\kappa_{3D}^T + Mk_{3D}^T - \alpha\kappa_{3D}^T - \mathcal{F}_X \cdot \mathbf{1} & \kappa - \mathcal{F}_X \cdot \mathbf{1} \\ \alpha\kappa_{3D}^T & -\kappa \end{bmatrix}}_{\mathbf{B}} \cdot \mathbf{T} + \underbrace{\begin{bmatrix} \mathcal{J}_X \\ 0 \end{bmatrix}}_{\mathbf{F}_X}$$

$$\frac{d\mathbf{T}}{dt} = \mathbf{B}\cdot\mathbf{T} + \mathbf{F}_X \ ,$$

this system is linear, non-homogeneous and easily solvable. Then, suppose $\mathbf{V}$ is the eigenvectors matrix of $\mathbf{B}$, and $\mathbf{V}$ is invertible since $\det \mathbf{B} \neq 0$. Since $\mathbf{V}$ is invertible it has an inverse matrix. Let $\mathbf{V}^{-1}$ this inverse, and define $\mathbf{T} = \mathbf{V}\cdot\mathbf{G}$. With some calculus and linear algebra techniques, we change our system to:

$$\frac{d\mathbf{G}}{dt} = \mathbf{D}\mathbf{G} + \mathbf{V}^{-1}\mathbf{F}_X \ ,$$

where $\mathbf{D}$ is the diagonal matrix of eigenvalues of $\mathbf{B}$ and $\mathbf{V}^{-1}\mathbf{F}_X \in \mathbb{R}^{2L}$. We solve the homogeneous solution ($\mathbf{V}^{-1}\mathbf{F}_X \equiv 0$) and then the particular solution as a product of the homogeneous and another function. From this system, we calculate the solution for $\mathbf{G}(t)$, which is

$$\mathbf{G}(t) = \mathbf{G}_0 e^{\mathbf{D}t} + \mathbf{V}^{-1}\mathbf{F}_X\left[\mathbf{D}^{-1}(e^{-\mathbf{D}t_0} - e^{-\mathbf{D}t})\right] \ .$$

In which $\mathbf{G}_0$ is a constant dependent on the initial conditions, and $\mathbf{D}^{-1}$ is the inverse of $\mathbf{D}$. Then, we recover the analytical solution for the TFs:

$$\begin{aligned} \mathbf{T}(t) = \mathbf{V}\mathbf{G}(t) &= \mathbf{V}\left(\mathbf{G}_0 e^{\mathbf{D}t} + \mathbf{V}^{-1}\mathbf{F}_X\left[\mathbf{D}^{-1}(e^{-\mathbf{D}t_0} - e^{-\mathbf{D}t})\right]\right) \\ \mathbf{T}(t) &= \mathbf{T}_0 e^{-\mathbf{D}t} + \mathbf{F}_X\left[\mathbf{D}^{-1}(e^{-\mathbf{D}t_0} - e^{-\mathbf{D}t})\right] \ , \end{aligned} \tag{3.6}$$

which $\mathbf{T}_0$ is the initial condition for the TF states.

3.2.2.1.2 Analytical Solution for the RNAP

By definition, the RNAP dynamics depend on a Bound TF to be recruited. Thus, we define in Eq. (3.7) the solution for the Bound TF:

$$T^b(t) = \mathbf{T^b}_0 e^{-\mathbf{D}_b t} + \mathbf{F}_X\left[\mathbf{D}_b^{-1}(e^{-\mathbf{D}_b t_0} - e^{-\mathbf{D}_b t})\right] \ , \tag{3.7}$$

where $\mathbf{T}_0^b$ is the initial conditions for the Bound TF and $T^b(t) \in \mathbb{R}^L$, and $\mathbf{D}_b$ is the diagonal matrix for the Bound TF. Then, we considered the equations for RNAP states using Eq. (3.7) and we defined $\mathbf{P} = [P^f P^b P^t]^T$, $\kappa_{3D}^P = k_{3D}^P \cdot Id$, $\kappa_I = k_I \cdot Id$, $\kappa_\varepsilon = k_\varepsilon \cdot Id$ in which Id the $L \times L$ identity matrix, and M the probabilities matrix form.

$$\frac{d\mathbf{P}}{dt} = \mathbf{B}(t)\cdot\mathbf{P} \ ,$$

in which

$$\mathbf{B}(t) = \begin{bmatrix} -\kappa_{3D}^P + Mk_{3D}^P - k_{3D}^P q\left(\mathbf{T^b}_0 e^{-\mathbf{D}_b t} + \mathbf{F}_X\left[\mathbf{D}_b^{-1}(e^{-\mathbf{D}_b t_0} - e^{-\mathbf{D}_b t})\right]\right) & 0 & \kappa_\varepsilon \\ k_{3D}^P q\left(\mathbf{T^b}_0 e^{-\mathbf{D}_b t} + \mathbf{F}_X\left[\mathbf{D}_b^{-1}(e^{-\mathbf{D}_b t_0} - e^{-\mathbf{D}_b t})\right]\right) & -\kappa_I & 0 \\ 0 & \kappa_I & -\kappa_\varepsilon \end{bmatrix} .$$

This system is non-linear and homogeneous, and the only way to obtain a solution of the form $\mathbf{P}(t) = \mathbf{P}_{t_0} e^{\int \mathbf{B}(s)ds}$ is if $\mathbf{B}(t)$ is a commutative matrix; i.e, let $t_1, t_2 \in \mathbb{R}$, we want to verify in Eq. 3.8 if $\mathbf{B}(t_1)\cdot\mathbf{B}(t_2) = \mathbf{B}(t_2)\cdot\mathbf{B}(t_1)$, i.e.:

$$\mathbf{B}(t_1)\cdot\mathbf{B}(t_2) - \mathbf{B}(t_2)\cdot\mathbf{B}(t_1) = \begin{bmatrix} 0 & 0 & c \\ a & 0 & c \\ b & 0 & 0 \end{bmatrix} , \tag{3.8}$$

where:

$$a = -\frac{\left(\mathbf{T^b}_0 e^{\mathbf{D}_b t_1} - \mathbf{T^b}_0 e^{\mathbf{D}_b t_2} + e^{-\mathbf{D}_b t_0}\left(e^{\mathbf{D}_b t_1} - e^{\mathbf{D_b t_2}}\right)\mathbf{F}_X\right) k_{3D}^P q\left(\kappa_{3D}^P - \kappa_I - k_{3D}^P \mathbf{M}\right)}{\mathbf{D}_b} ;$$

$$b = -\frac{\left(\mathbf{T^b}_0 e^{\mathbf{D}_b t_1} - \mathbf{T^b}_0 e^{\mathbf{D}_b t_2} + e^{-\mathbf{D}_b t_0}\left(e^{\mathbf{D}_b t_1} - e^{\mathbf{D_b t_2}}\right)\mathbf{F}_X\right) k_{3D}^P \kappa_I q}{\mathbf{D}_b} ;$$

$$c = \frac{\left(\mathbf{T^b}_0 e^{\mathbf{D}_b t_1} - \mathbf{T^b}_0 e^{\mathbf{D}_b t_2} + e^{-\mathbf{D}_b t_0}\left(e^{\mathbf{D}_b t_1} - e^{\mathbf{D_b t_2}}\right)\mathbf{F}_X\right) k_{3D}^P \kappa_\varepsilon q}{\mathbf{D}_b} .$$

Thus, the commutativity in this matrix is only possible if $t_1 = t_2$, i.e., $\mathbf{B}(t)$ is not a commutative matrix. Since $\mathbf{B}(t)$ is a function of time t, and both $\mathbf{B}(t)$ and $\int_0^t \mathbf{B}(s)ds$ are non-commutative, the analytical solution for RNAP is not obtained straightforwardly as we have for the TFs. However, since the entries of $\mathbf{B}(t)$ are constants in $\mathbb{R}_+^L$ and Eq. (3.7) is an exponential function, we have continuity in all entries of $\mathbf{B}(t)$. To solve such a system, one can think in terms of a time-oriented exponential matrix and use Magnus Expansion (Arnal; Casas; Chiralt, 2018; Bauer; Metzler, 2012).

Therefore, given the usual commutator: $[A, B] \equiv A\cdot B - B\cdot A$ (i.e., if matrices A and B commute then $[A, B] \equiv 0$) and our initial conditions $t_0 = t(0)$ and $P_0 = P(0)$, we write the analytical solution as

$$P(t) = P_0 \exp(\Omega(t_0, t)) ,$$

in which $\Omega(t_0, t)$ is defined by the following series:

$$\Omega(t_0,t) = \int_{t_0}^{t} \mathbf{B}(s)ds + \frac{1}{2}\int_{t_0}^{t}\int_{t_0}^{t_1} [\mathbf{B}(t_1), \mathbf{B}(t_2)]dt_2 dt_1 + \\ + \frac{1}{6}\int_{t_0}^{t}\int_{0}^{t_1}\int_{t_0}^{t_2} ([\mathbf{B}(t_1), [\mathbf{B}(t_2), \mathbf{B}(t_3)]] + [\mathbf{B}(t_3), [\mathbf{B}(t_2), \mathbf{B}(t_1)]])dt_3 dt_2 dt_1 + \dots . \tag{3.9}$$

The integration of $\mathbf{B}(t_i)$ for any i is made term by term, i.e.; considering $t_0 = 0$, $\int_0^t \mathbf{B}(s)ds$ is:

$$\int_0^t \mathbf{B}(s)ds = \begin{bmatrix} a & 0 & \kappa_\varepsilon t \\ b & -\kappa_I t & 0 \\ 0 & \kappa_I t & -\kappa_\varepsilon t \end{bmatrix} ,$$

with the values:

$$a = \frac{k_{3D}^P \left(\mathbf{T^b}_0 \mathbf{D}_b \left(-1 + e^{\mathbf{D}_b t}\right) q + (-1 + e^{\mathbf{D}_b t} \mathbf{F}_X q - \mathbf{D}_b t(\mathbf{D}_b(Id - \mathbf{M}) + \mathbf{F}_X q\right)}{\mathbf{D}_b^2} ;$$
$$b = \frac{k_{3D}^P q \left(\mathbf{T^b}_0 \mathbf{D}_b \left(-1 + e^{\mathbf{D}_b t}\right) + \mathbf{F}_X(-1 + e^{\mathbf{D}_b t} - \mathbf{D}_b t)\right)}{\mathbf{D}_b^2} .$$

In the second term of this series, we calculate using the commutator presented in Eq. 3.8 and integrate:

$$\frac{1}{2}\int_{t_0}^{t}\int_{t_0}^{t_1} [\mathbf{B}(t_1), \mathbf{B}(t_2)] dt_2 dt_1 = \frac{1}{2}\begin{bmatrix} 0 & 0 & c \\ a & 0 & c \\ b & 0 & 0 \end{bmatrix} ,$$

with:

$$a = -\frac{k_{3D}^P(\kappa_{3D}^P - \kappa_I - k_{3D}^P \mathbf{M} q(2 + \mathbf{D}_b t + e^{\mathbf{D}_b t}(-2 + \mathbf{D}_b t))(\mathbf{T^b}_0 \mathbf{D}_b + \mathbf{F}_X)}{\mathbf{D}_b^3} ;$$
$$b = -\frac{k_{3D}^P \kappa_I q(2 + \mathbf{D}_b t + e^{\mathbf{D}_b t}(-2 + \mathbf{D}_b t))(\mathbf{T^b}_0 \mathbf{D}_b + \mathbf{F}_X)}{\mathbf{D}_b^3} ;$$
$$c = \frac{k_{3D}^P \kappa_\varepsilon q(2 + \mathbf{D}_b t + e^{\mathbf{D}_b t}(-2 + \mathbf{D}_b t))(\mathbf{T^b}_0 \mathbf{D}_b + \mathbf{F}_X)}{\mathbf{D}_b^3}$$

We can calculate higher-order terms of this series, but the technique we used is the same one we used previously for our commutator. Trunking our solution in the first two terms of our series already shows how complex is the solution, which we present in the following equation:

$$\Omega(0,t) = \begin{bmatrix} \omega_{11} & 0 & \kappa_\varepsilon + \omega_{13} \\ \omega_{21} & -\kappa_I t & \omega_{13} \\ \omega_{31} & \kappa_I t & -\kappa_\varepsilon t \end{bmatrix} ,$$

with

$$\omega_{11} = \frac{k_{3D}^P \left(\mathbf{T^b}_0 \mathbf{D}_b \left(-1 + e^{\mathbf{D}_b t}\right) q + (-1 + e^{\mathbf{D}_b t} \mathbf{F}_X q - \mathbf{D}_b t(\mathbf{D}_b(Id - \mathbf{M}) + \mathbf{F}_X q\right)}{\mathbf{D}_b^2} ;$$
$$\omega_{13} = \frac{k_{3D}^P \kappa_\varepsilon q(2 + \mathbf{D}_b t + e^{\mathbf{D}_b t}(-2 + \mathbf{D}_b t)(\mathbf{T^b}_0 \mathbf{D}_b + \mathbf{F}_X)}{2\mathbf{D}_b^3} ;$$
$$\omega_{21} = \frac{k_{3D}^P q(-(\kappa_{3D}^P - \kappa_I - k_{3D}^P \mathbf{M})}{2\mathbf{D}_b^3}$$
$$\frac{(2 + \mathbf{D}_b t + e^{\mathbf{D}_b t}(-2 + \mathbf{D}_b t))(\mathbf{T^b}_0 \mathbf{D}_b(-1 + e^{\mathbf{D}_b t}) + \mathbf{F}_X(-1 + e^{\mathbf{D}_b t} - \mathbf{D}_b t)))}{2\mathbf{D}_b^3} ;$$
$$\omega_{31} = -\frac{k_{3D} \kappa_I q(2 + \mathbf{D}_b t + e^{\mathbf{D}_b t}(-2 + \mathbf{D_b} t)(\mathbf{T^b}_0 \mathbf{D}_b + \mathbf{F}_X}{2\mathbf{D}_b^3} .$$

This approximation solely considers two terms of the Magnus Expansion, as we discussed the need to trunk the solution at some point since it is a series. However, since $\Omega(0,t)$ has enough diagonal entries, this trunked solution can be numerically evaluated given a set of

parameters if needed since the calculation is less computationally heavy than it seems. In Eq. (3.10), we present the expression for the RNAP solution, in which $\mathbf{P}_0$ the RNAP's initial conditions for t_0 and the Magnus Expansion defined by function $\Omega(t_0, t)$.

$$\mathbf{P}(t) = \mathbf{P}_0 e^{\Omega(t_0,t)} \ . \tag{3.10}$$

3.2.2.2 Import/Export Flux Function

Some TFs have a translocation mechanism as they are not endogenous in the nucleus. For example, NF-κb is found mainly in the cytoplasm, and, once needed for transcription, NF-κb translocates into the nucleus. This translocation/re-accumulation process must be well-regulated as disruptions were observed in cancer pathways, for example (Xia; Shen; Verma, 2014). The translocation needs the help of nuclear pores to facilitate the TF entrance (and later mRNA export) (Allen et al., 2000). Hence, a TF must pass through specific pores to enter the nucleus.

Therefore, let the total number of TFs inside the cell, $T_{total} = [T] + [T]^C$, which depends on the nuclear concentration inside the nucleus, $[T]$, and outside the nucleus, $[T]^C$. We propose three steps for the translocation and represent it in Fig. 3.6:

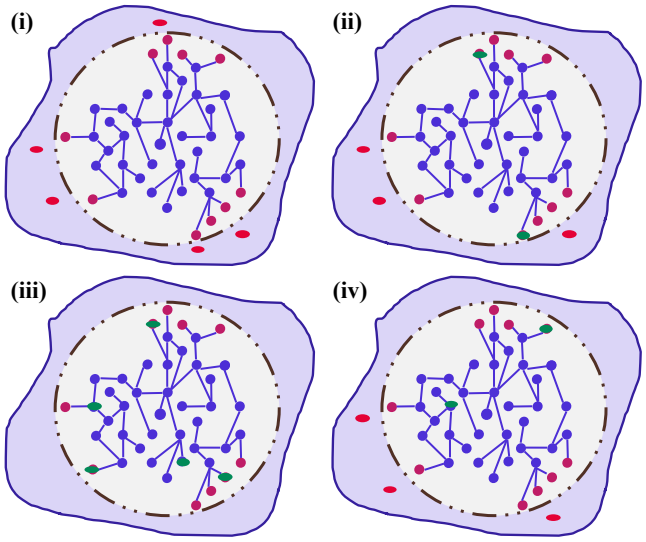

Figure 3.6 – **Cartoon Representation for Import/Export Flux Function.** Here, we consider cytoplasmic TFs as **red ellipses** and nuclear TFs as **green ellipses**, i.e., $T_{total} = [T_{\text{green}}] + [T_{\text{red}}]$. The network can be split into pore-connected nodes (**pink**) and non-connected to pores (**dark blue/purple**). Here, we present four-time points for a cell: **(i)** The initial time, $t = 0$, where the cell does not have nuclear TFs; **(ii)** The Import Process; **(iii)** Our system reaches the maximum nuclear concentration; **(iv)** The Export Process.

1. The nuclear TF concentration starts at zero, and the TFs start entering the nucleus.

2. The nuclear TF concentration reaches its maximum.

3. The export process begins to decrease the nuclear TF concentration.

From this total concentration, we model the translocation for our flux function in the Free TF equation:

$$\mathcal{F}_X^i(t) = \underbrace{K_X^i \left(k_{im} T_{total} e^{-k_{im} t}\right)}_{\textbf{import}} - \underbrace{K_X^i T_i^f E(t)}_{\textbf{export}} , \tag{3.11}$$

where K_X^i represents the presence (or absence) of a connection to a nuclear pore in the region i divided by the total number of pores, k_{im} is the TF import rate and $E(t)$ is the TF exporter function which depends on the Free TF density in the region i, and we evaluate by integrating:

$$\frac{dE}{dt} = \mu[T] - \delta E ,$$

with μ the export rate related to the nuclear TF concentration $[T]$ and δ, the exporter degradation rate. Similarly to the Import Flux Function, we split our solutions into two: one system for the TF and another for the RNAP.

3.2.2.2.1 Analytical Solution for the TF

To solve our system considering the Free TF in Eq.(3.11), first, we need to solve the expression for the exporter function, which is a separable ODE with the following solution:

$$E(t) = \frac{\mu[T] - e^{-\delta(t+c_1)}}{\delta} ,$$

with c_1 an arbitrary constant dependent on the initial conditions. We can substitute this solution in Eq.(3.11):

$$\begin{aligned} \frac{dT_i^f}{dt} = &-k_{3D}^T T_i^f + \sum_j k_{3D}^T M_{i \leftarrow j} T_j^f - \\ &- \alpha k_{3D}^T T_i^f + k_i T_i^b + K_X^i \left(k_{im} T_{total} e^{-k_{im} t}\right) - K_X^i T_i^f \frac{\mu[T] - e^{-\delta(t+c_1)}}{\delta} . \end{aligned}$$

Considering the same assumptions as the ones presented in the previous TF analytical solution, let L be the size of our network and consider $\mathbf{T} = [T^f T^b]^T$ and, given Id as the $L \times L$ identity matrix, $\kappa_{3D}^T = k_{3D}^T \cdot Id$ and $\kappa = k \cdot Id$. We rewrite our TF system in the matrix form:

$$\frac{d\mathbf{T}}{dt} = \underbrace{\begin{bmatrix} k_{3D}^T(-Id + M - \alpha Id) - K_X \left(\dfrac{\mu[T] - e^{-\delta(t+c_1)}}{\delta}\right) & \kappa \\ \alpha \kappa_{3D}^T & -\kappa \end{bmatrix}}_{\mathbf{B}(t)} \mathbf{T} + \begin{bmatrix} K_X k_{im} T_{total} e^{-k_{im} t} \\ 0 \end{bmatrix} .$$

Similar to the previous RNAP set of ODEs, this system is non-homogeneous so before we go any further with the ODE solutions, we need to verify if $\mathbf{B}(t)$ is a commutative

matrix (i.e, $\mathbf{B}(t_1)\mathbf{B}(t_2) - \mathbf{B}(t_2)\mathbf{B}(t_1) = 0$, for any $t_1, t_2 \in \mathcal{T}$):

$$\mathbf{B}(t_1)\cdot\mathbf{B}(t_2) - \mathbf{B}(t_2)\cdot\mathbf{B}(t_1) = \\ = \begin{bmatrix} 0 & \frac{e^{-\delta(2c_1+t_1+t_2)}(e^{\delta(c_1+t_1)} - e^{\delta(c_1+t_2)})\kappa K_X}{\delta} \\ \frac{\alpha e^{-\delta(2c_1+t_1+t_2)}(e^{\delta(c_1+t_2)} - e^{\delta(c_1+t_2)}\kappa K_X}{\delta} & 0 \end{bmatrix},$$

i.e., $\mathbf{B}(\mathcal{T})$ is a non-commutative matrix. However, $\mathbf{B}(\mathcal{T})$ is a continuous matrix and we can apply the Magnus Expansion, in this case, by using the commutator $[\mathbf{B}(t_1), \mathbf{B}(t_2)] = \mathbf{B}(t_1)\cdot\mathbf{B}(t_2) - \mathbf{B}(t_2)\cdot\mathbf{B}(t_1)$ to solve the homogeneous part of our problem and then find the particular solution. Once again, we remind that the Magnus Expansion, Eq. (3.9), is a series of integrals, and we choose to trunk our solutions in the first two terms because we aim to showcase the form of our analytical solution.

We calculate the first term and define $t_0 = 0$:

$$\Omega_1(t) = \int_0^t \mathbf{B}(s)ds = \begin{bmatrix} -k_{3D}^T(\alpha Id + Id - \mathbf{M})t + \frac{K_X\mu[T]t}{\delta} - e^{-\delta(c_1+t)}(-1+e^{\delta t})K_X & \kappa t \\ \alpha\kappa_{3D}^T t & -\kappa t \end{bmatrix}.$$

The second term of our series is obtained by integrating twice Eq. (3.12) and multiplying by 1/2:

$$\Omega_2 = \frac{1}{2}\int_{t_0}^{t}\int_{t_0}^{t_1}[\mathbf{B}(t_1), \mathbf{B}(t_2)]dt_2dt_1 = \\ = \frac{1}{2}\begin{bmatrix} 0 & \frac{e^{-\delta(c_1+t)}\kappa K_X(2+\delta t + e^{\delta t}(-2+\delta t))}{\delta^3} \\ -\frac{\alpha e^{-\delta(c_1+t)}\kappa_{3D}^T K_X(2+\delta t + e^{\delta t}(-2+\delta t))}{\delta^3} & 0 \end{bmatrix}.$$

The sum of these two first terms of the Magnus Expansion for the TF is:

$$\Omega_{TF}(t) = \\ = \begin{bmatrix} \frac{K_X\mu[T]t}{\delta} - \frac{e^{-\delta(c_1+t)}(-1+e^{\delta t})K_X}{\delta^2} - k_{3D}^T(\alpha+1)Id - \mathbf{M})t & \kappa t + \frac{e^{-\delta(c_1+t)}\kappa K_X(2+\delta t+e^{\delta t}(-2+\delta t))}{2\delta^3} \\ \alpha\kappa_{3D}^T - \frac{\alpha e^{-\delta(c_1+t)}\kappa_{3D}^T K_X(2+\delta t+e^{\delta t}(-2+\delta t))}{2\delta^3} & -\kappa t \end{bmatrix}.$$

Thus, the homogeneous part of our TF system is $\mathbf{T}_h(t) = \mathbf{T}_0 e^{\Omega_{TF}(t)}$, and the particular solution has the form $\mathbf{T}_p(t) = \mathbf{T}_h(t)\mathbf{v}(t)$.

$$\begin{aligned} \frac{d\mathbf{T}_p(t)}{dt} &= \frac{d}{dt}(\mathbf{T}_h(t)\mathbf{v}(t)) \\ &= \mathbf{T}_h'(t)\mathbf{v}(t) + \mathbf{T}_h(t)\mathbf{v}'(t) = \mathbf{B}(t)\mathbf{T}_h(t)\mathbf{v}(t) + \mathcal{F}_x(t) \\ \mathbf{T}_0 e^{\Omega_{TF}(t)}\mathbf{v}'(t) &= K_X k_{im} T_{total} e^{-k_{im}t} \\ \mathbf{v}(t) &= \int_0^t \mathbf{T}_0^{-1} e^{-\Omega_{TF}(s)} K_X k_{im} T_{total} e^{-k_{im}s} ds \\ \mathbf{v}(t) &= \mathbf{T}_0^{-1} K_X k_{im} T_{total} \int_0^t e^{-\Omega_{TF}(s)} e^{-k_{im}s} ds\ . \end{aligned}$$

We can apply the integration by parts in our expression for $\mathbf{v}(t)$, but it does not simplify the function, so we opted to present the solution for the TF in Eq. (3.12).

$$\mathbf{T}(t) = K_X k_{im} T_{total} e^{\Omega_{TF}(t)} \int_0^t e^{-\Omega_{TF}(s)} e^{-k_{im} s} ds \ . \tag{3.12}$$

3.2.2.2.2 Analytical Solution for the RNAP

We limit the $\mathbf{T}(t)$ to only the Bound state since it's the TF important for the RNAP dynamics, Eq. (3.13), and rewrite our system in its matrix form. First, we define $\mathbf{P} = [P^f P^b P^t]^T$, $\kappa_{3D}^P = k_{3D}^P \cdot Id$, $\kappa_I = k_I \cdot Id$, $\kappa_\varepsilon = k_\varepsilon \cdot Id$ in which Id the $L \times L$ identity matrix, and M the probabilities matrix form.

$$\mathbf{T}_b(t) = K_X k_{im} T_{total} e^{\Omega_{TF^b}(t)} \int_0^t e^{-\Omega_{TF^b}(s)} e^{-k_{im} s} ds \ . \tag{3.13}$$

$$\frac{d\mathbf{P}}{dt} = \underbrace{\begin{bmatrix} -\kappa_{3D}^P + M k_{3D}^P - k_{3D}^P q K_X k_{im} T_{total} e^{\Omega_{TF^b}(t)} \int_0^t e^{-\Omega_{TF^b}(s)} e^{-k_{im} s} ds & 0 & \kappa_\varepsilon \\ k_{3D}^P q K_X k_{im} T_{total} e^{\Omega_{TF^b}(t)} \int_0^t e^{-\Omega_{TF^b}(s)} e^{-k_{im} s} ds & -\kappa_I & 0 \\ 0 & \kappa_I & -\kappa_\varepsilon \end{bmatrix}}_{\mathbf{C}(t)} \cdot \mathbf{P}$$

$$\frac{d\mathbf{P}}{dt} = \mathbf{C}(t) \cdot \mathbf{P} \ .$$

The RNAP is an homogeneous system, we calculated $\mathbf{C}(t_1)\mathbf{C}(t_2) - \mathbf{C}(t_2)\mathbf{C}(t_1)$ for any $t_1, t_2 \in \mathcal{T}$ to verify if this matrix is commutative:

$$\mathbf{D}(t_1, t_2) = \mathbf{C}(t_1)\mathbf{C}(t_2) - \mathbf{C}(t_2)\mathbf{C}(t_1) = \begin{bmatrix} 0 & 0 & a \\ b & 0 & c \\ d & 0 & 0 \end{bmatrix},$$

with

$$a = -\kappa_\varepsilon k_{3D}^P K_X k_{im} T_{total} \left(e^{\Omega_{TF^b}(t_1)} \int_0^{t_1} e^{-\Omega_{TF^b}(s)} e^{-k_{im} s} ds - e^{\Omega_{TF^b}(t_2)} \int_0^{t_2} e^{-\Omega_{TF^b}(s)} e^{-k_{im} s} ds \right) \ ;$$

$$b = k_{3D}^P k_{im} K_X T_{total} (\kappa_{3D}^P + \kappa_I + k_{3D}^P \mathbf{M}) (e^{\Omega_{TF^b}(t_1)} \int_0^{t_1} e^{-\Omega_{TF^b}(s)} e^{-k_{im} s} ds -$$

$$- e^{\Omega_{TF^b}(t_2)} \int_0^{t_2} e^{-\Omega_{TF^b}(s)} e^{-k_{im} s} ds) \ ;$$

$$c = \kappa_\varepsilon k_{3D}^P K_X k_{im} T_{total} \left(e^{\Omega_{TF^b}(t_1)} \int_0^{t_1} e^{-\Omega_{TF^b}(s)} e^{-k_{im} s} ds - e^{\Omega_{TF^b}(t_2)} \int_0^{t_2} e^{-\Omega_{TF^b}(s)} e^{-k_{im} s} ds \right) \ ;$$

$$d = k_{3D}^P k_{im} K_X T_{total} \left(e^{\Omega_{TF^b}(t_1)} \int_0^{t_1} e^{-\Omega_{TF^b}(s)} e^{-k_{im} s} ds - e^{\Omega_{TF^b}(t_2)} \int_0^{t_2} e^{-\Omega_{TF^b}(s)} e^{-k_{im} s} ds \right) \ .$$

Once again, our system is a homogeneous non-commutative but with continuous terms. The solution for this system can be calculated using the commutator $[\mathbf{C}(t_1), \mathbf{C}(t_2)] = \mathbf{C}(t_1) \cdot \mathbf{C}(t_2) - \mathbf{C}(t_2) \cdot \mathbf{C}(t_1)$ and applying the Magnus Expansion, Eq. (3.9). We define

$\Omega_P(t)$ as the Magnus Expansion for our RNAP system and $\mathbf{P}_0$ its initial condition, the solution has the form:

$$\mathbf{P}(t) = \mathbf{P}_0 e^{\Omega_P(t)} \ , \tag{3.14}$$

with

$$\Omega_P(t) = \Omega_P(0,t) = \int_0^t \mathbf{C}(s)ds + \frac{1}{2}\int_0^t \int_0^{t_1} \mathbf{D}(t_1,t_2)dt_2 dt_1 + \dots$$

Even though it is possible to apply numerically the mathematical expressions in this system, it is clear the complexity of the analytical solution. Such complexity emphasizes the need for computational tools to analyze and understand these dynamics; and since our solutions are a series of integrals of commutators for a L-sized vector, numerical solutions are necessary to visualize our system. Next, we present numerical solutions for a dataset considering our flux functions.

3.2.3 Numerical Solutions for the Flux functions verify the dependency on the changes of transcriptional resources

We solved our model in Eq. (3.3) for the flux functions in Eqs. (3.5) and (3.11), with two types of implementations: (i) the **deterministic** in which we used Matlab's ode15s (a multistep algorithm (Gupta; Wallace, 1975)) for both functions to understand the likelihood of TF/RNAP states to occupy chromatin regions, and (ii) the **stochastic Gillespie Algorithm** (Gillespie, 1976) to represent the stochasticity of gene expression. Each simulation represents 180 minutes in time.

Except for the flux parameters, the other parameters (i.e., those from Eq. (3.3)) are the same for both functions: **(I) Import Flux Function** and **(II) Import/Export Flux Function**, with same RNAP total concentration and maximum nuclear TF concentration (i.e., $[P] = 400$ and $T_{max} = 200$). These fixed parameters are available in Table 2. Our network is the same as we considered in Chapter 2.

Similar to the represented in Fig. 3.4, just some regions are connected to pores. The regions connected to the nuclear envelope are selected by randomly picking 7% regions of the scaled network from the subset of the 50% regions with higher values of residence times. We designated these subnetworks this way because the active regions move towards the nuclear periphery during the transcription process (Kalverda; Röling; Fornerod, 2008).

3.2.3.1 Import Flux Function

In Eqs. (3.6) and (3.10), we showed how a continuous flux of free transcription factors would affect the organization of the molecules over time. Initially, we consider the concentration of TFs inside the nucleus to be zero, i.e., $T_{nuc} = 0$. Since to enter the Bound and Transcribing states for the RNAP, the Free RNAPs need a Bound TF in a region and RNAPs are only found in the nucleus, the only state we have for the RNAP initially

is the Free. It is clear that, as the TF nuclear concentration increases, the Free RNAP changes its state to Bound and, eventually, Transcribing. We can see the concentration changes in Fig. 3.7, we verified the Free RNAPs decrease for all regions because of the number of TFs in the Bound state, with a faster change in states since our system reaches the steady-state before the 5 minutes mark, i.e., our simulation converges to Eqs. (3.4).

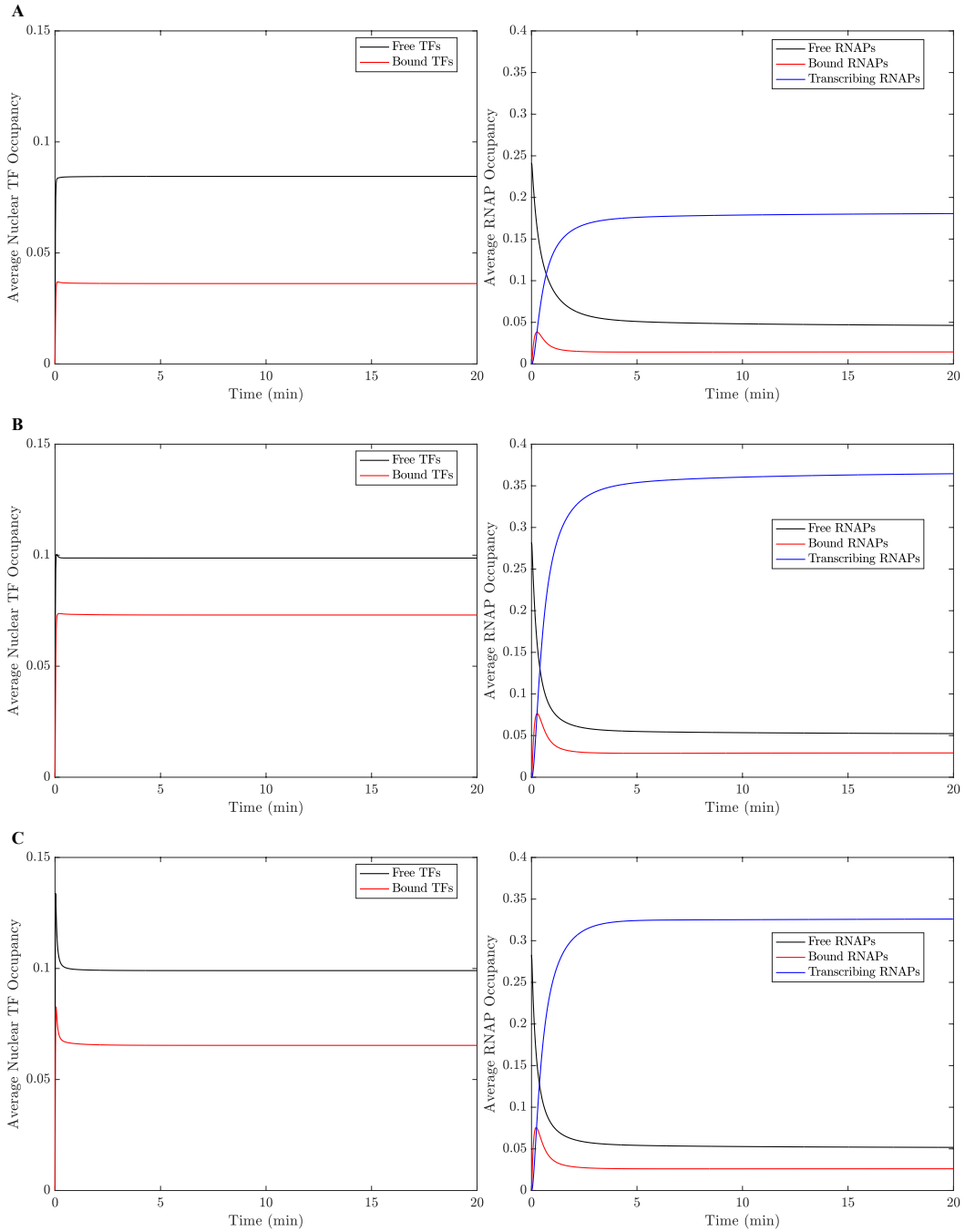

Figure 3.7 – **Average Concentration for TF/RNAP states for our Import Flux Function. A** All Chromatin Regions; **B** Active Regions not connected to pores; **C** Regions connected to pores. The overshoot of TFs for regions connected to pores is a direct consequence of the pore.

We present the global behaviour, calculating the average for all regions, Fig. 3.7 **A**

for TF and RNAP states, showing how the states organize themselves after 20 minutes (even if we simulated our model to represent 180 minutes). The low increase for Bound TF attests that most TFs in the nucleus remain in the free state by our assumptions (around 70% of free TFs). Since our network size, L, is bigger than the nuclear concentration of TFs and RNAPs, the residence times present a bimodality in inactive, and active regions (by construct) and our steady-states (Eqs. (3.4)), in Figs. 3.7 **B** and **C** we present studies for active regions, with **B** being those active but not connected to pores and **C** the regions connected to pores.

The pore connectivity causes an overshoot of TF states in regions connected to pores, but since our model reaches the equilibrium before reaching the 5 minutes mark and the fast diffusivity of Free TF/RNAP, we proved the pore connectivity is less important for the occupancy later in the implementations. More so, since the choice of pore-connected regions was random, we can see the effects of the parameters in our simulations and how active regions not connected to pores present on average more activity than regions connected to pores, when we compare RNAP states in Fig. 3.7 **B** and **C**.

However, this is an averaged result and we cannot affirm that all the active regions present higher activities than the pore-connected ones. To verify this, we present different subnetworks in Fig. 3.8 in form of heatmaps from 0 to 20 minutes for our deterministic solution for the Transcribing RNAP as it represents the actual transcription process (Roeder, 1996; Petrenko; Struhl, 2021). With this analysis, we compared the strength of the occupancy for all the nodes inside a subnetwork.

We present the following subnetworks: **A** All the chromatin regions - proving the inactivity expected in some regions; **B** Active regions not connected to pores; **C** Regions connected to pores (which are active regions). Comparing the patterns between **B** and **C**, we have seemingly more regions with reduced activity in **B** than **C**, which makes sense considering the **B** corresponds to 43% of our network and **C** to 7%, which means we only consider the half or the regions with higher chances of being active from Fig. 3.8 **A**.

Fig. 3.8 **D** shows the subnetwork for regions with more than one step away from a nuclear pore and **E** Regions connected to pores and their nearest neighbours. This means we calculated how many steps away each node is from a pore-connected node and then split our network between one step away maximum (Fig. 3.8 **E**) and more than one step away, Fig. 3.8 **D**. We also uncovered the RNAP clustering around some prolific regions a result justified by the transcriptional machinery as expected.

Besides, Bound TF is a limited resource and it affects the non-free RNAPs states. While comparing Fig. 3.8 **D** and **E**, we see most of the activity remains in the regions close to pores. Thus, the closeness to a pore is beneficial for Bound and Transcribing RNAPs states as their increase is more prominent. The closeness to a pore highlights the activation of the Transcribing state, which seems to corroborate experimental observations

that we studied more closely in Chapter 5.

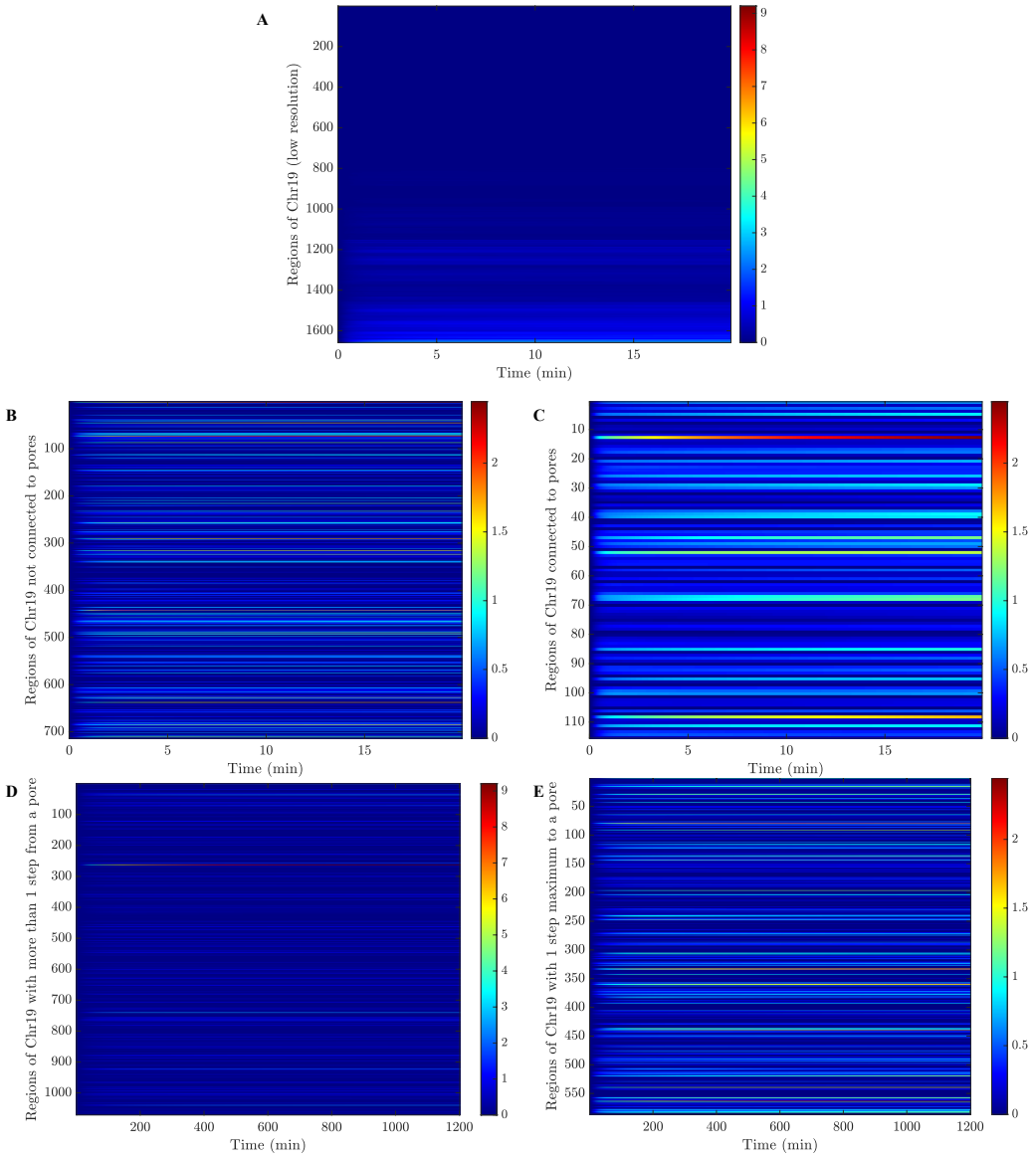

Figure 3.8 – **Heatmaps for our Import Flux Function for different subnetworks.** **A** All chromatin regions; **B** Active regions not connected to pores; **C** Regions connected to pores; **D** Regions with more than 1 step away from a nuclear pore; and **E** Regions with 1 step maximum from a nuclear pore (i.e., pore-connected regions and their nearest neighbours).

As a form of understanding the patterns our model can have, we propose a cluster analysis for the Transcribing RNAP. In fact, even if the connection to a pore influences how the occupancy occurs, we verified if other patterns emerge from the region, e.g., non-transcribing regions.

The cluster of our Transcribing RNAP solution is presented in Fig. 3.9 **A**. Most of our system clusters stabilize early and only two show a strong transcriptional activity increase, clearly accumulating resources in these clustered averages. The remaining averaged

clusters show a difference in activity levels, similar to the ones we presented in Fig. 3.8 with one cluster representing all inactive regions.

To understand how the solutions deviate from the average, we calculated the z-score for the Transcribing RNAP, clustering the results in Fig. 3.9 **B**. We demonstrated how some clusters show later activation while some present an overshoot in concentration which later stabilizes. This difference between clustered z-scores proves not all regions behave the same, showing how the transcriptional activities change, as we presented in Fig. 3.9 **A**.

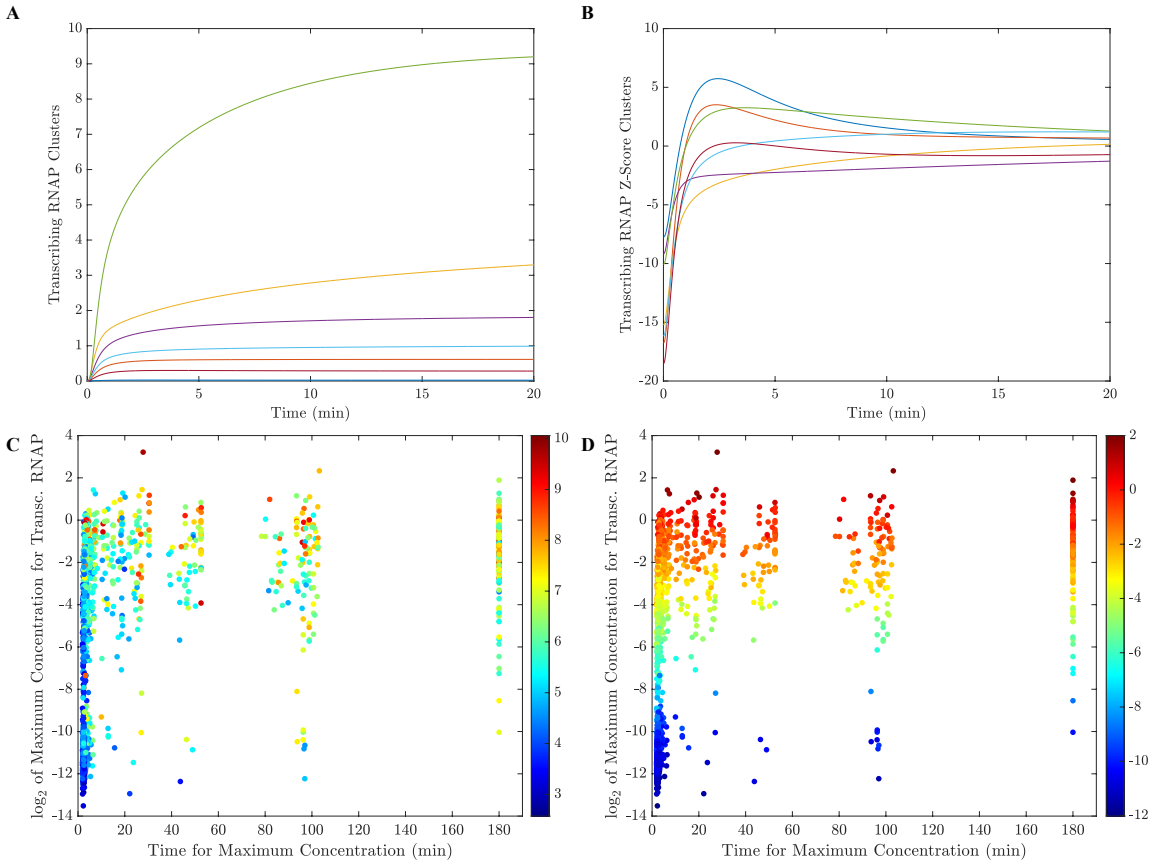

Figure 3.9 – **Behaviours for our Deterministic Solution. A** is the clusters for the Transcribing RNAP, with 7 different clustered averages. **B** The clustered z-score for our Transcribing RNAP solutions. **C**-**D** Time for reaching the maximum concentration over the log_2 of the maximum Transcribing RNAP concentration labelled with the number of connections (d) of the region, **C** and **D** $\log_2$ the residence times, $\log_2(\tau)$.

Our system seemingly reaches the equilibrium before 5 minutes (Figs. 3.7 and 3.8). Furthermore, in Eqs. (3.4), we defined the importance of the number of the connections, d_i, and the residence times $\tau_i's$. To check the effects of τ and d in our model, we present in Figs. 3.9 **C** and **D** the time for maximum concentration and $\log_2$ of the maximum Transcribing RNAP concentration for the number of connections, d (Fig. 3.9 **C**) and the $\log_2$ of residence times (Fig. 3.9 **D**).

From these results, we see how most of the less connected regions reach their maximum early and they present smaller Transcribing RNAP maximum occupancies. We also see how the residence times affect the occupancy by stratifying with its values, numerically proving how important is this parameter for transcription as the TF binding to a region is fundamental for RNAP activation. The maximum values being further than expected in our numerical solutions are explained by infinitesimal variations from our solver, but the different concentration values are a direct consequence of τ_i and d.

Gene regulation is a stochastic process (Elowitz, 2002; Kærn et al., 2005) and a deterministic solution allows us to have continuous values for Transcribing RNAP molecule allocation, which is not feasible in a cell since the transcriptional activity is a discrete process (more than that, it is a binary process - the cell is either transcribing or not). To represent this, we implemented a Gillespie Stochastic Algorithm for our model in Eq. (3.3) with the flux function Eq. (3.5) (Gillespie, 1976), and we simulated for 70 cells.

Our aim was to verify how the stochasticity and discreteness affect our system and on a deeper level, transcription. In Fig. 3.10, we present the fraction of active transcription for different subnetworks sorted by the number of connections (i) and residence times (ii). The subnetworks considered are the same present in Fig. 3.8, and by activity we mean the presence of at least one Transcribing RNAP in that region, ignoring the RNAP clustering.

From Fig. 3.10, we can see in **A** (i) and (ii) how the residence times have a strong influence on the fraction of active target sites than the number of connections and this can be explained by the fast diffusivity of TF and RNAP complexes. Besides, similar to the deterministic results, we verified how half of our network is inactive - which is explained by the bimodality from our exiting rates, Fig. 2.12. The network connectivity also impacts the transcriptional activity but on a small scale.

The subnetwork in Fig. 3.10 **B** represents all the active regions not connected to nuclear pores. Once again, the residence times are a stronger parameter to control transcription, but in this subnetwork, the connectivity is more relevant for active regions, as there is a need to be reached. However, in Fig. 3.10 **C** we can see how is a factor less relevant for regions connected to pores. Since the active regions and nearest neighbours are not the same subnetworks, in Fig. 3.10 **D** we present the regions with 1 step maximum to a nuclear pore and we recovered that almost half of these regions remain inactive, which is also present in Fig. 3.10 **E**, which we show regions more than 1 step away from a nuclear pore. With these results, we verified how the time a TF spends bound to a particular chromatin region impacts the transcription for this model and why even though the structure is fundamental to the whole transcriptional machinery, the residence time is a key factor to control gene expression.

We propose Fig. 3.11 **A** and **B** to analyze region-specific patterns: in **A**, we see how many transcriptions occur in a region per minute, and in most of our network we have

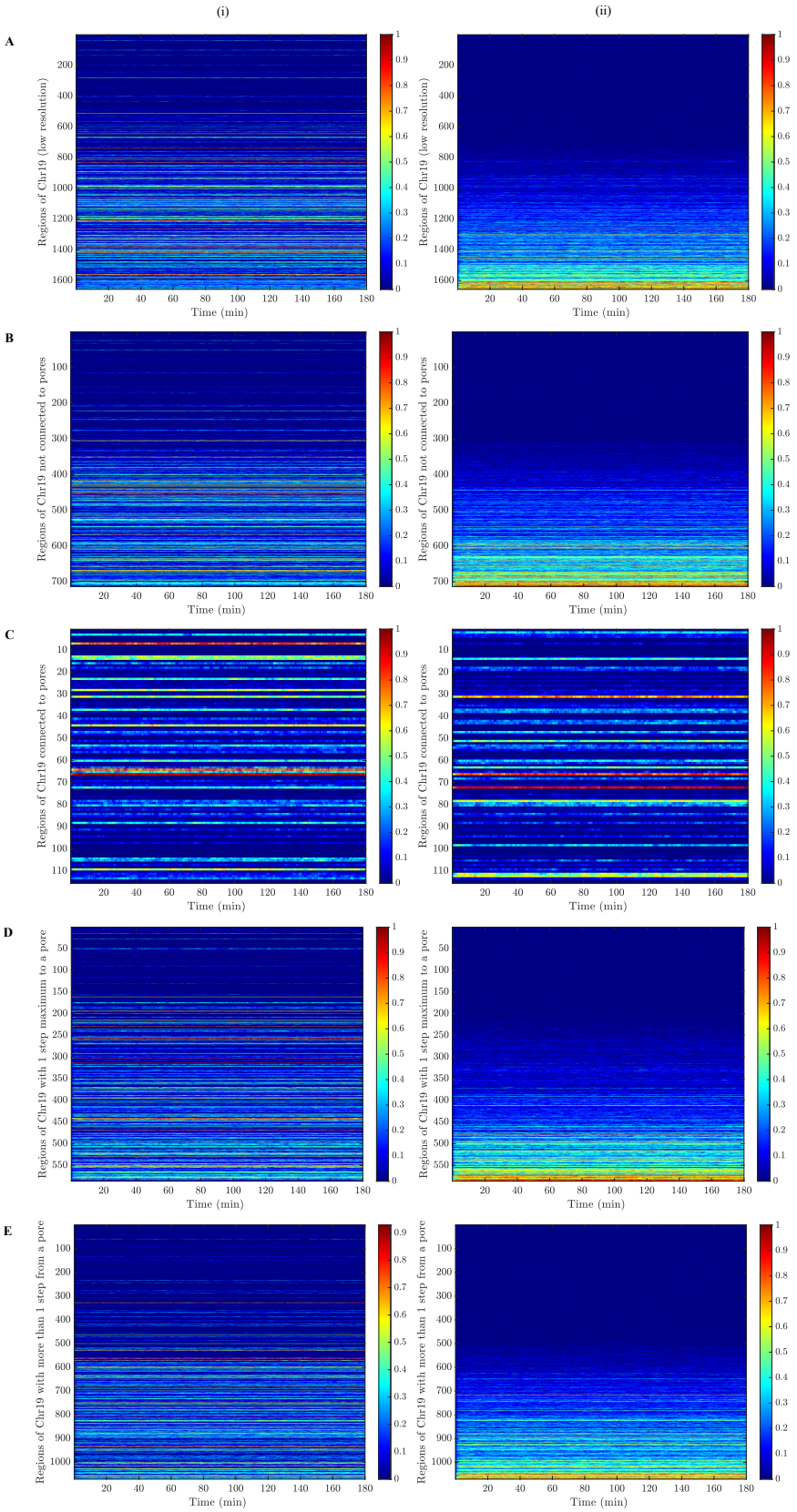

Figure 3.10 – **Fraction of Active Target Sites For Different Subnetworks sorted by (i) the number of connections and (ii) the residence times. A** All Chromatin Networks. **B** Active Regions not connected to pores. **C** Regions connected to pores. **D** Regions with a maximum of 1 step away from a nuclear pore. **E** Regions with more than 1 step away from a pore.

less than one transcription event per minute, mostly because the transcriptional resources are limited, thus the system occupies their preferential regions for transcription; in **B**, we present the On-time average in seconds for our model - i.e., the time spent between initiation and an elongation reaction. Both figures also recovered a block of completely inactive regions, which are regions closed to transcription. We also determined a set of prolific regions.

Since we expect variation between cells, in Fig. 3.11 **C**, we can see how our model predicts the transcriptional activity meaning we can predict an array of different likelihoods for all our network regions (as shown in Fig. 3.10). However, in Fig. 3.11 **C** facilitates the visualization of how transcription is stochastic but also consistent: i.e., once the RNAP gets activated the variation between time steps for each region is less than 20% for the region in small activities and around than 10% for regions with strong activity.

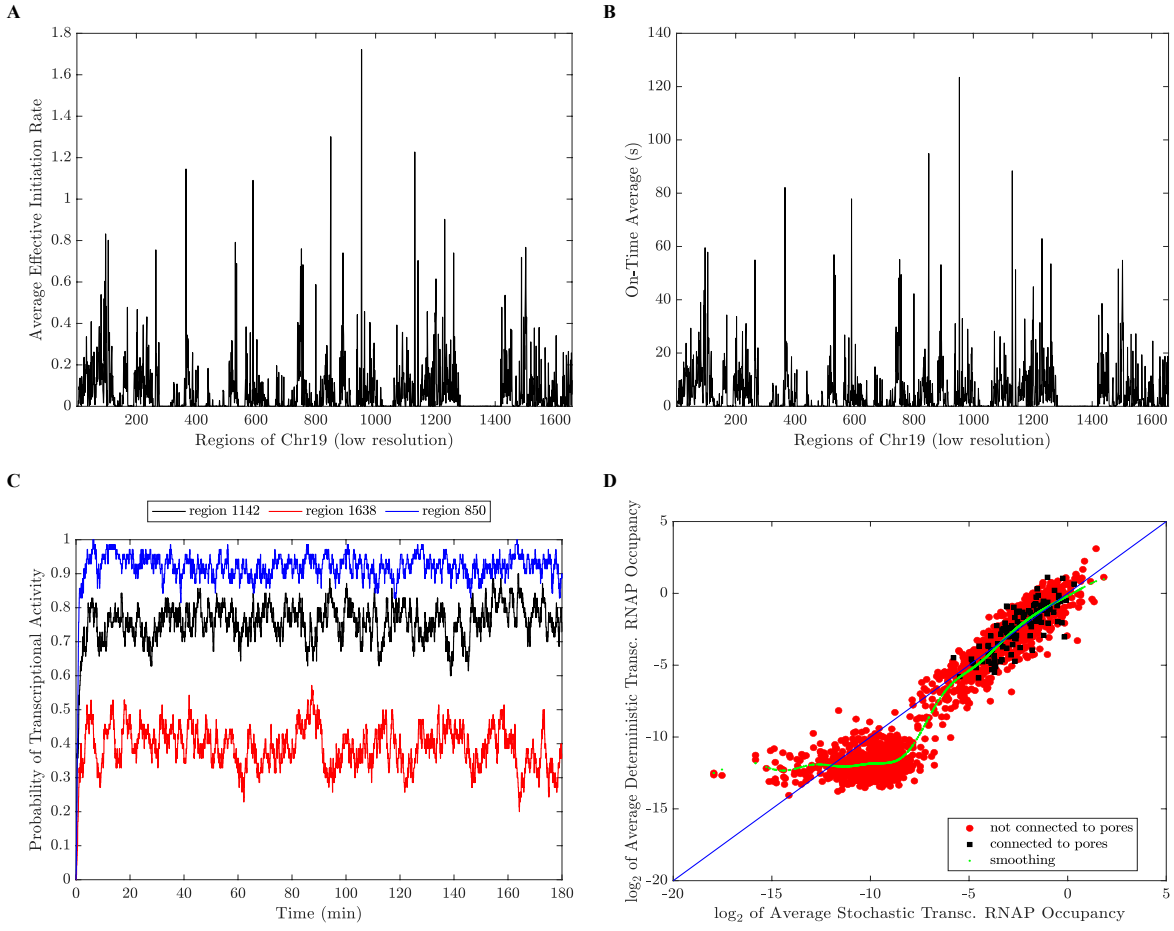

Figure 3.11 – **Behaviours for our model stochastic implementation. A** Effective Initiation Rate. **B** On-time Average (s). **C** Probability of Transcriptional Activity for three different regions. **D** Log-log plot of average of stochastic solution over deterministic solutions for connected to pores (black squares) and all the regions not connected to pores (red circles). In blue, we have the identity function and in lime green, we have the smooth function for our system.

Given Fig. 3.10, we see at least two patterns: (i) active (probability of being active is 50% or higher) or (ii) inactive (otherwise). Besides, our deterministic solution presents continuous values and our stochastic implementations have discrete expressions: meaning while deterministically it is possible to create a split evenly between two regions of the TF/RNAP complexes, it is impossible stochastically, forcing our system to choose between those regions. Yet, we did not uncover any tendency of pore connected regions to be more enriched than active regions not connected to pores, meaning the pore connectivity is not increasing activity.

So, we compared in Fig. 3.11 **D** the $\log_2$ of the averages of stochastic and deterministic solutions for all the regions given two different subsets: (1) **connected to pores**, the black squares, and (2) **not connected to pores**, the red circles. Here, we decided to consider all the regions not only the active ones, but the separation between active/inactive regions is easily distinguishable by the two separate clouds in the regions not connected to pores, the bottom one being the inactive regions (by the τ_i values). In the other cloud, we have the active regions and we checked how enriched and impoverished the regions get between the solutions, as the smooth function (lime green dots) split in Fig. 3.11 **D**.

This model showed the structure and the region's promiscuity influence the transcription. However, the translocation process in this model does not consider depleting the nuclear TF concentration and limiting resources impact the transcription as a whole. Thus, we propose next the simulations of our model in Eq. (3.3) with the flux functions in Eq. (3.11).

3.2.3.2 Import/Export Flux Function

We implement our numerical solution for the import/export flux function in Eq. (3.11), the model in Eq. (3.3), with the parameters in Table 2. Differently from Eq. (3.5), this new flux function is explicitly time-dependent. Because of this, implementing the classic version of the Gillespie Algorithm is not optimal. Thus, we implemented a hybrid Gillespie Algorithm (Vestergaard; Génois, 2015), by integrating the reactions to determine the following reaction and time.

More so, our model describes a TF translocation and NF-κb is a good candidate for this, since it is a well-studied TF in cancer pathways with well-defined translocation dynamics (Trask, 2012; Noursadeghi et al., 2008; Zambrano et al., 2020). Using the data available in (Zambrano et al., 2020), we fitted the parameters k_{im}, μ, δ and T_{total} from Eq. (3.11), and the results are available in Table 3, we present the TF flux dynamics in Fig. 3.12.

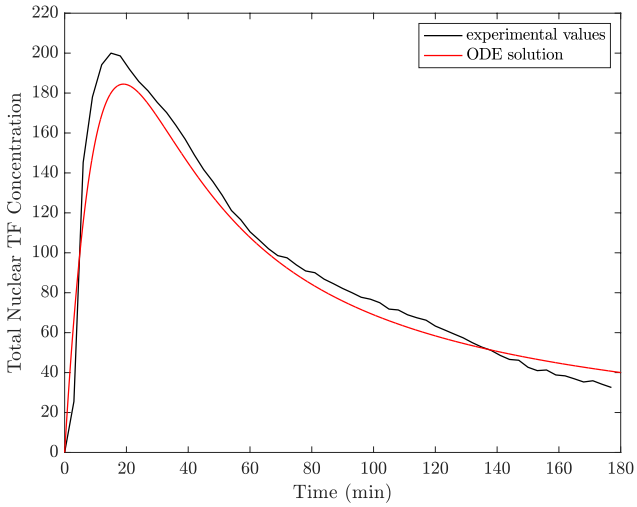

Figure 3.12 – **Fitting for Nuclear TF Concentration from Experimental Data and our flux function in Eq. (3.11).** From this, we can see the three steps for TF translocation: (1) The import begins, $(0 \leq t < 20)$; (2) The maximum is reached $t = 20$; and (3) The export occurs, $20 < t \leq 180$.

With all the parameters set, we implemented our deterministic and stochastic numerical solutions. First, we solved the deterministic solution and we calculated the averages concentrations for all the complex states and all regions, Fig. 3.13 **A**, showing that most of the TFs remain free, with a small peak of Bound states around the 20 minutes, following the experimental peak presented in Fig. 3.12. After this peak in concentration, the TF concentration decreases because of the cytoplasmic reaccumulation process.

When analysing the RNAP dynamics, Fig. 3.13 **A** presents a fast drop in the Free RNAP state and an increase in Transcribing RNAP with the same intensity with strong deactivation of Transcribing RNAP after the 20 minutes mark but delayed in comparison with TF states because of the whole transcription machinery. The Bound RNAP state also presents a slight overshoot around the same time but remains stable even after the re-accumulation dynamics start, which represents an intermediate state between being free and effectively transcribing.

We analyze the behaviours of the two active subnetworks: connected to a nuclear pore, C, and not connected, nC. We present average solutions for the nC subnetwork in Fig. 3.13 **B** and the average solution for the regions of subnetwork C in Fig. 3.13 **C**. Since both subnetworks represent active regions and the protein complexes are fast, the activation of RNAPs (from Free to Bound/Transcribing) is not delayed by being not connected to a nuclear pore, and we have similar behaviours for both Fig. 3.13 **B** and **C**. Here, we consider as activation of the RNAP complex as the presence of the Bound and Transcribing states and its deactivation as the decreasing Transcribing RNAP concentration.

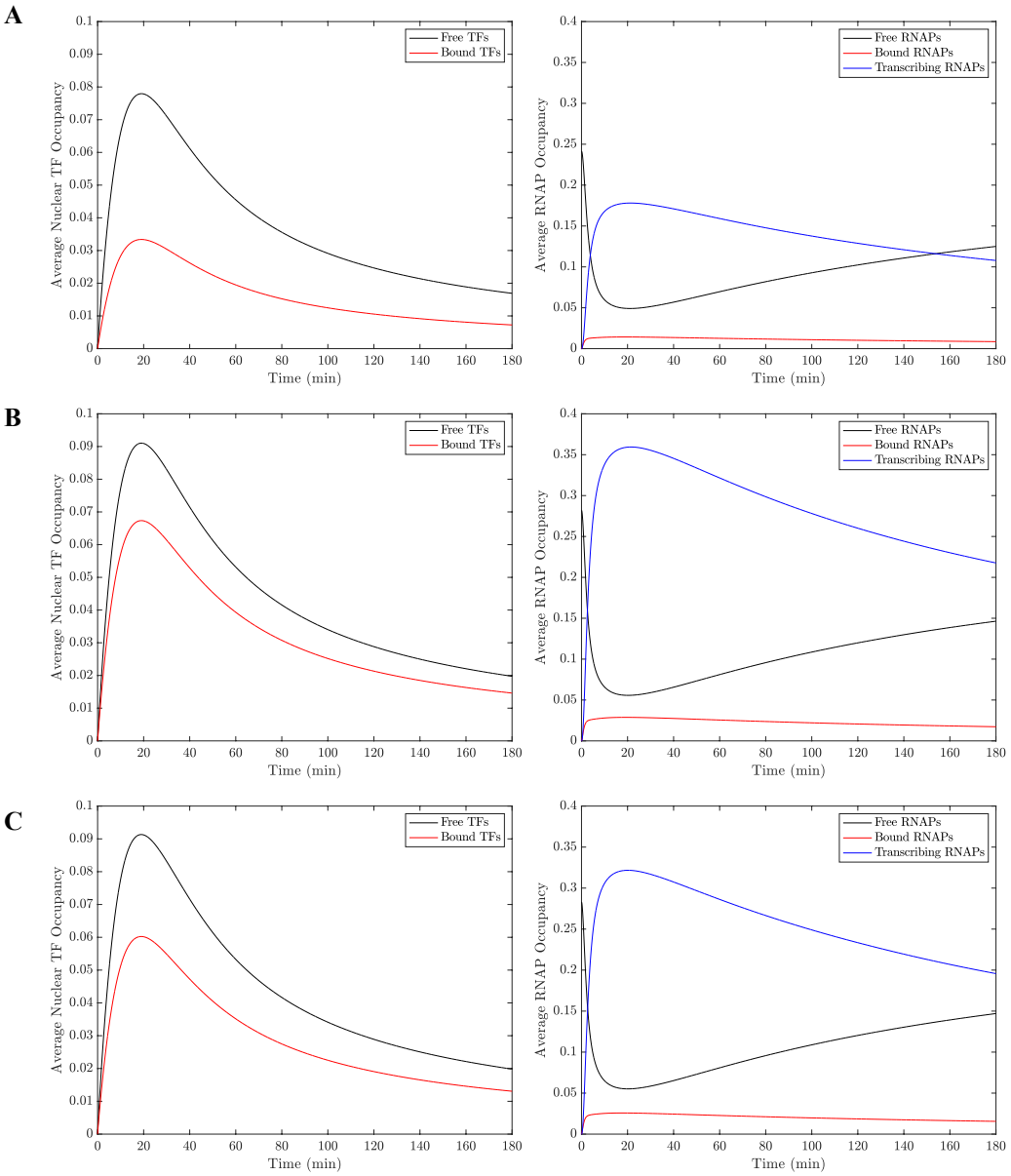

Figure 3.13 – **Average State Occupation in three different subnetworks for the flux dynamics in** (3.11). **A** All Regions with the average TF Occupancy and the average RNAP Occupancy; **B** Active Regions **not** connected to a pore; **C** Regions connected to pores.

Still, the activation was more robust in those regions close to nuclear pores, for our model with import flux function Fig. 3.8. Thus, we verified the global behaviour over our time interval in Fig. 3.14, focusing on how the non-constant transcriptional resources influence transcription, since we analyzed the Transcribing RNAP solution. In Fig. 3.14 **A**, we present all nodes in our network and we can see that the inactivity is more prominent in this flux function than in Eq. (3.5). We can predict three primary behaviours: (i) inactive regions, (ii) regions with fast activation/inactivation, and (iii) regions with fast activation that remain active. Note that our network does not reactivate (increase in concentration) after the deactivation starts. These behaviours can be easily verified in Figs. 3.14 **B** and

C, in which we analyze the active subnetworks, nC and C. And in those subfigures, we see how the exportation wave affects transcription over time and the intensity of this process.

We uncovered for our previous flux function that the closeness to a nuclear pore plays a role in transcription. So we calculated how many steps a region is from a nuclear pore and represented these subnetworks in Figs. 3.14 **D** and **E**. Looking closely, we demonstrated how being close to the nuclear pore increases transcription, with few exceptions.

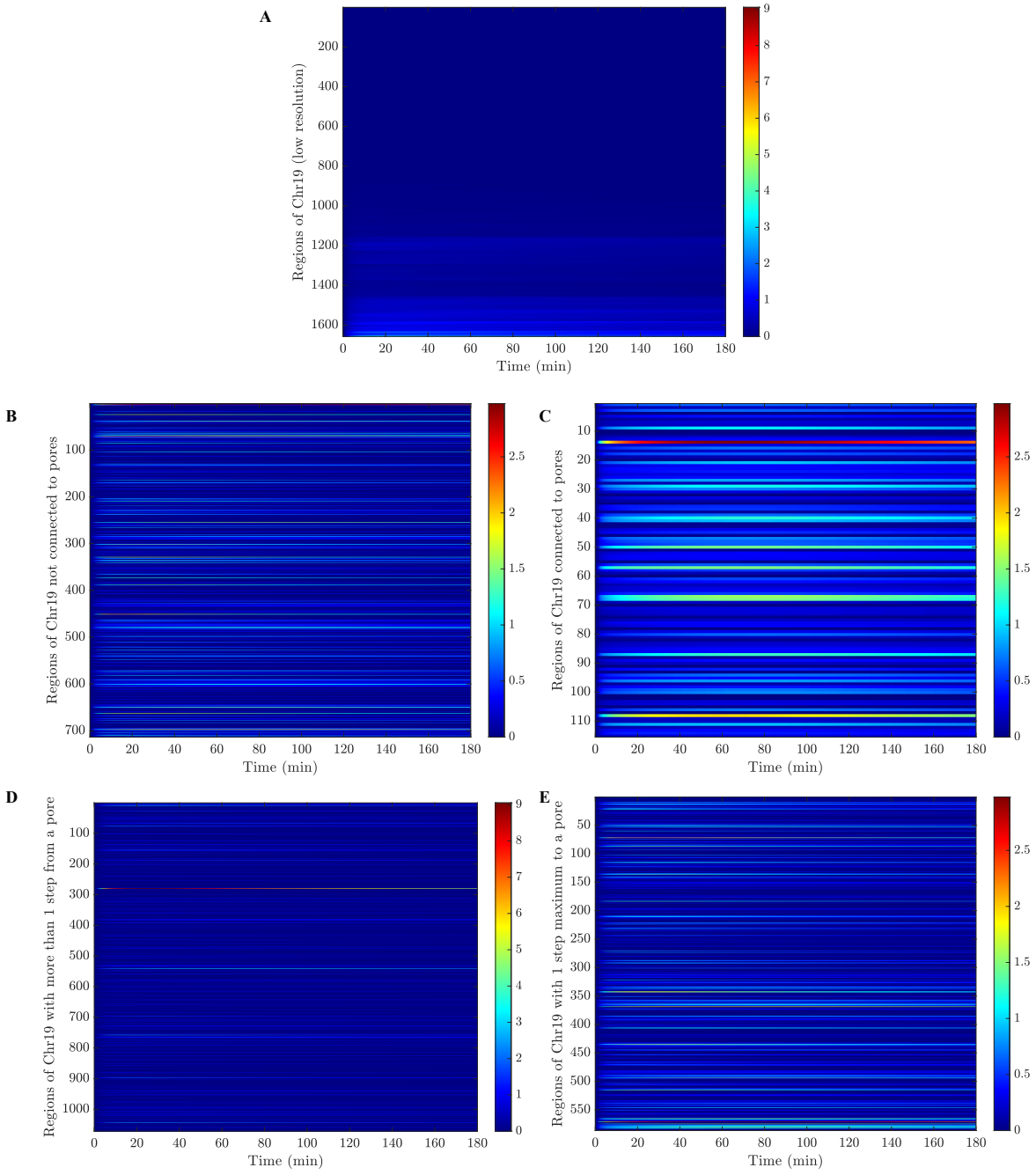

Figure 3.14 – **Heatmaps for our Import/Export Flux Function (Eq. (3.11)) for different subnetworks. A** All chromatin regions; **B** Active regions not connected to pores, nC; **C** Regions connected to pores, C; **D** Regions with more than 1 step away from a nuclear pore; and **E** Regions with 1 step maximum from a nuclear pore.

To understand how the Transcribing RNAP solutions behave, we clustered our solutions in Fig. 3.15 **A**, where we obtained overshoots around the TF maximum concentration ($t = 20$ minutes), without delays for all the clusters. Once the exportation process starts, we will have the deactivation process - i.e., the concentration of Transcribing RNAP decreases, which is a direct consequence of the TF nuclear concentration decreasing - and this deactivation pattern on average follows the same behaviour for all regions and this means the loss of available TFs affects the regions in similar ways. Another interesting result is the absence of reactivation which is a result consistent with the ones in Figs. 3.14.

The results in Fig. 3.13 represent average concentrations for the protein complex and we need to check how far our solutions deviate from the average, we calculated the z-score for our Transcribing RNAP and clustered the solutions in Fig. 3.15 **B**. In this subfigure, we verified how the regions behave, with the regions reaching their maximum in different time points (but no later than the 80 minutes mark), but we can see that the activation is not delayed anywhere. Interestingly, the deactivation pattern is similar for all regions.

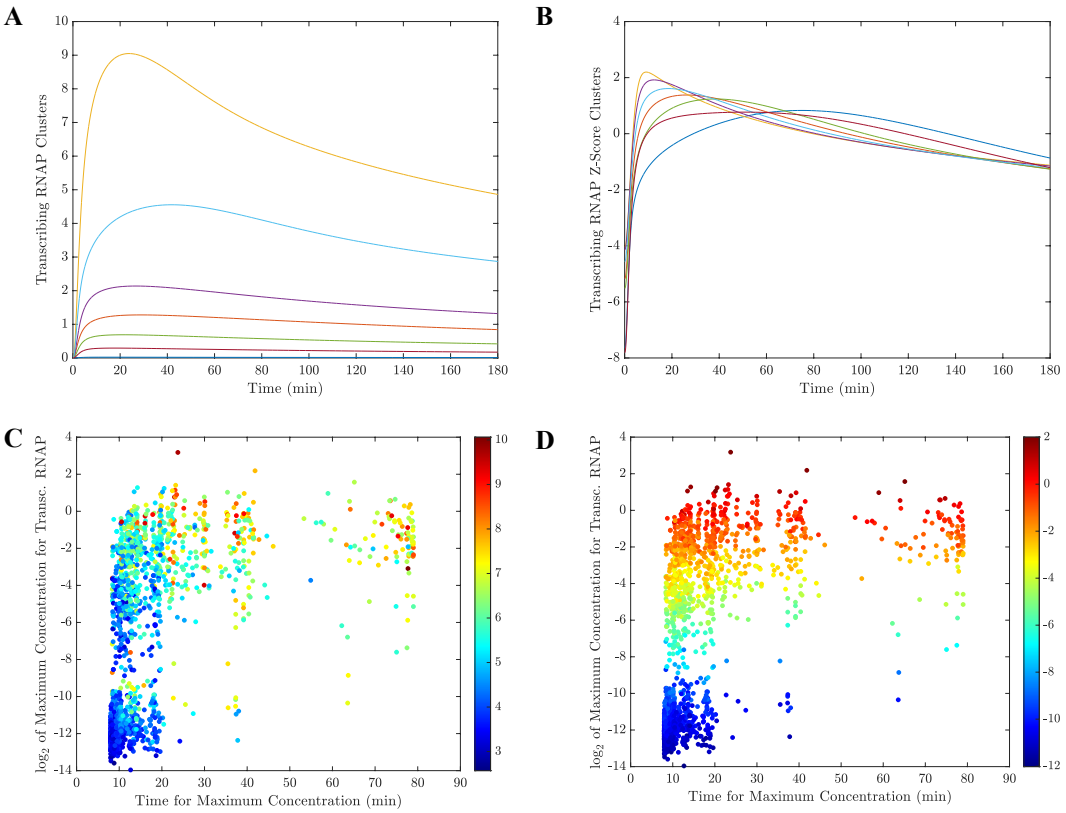

Figure 3.15 – **Behaviours for our Deterministic Solution. A** is the clusters for the Transcribing RNAP, with 7 different clustered averages. **B** The clustered z-score for our Transcribing RNAP solutions. **C-D** Time for reaching the Maximum concentration over the log_2 of the Maximum Transcribing RNAP concentration sorted by the number of connections (d) of the region, **C**, and the $\log_2$ the residence times, $\log_2(\tau)$, **D**.

To verify this prediction of maximum values over time, we propose Figs. 3.15 **C** and **D**, in which we plot the $\log_2$ of maximum Transcribing RNAP values over the time spent to reach such concentration sorted by the number of connections of the network, **C**, and the $\log_2$ of the residence times, **D**. We can see the lowly connected regions reach their maximum earlier than highly connected regions (Fig. 3.15 **C**) and how the values of d influence the concentration, a result predicted in the steady-states (Eqs. (3.4)) even if our model does not stay in equilibrium.

However, the residence time is an active player in import/export dynamics and transcriptional activity and our system relates strongly to higher values of τ, as shown in Fig. 3.15 **D**, which is more stratified by the values than d. With this result, we can affirm that active histone marks and good binding motifs might affect the concentration levels of the transcriptional machinery but they do not impose the system to be reached early. Besides, most of the regions attain their maximum until the 30 minutes mark and no earlier than the 10 minutes mark.

The deterministic solutions proved how fast the exploration mechanism for TF/RNAP molecules is, how this affects the transcription activation, and how the parameters play a role in transcription. Likewise, the depleting level of TFs inside the nucleus shows how the deactivation gradually occurs. However, we still need to consider the stochasticity in our model; after all, transcription is a stochastic process so next, we present our Hybrid-Gillespie Algorithm.

Since the exporter reaction is time-dependent, we implemented an adapted Gillespie Algorithm. In this case, we integrate the exporter function and calculate the next time, t_{i+1}, from the sum of the reactions at time t_i. This algorithm is implemented 70 times, i.e., we generated a set of 70 cells to analyze, in which we verified how our stochastic solutions correlate with the deterministic one in Fig. 3.13. Even so, our algorithm recovers the Nuclear TF Concentration from the flux function considered which we present in Fig. 3.16.

Considering the Transcribing RNAP state describes the transcription process in our model, we focus our analysis on this state to understand how the activity levels change in a system with a translocation/reaccumulation process. Our flux in this model is non-constant and we have three different stages concerning the transcription process similar to a response to a proinflammatory trigger (Meier-Soelch et al., 2021): (i) Fast increasing of nuclear TF concentration; (ii) Maximum concentration inside the nucleus, $T_{max} = [T]$; and (iii) Fast decreasing TF concentration.

These three steps are present in Fig. 3.16 and the transcription activity pattern follows these three stages: first, transcription increases proportional to the number of available TFs, up to its maximum. Then, we see a reduction in transcriptional activity due to the depletion of resources.

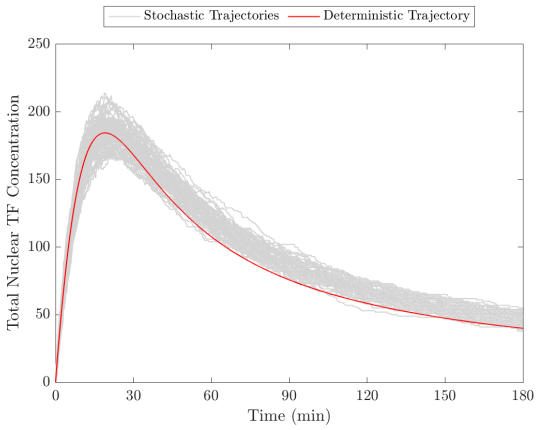

Figure 3.16 – **Comparison between our deterministic and stochastic solutions for the Nuclear TF Concentration.** Proving our Hybrid-Gillespie simulation on average recovers the same concentration behaviour as a deterministic solution.

We analyze the fraction of active cells for our model over time considering different subnetworks and sorting by the main parameters for gene expression, d and τ in Fig. 3.17. The subnetworks we considered are the same ones from Fig. 3.14, i.e., in **A**, we have our whole network; in **B**, the active regions not connected to pores; in **C**, the regions connected to pores; in **D**, the regions in **C** and their immediate neighbours; in **E**, regions with more than one step away from the pore. In Fig. 3.17 A, the fraction of activity for all the network shows that at least half of our network is inactive through all our simulations, i.e., the probability $p(i)$ of the region i being active is smaller than 20%. Then, 2/3 of the remaining regions represent what we can call middle-range active, i.e., $20\% \leq p(i) < 50\%$. High-active regions correspond to the rest of our system, $p(i) \geq 50\%$. Again, the residence times correlate with the activity more than the connectivity. Overall, the regions present maximum activity of around 20 minutes before starting their deactivation process.

We verified how the translocation affects the active regions in Figs. 3.17 **B** and **C**, with the subnetworks nC and C, respectively. In Fig. 3.17 **B**, we again see the importance of τ in our system activity and also that more than 1/3 of those regions present an inactive behaviour, for example and it is easily identifiable in column **(ii)**. For the regions connected to nuclear pores, Fig. 3.17 **C**, we can see that while the activity is higher in those regions there is no correlation between parameter and activity. This is a consequence of these particular regions not having to attract TFs to themselves, because in the import and export processes they receive Free TFs, which bind them because they are by definition prolific regions, and then they can recruit RNAPs to start transcription. Figs. 3.17 **D** and **E** show that while we do not have a huge difference in numbers of active/inactive regions for these complementary subnetworks, we verified that regions with more than 1 step away from a nuclear pore deactivate faster and also regions close to a pore show more regions in a high probability of being active.

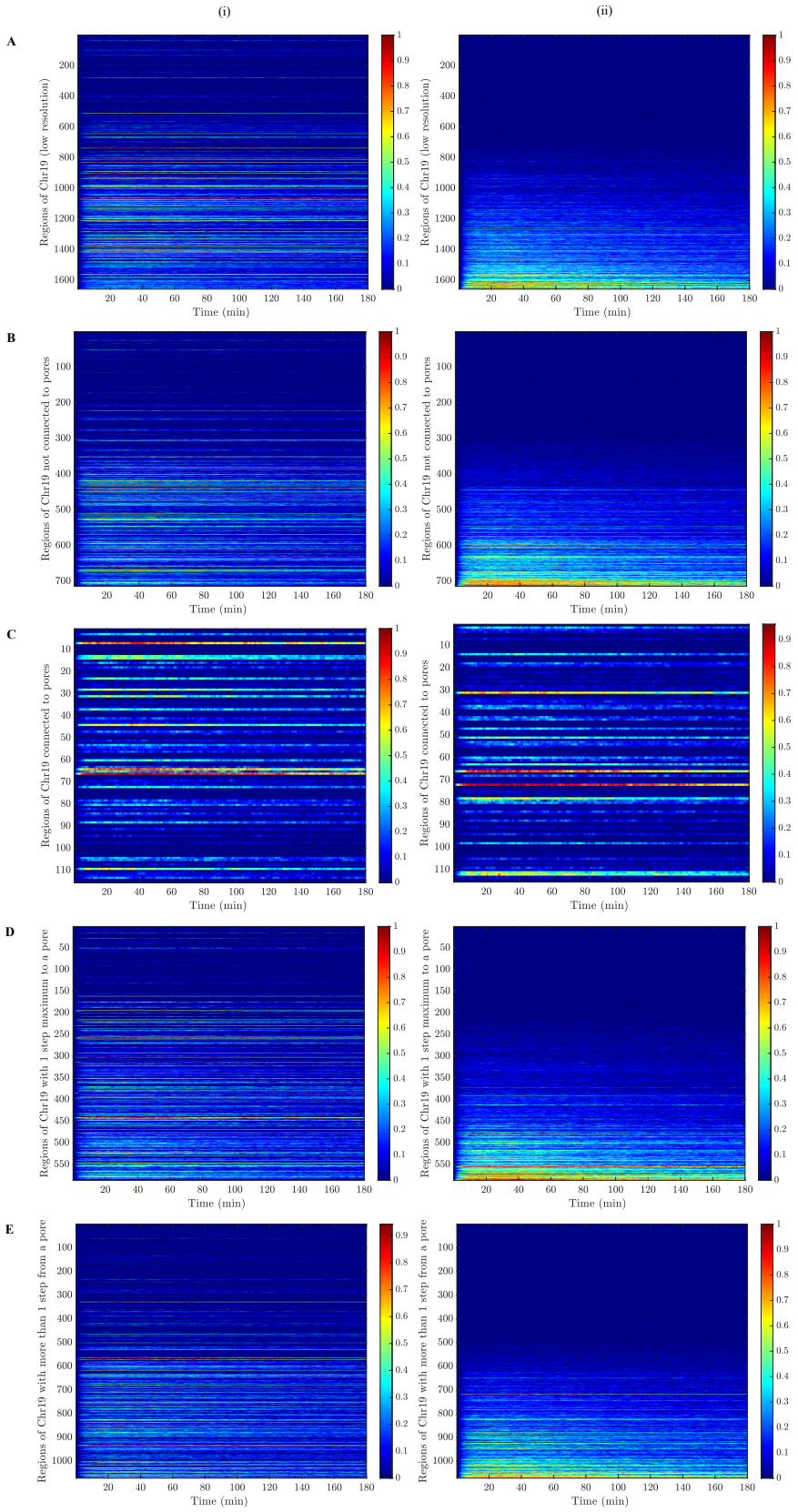

Figure 3.17 – **Fraction of Active Transcription for our Flux Function in Eq. (3.11) sorted by (i) number of connection, d and (ii) residence times, τ. A** All Chromatin Networks. **B** Active Regions not connected to pores. **C** Regions connected to pores. **D** Regions with a maximum of 1 step away from a nuclear pore. **E** Regions with more than 1 step away from a pore.

We present global behaviours for our stochastic implementations in Fig. 3.18 **A** and **B**, as a way to verify how limiting resources influence transcription, if we compare with the results in Fig. 3.11 **A** and **B**. In Fig. 3.18 **A**, we evaluate the Average Effective Initiation Rate profile to verify how many initiations per minute we have in our model, which decreased in comparison with Fig. 3.11 **A**, which only one region has one initiation per minute, which is a consequence of the smaller concentration of TFs available in this model: after 60 minutes, we have less than half of the maximum TF concentration for our model Figs. 3.12 and 3.16. Of course, the activation is a consequence of these numbers, proving how an ongoing limitation of resources affects the transcription volume. In Fig. 3.18 **B**, we calculate the average on-time interval between the initiation and the elongation reactions for all the regions and the time does not change much between the two flux functions since we considered the same elongation time for both, being less than 2 minutes on average, i.e., transcription is a fast process.

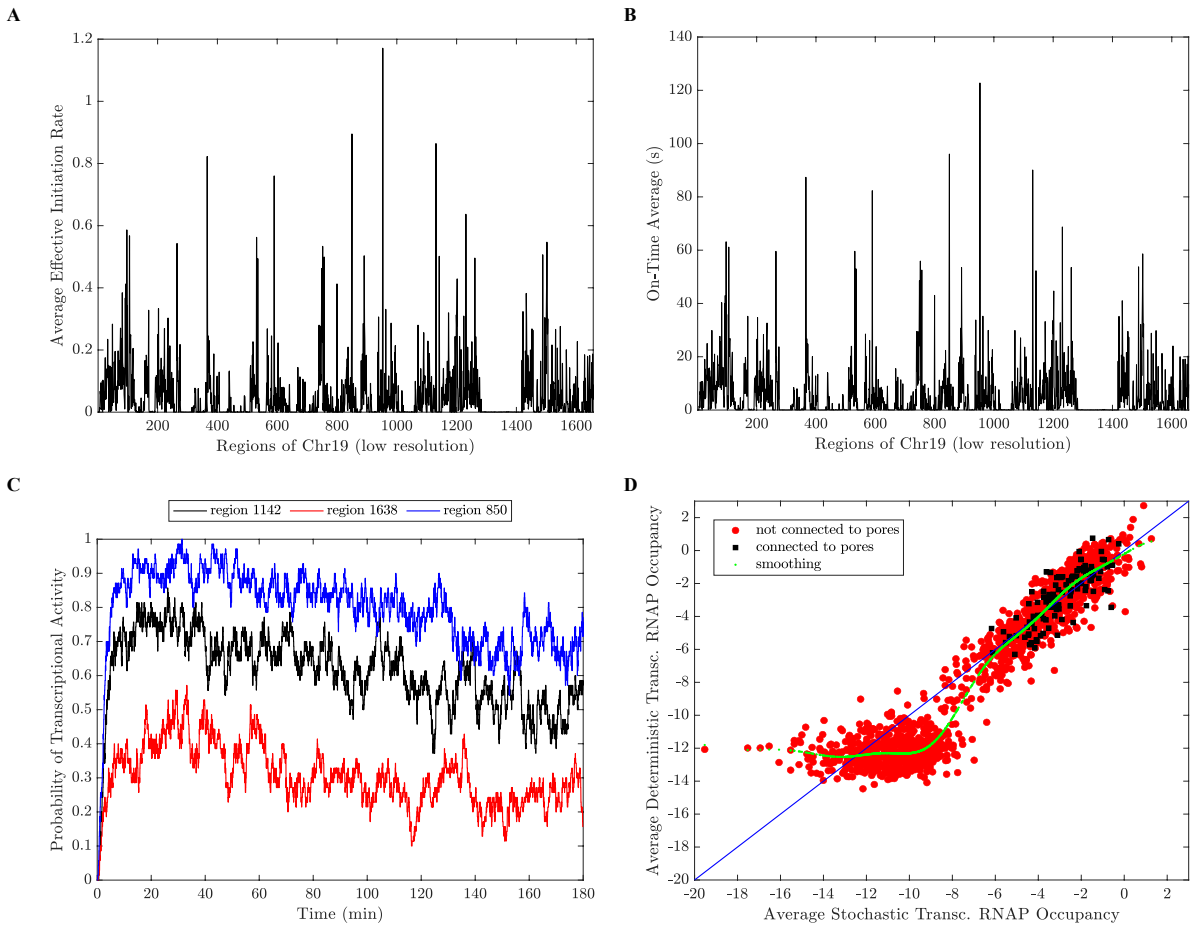

Figure 3.18 – **Behaviours for the stochastic implementation of our model considering the flux function in Eq. (3.11). A** Effective Initiation Rate. **B** On-time Average (s). **C** Probability of Transcriptional Activity for three different regions. **D** Log-log plot of average of stochastic solution over deterministic solutions for connected to pores (black squares) and all the regions not connected to pores (red circles). In blue, we have the identity function and in lime green, we have the smooth function for our system.

Therefore, our model proves that the sequence motif and active histone marks are fundamental to transcription, mostly for regions not connected to pores. As experimental results have shown, some genes (in our model, chromatin regions) are more likely to transcribe than others (Zambrano et al., 2020), but experiments only can show the specific genes studied and our model can predict a continuous array of different transcriptional activities, Fig. 3.17. To facilitate the visualisation of some activity probability, we have Fig. 3.18 **C** which we can see that the decreasing TF concentration affects transcription, as expected.

The short activity periods and strong deactivation from our simulations guarantee that while some regions will have enriched transcription, many others will present smaller-than-expected activity. We may not call this pattern explicitly upregulation/downregulation in the strict biological sense as this behaviour is a consequence of some regions having more transcriptional resources, hoarding the machinery impairing other regions instead of being a direct consequence of the transcribed genes. We decided to call the consequence of the lack of resources for all regions favouring a few specific nodes of our network an enrichment/impoverishment pattern. We verify the stochasticity effects by comparing the average stochastic with the deterministic solution in Fig. 3.18 **D** for regions connected to pores (black squares) and regions not connected to pores (red circles). The separation between active and inactive regions can be seen by the two different clouds of values. We can see by the smoothing (lime green points) of our results how some regions get enriched or impoverished.

Our model predicts the presence of flux input affects the transcription, and the volume of transcription decreases proportionally to the nuclear concentration of TFs which is derived from experimental results. The limited resources autoregulate themselves, enforcing different activity levels in all the target sites.

We have experimental results to verify the existence of pore-placed regions, i.e., regions near the nuclear envelope and we analyze the microscopy data in Chapter 5. Another point is that we only considered transcription in this Chapter without evaluating the mRNA concentration each region might synthesize. Since the central dogma of molecular biology states two processes: (1) Transcription and (2) Translation, next chapter, we present an extension of our transcription model to consider the changes in mRNA concentration per gene and how it affects the eventual translation process.

4 Incorporating mRNA export into gene regulation and RNA Velocity model

In the previous chapter, we proposed a model for transcription and showed how some regions are more prolific than others. However, our model did not consider the volume of transcription each region can achieve, i.e., while we could predict how promiscuous a node in our network and where the transcription is more likely to occur, we did not show the mRNA concentrations in the model, and it is clear the mRNA cytoplasmic concentration is one of the mechanisms behind translation regulation.

As discussed previously, the central dogma of molecular biology, we have two processes for protein production: (i) **Transcription**, where the mRNA is produced, and (ii) **Translation**, where this mRNA is coded into a gene by a ribosome (Carmody; Wente, 2009; Livingstone et al., 2010; Grünwald; Singer; Rout, 2011; Cobb, 2017). The presence of the nuclear envelope in eukaryotes is a fail-safe mechanism to separate transcription (nucleus) from the translation (cytoplasm), being a fundamental element for gene regulation (Magistris, 2021; Vargas et al., 2005; Wickramasinghe; Laskey, 2015). We present a simplified schematics for this separation between Transcription and Translation in Fig. 4.1.

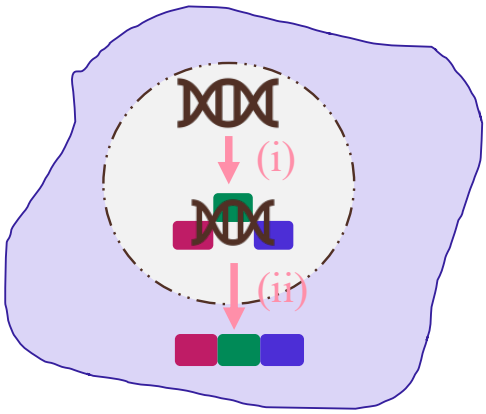

Figure 4.1 – **Cartoon Representation for Gene Expression Mechanisms.** Here, a DNA strand is transcribed in step (i) and translated in step (ii).

Previous results predicted the TF clustering and the RNAP transcription activation pattern, Chapters 2 and 3 respectively. Both models showed how chromatin connectivity and accessibility create differences in promiscuity and TF/RNAP allocation. Yet, we did not consider the actual volume of the mRNA produced per region. Thus, explicitly incorporating the mRNA in our mathematical model creates a more mechanistic way to understand those biological processes and predict *in silico* the concentration of each protein from our network. We proposed two different mRNA export models: (i) Incorporating

the mRNA export in our model from Chapter 3 (Eq. (3.3) considering the translocation function from Eq. (3.11)) and (ii) an ODE model for unspliced/spliced mRNA export based on the RNA velocity model (Manno et al., 2018) which we split the spliced state between nuclear-spliced mRNA and cytoplasmic spliced mRNA, in a collaboration with Mendoza's lab from IGBMC.

4.1 Incorporating mRNA export to our translocation/reaccumulation flux model from Chapter 3

Given our model with import/export flux function in Eq. (3.11), we proposed the Transcribing RNAP in a node i produces a nuclear mRNA$_i$ that diffuses with an effective diffusive rate, k^r_{3D} through the network using the same probability of movement of TF/RNAP, until it finds a pore-connected region to be exported, becoming a cytoplasmic mRNA$_i$. We represent our mRNA export dynamics in Fig. 4.2.

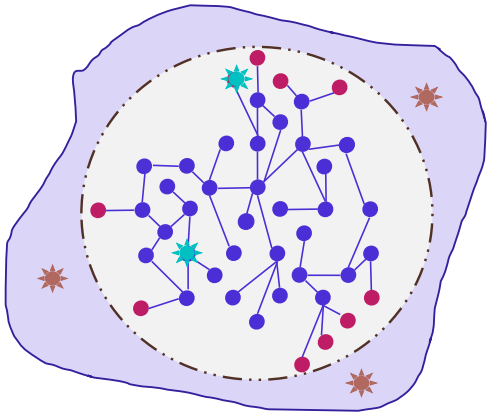

Figure 4.2 – **Cartoon Representation for mRNA exportation**. Where nuclear mRNAs are represented as the green suns and the cytoplasmic ones brown suns and we used the same network and pore connectivity from Chapter 3, with the equations from Eq. (4.1).

From Fig. 4.2 and Eqs. (3.3) and (3.11), we incorporate the Eqs. (4.1) to our TF/RNAP model. We used the variable m to represent the mRNA, δ_{ig} is the Dirac-delta for gene g in the region i, and K is the number of nuclear pores. Here, a nuclear mRNA (m^N_{gi}) that generates gene g in the region i is produced after the elongation process for the Transcribing RNAP and it is region-specific. This nuclear mRNA explores our network, being exported to the cytoplasm, entering the cytoplasmic state, m^C_g, which can be degraded with a rate γ. Here, we assume the mRNAs from all the regions have the same parameters.

$$\begin{cases} \dfrac{dm^N_{gi}}{dt} = -k^r_{3D} m^N_{gi} + \sum_j k^r_{3D} M_{i\leftarrow j} m^N_{gj} + \delta_{ig} k_\varepsilon P^T_i - (k^r_{3D} K) K^i_X m^N_{gi} \ ; \\ \dfrac{dm^C_g}{dt} = \sum_i k^r_{3D} K^i_X m^N_{gi} - \gamma m^C_g . \end{cases} \tag{4.1}$$

As this model is an extension of our previous model from the Chapter 3, we know the analytical solutions are determined by a Magnus Expansion and are represented by a series of integrals, we decided to omit the analysis for this model and only proposed the deterministic numerical solutions using ode15s from Matlab (Gupta; Wallace, 1975) for all genes g in our network, i.e., we implemented L times our model. We did not implement stochastic simulations as the number of cells necessary to obtain a significant result is linearly dependent on L.

4.1.1 Deterministic Solution for our model

We implemented the deterministic solution considering the parameters from Tables 2, 3 and 4. We present the global effects for different mRNAs produced in Fig. 4.3 - i.e., we verified how different regions have different mRNA concentration levels. This also shows the regions each mRNA moves - proving the preference of moving inside the TAD and visiting the nearest neighbours.

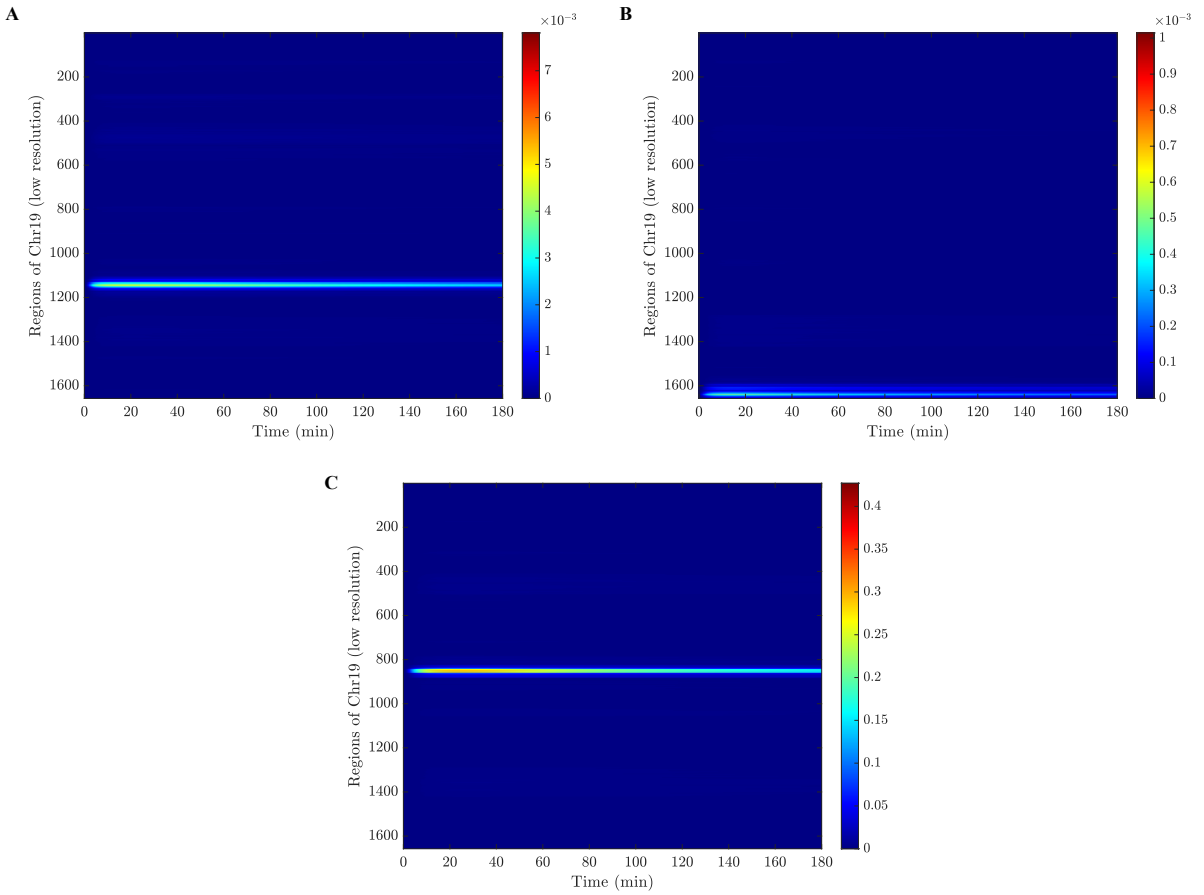

Figure 4.3 – **Nuclear mRNA concentration for 3 different regions of our network.** Here, we considered the same regions from Figs. 3.11 and 3.18 **C**. Where **A** is region 1142, **B** 1638 and **C** 850. The concentration changes per mRNA type at different levels.

We are using the flux function that represents the NF-κb translocation process,

and we verified the spike in concentration around 20 to 40 minutes and a decrease in mRNA concentration as the nuclear TF concentration decreases. The reason behind the mRNA movement through the nearest neighbours is explained by the fact those regions are more likely to be connected. We also remember we considered mRNAs to be slower than RNAPs, so they diffuse slower. The average nuclear concentration for those regions can be found in Fig. 4.4 and we can see how some regions have a more strong production of mRNA than others.

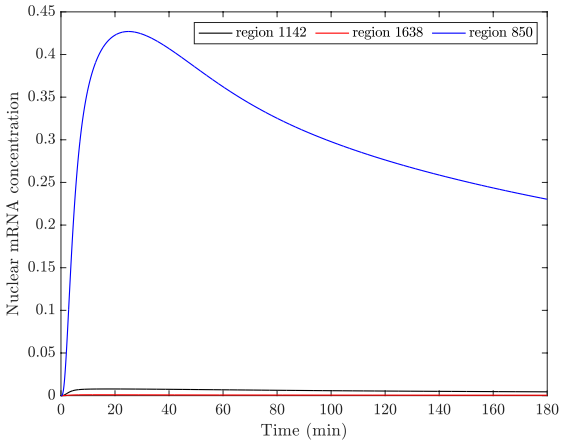

Figure 4.4 – **Average Nuclear mRNA concentration for the different regions of our network from Fig. 4.3.** The volume of mRNA produced changes given the promiscuity of a node.

In Chapter 3, we showed how the Transcribing RNAP has different occupancy patterns for different subnetworks, and how being close to a pore increases the chances of being an active region even if the TF/RNAP complexes diffuse fast and we proved the absence of late activation/reactivation. Since mRNA is modelled as a product of the elongation process in a region i, we verified in Fig. 4.5 the occupancy of the mRNA from region i in heatmaps for different subnetworks: **A** all chromatin regions; **B** active regions not connected to pores; **C** regions connected to pores; **D** Regions with more than 1 step from a nuclear pore; **E** Regions connected to pores and their immediate neighbours. Here, we only considered the concentration in all the regions per gene even if mRNA is moving in our network since we want to understand the mRNA production and how the network affects it. In Fig. 4.5 **A** we show the concentration of nuclear mRNA is low in most of the regions (> 1000 different regions show $[m_i^N] < 0.1$), but the colour bar shows we have regions with higher concentrations of mRNA we do not see in Fig. 4.5 **A** because we have few regions with higher concentrations not identified.

We verified this is not the case in Fig. 4.5 **B** and **C**, where we studied the subnetworks of active regions not connected and connected to pores, respectively. Those active regions show the expected heterogeneity of mRNA production and also the peaks of production around the peak of transcriptional resources and a decrease in consequence

of limiting TF concentrations. Our heatmap proves how the residence times increase the volume of transcription per region.

Again, to check how being close to a pore affects transcription, we propose the last subnetworks: **D** Regions with more than 1 step away from a nuclear pore and **E** Regions with 1 step maximum from a pore in Fig. 4.5. From those two complementary subnetworks, we can see how the proximity to a nuclear pore increases the chances of a region producing more mRNA, with few exceptions. From all subnetworks present in Fig. 4.5, we can see how the activation/deactivation process of the TF affects the transcription machinery.

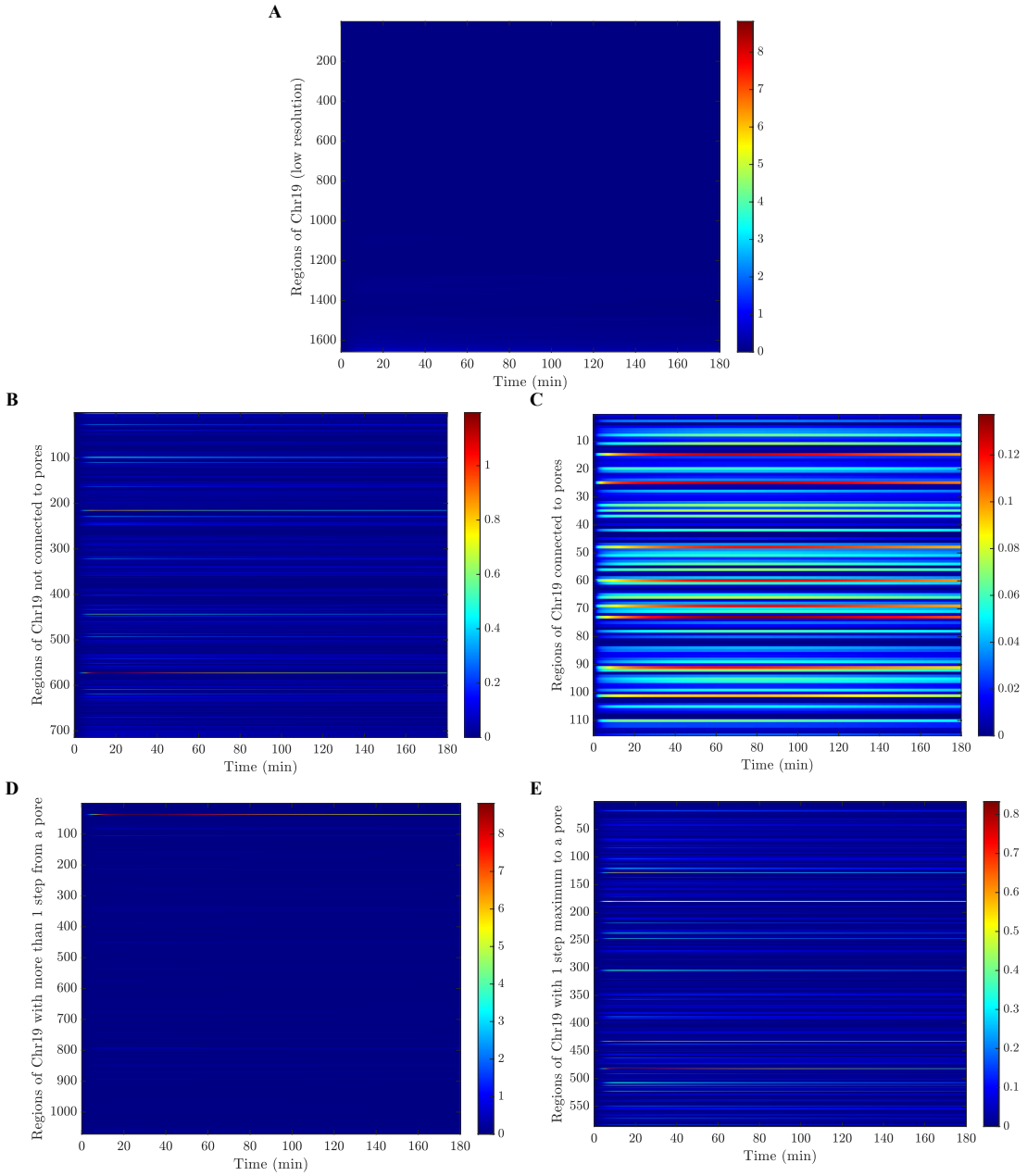

Figure 4.5 – **Heatmap for Nuclear mRNA Concentrations for different subnetworks. A** All regions. **B** Active regions not connected to pores. **C** Regions connected to nuclear pores. **D** Regions with more than 1 step way from a pore. **E** Regions with 1 step maximum from a pore.

The global view of deterministic behaviours can difficult the understand intrinsic patterns from our model. Therefore, we proposed clustered solutions to help us with the visualization of changes in the nuclear mRNA concentration in Fig. 4.6: in **A** the averaged cluster of nuclear mRNA solutions and **B** the cluster of the z-score of the variable.

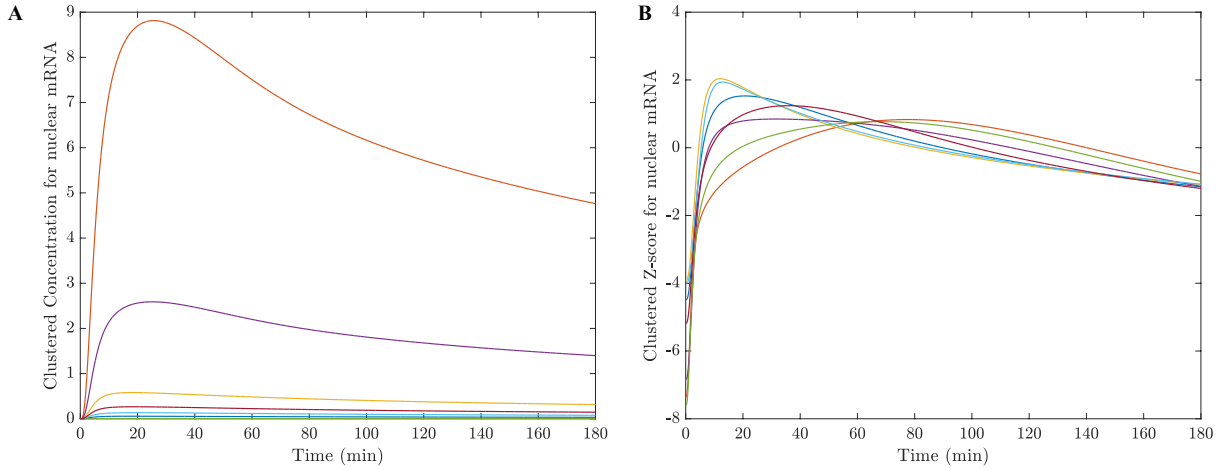

Figure 4.6 – **Cluster Analysis for the nuclear mRNA concentration. A** Clustered values for the nuclear mRNA concentration. **B** Clustered z-score of the nuclear mRNA.

Analyzing Fig. 4.6 **A**, we can see highly-expressed genes grouped with a maximum before the 30-minute mark and regions with somewhat stable production of mRNA over time in different levels of concentration. Similar to previous studies in Chapter 3, we proposed seven different clusters. To verify how skewed the average mRNA concentrations per region are, we calculated the z-score and clustered the results in Fig. 4.6 **B**. We can see how the nuclear mRNA z-score correlates with the Transcribing RNAP in Fig. 3.15 **B** in Chapter 3, and we can see the mRNA production follows the Transcribing RNAP behaviour, as expected.

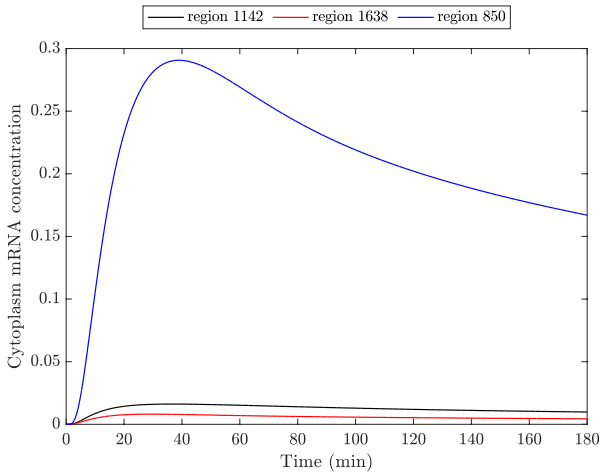

Figure 4.7 – **Concentration of Cytoplasmic mRNA for different regions of our network, the ones presented in Fig. 4.4.** We can see the delay in accumulation from nuclear to cytoplasmic mRNAs.

This analysis only follows the nuclear mRNA concentration and one of the fundamental functions of mRNA is to be exported to the cytoplasm, starting translation. Thus, we analyzed the concentration in Fig. 4.7, for the same regions presented in Figs. 4.3 and 4.4. The concentration reaches the maximum after the reaccumulation process started, as shown in the previous chapter. This means there is a delay between the mRNA produced in the nucleus and its accumulation in the cytoplasm because of the degradation of cytoplasmic mRNA and the diffusivity of nuclear mRNA.

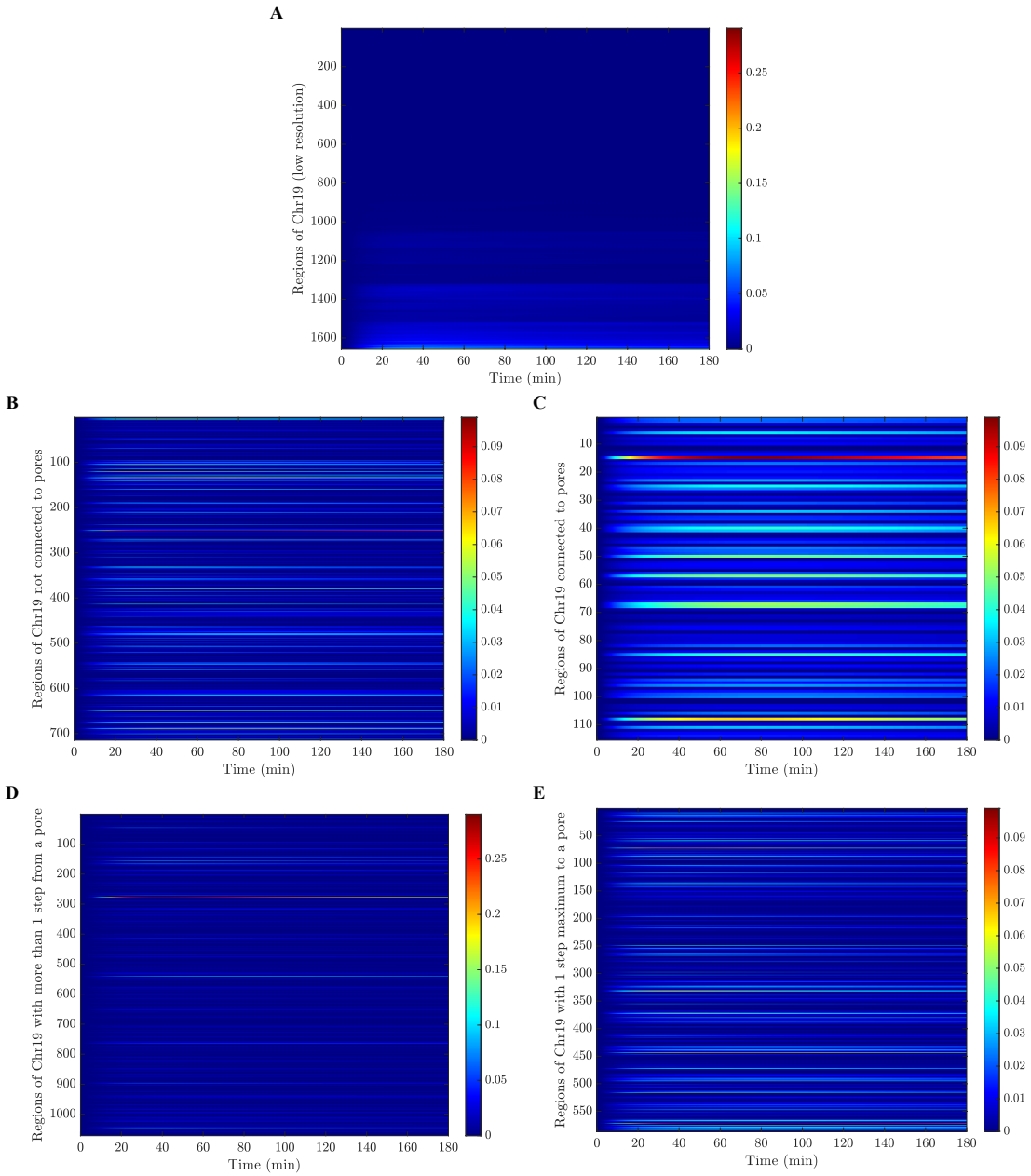

Figure 4.8 – **Heatmaps for the cytoplasmic mRNA concentrations and different subnetworks. A** All chromatin regions; **B** Active regions not connected to pores; **C** Regions connected to pores; **D** Regions with more than 1 step away from a nuclear pore; and **E** Regions connected to pores and their immediate neighbours.

Since those three regions do not represent the array of different behaviours we can see in our model, we proposed in Fig. 4.8 for different subnetworks the same from Fig. 4.5: **A**, all the nodes of our network; **B**, Active regions not connected to pores; **C**, Regions connected to pores; **D**, Regions with more than 1 step away from a pore, which we can call $\mathbf{S}_1$; and **E**, Regions with 1 step maximum from a pore - the pore connected regions and their neighbours, $\mathbf{S}_0$. Similar to Fig. 4.5 **A**, most of the regions present small mRNA concentrations, but approximately 1/3 of the regions show $[m_i^C] > 0.05$, meaning the inactive/low-activity regions are a bigger portion of our system.

From previous results, we know active regions produce different concentration patterns as some genes must be produced more frequently than others, verifying these results in Figs. 4.8 **B** and **C**. Most regions with higher mRNA production present a strong peak in production which is reduced once the TF exportation is amplified reducing the overall transcription by limiting the resources. Hence, we proved some regions might produce in smaller volumes but more perennially, even if infinitesimal.

In the previous chapter, we discussed how positioning close to a nuclear pore optimizes transcription. Thus, once we analyzed the complementary subnetworks $\mathbf{S}_1$ and $\mathbf{S}_0$ and the more prolific regions are found in $\mathbf{S}_0$ but the highest concentration found are in the $\mathbf{S}_1$ subnetwork.

Given that we have limited transcriptional resources and a non-constant TF concentration for our model, we cluster the m_i^C solutions into seven different clusters to verify the concentration patterns that might emerge in the cytoplasmic mRNA or even if there is an unexpected late accumulation due to being in not optimal nodes and a result for delayed transcription, which is not expected from Fig. 3.15 **A**. These clustered values are shown in Fig. 4.9 **A**.

Once again, we can see there is no delayed accumulation of mRNAs, which means they are fast diffusing proteins, and different levels of concentration follow the same peak around the 30 minutes mark, with the system being inactivated. Besides, mRNAs with smaller concentrations are more constant over time than highly-active ones.

From the previous chapter, we know some regions present a delayed maximum and overshoots in concentration that stabilizes after the reaccumulation process starts (Fig. 3.15 **B**). Since this translocation pattern affects transcription, we expected to see similar behaviour in Fig. 4.9 **B**. In this subfigure, we show how most of the nodes peak at 30 minutes (as expected from Fig. 4.9 **A**) except for one delayed subset of more lasting mRNA production than the others, meaning this particular cluster proposes a lasting transcription in low concentrations, in which is true for the continuous space of a deterministic model since in a discrete space (a more biologically feasible space) this means transcription is not occurring in those particular regions.

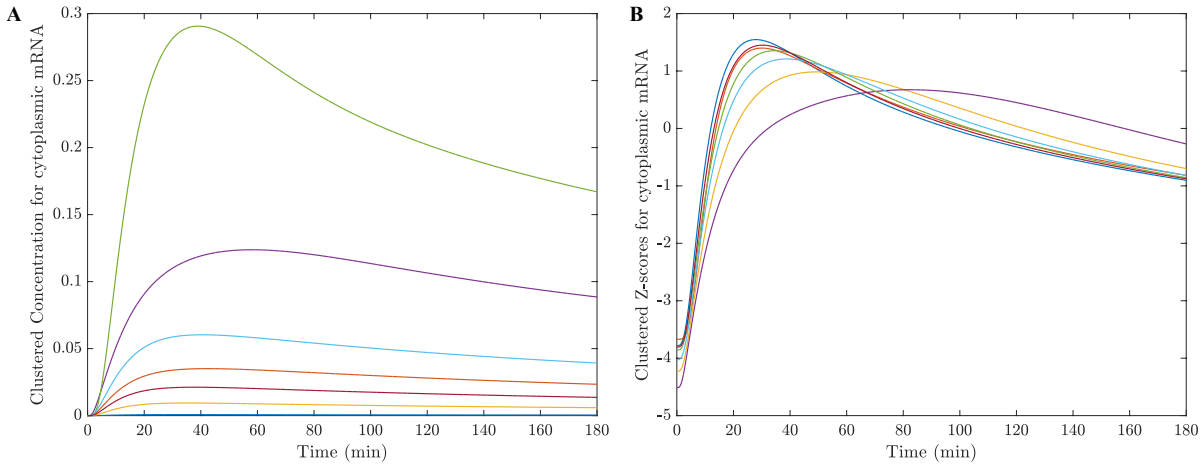

Figure 4.9 – **Cluster analysis for cytoplasmic mRNA. A** Clustered values for $m^C(t)$ and **B** Cluster of the z-score for the same variable. Each subfigure has 7 clusters and they are not correlated.

Furthermore, we need to compare the nuclear and cytoplasmic mRNA concentrations to check the emergence of a pattern from the export process. Thus, we present the difference between nuclear and cytoplasmic mRNA in a $\log_2$ scale and we labelled with the number of connections, d, and residence times, τ, in Fig. 4.10 **A** and **B**, respectively. From both subfigures, we can see a cloud of inactive regions (lower dots in dark blue). In Fig. 4.10 **A**, we proved that regions with more connections produce higher levels of mRNA, which is verified in Fig. 4.10 **B**. From those subfigures, we proved the mRNA synthesis occupancy is also determined by the parameters, which is explained by the Transcribing RNAP occupancy, as described in Chapter 3. The outliers in both images are the pore-connected regions, the export process is facilitated in those regions, increasing their cytoplasmic concentration in comparison with the nuclear one.

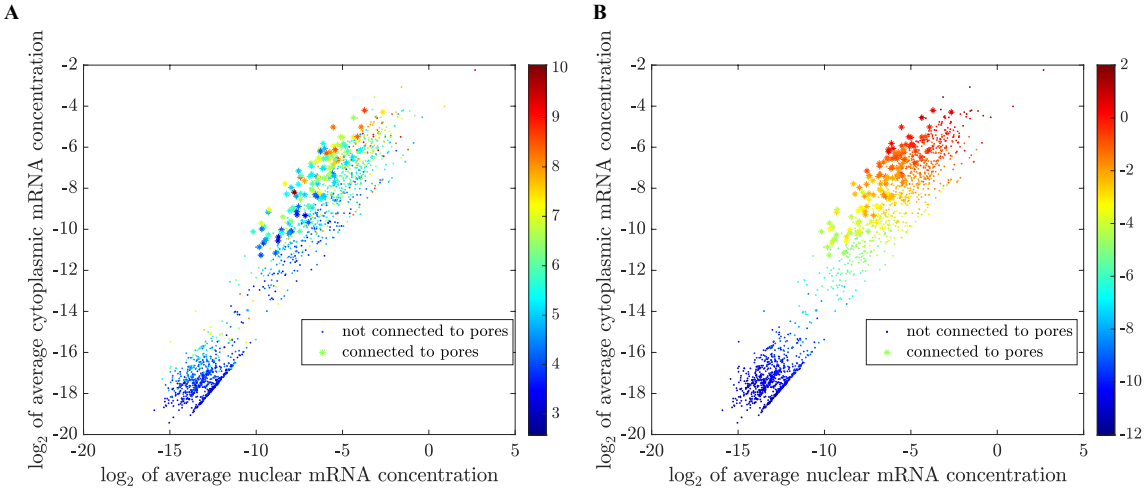

Figure 4.10 – **Comparison between nuclear and cytoplasmic mRNA concentrations labelled with A** number of connections, d and **B** residence times, τ.

This expansion of our import/export model from Chapter 3 incorporating the mRNA expressions in Eq. (4.1) did not consider different gene lengths, which can affect the mRNA, for example. Yet, our model predicted interesting results for regions connected to pores, which we believe should be further explored experimentally and theoretically.

The difference in the size of an mRNA produced also affects the diffusive rate of a protein. Since those parameters are gene-specific, we proposed next a model to describe the unspliced/spliced dynamics of mRNA.

4.2 Model for RNA velocity with mRNA exportation

Disclaimer: *This model is part of a collaboration with Mendoza's lab from IGBMC.*

As Fig. 4.1 represented, the protein synthesis has two physically separated processes: **Transcription**, in which the transcriptional machinery located in the nucleus produces a non-mature mRNA, i.e., the mRNA was transcribed together with long noncoding sequences, called **introns** (the coding sequences are called **exons**), and **Translation**, which the mature mRNA (an mRNA without introns) is translated into a protein in the cytoplasm (Roy; Gilbert, 2006; Alberts et al., 2002; Alberts, 2004; Lee et al., 2020). The process of removing the introns from non-mature mRNAs is called splicing. Hence, a non-mature mRNA is called unspliced and a ready-to-translation mRNA is a spliced mRNA.

One way to predict the unspliced/spliced dynamics is by using RNA velocity models (Manno et al., 2018; Gorin et al., 2022; Bergen et al., 2019). Those models however do not consider the mRNA export process and how it can impact translation. To incorporate the export dynamics, we proposed a model with three different mRNA states: (1) **Unspliced**, i.e., the mRNA is non-mature and nuclear; (2) **Spliced Nuclear** in which the mRNA was spliced but not exported yet; and (3) **Spliced Cytoplasmic**, where the mRNA is ready to be translated. We represent these dynamics in Fig. 4.11.

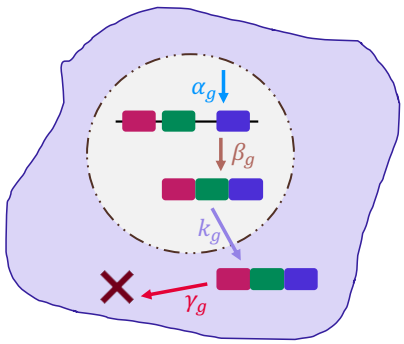

Figure 4.11 – **Cartoon representation for our RNA velocity model.** Here, we can see how the three states of our model are related, in which a produced unspliced mRNA is spliced inside the nucleus gets exported and eventually degraded.

Therefore, a specific mRNA that codes gene g has the following dynamics: the unspliced mRNA, u_g^N is produced by a transcription rate, α_g. Then, with a splicing rate β_g, u_g^N is spliced and enters the spliced state, remaining in the nucleus, which we defined as s_g^N. This s_g^N is exported with an export rate, k_g, entering the cytoplasmic spliced state, s_g^C, which is degraded with a degradation rate, γ_g.

For our RNA velocity model, we need to consider gene-specific parameters since from the models presented in Chapter 2 and 3, transcription is a region-specific process. We opted for constant parameters instead of time-dependent functions, thus making our model a linear ODE-system, proposed in Eq. (4.2), in which all the parameters are non-negative, i.e., $\alpha_g, \beta_g, k_g, \gamma_g \geq 0$.

$$\begin{cases} \dfrac{du_g^N}{dt} = \alpha_g - \beta_g u_g^N \ ; \\ \dfrac{ds_g^N}{dt} = \beta_g u_g^N - k_g s_g^N \ ; \\ \dfrac{ds_g^C}{dt} = k_g s_g^N - \gamma_g s_g^C \ . \end{cases} \tag{4.2}$$

The state of the art of this model relies on the experimental sequencing for thousands of genes (precisely, different 55400 genes) using three different techniques: (i) RNA-Sequencing (RNA-Seq); (ii) Single-cell RNA-Sequencing (TT-Seq); and (iii) Fractionation RNA Sequencing (Frac-Seq), (Saliba et al., 2014; Sterne-Weiler et al., 2013; Stark; Grzelak; Hadfield, 2019). We aim to understand how the gene expression changes from the control and auxin-treated cell lines, a plant hormone known for reprogramming the pluripotency in mammalian cells (Palomo et al., 2014).

We used our model in Eq. (4.2) to estimate the parameters for both Control and Auxin-Treated cell lines and then we verified how the treatment affects the cell type. Since time evolution is not possible in sequencing, first we need to estimate our system's steady states.

4.2.1 Steady-states, transcription rate dependency and stability

Our model in Eq. (4.2) is a simple three-state model for each gene g is independent of each other. We estimated the steady-states for our model in Eq. (4.3), and we can see all the states are directly dependent on the transcription rate α_g. Considering α_g shows how prolific is the gene, we can see how the equilibrium interacts with this value: more mRNAs produced mean more proteins synthesized.

$$U_g^N = \frac{\alpha_g}{\beta_g} \ ; \ S_g^N = \frac{\alpha_g}{k_g} \ ; \ S_g^C = \frac{\alpha_g}{\gamma_g} \ . \tag{4.3}$$

From Eq. (4.3), we assume β_g, k_g and γ_g should be more than just non-negative. Those parameters have to be strictly positive.

More so, since we use steady-states to estimate the parameter set for each gene g in our sequencing data, we need to prove the stability of our model as the sequencing techniques assume a stable accumulation of RNAs. Next, we present the stability analysis for our model in Eq. (4.2) considering the steady states in Eq. (4.3).

4.2.1.1 Characteristic Polynomial and Stability

We calculated the stability of our model in Eq. (4.2) by first calculating its Jacobian Matrix and then studying its characteristic polynomial (Murray, 2007; Edelstein-Keshet, 2005; Strogatz, 2015). Next, we present the Jacobian Matrix, $\mathbf{J}(U_g^N, S_g^N, S_g^C)$:

$$\mathbf{J}(U_g^N, S_g^N, S_g^C) = \begin{bmatrix} \frac{\partial}{\partial u_g^N}\frac{du_g^N}{dt} & \frac{\partial}{\partial s_g^N}\frac{du_g^N}{dt} & \frac{\partial}{\partial s_g^C}\frac{du_g^N}{dt} \\ \frac{\partial}{\partial u_g^N}\frac{ds_g^N}{dt} & \frac{\partial}{\partial s_g^N}\frac{ds_g^N}{dt} & \frac{\partial}{\partial s_g^C}\frac{ds_g^N}{dt} \\ \frac{\partial}{\partial u_g^N}\frac{ds_g^C}{dt} & \frac{\partial}{\partial s_g^N}\frac{ds_g^C}{dt} & \frac{\partial}{\partial s_g^C}\frac{ds_g^C}{dt} \end{bmatrix} = \begin{bmatrix} -\beta_g & 0 & 0 \\ \beta_g & -k_g & 0 \\ 0 & k_g & -\gamma_g \end{bmatrix}.$$

Since our system is linear, we do not have a dependency on any equilibrium point, which is uniquely defined in Eq. (4.3). We calculated the characteristic polynomial for our model, i.e., $p(\lambda) = \det(\mathbf{J} - \lambda Id)$, where Id represents the identity matrix:

$$p(\lambda) = (-\beta_g - \lambda)(-k_g - \lambda)(-\gamma_g - \lambda) \ ,$$

with three distinct polynomial roots $\lambda_1 = -\beta_g,\ \lambda_2 = -k_g$ and $\lambda_3 = -\gamma_g$. The roots of this system define a stable node, and this means that given enough time, the system will reach its steady state and remain in those evaluated points. This result guarantees the feasibility of using Eq. (4.3) to estimate the parameters of our model.

We verified our model in Eq. (4.2) and its steady state by applying the ode45 function from Matlab based on the Runge-Kutta method (Shampine; Reichelt, 1997). As an initial test, we proposed random parameters for our parameter set, P, in Table 5. The result of this simulation is present in Fig. 4.12.

From this simulation, we can see our model reaching stability even if it takes longer (around 120 minutes for the cytoplasmic spliced). The unspliced reached the equilibrium first, followed by the nuclear spliced and then the cytoplasmic mRNAs, which makes sense by the mRNA production steps (as we described in Fig. 4.11).

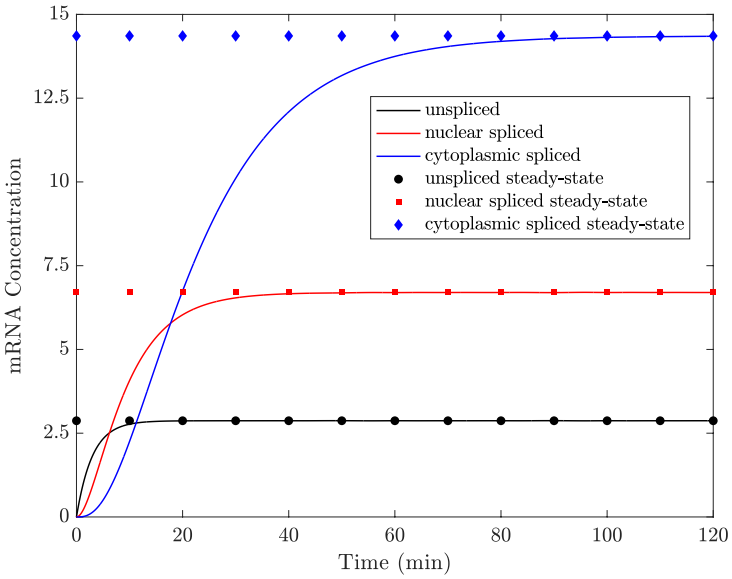

Figure 4.12 – **Simulation for our model, Eq. (4.2), and our steady-states, Eq. (4.3).** Our model reaches the steady state (which is a stable node) and the mRNA concentrations remain constant once we reach the equilibrium.

Since our steady-state is a stable node and we have experimental values from sequencing to propose parameters for specific genes, we can use them as a tool to estimate the parameters. The sequencing experiments were done to understand mRNA splicing are:

- **RNA-Seq**, which evaluates the unspliced and spliced mRNAs in equilibrium, i.e., $[U^{\text{RNA-Seq}}] = U_g^N$ and $[S^{\text{RNA-Seq}}] = S_g^N + S_g^C$;
- **TT-Seq**, which proposes the unspliced/spliced nuclear mRNA concentration after 15 minutes of stimulation, $[U^{\text{TT-Seq}}] = u_g^N(15)$ and $[S^{\text{TT-Seq}}] = s_g^N(15)$;
- **Frac-Seq**, unspliced and spliced (nuclear and cytoplasmic) mRNAs in equilibrium, i.e., $[U^{\text{Frac-Seq}}] = U_g^N$, $[S_N^{\text{Frac-Seq}}] = S_g^N$, and $[S_C^{\text{Frac-Seq}}] = S_g^C$.

As the TT-Seq experiments need the evaluation of our model for $t = 15$ minutes and our system is a linear ODE system, we solved the ODE system in Eq. (4.2). Next, we present the analytical solutions for our model.

4.2.2 Analytical Solution and Parameter Set

The model in Eq. (4.2) is a first-order linear ODE system, which is solvable. Different from our models in the previous chapters, this model is easily solvable with basic calculus. One way to solve it is by solving the unspliced equation first, then applying the solution to the nuclear-spliced equation to finally obtain the cytoplasmic equation.

To obtain a specific expression for our model, we need the initial conditions for our system. Since the experimental results consider the absence of mRNA concentration before the experiment, we assumed $u_g^N(0) = 0$, $s_g^N(0) = 0$ and $s_g^C(0) = 0$.

4.2.2.1 Unspliced Equation

The unspliced equation for our model in Eq. (4.2), with the proposed initial condition $u_g^N(0) = 0$ is:

$$\frac{du_g^N}{dt} = \alpha_g - \beta_g u_g^N \ .$$

Unfortunately, this equation is not separable. However, we can solve it by assuming $h = \alpha_g - \beta_g u_g^N$, with $dh = -\beta_g du_g^N$. Thus, our unspliced equation is separable and we can integrate our system:

$$\begin{aligned}
\int -\frac{1}{\beta_g h} dh &= \int dt \\
e^{\ln(\alpha_g - \beta_g u_g^N)} &= e^{-\beta_g (t + C_u)} \\
\alpha_g - \beta_g u_g^N &= e^{-\beta_g (t + C_u)} \\
u_g^N(t) &= \frac{\alpha_g - e^{-\beta_g (t + C_u)}}{\beta_g} \ .
\end{aligned}$$

Since $u_g^N(0) = 0$, we found that $e^{-\beta_g C_u} = \alpha_g$. Thus, in Eq. (4.4) we have our solution.

$$u_g^N(t) = \frac{\alpha_g}{\beta_g} \left(1 - e^{-\beta_g t}\right) \ . \tag{4.4}$$

From that, we find the first condition which is $\beta_g \neq 0$. With the solution in Eq. (4.4), we can apply the solution in the nuclear-spliced equation.

4.2.2.2 Nuclear-Spliced Equation

The nuclear-spliced equation from the model in Eq. (4.2) has the initial condition $s_g^N(0) = 0$. Since this equation depends on the unspliced expression, Eq. (4.4), we write the nuclear-spliced ODE as:

$$\frac{ds_g^N}{dt} = \alpha_g \left(1 - e^{-\beta_g t}\right) - k_g s_g^N \ .$$

To solve this equation, we proposed the integrating factor, $\mu(t)$ as $\mu(t) = e^{\int k_g dt} = e^{k_g t}$. We multiplied both sides by $\mu(t)$ and obtain:

$$e^{k_g t} \left(k_g s_g^N + \frac{ds_g^N}{dt}\right) = e^{k_g t} \left(\alpha_g (1 - e^{-\beta_g t})\right) \ .$$

With that and the product rule (i.e., $\frac{d(fg)}{dt} = g\frac{df}{dt} + f\frac{dg}{dt}$) and the fundamental theorem of calculus (Stewart, 2015), we obtained:

$$\begin{aligned}\frac{d}{dt}\left(e^{k_g}s_g^N\right) &= e^{k_g t}\left(\alpha_g(1-e^{-\beta_g t})\right)\\ e^{k_g}s_g^N &= \alpha_g\left(\int e^{k_g t}(1-e^{-\beta_g t})\right)\\ s_g^N(t) &= \frac{\alpha_g}{(k_g-\beta_g)k_g}\left(k_g(1-e^{-\beta_g t})-\beta_g\right)+C_{sn}e^{-k_g t}\ .\end{aligned}$$

Since $s_g^N(0)=0$, we can find the expression for C_{sn}, which is $C_{sn} = \frac{\alpha_g\beta_g}{k_g(k_g-\beta_g)}$. Then, the solution for our nuclear-spliced equation is in Eq. (4.5).

$$s_g^N(t) = \frac{\alpha_g}{(k_g-\beta_g)k_g}\left(k_g(1-e^{-\beta_g t})+\beta_g(e^{-k_g t}-1)\right)\ . \tag{4.5}$$

From this solution, two other conditions emerge to obtain a determined solution: $k_g \neq 0$ and $\beta_g \neq k_g$. Thus, we applied Eq. (4.5) to the cytoplasmic spliced equation.

4.2.2.3 Cytoplasmic Spliced Equation

Last, we used the same techniques described above to calculate our cytoplasmic spliced equation, i.e., we substituted Eq. (4.5) in our ODE from Eq. (4.2), considering the initial condition $s_g^C(0)=0$. Thus, the ODE expression for the cytoplasmic spliced mRNA is:

$$\frac{ds_g^C}{dt} = \frac{\alpha_g}{(k_g-\beta_g)}\left(k_g(1-e^{-\beta_g t})+\beta_g(e^{-k_g t}-1)\right)-\gamma_g s_g^C\ .$$

Given the same techniques we used for the nuclear-spliced equations, i.e., we started by defining the integrating factor $\mu(t) = e^{\int \gamma_g dt} = e^{\gamma_g t}$. Then, we used the product rule and finally the fundamental theorem of calculus to obtain the solution for this ODE, which we show the steps in the following:

$$\begin{aligned}e^{\gamma_g t}\left(\gamma_g s_g^C + \frac{ds_g^C}{dt}\right) &= \frac{\alpha_g e^{\gamma_g}}{k_g-\beta_g}\left(k_g(1-e^{-\beta_g t})+\beta_g(e^{-k_g t}-1)\right)\\ e^{\gamma_g t}s_g^C &= \frac{\alpha_g}{k_g-\beta_g}\left(\int e^{\gamma_g t}(k_g(1-e^{-\beta_g t})+\beta_g(e^{-k_g t}-1))dt\right)\\ s_g^C(t) &= \frac{\alpha_g}{\gamma_g}+\frac{\alpha_g}{k_g-\beta_g}\left(\frac{k_g e^{-\beta_g t}}{\beta_g-\gamma_g}+\frac{\beta_g e^{-k_g t}}{\gamma_g-k_g}\right)+C_{sc}e^{-\gamma_g t}\ .\end{aligned}$$

Given the initial condition, $s_g^C(0)=0$, we calculated the expression for C_{sc} as being:

$$C_{sc} = -\alpha_g\left(\frac{1}{\gamma_g}+\frac{1}{k_g-\beta_g}\left(\frac{k_g}{\beta_g-\gamma_g}+\frac{\beta_g}{\gamma_g-k_g}\right)\right).$$

Thus, we present in Eq. (4.6) the solution for our ODE. We also found another set of constraints for our system. From this equation, our model is only solvable if $\gamma_g \neq 0$, $\beta_g \neq \gamma_g$ and $k_g \neq \gamma_g$.

$$s_g^C(t) = \frac{\alpha_g}{\gamma_g} + \frac{\alpha_g}{k_g - \beta_g}\left(\frac{k_g e^{-\beta_g t}}{\beta_g - \gamma_g} + \frac{\beta_g e^{-k_g t}}{\gamma_g - k_g}\right) - \\ - \alpha_g\left(\frac{1}{\gamma_g} + \frac{1}{k_g - \beta_g}\left(\frac{k_g}{\beta_g - \gamma_g} + \frac{\beta_g}{\gamma_g - k_g}\right)\right) e^{-\gamma_g t} \ . \quad (4.6)$$

We verified the analytical solutions by evaluating Eqs. (4.4), (4.5) and (4.6) with the parameters from Table 5 and compared with the built-in ODE solver (ode45) from Matlab. This test is present in Fig. 4.13, and it is possible to verify how both the solver and our equations have the same values.

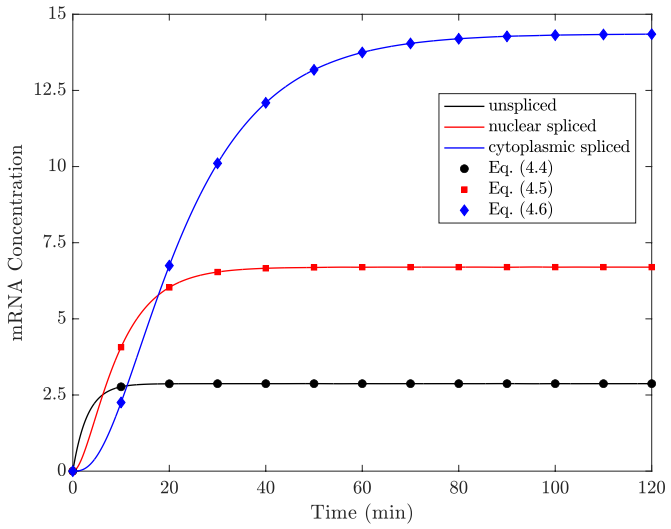

Figure 4.13 – **Comparison between Eqs. (4.4), (4.5) and (4.6) and the numerical solution for our model Eq. (4.2).** With the same parameters from Fig. 4.12, just to prove the analytical solution found is right.

From our analytical solutions (Eqs. (4.4), (4.5) and (4.6)) we found six different constraints for the parameters of our model, for the splicing rate (β_g), the export rate (k_g) and degradation rate (γ_g): all those parameters must be non-zero and different between each other.

Moreover, we proposed non-negative values for our parameters because our model deals with real experimental sequencing data, which means each gene is counted in the batch, we can rethink our constraints as:

- $\alpha_g \geq 0$;
- $\beta_g, k_g, \gamma_g > 0$;
- $\beta_g \neq k_g \neq \gamma_g$.

With those conditions and analytical solutions, we can evaluate our parameters from sequencing data.

4.2.2.4 Bayes' Theorem and Parameter Finding

Again, our model aims to predict parameters for 55400 different genes that were sequenced from three different experiments with replicas. Since it is a *count* of gene reads, our set is a positive enumerable set of how many times each gene was counted and the volume of reads per gene is gene-dependent. Thus, we can understand the number of reads of gene g normalized by the total number of reads as the probability of randomly selecting gene g. However, we still need to find a way to correlate our parameters with the experimental data from the sequencing.

From the steady-states equations, Eq. (4.3), all the experimental values are directly dependent on α_g. To estimate α_g, we used Eq. (4.4) and supposed $\beta_g << 15$ minutes. From this assumption, we obtained $u_g^N(15) \approx \alpha_g$, and we used the **TT-Seq** data to obtain the values of α_g considering Bayesian Inference (Viertl, 1987; Bayesian. . . , 2007; Breda; Zavolan; Nimwegen, 2019).

Let u_i^T the number of TT-Seq reads for gene i, α_i the transcription rate of gene i and N^T the total number of reads from a single TT-Seq experiment. Since they are experimental reads, N^T and u_i^T are natural numbers.

Therefore, from the Bayes' theorem we have $P(\alpha_i|u_i^T) = \frac{P(u_i^T|\alpha_i)P(\alpha_i)}{P(u_i^T)}$. Then, since we can either select gene i or any other gene that is not i and gene i depends on its transcription rate, α_i, we assumed $P(u_i^T|\alpha_i)$ is a beta distribution, i.e.,

$$P(u_i^T|\alpha_i) = \binom{N^T}{u_i^T} \alpha_i^{u_i^T} (1-\alpha_i)^{N_T - u_i^T},$$

which the number of reads for i, u_i^T are related to the transcription rate and $N^T - u_i^T$ is all the reads that are not for gene i from the sequencing experiment. Besides this probability, we need to define both $P(u_i^T)$ and $P(\alpha_i)$. For $P(u_i^T)$, we have:

$$\begin{aligned}
P(u_i^T) &= \int_0^1 P(u_i^T|\alpha_i) d\alpha_i \\
P(u_i^T) &= \binom{N^T}{u_i^T} \int_0^1 \left(\alpha_i^{u_i^T} (1-\alpha_i)^{N_T - u_i^T} \right) d\alpha_i \\
P(u_i^T) &= \binom{N^T}{u_i^T} \left(\frac{\Gamma(u_i^T + 1)\Gamma(N^T - u_i^T + 1)}{\Gamma(N^T + 2)} \right),
\end{aligned}$$

in which $\Gamma(X)$ is the gamma function in X. Since both u_i^T and N^T are enumerable natural numbers, we have $\Gamma(X) = (X-1)!$, for u_i^T and N^T. Thus,

$$P(u_i^T) = \frac{(N^T)!}{(u_i^T)!\,(N^T - u_i^T)!} \left(\frac{(u_i^T)!\,(N^T - u_i^T)!}{(N^T + 1)!} \right).$$

Since N^T is a large number, we assumed $(N^T + 1)! \approx N^T!$. Therefore, $P(u_i^T) = 1$. From this, we obtained $P(\alpha_i|u_i^T) = \frac{P(u_i^T|\alpha_i)P(\alpha_i)}{P(u_i^T)} = P(u_i^T|\alpha_i)P(\alpha_i)$. Then, we estimated $P(\alpha_i)$ by

assuming, again, it is a beta distribution, i.e., let $m, n \in \mathbb{N}$,

$$P(\alpha_i) = f_B(\alpha_i; m, n) = \frac{\alpha_i^{m-1}(1-\alpha_i)^{n-1}}{\mathbf{B}(m,n)},$$

which $\mathbf{B}(m,n)$ is the beta function.

Since $m, n \in \mathbb{N}$, we can write it as $\mathbf{B}(m,n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$. Besides, we assumed $m = n = 1$, thus $P(\alpha_i) = 1$. From these results, we have $P(\alpha_i|u_i^T) = P(u_i^T|\alpha_i) = f_B(\alpha_i; u_i^T + 1, N^T - u_i^T + 1)$. With this expression in our hands, we could evaluate the mean and variance for α_i. As the measurement from different experiments is independent of each other, we consider all the replicas as parts from the same experiment - i.e., $N^T = N_1^T + N_2^T + ... + N_r^T$ and $u_i^T = u_{i_1}^T + u_{i_2}^T + ... + u_{i_r}^T$. In Eqs. (4.7) and (4.8), we present the expressions for our parameters.

$$\alpha_i = \mu = \mathbf{E}[\alpha_i] = \frac{u_i^T + m}{N^T + m + n}\ , \tag{4.7}$$

$$Var(\alpha_i) = \mathbf{E}[(\alpha_i - \mu)] = \frac{(u_i^T + m)(N^T - u_i^T + n)}{(N^T + m + n)^2(N^T + m + n + 1)}\ . \tag{4.8}$$

For our fittings, we assumed $m = n = 1$. Thus, those are the expressions for evaluating α_i.

Similarly, we can approximate the other parameters by their analytical expressions, Eqs. (4.3). From the steady-states, α_i values, and the **Frac-Seq** experiments we found our splicing rate for gene i, β_i, the export rate k_i and the degradation rate γ_i. Since each parameter depends only on a specific sequence read and α_i, we can describe the mechanics behind these parameters by using a dummy variable, η_i, which represents the inverse of either β, k or γ for gene i.

Let Θ_i be any steady-state from Eq. (4.3) such that $\Theta_i = \alpha_i\eta_i$ and N_ρ^F as the total number of reads from **Frac-Seq** experiments for each type of genes read. We define the number of reads for gene i from **Frac-Seq** as ρ_i^*, which $*$ denotes either nuclear or cytoplasmic reads, and we assume $\Theta_i \approx \rho_i^*$. From this expression, we also assumed beta distribution and obtained similar expressions from Eqs. (4.7) and (4.8) for η_i. Assuming $m, n \in \mathbf{N}$, the expressions are:

$$\eta_i = \mu = \mathbf{E}[\eta_i] = \frac{\rho_i^* + m}{N_\rho^F + m + n}\ ,$$

$$Var(\eta_i) = \mathbf{E}[(\eta_i - \mu)] = \frac{(\rho_i^* + m)(N_\rho^F - \rho_i^* + n)}{(N_\rho^F + m + n)^2(N_\rho^F + m + n + 1)}\ .$$

The expressions for β_i, k_i and γ_i are found by dividing the α_i by η_i. However, the standard deviation expression is found by applying the following expression:

$$Std(*) = \sqrt{\frac{1}{(\eta_i^*)^2}Var(\eta_i) + Var(\alpha_i)\frac{\alpha_i}{((\eta_i^*)^2)^2}}\ .$$

Given the expressions, we can estimate all the parameters from the experimental data. Some genes are histone genes (non-intronic), i.e., some genes do not require splicing (Fedorov, 2001; Volanakis et al., 2013) and for those genes, we assumed β_i is a non-number, and we evaluated α_i by using the spliced **TT-Seq**. In Fig. 4.14, we present the parameter space for control and auxin-treated cells. The behaviours from Control and Auxin-treated cells are similar, where smaller genes are closer to the outliers from the measurements and more spread for longer genes, even if more randomly spread.

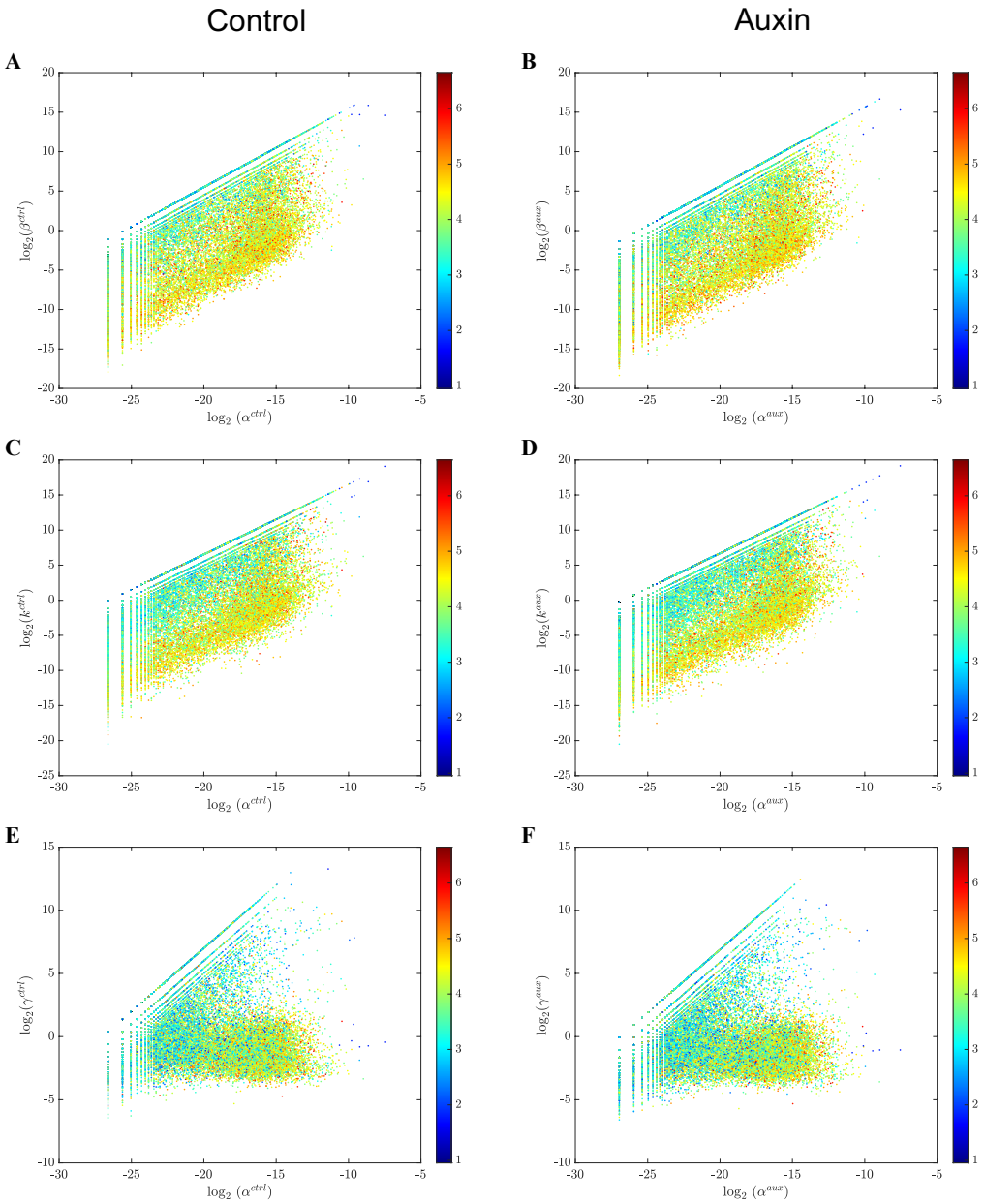

Figure 4.14 – **Parameters Space for Control and Auxin-treated cells, coded by the gene lengths** ($\log_{10}$)**. A** and **B**, $\log_2$ of the transcribing rate, α_g, over $\log_2$ of the splicing rate, β_g, where **A** Control and **B** Auxin. **C** and **D**, $\log_2$ of the transcribing rate, α_g, over $\log_2$ of the export rate, k_g, with **C** Control and **D** Auxin. **E** and **F**, $\log_2$ of the transcribing rate, α_g, over $\log_2$ of the degradation rate, γ_g, in which **E** Control and **F** Auxin.

Fig. 4.14 shows how splicing is a process unaffected by gene lengths as export and, more strongly, degradation is more affected. The export of longer genes has an apparent separation for gene lengths, as bigger genes occupy more space, have bigger masses, and, as a consequence, are slower than smaller genes. Yet, the separation by the length of a gene is more prominent in small genes than for the other genes, implying the length of a gene is less important for the parameter set.

The agglutination of values for γ creates a separation between smaller genes and the rest, which can be explained as longer genes take longer to be fully read and its production must be optimized, thus longer times to be degraded. Those results proved that even if the values are more spread for producing and splicing, the size is important for the export (where the mRNA is made available for use) and degradation (after the use) processes when we compare the small genes ($\log_{10}$(gene length) < 4) with the rest.

Next, we need to verify if gene length plays a role in our parameters and compare the effect of the auxin treatment on the cells, to verify if the treatment affected the cells, as Fig. 4.14 shows the same behaviour for both control and auxin.

4.2.3 Comparison between Control and Auxin-treated cells

To compare both treatments, we present a comparison between the parameters in both control and auxin in Fig. 4.15. We compared the $\log_2$ for all the parameters, evaluating the correlation between the parameter sets - if the highest correlation between the values and the treatments close to the identity function is the system. Besides, since our dataset is big enough to make it difficult to differentiate between values, we also evaluated the correlation between control and auxin-treated parameters. The gene length was used as means to check the correlation between parameters, but no strong correlation was found as expected from Fig. 4.14.

In Fig. 4.15 **A**, we have a high correlation between the parameters, even if the plot does not look similar to a line, as compared to the other parameters. The length of a gene seems to amplify the differences between control and auxin, as the longer genes seem to be in the periphery of the plot, but given the size of the parameter set this result is not significative enough to affirm the importance of the gene length in our model.

The correlation decreases in Fig. 4.15 **B** for the splicing rates even if it is high. The parameters created an ellipsoid pattern, with longer genes having smaller rates, implying for those longer genes are more effective in splicing. The random distribution of the gene length values persists in this parameter as expected by the definition of β.

Fig. 4.15 **C** shows again a high correlation between parameters and a pattern closer to the identity function. Again, longer genes were found in the pattern periphery proving the randomness of these values. The same result is found in Fig. 4.15 **D**, which has a more

spread behaviour than the others and the lowest correlation. Thus, Fig. 4.15 proved while some parameters are not affected by the auxin treatment (α_g and k_g), other parameters were affected (β_g and γ_g)

Such a result implies a strong effect of the auxin on splicing and degrading mRNA. More than that, we concluded the gene length does not influence our parameters, meaning the size of a gene does not impact the mRNA dynamics for this cell type - an impressive result per se.

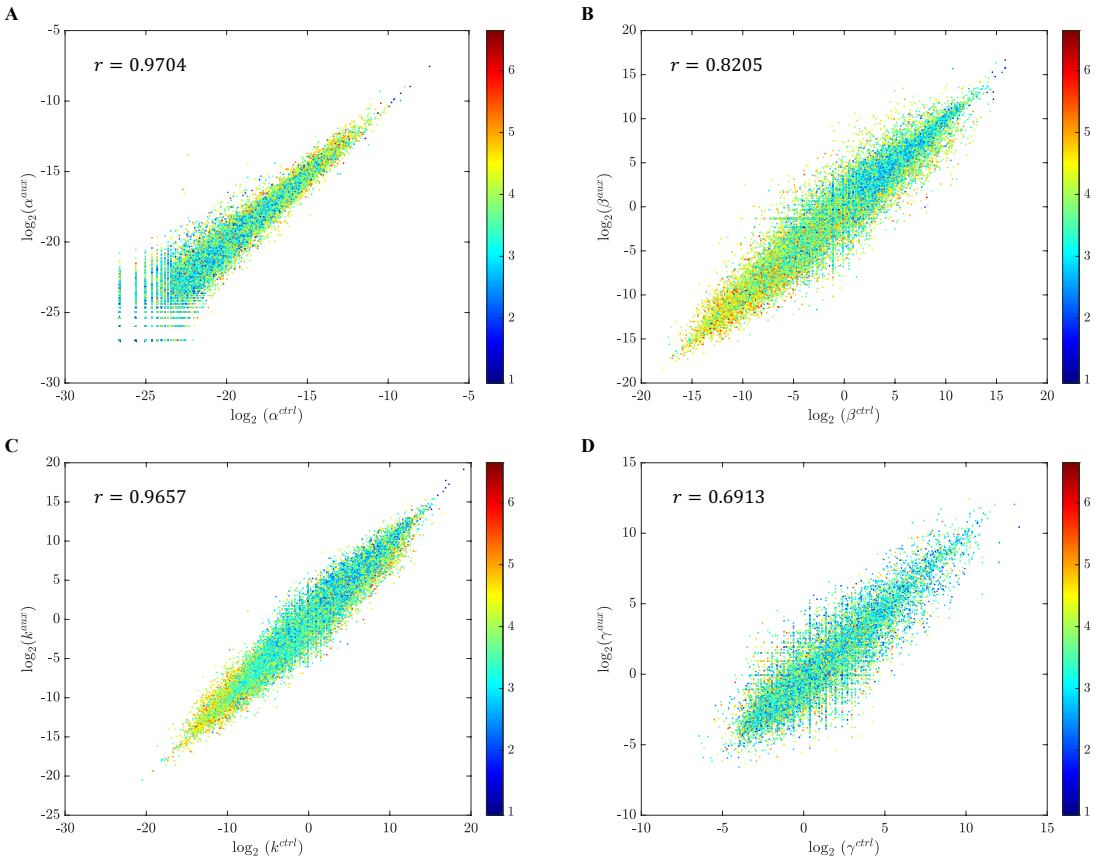

Figure 4.15 – **Comparison between control and auxin for our model parameters in log_2 scale and coded by its gene lengths, and the correlation between the parameters. A**, transcription rate, α_g; **B**, splicing rate, β_g; **C**, export rate, k_g; and **D**, degradation rate, γ_g.

To check how auxin affected the genes, we studied our system $\log_2$ fold-change in Fig. 4.16 for all the parameters in which the correlation between the experimental reads from the sequencing and our model analytical values is bigger or equal than 40%, reducing our dataset to 9094 genes. This analysis helped us to understand how the auxin treatment influenced the dynamics of the mRNA. Again, for all the parameters the gene lengths have no impact whatsoever.

In Fig. 4.16 **A**, we can see how auxin influences the z-score for the transcription rate, α. Thus, the number of genes downregulated in auxin-treated cells (i.e., $\log_2$ (fold-change

for α) < -1) represents 879 genes, with average z-score in this subset ($\left\langle Z^{\alpha}_{fc<-1} \right\rangle$) 6.7110, while the upregulated genes ($\log_2$ (fold-change for α) > 1) represents 703 genes, and $\left\langle Z^{\alpha}_{fc>1} \right\rangle = 5.0678$, meaning the z-score is higher on the downregulated subset. By the total number of actual changes in behaviour from the treatment, we verified around 82% of the genes remain unchanged by auxin. We conclude the auxin treatment downregulates more genes than the genes it upregulates, meaning an auxin treatment decreases transcription rates more than increases them. Note $\max(z_{score}(\alpha)) >> 1$, meaning α has parameters that deviate greatly from the standard deviation.

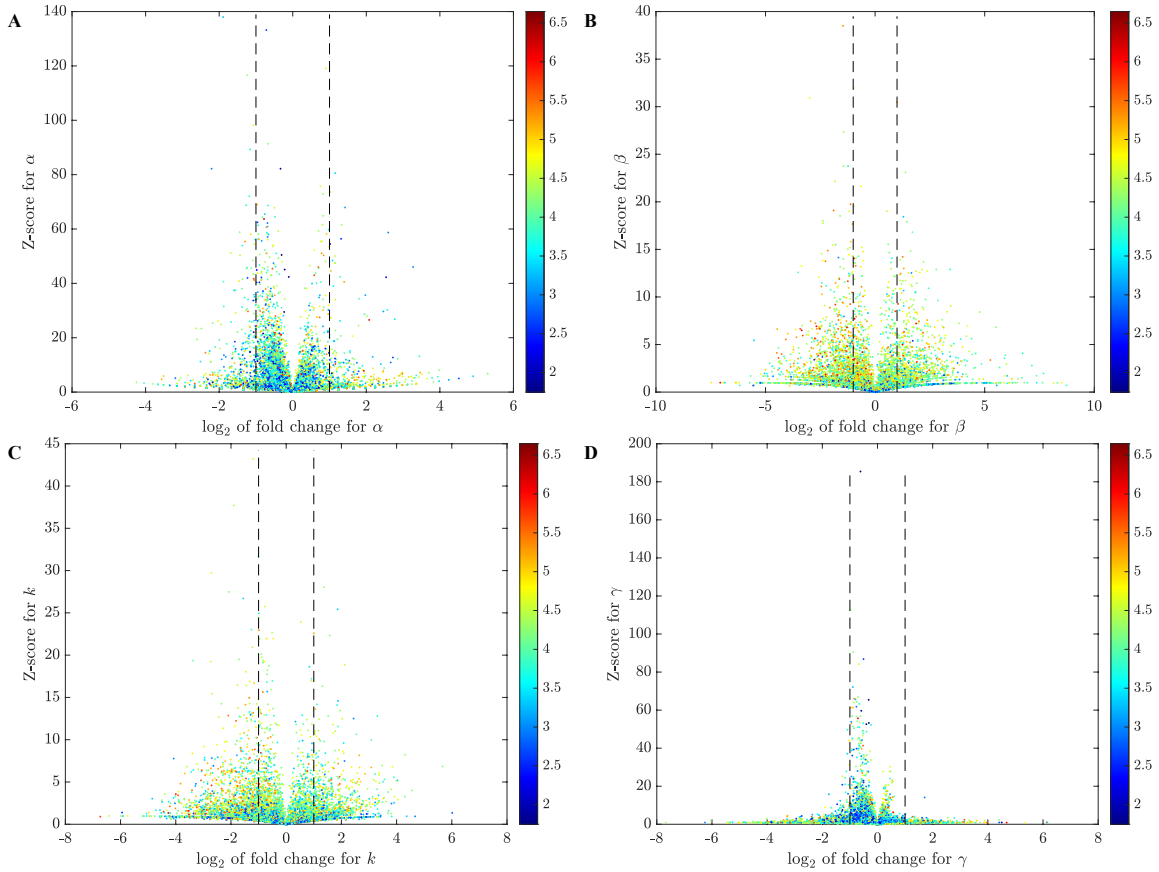

Figure 4.16 – **Volcano plot for all our model parameters.** Here, we calculated the z-score for each gene and the $\log_2$ of auxin-control fold change, coded by gene length. **A**, transcription rate, α_g. **B**, splicing rate, β_g. **C**, export rate, k_g. **D**, degradation rate, γ_g.

In Fig. 4.16 **A**, we can see how auxin influences the z-score for the transcription rate, α. Thus, the number of genes downregulated in auxin-treated cells (i.e., $\log_2$ (fold-change for α) < -1) represents 879 genes, with average z-score in this subset ($\left\langle Z^{\alpha}_{fc<-1} \right\rangle$) 6.7110, while the upregulated genes ($\log_2$ (fold-change for α) > 1) represents 703 genes, and $\left\langle Z^{\alpha}_{fc>1} \right\rangle = 5.0678$, meaning the z-score is higher on the downregulated subset. By the total number of actual changes in behaviour from the treatment, we verified around 82% of the genes remain unchanged by auxin. We conclude the auxin treatment downregulates more genes than the genes it upregulates, meaning an auxin treatment decreases transcription

rates more than increases them. Note $\max(z_{score}(\alpha)) >> 1$, meaning α has parameters that deviate greatly from the standard deviation.

The other parameters - namely, β, k and γ - were found by evaluating α, meaning those parameters are influenced by α. Fig. 4.16 **B**, the interval of the fold-change is bigger and the values are more spread than we showed in Fig. 4.16 **A**. We emphasize that intronic genes were removed from this figure since they do not have real values for splicing, as they do not splice. Similar to α, more genes are downregulated due to auxin treatment than upregulated, i.e., let D_β be the number of genes downregulated in β ($\log_2$ (fold-change for β) < -1) and U_β the number of genes upregulated in β ($\log_2$ (fold-change for β) > 1), we have $D_\beta = 2015$ (in which ($\left\langle Z^\beta_{fc<-1} \right\rangle = 2.4034$) and $U_\beta = 1392$ (with ($\left\langle Z^\beta_{fc>1} \right\rangle = 2.3932$). From the averages, we can see the z-score in β does not change much between the downregulated and upregulated genes. Thus, the percentage of genes unaffected by auxin decreases to 62%.

For Fig. 4.16 **C** the export rate parameter k, the z-score is lower than. However, the standard deviation for k is very high when compared with the other parameters, thus this small change within the standard deviation does not mean there is no variation between genes. Interestingly, some longer genes are more prominent in the downregulated fold-change, $D_k = \{i : (\log_2 \text{(fold-change for } k_i) < -1)\}$ than in the upregulated fold-change, $U_k = \{i : (\log_2 \text{(fold-change for } k_i) > 1)\}$. More than that, D_k is almost two times bigger than U_k (i.e., $D_k = 2276$ and $U_k = 1141$), with a similar to the β values of the proportion of the non-changing genes (more than 62% of unaffected genes). Their averages are also similar between themselves and the values of β: $\left\langle Z^k_{fc<-1} \right\rangle = 2.2456$ and $\left\langle Z^k_{fc>1} \right\rangle = 2.3263$, but for k the z-score averages for upregulated are slightly bigger.

Finally, in Fig. 4.16 **D**, we saw a more confined fold-change with a small deviation from the standard deviation for the degradation rate γ. Even if the fold-change is more confined, we saw more evenly spread values for γ, with no visual influence from the gene length. We evaluated the number of downregulated and upregulated genes, $D_\gamma = 1262$ and $U_\gamma = 479$ with two times more genes being downregulated than upregulated. The z-score for the degradation rate is lower than the other parameters, with the following averages: $\left\langle Z^\gamma_{fc<-1} \right\rangle = 1.8362$ and $\left\langle Z^k_{fc>1} \right\rangle = 1.0003$. Again, the percentage of genes that remain unchanged is approximately 80%.

Fig. 4.16 proved most of the genes were not strongly affected by auxin treatment since the minimum percentage of fold-change between -1 and 1 is 62%, but the genes that were affected can be split between $56 - 72\%$ downregulated to $28 - 44\%$ upregulated. This means most of the genes were downregulated with the auxin treatment. We also verified the z-score averages for all parameters and subsets and α, β and γ the downregulated subsets have higher z-score averages compared with the upregulated genes (the export rate had the opposite behaviour). Therefore, we conclude auxin might decrease transcription

in most of the transcriptional sites.

From the theoretical front, our model is an improvement from the RNA velocity models by incorporating the delay of the mRNA export and differing between mRNA in the nucleus and cytoplasm. Our model also proved to be a great tool to understand experimental data and predict the behaviours of thousands of genes.

Besides, incorporating mRNA synthesis into our model from Chapter 3 helped us to uncover concentration differences per chromatin region, even in simulations with the same parameters everywhere, which is not true from experimental results. Thus, more than just incorporating structure and residence times, we ought to consider the specificity of each gene to understand gene regulation.

5 Image Analysis: Exploring Theoretical predictions via Experimental results

Disclaimer: Some parts of this chapter can be found in my paper *"Modelling Transcription Factors Search and Polymerase Recruitment Dynamics within a complex chromatin structure"*.

In Chapters 3 and 4, we predicted how the closeness to a nuclear pore increases transcriptional activity. However, those results were only from mathematical models and we still need experimental results corroborating our claims.

Therefore, in this chapter, we present all the Image Analyses we did for different experiments realized either by the PhD student during her secondment at Ospedale San Raffaele in Milan, Italy or by collaborators from the same institute. We decided to split our results here by the experimental setup used, i.e., we present two different image analyses to confirm our theoretical predictions: (i) Single-Molecule FISH for HeLa-MS2 cells for the transcription factor $p65$ (RELA); (ii) Single-Molecule Tracking for the same cell line and TF.

5.1 Single-Molecule FISH (HeLa Cells)

Single-Molecule fluorescence *in situ* hybridization (smFISH) is a more and more common technique to visualize gene expression for a single cell, as it presents a way to detect mRNA molecules individually. Ignoring in our analysis the *count* of mRNA per probe, we can detect the localization of those mRNA spots and transcriptional sites (Pharris et al., 2017; Femino et al., 1998). Different algorithms to obtain smFISH image segmentation are proposed yearly, e.g. (Tsanov et al., 2016; Imbert et al., 2021), however, we opted for doing our own code for image analysis as our dataset was small enough to allow us to do it by ourselves.

Our models predict activity in specific regions (Chapter 3) and different volumes of mRNA in the cytoplasm (Chapter 4). Thus, this technique was considered to test some of our theoretical results. By using the facilities from our collaborators in Ospedale San Raffaele in Milan Italy, we obtained the smFISH for a translocating TF family, NF-κb, more specifically, $p65$ (RELA), which reactions in this TF are required for NF-κb activation (Chen; Greene, 2004).

Besides, the closeness to the nuclear pore complex (NPC) seems to optimize transcription, as we predicted in Chapters 3 and 4. Yet, we needed empirical proof

to confirm our hypothesis. Given the NF-κb is an inducible TF, we realized smFISH experiments for HeLa-MS2 cells in different time points of TNF-α treatments and for different NF-κb genes, with replicas of the stacks.

To represent the smFISH and our generated masks, we present Fig. 5.1, in which different time points for the same NF-κb gene, NF-κbia. The choice for NF-κb as our TF for this analysis is due to its translocation pattern and correlation with both Chapter 3 and (Zambrano et al., 2020). The time points are dependent on times after the TNF-α treatment our cells were fixed, i.e., in $t = 0$ (no TNF-α treatment), $t = 20$ minutes after treatment (the expected time for NF-κb) to reach the maximum TF concentration), $t = 30$ (where the exportation process is occurring), and $t = 60$ minutes (i.e., the expected time to finish translocation in some cell lines) (Trask, 2012). As this specific TF is endogenous to the cytoplasm, we expected not to find many active TSs, as opposed to the other time points.

In Fig. 5.1, we used the smFISH maximum projection to create one image to represent all the stack of the data. Here, we have two channels: the nucleus channel (DAPI) in blue and the mRNA channel (mCherry) in red. We emphasize that all the image analyses were done by using the built-in functions of Matlab.

From those 2D smFISH images, we segmented the nucleus by using a gaussian filter first to reduce our image noise due to the image acquisition (Haddad; Akansu, 1991). Later, we thresholded the image, removing the background noise and filling any *holes* the connected pixels might have.

Then, we proceeded to the nuclei segmentation by applying the water-shedding algorithm, which is a known algorithm for separating two nuclei (i.e., basins) (Kowal et al., 2019; Malpica et al., 1998; Meyer, 1994). Once we segment the different nuclei, we give them different masks to differentiate between nucleus #3 and #17, for example. This process is only for segmenting the nucleus (blue channel). After the water-shedding, we consider our nuclei segmented.

By hybridizing our probes with mCherry, we obtain numerous mRNA spots, as each transcript is also fluorescent. However, we want to detect only the brighter spots in the cell, since those are the transcription sites (TS). To detect those specific spots, we binarize the image and select all the spots.

Then, we evaluated the pixel size of the detected spot, their mean intensity, and if they are found inside a nucleus or not, because a TS should be inside the nucleus and it accumulates mRNAs, so we expected big and bright spots, our choice for detecting only the TSs is to find where transcription of a gene g occurs in the nucleus for a given cell. Another condition we imposed for our spot identification code was no nuclei should have more than four TSs, and this requirement was derived from the fact each TS should be

doubled by the number of chromosome pairs, but since our cell line is a non-healthy one, we accept more than two TSs.

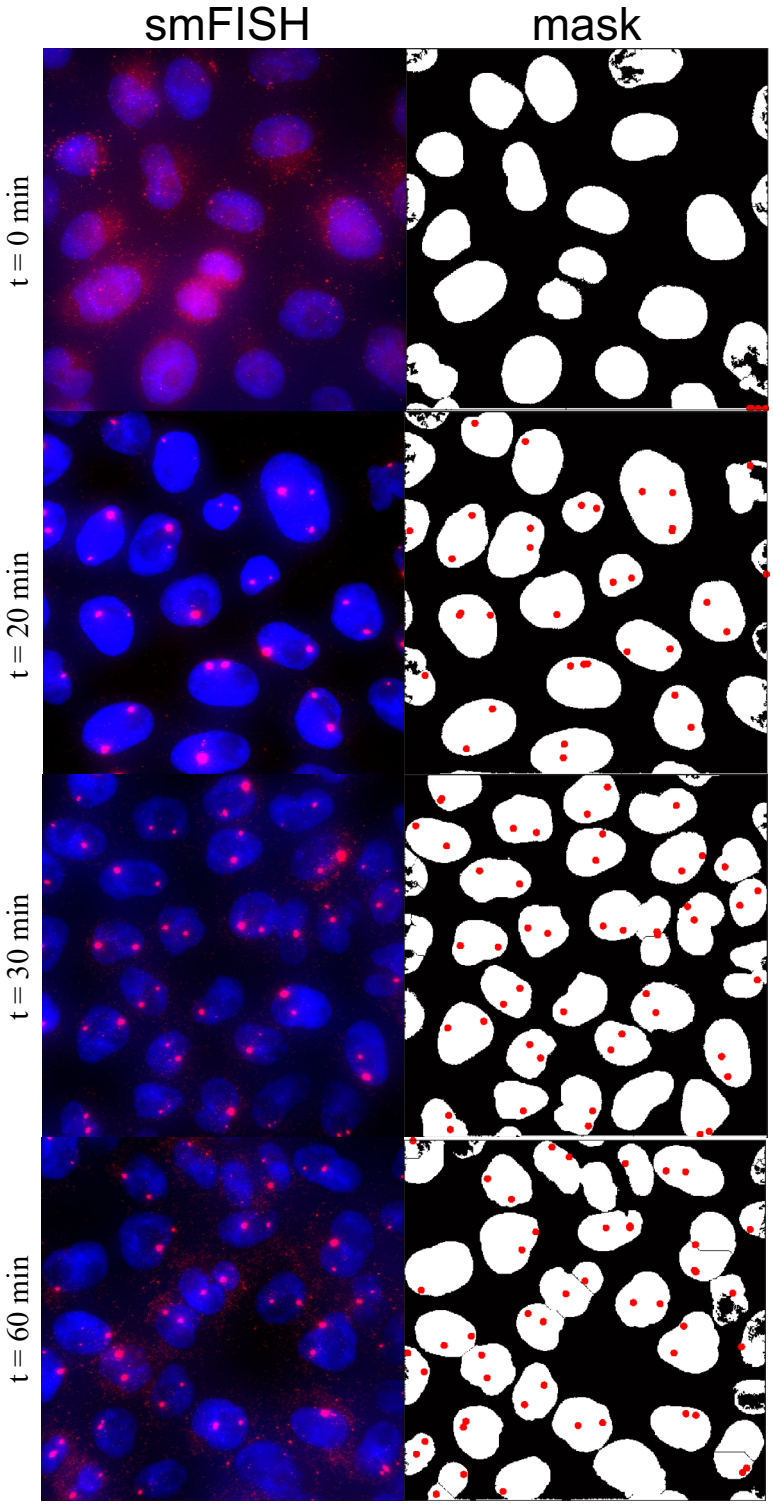

Figure 5.1 – **Image Segmentation for smFISH in different time-points of treatment for NF-κbia gene.** Here, we proposed the maximal projection of smFISH stacks and compared it with our generated mask from the experimental data collected, and the centroid of the spots identified. The window size represents 1024×1024.

Fig. 5.1 shows how effective is our segmentation by comparing our segmented masks to our smFISH data. For such a simple pipeline we obtained good results, even if some small TS can be detected in the smFISH image and not found in our masks. This is a direct consequence of the pixel size of the spot, as we deleted any detected spot with less than 12 pixels in size, as we considered those spots *too* small to be a TS. Of course, we considered *only* big spots as TSs, so we determined their position by calculating their centroid. Here, most of the found spots were not considered due to being too faint or small for our image analysis (e.g., $t = 0$ min smFISH image, Fig. 5.1), which we consider transcripts floating in the environment.

Our aim with this segmentation is to uncover how far from the nuclear border are our TSs, as our model in Chapter 3 predicted the closeness to the nuclear pore complex as an optimal feature for transcription. To evaluate the distance from the nuclear border, we create a distance map, where we calculate how far the pixels from our mask are from the edge of the mask, Fig. 5.2.

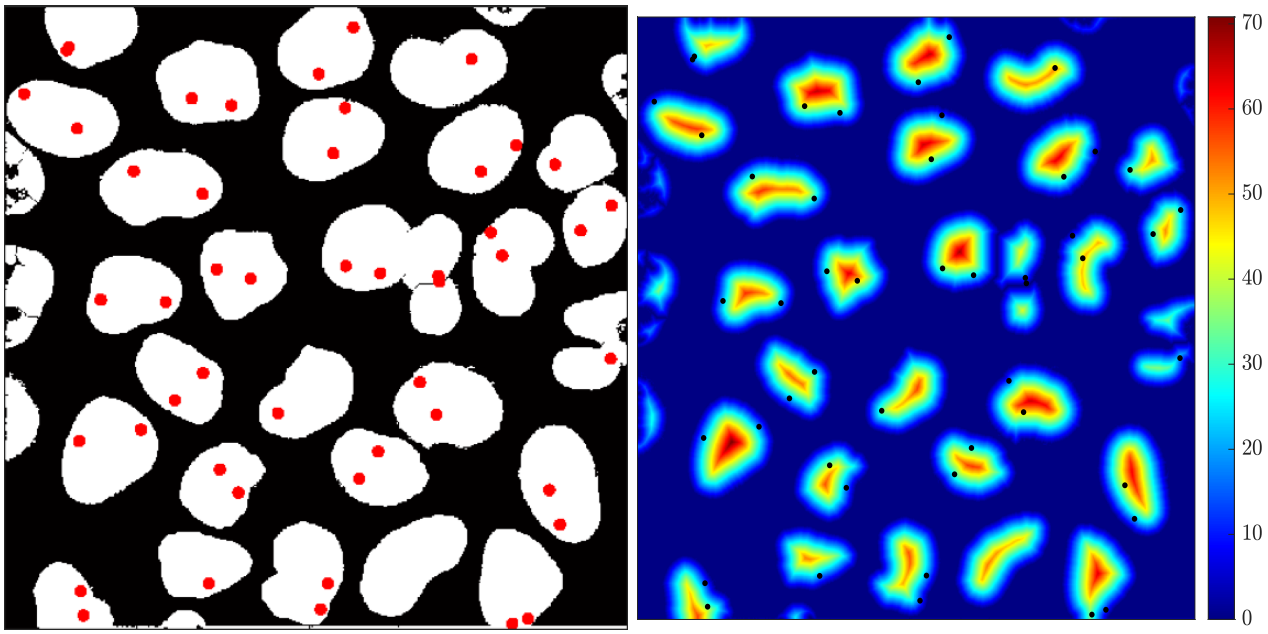

Figure 5.2 – **smFISH mask and its distance map.** From our mask, we calculated how far from the border is all nucleus pixels with our TS spots. Here, we present the $t = 30$ minutes mask for NF-κbia from Fig.5.1.

From those distance maps from all our smFISH images, we evaluated the distance from the edge for all the spots found. As the stochasticity of a gene influences the volume of transcription, we tested our pipeline for the data set from (Zambrano et al., 2020), since they proposed a set of four different genes (namely, NF-κbia, IL6, TNF and CCL5) and in four different time points after TNF-α induction (0 (no treatment), 20, 60 and 180 minutes, time points that represent NF-κb translocation dynamics).

As the maximum TF nuclear concentration is found around 20 minutes after

translocation induction, e.g. Fig. 3.12, we analyzed how the genes were localized at this time point and verified how the target sites are more likely to be close to the nuclear border than random positions inside the cell. We defined random positions as the generated spots with the same size as the TS spots we run inside the nuclei, only calculating the distance if those generated spots were completely inside a nucleus.

Given a spot and nuclei, we can estimate how far the spot is from the edge, for all the spots and nuclei from a smFISH picture. However, this is a quantitative measure and we want to understand the behaviour more qualitatively. Thus, we evaluated the probability density function for the spots' distances. In Fig. 5.3 **A**, we present the probability density of the TS localization for different genes.

As expected by the differences in mRNA produced we found in Chapter 4, we wanted to check how active genes occupy the nucleus: some genes are more spread inside the nucleus (e.g., NF-κbia) than others (e.g., TNF). This proves the differences in positioning between genes but also that the activity is still found near the border and this experimental result corroborates our model prediction. In Fig. 5.3 **B**, we verify how the random spots (i.e., spots from the same size of the transcription sites spots fitted in all cells from a smFISH picture) calculated inside all cells from each cell mask. There is no bias from the random positions.

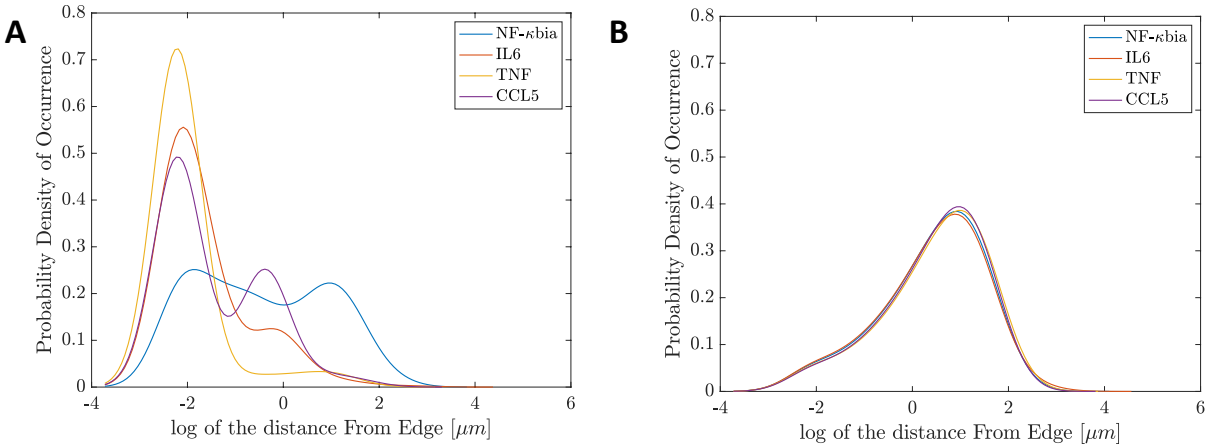

Figure 5.3 – **Experimental data corroborate our assumption of closeness to the nuclear envelope.** smFISH data for HeLa cells with NF-κb for different genes (i) NF-κbia (ii) IL6; (iii) TNF; and (iv) CCL5 after 20 minutes of treatment. Here, we calculate the log of the distance from the nuclear border for **A** transcription sites (spots) and **B** random positions. We see that there are no changes per gene for random positions, which is verified in **A**, proving the distance from the nuclear envelope is not an artefact.

Those results are a comparison between genes in the maximal NF-κb concentration. To understand the translocation process and the activation for our data, we proposed the probability density function for all the time points, obeying gene-specificity, in Fig. 5.4.

Each gene has a different translocation pattern. The random spot studies proved the absence of artefacts from the cell segmentation.

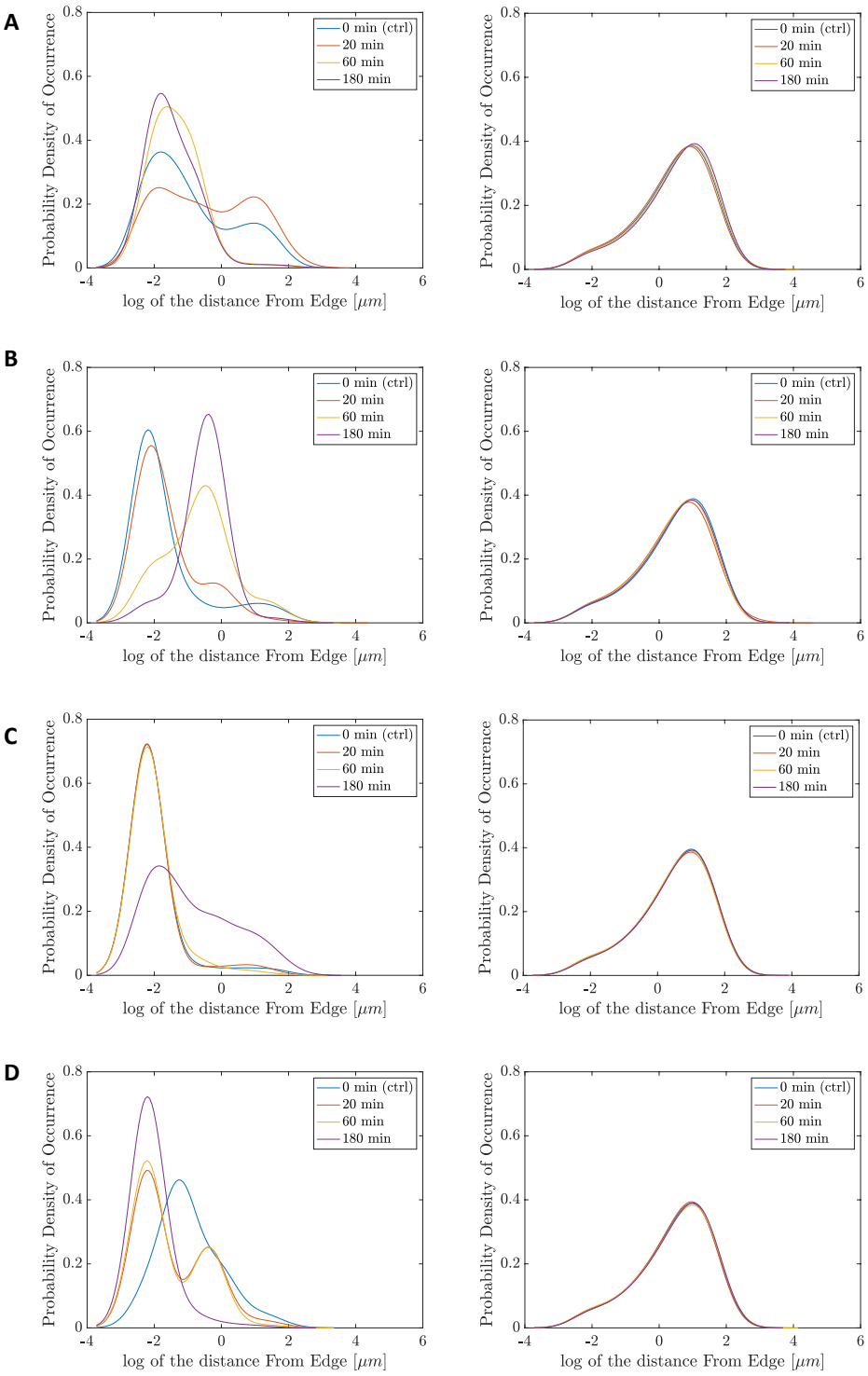

Figure 5.4 – log **of the distance from the nuclear border calculated from our watershed algorithm for spots from different genes and time points in μm.** In which: **A** NF-κbia; **B** IL6 **C** CCL5 **D** TNF. From this figure, we can recover different behaviours from the genes over time: some genes present a late activation (TNF, for example), the movement inside/outside the nucleus (NF-κbia for example) and even potential reactivation (IL6).

The gene from Fig. 5.4 **A**, NF-κbia (NF-κb inhibitor alpha) is a known gene for its high-activity upon NF-κb activation, as it helps the binding of the NF-κb in the IkB kinase (IKK) complex (Courtois et al., 2003). Thus, its fast activation is expected. Besides, the translocation pattern for this gene shows how this gene is found the furthest in time $t = 20$ minutes, where the NF-κb peaks its concentration in the nucleus and re-accumulates near the nuclear border in later points.

Interleukin-6 (IL6) is a gene involved in inflammation (either acute or chronic) and maturation of B cells (Chou et al., 2010). As we analyzed its localization pattern, Fig. 5.4 **B**, we found a late activation pattern, in which in later time points we find more TS spots inwards the cell, showing a late activation and re-activation behaviours.

In Fig. 5.4 **C**, we analyzed the translocation for the Tumor Necrosis Factor (TNF, also known as TNF-α), a gene known for encoding proinflammatory cytokines ranging from cell proliferation, differentiation and apoptosis (Rzeszotarska et al., 2021). This gene presents late activation and a stronger tendency to be found deeper inside the nucleus.

Finally, the C-C Motif Chemokine Ligand 5 (CCL5 or RANTES: **R**egulated on **A**ctivation, **N**ormal **T** cell **E**xpressed and **S**ecreted) gene and chemokines are also involved in inflammatory processes, being involved in transplantation and tumour development, for example (Selvaraj et al., 2011; Krensky; Ahn, 2007). From the control ($t = 0$), we can see how more accumulated near the nuclear membrane this gene is after TNF-α treatment.

Those gene-specific results proved how despite the fact a gene has its specific activation pattern, the closeness to a nuclear pore facilitates transcription. This result corroborates both our model from Chapter 3 and its expansion from Chapter 4.

5.2 Single-Molecule Tracking (HeLa Cells)

As presented previously, single-molecule techniques are good techniques to understand gene regulation, as we can track small volumes in temporal resolution, revealing patterns and cell stochasticity. Single-molecule tracking (SMT) is a technique based on total-internal-reflection fluorescence (TIRF) microscopy, which is different from fluorescence microscopy (widefield) where the sample is fully excited, the samples are excited in smaller lengths, remaining on a surface level, which increases the quality of the molecules tracked (Vrljic; Nishimura; Moerner, 2007; Moerner, 2015). The use of TIRF microscopy is the first generation SMT. Since the excitation length is smaller than the widefield microscopy, SMT techniques only recover 2D information from the samples (Liu et al., 2016; Fish, 2009).

To create a widefield for the sample while tracking the molecules, one of the techniques is to use a highly-inclined and laminated optical (HiLO) sheet, in which the

laser has a sharp angle than the traditional light beam, and improving the SMT (Garcia et al., 2019; Tokunaga; Imamoto; Sakata-Sogawa, 2008). Incorporating HILO into SMT, we have a second-generation SMT.

In smFISH experiments, we can quantify the mRNA produced and identify the TSs, but the experiment has to fix the cell, meaning we do not have temporal changes inside the same cell. Thus, we wanted to verify the translocation process in one cell over some time. However, we cannot acquire images for longer intervals (more than a few seconds at a time) because the fluorescent components are photobleached - i.e., the fluorescence is damaged by the light.

In Fig. 5.5, we present reference pictures for the HeLa cell line we used in our smFISH experiments. Here, we can see how the TFs (magenta) in the cell are plenty and how the nucleus (blue) has different TF concentrations at different time points, starting without nuclear TFs, to a maximum concentration during the exportation process. The colours were digitally added to the reference channels.

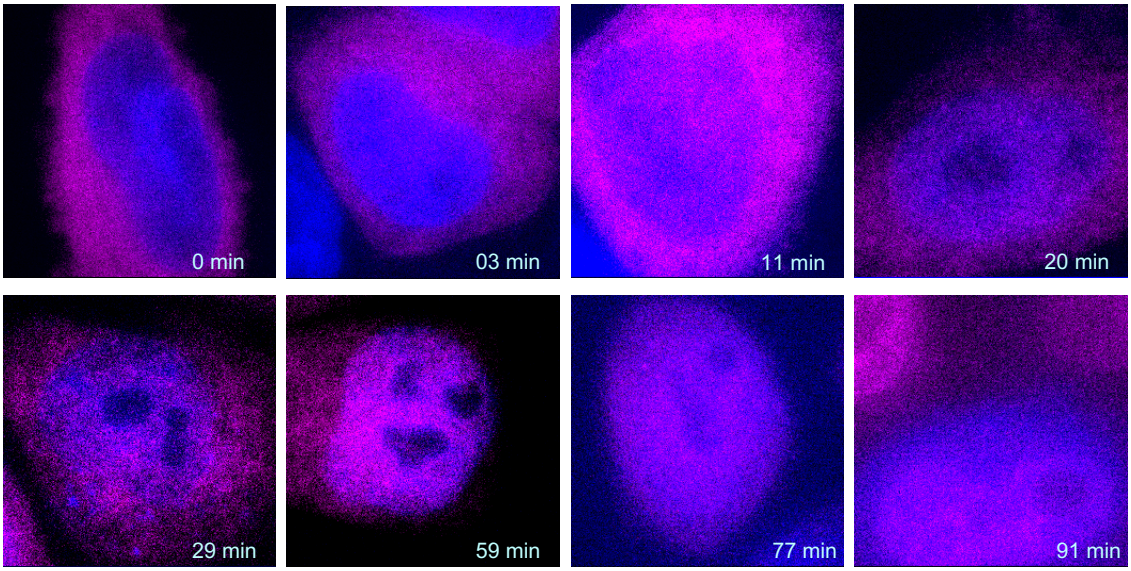

Figure 5.5 – **Merged channels for the references pictures for HeLa cells and different time points and $p65$ as our target TF.** Here, we consider the TFs in magenta and the nucleus in blue. We can see the different concentrations of TFs at different time points, following the translocation-reaccumulation pattern proposed for NF-κb. The window size is 256×256.

We aim to determine the distance between all TFs in the nucleus and the nuclear membrane. Thus, we need to create an image analysis algorithm that takes into account two steps: (i) **Nuclei Definition** and (ii) **Spot Detection**. We exemplified our image analysis for SMT experiments in Fig. 5.6.

For **Nuclei Detection**, we started detecting the nucleus by the brightest pixels, binarizing them, and creating a connected-component label then filling the gaps between

those components, which represents the nucleus (Rosenfeld; Pfaltz, 1966; Dillencourt; Samet; Tamminen, 1992; Cloppet; Boucher, 2010). However, this analysis by the brightness of a pixel is not fully efficient as the microscopy images might have brighter debris than the nucleus or even another brighter nucleus out of focus. In most of the experimental image analyses, we can delete the debris by thresholding the minimum size of a nucleus, which is not an optimal solution for some of our SMT images.

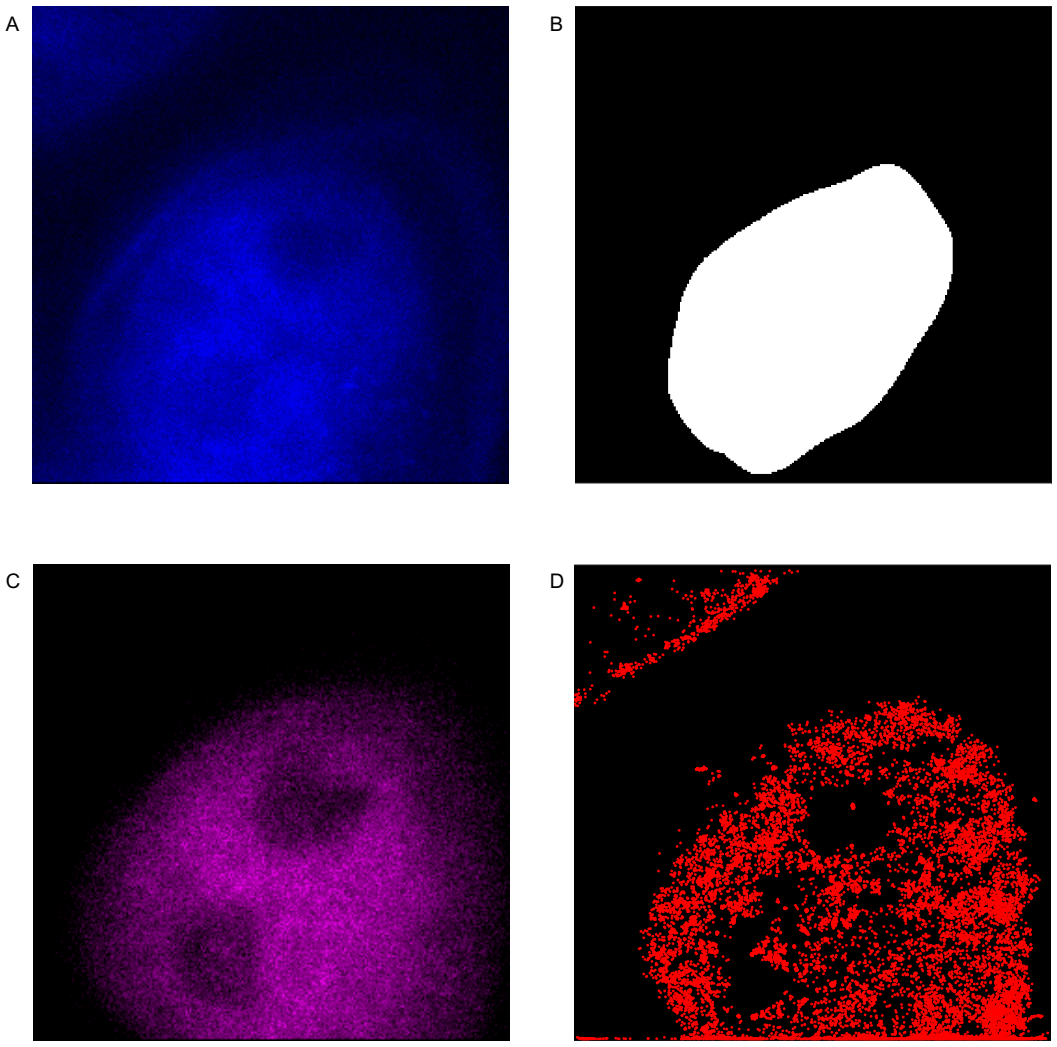

Figure 5.6 – **Steps of our image analysis. A**, reference of the nucleus channel (blue). **B**, nucleus mask. **C**, TF spots in the reference channel (magenta). **D**, spots found from the SMT stacks. Here, we can see how the debris/not-centralized nucleus affects our image analysis. This figure represents our data after 40 minutes of TNF-α treatment.

Since we run all our image analysis in Matlab, which already has enough image analysis tools built-in and the number of images is small enough (48 nuclei split between 12 before TNF-α treatment and 26 different after induction), we used a built-in tool to draw by hand a region of interest (ROI) from each SMT image, which helps us to define the nucleus. To facilitate the detection, we created two different ROIS from the two reference

images (e.g. Fig. 5.5): the nuclei channel (blue, Fig. 5.6 **A**) and the TF spots (magenta, Fig. 5.6 **C**). We combine both ROIs and generate our initial nucleus, to which we apply a gaussian filter, to minimize the noise and filled the eventual holes our binarized images may have. With that, we have our nuclear mask, Fig. 5.6 **B**.

The **Spot Detection** part considers all the 4000 image slices (2 reads per cell, with 2000 milliseconds reads) from our SMT data. Then, our first step is to apply a gaussian filter to the image, to reduce the noise and found the maximum pixel in all rows of our image, and average those values to obtain our threshold for the signal. As this binarized image is dependent on the threshold, we find the number of connected components for the slice, calculating the number of pixels for each connected component. Again, we opted to remove all spots without a minimum number of pixels (in this case, the minimal spot size is 20 pixels). From those selected spots, we calculated their centroids, the volume of found spots is in Fig. 5.6 **D**.

From our segmentation pipeline, we can create a distance map for each nucleus, i.e., we evaluate how far a pixel from a nucleus is from the nuclear envelope. In Fig. 5.7, we show a nucleus mask with all the spots (in red) and its distance map with the same spots (in black). At this time point, most of the found spots were outside the nucleus (i.e., their distance $d(x, y) \equiv 0$).

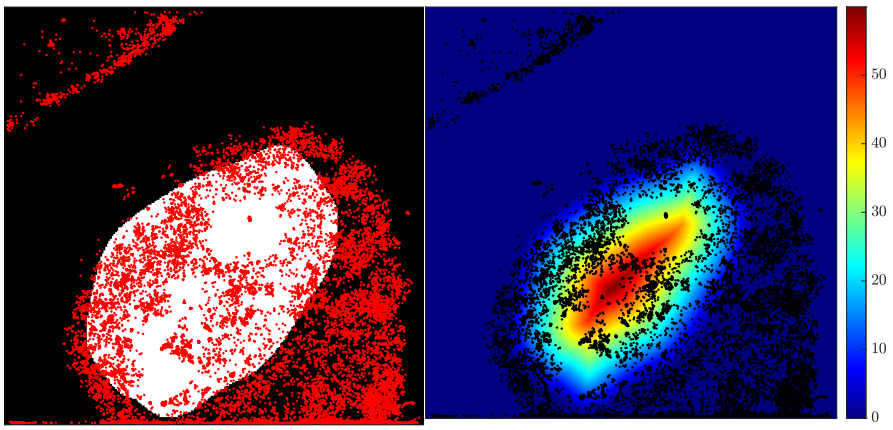

Figure 5.7 – **Distance Map from the nucleus mask.** In the generated mask, the nucleus is white and the spots are red, but in the distance map, since the colour red represents a big distance from the NPC edge, we opted to represent the spots as black dots. Most of the found spots are outside of the nucleus for this time of treatment.

The smFISH data provided us insights about how close to the NPC is a TSs. Hence, we used the SMT experiments to calculate the distance from the edge as represented by Fig. 5.8, for both the spots found and the random positions inside the nucleus. Since the TFs are fast diffusing, we analyzed the early treatment SMT images. Note those measurements were made in different live cells and this means we have cell-to-cell variability in volume and size, affecting the maximum distances inside a nucleus. Another important detail to

consider is the live cells allow TF diffusion, which is faster than our time points (Izeddin et al., 2014).

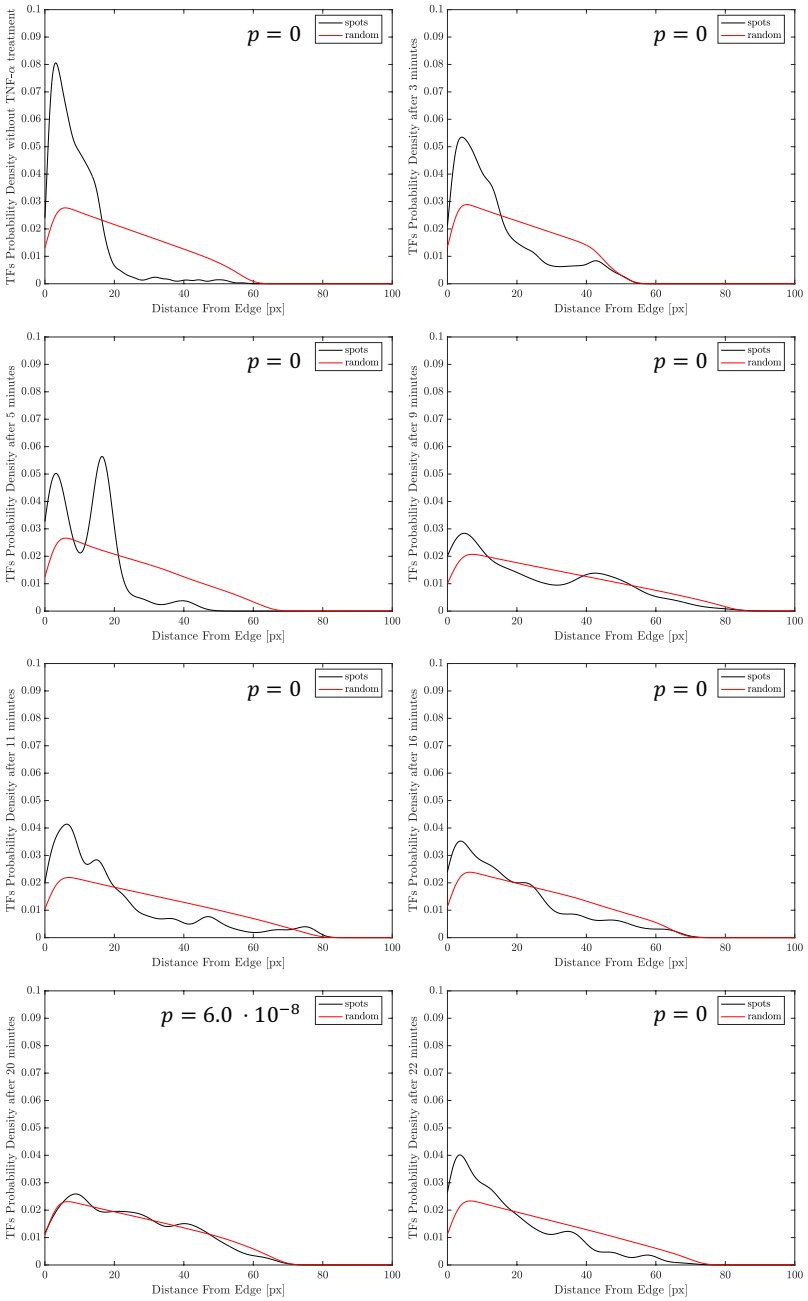

Figure 5.8 – **Probability density for the earlier time points of our SMT experiment with their respective p-values.** As a time progression, we can see how at the beginning most of the spots are found near the nuclear membrane ($t = 0$ (untreated cell) and $t = 3$ minutes after induction). Later ($t = 5$ minutes), we had an increase in TFs in the interior of the nucleus and, in later times ($t = 9$, $t = 11$ and $t = 16$ minutes), the TFs are completely inside the nucleus and randomly occupying positions inside the nucleus, until $p65$ reaches its expected maximum in $t = 20$ minutes and the TFs found to correlate with the random pixels from the distance map, which shows its randomized occupancy. Finally, we can see the re-accumulation in $t = 22$ minutes as the TFs accumulate near the nuclear border.

For each cell and spot, we evaluated their Wilcoxon Rank sum test (Or Mann-Withney test), which is a non-parametric version of the Student's t-test and a tool to compare the location of two independent populations, and the size of the samples can be different (Heiberger; Holland, 2015). We used a function from Matlab to estimate the values from the TFs and the random spots, to verify if the patterns found are significant, and in most of the sets (the whole cell, the localized spots) the p-values were close to 0. This means the results are significant, and the distances (the random and the detected TFs) are independent. The p-values from the probabilities densities show how the random behaviour found in $t = 20$ minutes where the values between the random pixels and the detected TFs showed how correlated both results are, meaning the nuclear TFs around 20 minutes are localized randomly.

We started estimating the behaviour for a condition before the TNF-α treatment, i.e., $t = 0$ minutes. Meanwhile, the random spots distribute themselves between 0 and 60 pixels (red line), and most of the spots (black line) found are closer to the edge, between 0 and 20 pixels, meaning the real spots are more localized in nuclear pore complex edges. The TFs remain in the periphery of the NPC, an expected behaviour for $p65$, as it is a TF from NF-κ b family, which are known signal-dependent TFs.

After 3 minutes of TNF-α induction, the occupancy behaviour changes from being more localized in the border to entering the nucleus efficiently. This significant increase in the spot localization is explained by the fast TF dispersion upon activation, even if most of the spots remain closer to the nuclear border, a result we discussed in depth for our smFISH image analysis. The next cell recorded, 2 minutes later, showed an increased probability of finding a TF in the interior of the nucleus, splitting the probability into two peaks: one very close to the nuclear pore and the other (higher valued) inward of the NPC. It should be noted that this cell has a bigger nucleus than the previous ones and its maximum probability is still around 20 pixels.

After 9 minutes of treatment, we can see a more evenly TF occupation pattern, with the two peaks separating themselves close to the edge and the other to the nucleus centre. This cell is also a bigger cell than the previous ones, which means the TFs have more space to occupy. As the TFs diffuse inside the nucleus, we can see more peaks of probability density formed. For example, in time point $t = 11$ minutes, we have three different expected occupancies: (i) near the border (less than 20 pixels) which has a stronger probability, (ii) middle of the way (more than 20 pixels but less the 60 pixels), and (iii) in the centre (more than 60 pixels). The occupancy spread somewhat follows the randomized pattern (red line).

This randomized localization for the TFs is verified in $t = 16$ minutes as a potential occupancy pattern with the five detected peaks in the previous cell seeming to smooth towards the random positions of the distance map. The predicted behaviour occurs in the

$t = 20$ minutes video, a time point known for being when our system reaches its maximum TF nuclear concentration (Zambrano et al., 2020). In this time point, the TFs are found in all the areas of the nucleus with the same occupancy pattern of the random positions, meaning at this point all the nucleus is fully occupied with $p65$. The re-accumulation process is shown in $t = 22$ minutes after induction, since after the signal-induced TF reaches its maximum the exportation process occurs. Thus, the TFs are found closer to the nuclear border.

We proposed in Fig. 5.9 a comparison between all the probability densities from Fig. 5.8, where all the random of the spread is achieved in the time-lapse and the eventual re-accumulation process. We colour-coded the probabilities to facilitate the visualization of the localization patterns the TFs have, showcasing the flux to the interior of the nucleus and the re-accumulation in the cytoplasm.

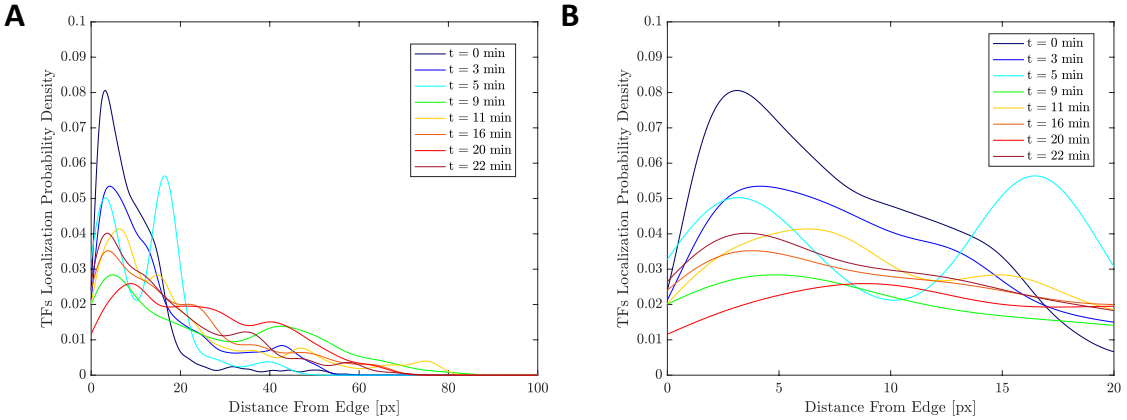

Figure 5.9 – **Probability density for all the time points of our SMT experiment.** All the time densities from Fig. 5.8 for **A** All the Distance from Edge considered (100 pixels) and **B** the 20 pixels from the border, to facilitate visualization of the behaviour around the NPC.

The SMT image analysis helped us to understand more about our system: for example, we verified that even if the TF disperses randomly in the nucleus, most of the TFs found remain closer to the nuclear edge. The experimental results proved the claims from our model in Chapter 3 about the influence of the position of a transcription site and its distance from the nuclear membrane and how being close to a nuclear pore increases transcriptional activity.

From our experimental data, we can use our image analysis algorithms to differentiate, for example, if the mRNA floating in smFISH experiments are inside or outside the nucleus and, with this, we can have an estimate of the cytoplasmic and nuclear mRNA concentrations, for example. However, since we do not have a current interest in such results, we did not present this analysis. Moreover, our image analysis was a great tool for validating our mathematical models, being a great starting point for further experiments.

6 Discussions

Chromatin is a compact environment for any molecule to explore, e.g., the process of a TF searching its target site, i.e., there are many key factors to consider in gene expression than just a gene being *on* or *off*, some of them yet to be discovered. TFs are motif-specific, so they must bind specific DNA sequences in order to transcribe and this sequence specificity influences the TF search mechanism as TFs do not represent a diffusive process. Thus, most of the models of this mechanism assume the facilitated diffusion process (3D/1D) (Zabet; Adryan, 2012; Mirny et al., 2009; Woringer; Darzacq; Izeddin, 2014).

We started Chapter 2 by presenting our own analysis of the model proposed by Molina's lab in Eq. (2.3), (Avcu; Molina, 2016). To incorporate the facilitated diffusion mechanism, we supposed the search through the chromatin network (which we constructed using Hi-C data to estimate the connection between two regions) represented the 3D diffusion process, and the accessibility/good binding motifs to estimate the TF residence times, which represent the 1D sliding process. We based our studies on the Hi-C data for chromosome 19 in $5kb$ resolution from (Rao et al., 2014) and the residence times based on $p65$ binding motifs. We presented the probability density for our network from our model is present in Fig. 2.3 **A** and the residence times in Fig. 2.3 **B**, in which we demonstrated the constructed bimodality of active/inactive regions.

Even if the model in Eq. (2.3) is easily solvable, we opted for studying the steady states (Eq. (2.5)) to test how the structure and the residence times influence the TF occupancy pattern. First, with a biologically defined parameter set, we obtained Fig. 2.4, where we uncovered its occupancy pattern and also two separate clouds of allocated TFs that represent active/inactive chromatin regions (Fig. 2.4 **C**), defined by the values of τ.

Then, we generated two artificial conditions and two randomized ones to understand Eq. (2.3). The artificial conditions helped us to understand the separate roles of the residence times and structure. First, we proposed a fully-connected network, which facilitates the TF diffusion, as any region is one step away, i.e., τ_i is the only parameter that affects the TF occupancy Fig. 2.5 **A** (the linearity due to the values of τ was shown in Fig. 2.5 **B**).

Later, we considered all regions had the same residence time (i.e., $\forall i \in \mathbf{N}, \tau_i = \langle\tau\rangle$), thus the search process is a 3D diffusion through the chromatin network and the TF occupancy depends on the connectivity of a node. This condition spreads more of the TFs in the chromatin network than the results we had previously. By comparison, we did not obtain transcription hubs in specific - more attractive to the TFs - regions, as

one might conclude by comparing Figs. 2.4 **A** and Fig. 2.5 **A** with Fig. 2.6 **A**. Again, we verified the linearity in the TF allocation with the remaining region-specific parameter, the connectivity, Fig. 2.6 **B**.

We assumed a randomized condition as rearranging the positions inside **N** to obtain a different system but the same values from Fig. 2.3. Then, we fixed one of the values to explore the consequences of the other in our system.

First, we randomized the nodes of our network (Fig. 2.7 A), creating a different network with the same connectivity. Since the chromatin moves inside the nucleus and is not fixed as our model assumes, a change in the network with the same residence time values can be understood as a different chromatin network but the same TF, as the TF occupancy pattern in Fig. 2.7 **B**, where we had a similar split between inactive and active regions from Fig. 2.4 **C** in Fig. 2.7 **D**.

Then, we randomized the values of τ maintaining our initial network, Fig. 2.9 **A**, which one might understand as a different TF - with a different motif in the cell. The TF occupancy pattern (Fig. 2.9 **B**) showed a decrease in concentration for the highly occupied regions when we compare with Figs. 2.4 **A** and 2.7 **B**.

In addition, when we evaluated the $\log_2$ of Fig. 2.4 **A** over $\log_2$ Fig. 2.9 **B** (Fig. 2.10), we obtained the same pattern from the randomization, Fig. 2.9 **A**, which it did not occur for the randomized network. Thus, from our steady-state studies, we conclude that while both structure and residence times play important and complementary roles in the TF search process, accessibility is a key feature for gene expression, further confirming the need for a facilitated diffusion process to represent TF searches in the nucleus.

As the results from our steady states showed, some regions are more prone to accumulate TFs than others, and TFs in such regions might interact with each other in weak protein-protein interactions (PPI), which facilitates the recruitment of more TFs. Therefore, to understand the TF occupation process in the chromatin, we assumed the TFs can form PPI between themselves. However, since chromatin is limited in space, we also opted to include the presence of volume exclusion in our model, Eq. (2.6), and, despite accessibility being an important factor for the search for target sites, we considered all regions allow the same maximum TF concentration, C. We are aware open regions have more space to fit a TF than closed ones but we decided to maintain the constant values for C and leave this region-specific volume exclusion for a later extension of our model.

To understand cluster formation, we implemented deterministic solutions for Eq. (2.6), fixing the dissociation rate ($K_d = 1s^{-1}$), and assuming a reduced network, which we presented the probability density in Fig. 2.12. We run Matlab simulations for $t = 400$ seconds, in which the system is still far from the steady-states (Eq. (2.8)) for different association rates, K_a, carrying capacities, C, and total TF concentrations, $[T]$.

By analyzing the association rate with a fixed C, we verified how the increase in the K_a values (Fig. 2.14) creates well-connected (clusters with higher PPI, $I = 3$) clusters around regions with better residence times. We deepened our analysis in three different regions, all active but with different values of τ and d (Table 1): (i) highly-connected regions with a low residence time; (ii) a region with relatively low connectivity but a high residence time; and (iii) a region with high values of both τ and d, Fig. 2.15 **A**, **B** and **C**, respectively. Here, we restricted our analysis to the highest PPI, $I = 3$, as it represents the strongest cluster formation our system can achieve.

In Fig. 2.15, we established that for regions with good residence times, non-negative values of K_a impact the TF concentration in $I = 3$, yet the number of connections can accelerate the maximum formation of protein-protein interactions. Meanwhile, regions with smaller residence times but high connectivity showed a strong increase in TF concentration that, with the increase in the efficiency of K_a in forming clusters and allowing a greater concentration of TFs in all regions (by expanding the carrying capacity), this region has its concentration reduced, as TFs occupy other, more attractive, regions.

Since the maximum allowed occupancy influences TF activity, we also studied the effects of C in our system, Fig. 2.17, in which we verified again that the clustering around prolific regions is enhanced depending on the values of C, even if the region does not reach its upper bound, Fig. 2.18. We considered the same regions from Fig. 2.15 to analyze how changes in C affect our system and obtained Fig. 2.19.

We could exemplify how the K_a influences the TF allocation for different values of C: as expected high values of K_a enhanced the regions with good motifs (higher $\tau's$). However, once we analyzed the connected regions, we found that lower values of K_a are not enough to obtain the $I = 3$ as the biggest TF concentration, which is a result depending on d. Therefore, structure and residence times continue to be key factors for the TF occupancy that are later regulated by the carrying capacity and association rate.

Finally, since the number of TFs available affects the contact between two or more TFs, we decided to verify the cluster formation for different values of $[T]$ as present the global effects of the total TF concentration in Fig. 2.22 for different values of K_a and C. Once we analyzed $[T]$, we found how by changing the concentration we improve or worsen the clustering and how the benefit or harm is also region-specific depending on $[T]$, Fig. 2.24. Thus, we can consider the total nuclear TF concentration as an internal mechanism to control gene expression.

By analyzing the same regions from Figs. 2.15 and 2.19 in Fig. 2.24, we once again obtained that smaller values of K_a favour regions with better residence times, which must be bigger for regions with smaller $\tau's$. Interestingly, the TF concentrations drop for active regions as other regions organize their clusters. Once more, our simulations were stopped before reaching stability, so the rearrangement pattern just showed our model tendencies,

and not the equilibrium.

Chapter 2 helped us to comprehend how the facilitated diffusion process models TF searches and how by limiting the maximum concentration allowed in a region we influence the TF occupancy patterns increasing the number of TFs on less attractive regions. More than that, by assuming the PPI as an inherent condition for the TF to form transcription hubs and influence transcription, we uncovered that while higher values of K_a seemed to influence cluster in prolific nodes, regions with less expected TFs when $K_a \equiv 0$ benefit more from smaller values of K_a. Still, the volume of TFs available also impacts the TF organization in the chromatin.

Thus, even with the limitations in scale, Eq. (2.6) is a good model to explore TF occupancy conditions and potential interactions inside chromatin. Yet, the TF search process is the first step in gene regulation and this TF still needs to recruit polymerase to start transcription and produce mRNA. We studied this recruitment in Chapter 3 and the mRNA export, which eventually leads to protein production in Chapter 4.

In Chapter 3, we regulated the TF occupancy by limiting the transcriptional resources; thus, our model considered different states for the TF/RNAP to represent their interactive dynamics and mathematically understand transcription. The TF two-state model (Free and Bound) and three states for RNAP (Free, Bound and Transcribing) seem sufficient to represent the complexities of this system in the same scaled network from Chapter 2 and with our working model present in Eq. (3.3).

We expressed our model steady-states in Eqs. (3.4), which intrinsically depends on the number of connections between regions, d and the residence times, τ. Such dependency means transcription should be more prolific in active regions (i.e., regions with higher values of τ), which are usually more connected, and thus easily reached, even if TF/RNAP molecules are fast diffusing, a similar result from previous models.

It is clear from Eqs. (3.4) that the Bound TF state influences both Bound and Transcribing RNAP states because of the nonlinearity present in the RNAP Free equation. This non-linearity also increased the structural effects on the other states of RNAP, as we showed in Fig. 3.3. The Bound TF state is fundamental to our system equilibrium as we showed in our characteristic polynomial which is the only explicit value from our variables.

Eq. (3.3) is good to predict the occupancy pattern from a given fixed network without degradation of TF/RNAP. However, some transcription factors remain outside the nucleus and need an activation mechanism to enter the nucleus and start transcription. NF-κb, for example, is a family of TFs with a translocation mechanism and its non-constant concentration inside the nucleus affects transcriptional volume for their target genes, as it influences the TF binding and RNAP recruitment.

To incorporate the change in TF concentration in our model, first, we proposed

the presence of regions connected to the nuclear pore, i.e., regions from our network that are so close to the nuclear pore complex (NPC) we can consider to be connected. Since RNAP does not translocate to the nucleus because it is found only in the nucleus during the cell-growth phase, we only consider TFs can leave cytoplasm and enter the nucleus. Once a TF enters the nucleus, it starts exploring our network in the Free state before binding to a region and recruiting a Free RNAP.

To verify how the depleting TF concentration affects the transcriptional activity, we proposed two different flux import functions: **Import Function**, in which the TF translocates into the nucleus and remains inside, eventually reaching its equilibrium (Eqs. (3.4)), as represented in Fig. 3.5 and Eq. (3.5); and **Import/Export Function**, where a TF enters the nucleus, its maximum concentration is reached and then the exportation process starts, proposed in Eq. (3.11) and represented in Fig. 3.6.

For both flux functions, we calculated their analytical expressions, which is not a straightforward result given our system is a non-linear and non-homogeneous ODE system with $5L$ equations (L being our network size), but we can split our solutions in TF solution and RNAP solution. The solution depends on the flux function considered.

For import-only flux, the TF equations are linear and non-homogeneous, and the linearity of the system guarantees the existence of a well-posed analytical solution, Eq. (3.6). This is not a possibility for the RNAP solution, given our system is a non-linear homogeneous ODE system, we depended on the commutativity of our system to solve it. Unfortunately, our system is not commutative (Eq. (3.8)). To solve this commutativity problem, we proposed the usual commutator and used the Magnus Expansion to obtain an infinite series of integrals that represents our RNAP solution, Eq. (3.10). Therefore, the solution depends on the solution convergence, which is dependent on the parameters considered.

However, to calculate the TF/RNAP analytical solutions for the import-export flux, we encountered non-linearity in both systems. Besides being non-linear, the TF equations are also non-homogeneous, meaning the expression for its solution is more complex than just a Magnus Expansion expression, Eq. (3.12). Since we need the TF equation to obtain the RNAP expression and how complex is the solution for the TF, Eq. (3.14) is expected to be very complex.

Again, it is numerically possible to evaluate our model in higher resolution networks by using our proposed solutions, if we can prove the convergence of the series $\Omega_X(0,t)$, for both TF and RNAP, where it fits. The analytical expressions are a tool to exemplify why numerical solutions are needed in mathematical modelling: a seemingly simple model having such complex analytical expression.

Yet, we opted to numerically evaluate our model using both flux functions with two

different methods, first with an ODE solver (ode15s from Matlab) and then implementing a Gillespie Stochastic Algorithm. Both implementations used the same parameters (Table 2).

For the first case, the Import Flux, our system reaches the steady state early on, Fig. 3.7, and regions connected to pores present an overshoot of TF states before stabilising. The activity is higher in the active regions connected to pores. Most of the regions showed inactivity, i.e., the concentration is considered too low to be significant. Comparing different subnetworks, we can see the regions close to a nuclear pore (i.e., regions connected to a pore and their immediate neighbours) are more prolific than other regions.

Since our model reaches the steady state, we proved no regions present a late-activation or deactivation in these conditions, a result that proves TF/RNAP diffusive rates are fast as expected Fig. 3.9 and its ability to reach the equilibrium. After we analyzed how long it takes for each node in our network to reach its maximum concentration, we showed some regions reach their maximum at the end of our simulation, which is a consequence of the infinitesimal differences between the time points of implementation from an ODE solver continuity.

The importance of the τ and d are explicitly shown and our model proved the residence time is a more fundamental component to transcription than the number of connections a node has, an expected result in fast-diffusing molecules. Those results were continuous and gene regulation (and transcription for that matter) are stochastic in nature. Thus, we implemented a classical Gillespie SSA 70 times, to mimic 70 different cells and from this implementation, we verified the fraction of active target sites in our network, Fig. 3.10.

The importance of the τ values in the activity is shown for all the subnetworks except for the pore-connected ones, which means that regions connected to pores will receive enough TFs to sustain their transcriptional machinery. However, this importance on the residence time of a region implies that the facilitated diffusion model was correct and there are different components in transcription that a 3D diffusion model will not capture.

When we compared the deterministic (continuous) with the stochastic (discrete) implementations Fig. 3.11 **D**, we found most of the regions are impoverished by the limited resource of TF/RNAP molecules in our system. Active regions, however, are regions enriched by the stochasticity of our system. Therefore, our model can represent the stochasticity found in gene expression and explain well why some genes are more likely to be transcribed than others.

It is expected that the volume of TF/RNAP molecules present in the nucleus influences the volume of transcription. To simulate the TF translocation/re-accumulation

process, we used NF-κb translocation dynamics to understand how a non-constant TF concentration affects transcription. For this flux function, our system does not reach stability since there is either an import or export process occurring.

In the analysis of the averages of TF/RNAP molecules in our system, Fig. 3.13, we verified that while on average our system is deactivated (i.e., the concentration of Free RNAPs surpasses the Transcribing again), active regions remain activated after 180 minutes. From our model, we predicted that around the 80-minute mark, the TF nuclear concentration is half of its maximum, which is verified in a change of behaviour for all the subnetworks with a stronger decrease in activity. However, our model allows the proportion of $1:4$ for Free and not Free (Bound and Transcribing) RNAPs. As a consequence, our system is highly prolific even with smaller concentrations of TFs.

Once more, active regions produce stronger activities with regions not connected to pores presenting higher averages. Thus, we verified different subnetworks as they showed the optimal choice of being close to a nuclear pore for the import-only flux function. Fig. 3.14 exemplified how is preferential for a region to be close to a nuclear pore (i.e., one step away maximum from a pore-connected region), even if, as we discussed exhaustively, TF/RNAP are fast diffusing through chromatin to find good binding sites.

Differently from the previous model, we encountered a deactivation process in our simulations due to the decreasing TF concentration. In an import-only condition, we found an absence of late activation in our model, which is replicated for this flux function. Since this model has a deactivation process, the re-activation was a possibility but we did not find it, probably because the active regions secure the available TFs to themselves in order to maintain their transcription function. We uncovered the importance of a good binding site from the residence times values, i.e., attractive regions are more likely to be transcribed.

Deterministic solutions are useful to obtain fast results and to understand averaged patterns from our model, but the solutions assume the possibility of splitting molecules between regions which are not biologically feasible. Therefore, given the discrete occupancy of TFs and the stochasticity experimentally observed in cells, we proposed a Gillespie SSA. However, Gillespie assumes the next reaction is time-dependent and our flux function (Eq. (3.11)) is explicitly time-dependent. To solve this time dependency, we proposed a Hybrid Gillespie SSA.

Since we proposed a model for transcription, we analyzed the fraction of active target sites in our stochastic simulations, Fig. 3.17 for different subnetworks and sorted by both the network's connectivity and residence times. As the previous results showed, the residence times impact more the transcriptional activity than the connectivity for all the subnetworks *except* the pore-connected subnetwork, C.

This seeming lack of importance for the parameters is explained by the definition of a pore-connected region: by construct, we randomly selected active regions, which means transcription has a strong probability to occur in those particular regions. Since those regions receive the Free TFs on at least two different occasions, the import and the export processes and they are attractive regions by definition, transcription is expected. However, the presence of inactivity is a consequence of fast diffusing TFs and other active regions available to transcription.

When we compared Figs. 3.11 and 3.18, we saw the on-time averages (the interval between initiation and an elongation reaction) were similar, but the effective initiation rate is reduced in the import/export flux model as a consequence of less available transcriptional resources. Another interesting result is the lack of variation of active regions for the import-only function, which is explained by its lack of late-activation/ deactivation. Even if late activation is not found in our import-export function, the effects of deactivation are visible even in prolific regions. The pattern of enrichment/impoverishment for the regions is also found in Eq. (3.11).

Therefore, our model reproduces well gene expression with the given conditions but fails to consider the changing in connections and accessibility a region might present during interphase. Our simulations considered a small time frame to ignore those changes, thus forgoing chromatin condensation and bookmarking, both important for gene regulation (Luo et al., 2017; Engeland, 2017; Schmitz; Higgins; Seibert, 2020). Thus, a model that considers longer periods of the cell life cycle, must consider the chromatin remodellation process.

In Chapter 2, we proposed the presence of volume exclusion in the chromatin regions, which we ignored in our model, allowing regions to hoard as many TFs as Eq. (3.4) allowed since we restrained our system by limiting the number of available TFs. However, even if the clustering in specific regions was allowed, the depleting TF concentration was the only explicit mechanism to regulate transcription in the flux function in Eq. (3.11).

Another important point is our model in Chapter 3 predicts transcription activity but not how much mRNA each region is producing. Thus, we proposed an extension of our model to admit mRNA synthesis and exportation to the cytoplasm, incorporating translation into our model in Eq. 4.1 in Chapter 4.

There is no cell maintenance without the translation process and, in eukaryotes, this part of protein synthesis occurs in the cytoplasm while transcription is a nuclear process, which is a mechanism to troubleshoot protein synthesis, and involves the active transport of mRNA from the nucleus to the cytoplasm. In Chapter 4, our focus was the comprehension of the mechanisms behind mRNA export. We proposed two different types of modelling in Chapter 4: (i) incorporating mRNA export to our model in Chapter 3 and (ii) an expansion of the RNA Velocity model in collaboration with Manuel Mendoza's lab from IGBMC.

First, we want to understand how the structure and pore-connectivity influence the mRNA export. Therefore, our model in Chapter 4 is an extension of our model in Eq. (3.3) with the flux dynamics from Eq. (3.11) incorporating the mRNA synthesis and export dynamics with Eq. (4.1), which we implemented deterministic solutions for all the genes our network. Since this model is an extension of the one proposed in Chapter 3, solving it analytically requires the use of the Magnum's expansion for the Transcribing RNAP state, a mathematically heavy system to solve by itself. Thus, we implemented the deterministic solutions for all genes g in our network with the use of the ode15s tool from Matlab.

The nuclear mRNA from any region must be exported to the cytoplasm which means the chromatin network plays a fundamental role in the mRNA dispersion in the nucleus leading to its exportation. As the regions have bigger averages of connectivity with their neighbours because of the TAD formation, if we analyzed the concentration of nuclear mRNA for all genes g in our network in Fig. 4.3. The results in Fig. 4.3 showed how different regions are active in different ways, a result we uncovered with distinct mechanisms in Chapters 2 and 3, meaning our model captures the differences in transcriptional activity per region even if our current resolution does not allow us to predict specifically which gene is active. The array of different activities is more easily verified in Fig. 4.4.

Globally, we verified how each subnetwork behaved with its nuclear mRNA concentration for each gene g, Fig. 4.5. There, we obtained an interesting result: while the regions connected to a nuclear pore showed activity, high mRNA concentration was not an exclusivity of those regions. Once we clustered our results, Fig. 4.6 **A**, we found that despite the distinct concentration levels, there are no changes in behaviour, a result expected given Fig. 3.15 and the mRNA dependency on the Transcribing RNAP allocation patterns. The z-score, Fig. 4.6 **B** showed all the regions activate around the same time, reaching their maximum at distinct points, but deactivating similarly.

We proposed the same analysis for the cytoplasmic mRNA. Of course, the peak in mRNA concentration is shifted as the mRNA must travel through chromatin to find a nuclear pore and get exported, in Fig. 4.7, for example, the peak of concentration is found around 40 minutes. This delay in maximum concentration for the cytoplasmic mRNA is verified in Fig. 4.8 for different subnetworks. The accumulation in the cytoplasm is increased for regions connected to pores since once an mRNA is produced, it is exported.

Once more, the cluster analysis (Fig. 4.9 **A**) showed that we obtained a difference in concentration levels. The z-score (Fig. 4.9 **B**) showed that the activation is not delayed but each clustered z-score reaches its maximum at different time points. Yet, the deactivation process is similar, as we obtained previously in Fig. 4.6 **B**.

To conclude our analysis in Eq. (4.1), we compared the mRNA concentration inside and outside the nucleus in Fig. 4.10, comparing with values of d and $\log_2 \tau$. As we extensively discussed in Chapters 2 and 3, the structure and residence times control the

occupancy pattern for TF/RNAP that leads to transcription. The structure and DNA sequence influence remains for the mRNA. This analysis showed an outlier of regions connected to pores accumulating more strongly in the cytoplasm. Therefore, our model helped us to understand mRNA export and, given only cytoplasmic mRNAs get translated, gene regulation. More so, we proved how pore connectivity improves the transcript levels in the cytoplasm. We believe further experiments confirming this result are necessary and welcome.

However, our model in Eq. (4.1) does not consider region-specific parameters, and we know some genes are more prolific than others - some are necessary in bigger volumes for example, we need to consider gene-specific parameters for mRNA export. To do so, we opted for ignoring the chromatin structure and binding/unbinding processes and focusing on RNA velocity models, Eq. (4.2). This model was used to understand sequencing data from our collaborator, Manuel Mendoza.

Based on RNA velocity models, we proposed a three-state model for mRNA production to represent the **unspliced mRNA**, the **nuclear-spliced mRNA** and the **cytoplasmic spliced mRNA**, Eq. (4.2). We used this model to analyze sequencing data from 55400 genes and three different sequencing methods:(i) RNA-Sequencing (RNA-Seq) - represents unspliced and spliced mRNA in the equilibrium; (ii) single-cell RNA-sequencing (scRNA-Seq) - unspliced/spliced mRNA after 15 minutes of stimulation; (iii) Fractionation RNA-sequencing (Frac-Seq) - nuclear and cytoplasmic unspliced/spliced reads. To obtain significative results, all the experiments have at least three replicas of both control and auxin-treated cells.

In Eq. (4.3), we proposed the steady-states values for our model, in which the steady-states are global attractors, i.e., it describes a stable node, which means that given enough time, all solutions converge to Eq. (4.3). We verified this behaviour in Fig. 4.12 with random parameters from Table 5.

Our model is analytically solvable if we solved it in the following order: the unspliced, the nuclear spliced and the cytoplasmic spliced. We also admitted there are zero mRNAs at time $t = 0$. Therefore, we organized the solution as (i) **unspliced mRNA**, Eq. (4.4); (ii) **nuclear spliced mRNA**, Eq. (4.5) and (iii) **cytoplasmic spliced mRNA**, Eq. (4.6). Using the same parameters in Table 5, we verified our solutions and the numerical solution we implemented by using the Matlab tool ode15s, the results available in Fig. 4.13.

Given the sequencing data, we used the reads to estimate the parameters using Bayesian Inference, we obtained the values for the transcription rate for gene g, α_g in Eq. (4.7) and the unspliced scRNA-Seq data. Given the steady-states stability proposed in Eq. (4.3) and what each sequencing experiment represents, we verified the reads for β_g, k_g and γ_g are inverse to the values of U_g^N, S_g^N and S_g^C respectively. Thus, we obtained the expression for η_g for each parameter by applying again Bayesian inference and obtained

the desired values by dividing α_g by η_g. An important distinction to make is that some genes do not require splicing; thus, we consider the spliced scRNA-seq values to evaluate α_g and considered $\beta_g \to \infty$ and, as a consequence, $U_g^N \to 0$. We presented the parameter space in Fig. 4.14.

The experimental setup was to comprehend how the auxin treatment affects the transcription for different genes. Thus, in Fig. 4.15, we proposed a comparison of the parameters in control cells and auxin-treated ones. The correlation between auxin-treated and control cells is higher than 0.5 and we also proved how auxin affected each parameter: the transcription and export rate, α_g and k_g, for example, were not as affected as the splicing and degradation rates, β_g and γ_g. This means that while the transcription and the export rate remain unchanged - i.e., the cells produce similar values of gene g and such gene is exported with the same efficiency. However, auxin affects the splicing of exonic genes (not evaluated for intronic genes, as they do not require splicing) and gene degradation.

Therefore, we verified how the auxin influenced our genes by evaluating its z-score and fold-change, Fig. 4.16. However, to obtain high-quality values for better-fitted genes, we evaluated the correlation between the reads from sequencing experiments (experimental data) and the values we obtained from the steady-states (analytical data) and selected the values with the correlation between analytical and experimental values with higher than 0.4. As a result, we obtained 9094 genes to analyze in Fig. 4.16.

Corroborating our correlation between control and auxin-treated cells in Fig. 4.15, we found that the percentage of genes that were not affected by the auxin treatment is more than 60% for all the parameters. More than that, most of the affected genes are downregulated instead of upregulated (between $56-72\%$ against $28-44\%$) by the auxin treatment. Thus, we conclude that by treating a cell line with auxin we can arrest the mRNA synthesis for most of the genes. Another interesting result was the gene length independence in mRNA export as we did not find any correlation between the parameters and the gene length.

From Chapter 4, we obtained insights about gene expression from Eqs. (4.1) and (4.2). Even with simple deterministic models, we verified how the volume of mRNA synthesis is gene-specific and should be considered as such and how with simple inferences and data analysis we can understand the effects of treatment in a cell line. Therefore, merging both experimental and theoretical research is the best path to understanding the mechanisms behind gene regulation.

To incorporate experimental results in our theoretical claims, we considered our results in Chapter 3, where we had two main hypotheses to experimentally verify with image analysis: (1) Active target sites remain in the periphery of the nuclear envelope and (2) The expected localization for TFs in the nucleus change over time because of the

translocation process. We used two different microscopy experiments to corroborate our claims, smFISH and SMT both for MS2 fused HeLa cells and $p65$ (RELA) as the target TF. The choice of TF was due to the fact $p65$ is a TF from the NF-κb family, which are inducible TFs that remain in the cytoplasm if not activated, giving us the control to know how much time elapsed after the activation.

The first experiment was smFISH for the cell line in which we fixed the cells after different moments of TNF-α treatment - a cytokine that induces NF-κb translocation to the nucleus - and different probes to target different genes. Thus, we had different genes at different time points (with replicas) to analyze the distance between a target site and the nuclear membrane of a cell. To segment the maximum projection of our cell, we used the watershed algorithm from Matlab to separate the nuclei and selected the target site as the maximum four bigger and brighter spots (as HeLa might have more than two chromosomes in the nucleus), since a TS is crowded with transcripts, increasing its size and intensity in comparison with a sole transcript, as we exemplified in Fig. 5.1.

With the nucleus and target sites properly detected, we evaluated the distance of a pixel from the edge of this nucleus and positioned our the centre of the TS inside of the cell, Fig. 5.2. From this result, we saved all the distances from the edge and calculated the probability density (Fig. 5.3 **A**) and, to eliminate the possibility of our result being an artefact, we also found the distance for generated (random) spots with the same size of the TSs (Fig. 5.3 **B**).

In Fig. 5.4, we proposed the time evolution for different genes and random spots. We verified the localization patterns for the genes in which we found different activation patterns being early or late depending on the gene function. However, most of the detected spots were found in the periphery of the nucleus, which indicates the conclusion in Chapter 3 (and its later expansion in Chapter 4) are experimentally justified.

Thus, we conclude that for an inducible TF, the target sites can be found closer to the nuclear envelope, as it facilitates the produced mRNA to export to the cytoplasm and accelerates the translation. For other TFs, further experimentation is necessary.

We used the single-molecule tracking (SMT) for the same TF and cell line to understand the translocation pattern and TF occupancy inside the nucleus, Fig. 5.5. We confirmed the changes in the TF density gradient between cytoplasm and nucleus: initially, there is a separation between the TFs (magenta) and the nucleus (blue) which disappears due to TF import of the translocation process and reappears because of the re-accumulation process.

For the image analysis, since we dealt with short videos of live cells, and we can only track one cell per time point - we exemplified the nucleus definition and spot tracking in Fig. 5.6. As this experiment is live-imaging for the TF and each stack has 2000 images,

we can see how many spots are detected from one stack in Fig. 5.6 **D**. Another important factor is that we calculated the distance from inside the nucleus, i.e., any TF found outside the nucleus is not considered to not create a bias towards the nuclear border.

Once more, we evaluated the distance in pixels from the nucleus towards the nuclear centre (Fig. 5.7 **B**) and then analyzed the TF localization for different time points with the detected spots and random spots, Fig. 5.8. From this result, we verified the absence of artefacts influencing the TFs to be localized near the nuclear border and we also uncovered the changing pattern of TF accumulation.

Before the TNF-α treatment, the TFs are more likely to be found around the nuclear membrane, with the translocation activated, the probability to find TFs in central parts of the nucleus increases until the random spots and the real TFs have virtually the same probability density (around the known maximum TF concentration for $p65$); then, the cytoplasmic re-accumulation starts and the probability indicates a tendency of being near the edge once again. This description was found in a time evolution in Fig. 5.9.

Through this thesis, we tried to understand gene regulation not only as a biological process but rather as a mechanistic phenomenon. As far as modelling goes, we are confident our research brought to light key factors for transcription, even if we could not implement it genome-wide or even in higher resolutions. We also proved how *in silico* models can be good exploratory tools for complex systems as biology often shows itself. Furthermore, this project was the result of both an internal and external search for more interdisciplinarity in the sciences and can indeed be considered a success within this scope.

References

Alberts, B. (Ed.). *Essential cell biology*. 2. ed. ed. New York [u.a.]: Garland, 2004. CD-ROM u.d.T.: Essential cell biology 2 interactive. ISBN 0815334818.

Alberts, B. et al. *Molecular Biology of the Cell*. 4th. ed. : Garland, 2002.

Allen, T.; Cronshaw, J.; Bagley, S.; Kiseleva, E.; Goldberg, M. The nuclear pore complex: mediator of translocation between nucleus and cytoplasm. *Journal of Cell Science*, The Company of Biologists, v. 113, n. 10, p. 1651–1659, may 2000.

Almassalha, L. M. et al. The global relationship between chromatin physical topology, fractal structure, and gene expression. *Scientific Reports*, Springer Science and Business Media LLC, v. 7, n. 1, jan 2017.

Anink-Groenen, L. C. M.; Maarleveld, T. R.; Verschure, P. J.; Bruggeman, F. J. Mechanistic stochastic model of histone modification pattern formation. *Epigenetics & Chromatin*, Springer Science and Business Media LLC, v. 7, n. 1, p. 30, 2014.

Antikainen, A. A.; Heinonen, M.; Lähdesmäki, H. Modeling binding specificities of transcription factor pairs with random forests. *BMC Bioinformatics*, Springer Science and Business Media LLC, v. 23, n. 1, jun. 2022. Available on: <https://doi.org/10.1186/s12859-022-04734-7>.

Aptekmann, A. A.; Bulavka, D.; Nadra, A. D.; Sánchez, I. E. Transcription factor specificity limits the number of dna-binding motifs. *PLOS ONE*, Public Library of Science, v. 17, n. 1, p. 1–13, 01 2022. Available on: <https://doi.org/10.1371/journal.pone.0263307>.

Arensbergen, J. van; Steensel, B. van; Bussemaker, H. J. In search of the determinants of enhancer–promoter interaction specificity. *Trends in Cell Biology*, Elsevier BV, v. 24, n. 11, p. 695–702, nov. 2014. Available on: <https://doi.org/10.1016/j.tcb.2014.07.004>.

Arnal, A.; Casas, F.; Chiralt, C. A general formula for the magnus expansion in terms of iterated integrals of right-nested commutators. *Journal of Physics Communications*, IOP Publishing, v. 2, n. 3, p. 035024, mar 2018. Available on: <https://doi.org/10.1088/2399-6528/aab291>.

Avcu, N.; Molina, N. Chromatin structure shapes the search process of transcription factors. Cold Spring Harbor Laboratory, apr 2016.

Ay, A.; Arnosti, D. N. Mathematical modeling of gene expression: a guide for the perplexed biologist. *Critical Reviews in Biochemistry and Molecular Biology*, Informa UK Limited, v. 46, n. 2, p. 137–151, mar. 2011. Available on: <https://doi.org/10.3109/10409238.2011.556597>.

Azpeitia, E.; Wagner, A. Short residence times of DNA-bound transcription factors can reduce gene expression noise and increase the transmission of information in a gene regulation system. *Frontiers in Molecular Biosciences*, Frontiers Media SA, v. 7, apr 2020.

Bancaud, A.; Lavelle, C.; Huet, S.; Ellenberg, J. A fractal model for nuclear organization: current evidence and biological implications. *Nucleic Acids Research*, Oxford University Press (OUP), v. 40, n. 18, p. 8783–8792, jul. 2012. Available on: <https://doi.org/10.1093/nar/gks586>.

Banks, D. S.; Fradin, C. Anomalous diffusion of proteins due to molecular crowding. *Biophysical Journal*, Elsevier BV, v. 89, n. 5, p. 2960–2971, nov. 2005. Available on: <https://doi.org/10.1529/biophysj.104.051078>.

Barboric, M.; Nissen, R. M.; Kanazawa, S.; Jabrane-Ferrat, N.; Peterlin, B. NF-κb binds p-TEFb to stimulate transcriptional elongation by RNA polymerase II. *Molecular Cell*, Elsevier BV, v. 8, n. 2, p. 327–337, ago. 2001. Available on: <https://doi.org/10.1016/s1097-2765(01)00314-8>.

Bauer, M.; Metzler, R. Generalized facilitated diffusion model for DNA-binding proteins with search and recognition states. *Biophysical Journal*, Elsevier BV, v. 102, n. 10, p. 2321–2330, may 2012.

Bauer, M.; Metzler, R. In vivo facilitated diffusion model. *PLoS ONE*, Public Library of Science (PLoS), v. 8, n. 1, p. e53956, jan. 2013. Available on: <https://doi.org/10.1371/journal.pone.0053956>.

Bauer, M.; Rasmussen, E. S.; Lomholt, M. A.; Metzler, R. Real sequence effects on the search dynamics of transcription factors on DNA. *Scientific Reports*, Springer Science and Business Media LLC, v. 5, n. 1, jul 2015.

Bayesian Core: A Practical Approach to Computational Bayesian Statistics. : Springer New York, 2007.

Becskei, A.; Mattaj, I. W. Quantitative models of nuclear transport. *Current Opinion in Cell Biology*, Elsevier BV, v. 17, n. 1, p. 27–34, fev. 2005. Available on: <https://doi.org/10.1016/j.ceb.2004.12.010>.

Behjati, S.; Tarpey, P. S. What is next generation sequencing? *Archives of disease in childhood - Education & practice edition*, BMJ, v. 98, n. 6, p. 236–238, ago. 2013. Available on: <https://doi.org/10.1136/archdischild-2013-304340>.

Belton, J.-M. et al. Hi–c: A comprehensive technique to capture the conformation of genomes. *Methods*, Elsevier BV, v. 58, n. 3, p. 268–276, nov. 2012. Available on: <https://doi.org/10.1016/j.ymeth.2012.05.001>.

Bénichou, O.; Chevalier, C.; Meyer, B.; Voituriez, R. Facilitated diffusion of proteins on chromatin. *Physical Review Letters*, American Physical Society (APS), v. 106, n. 3, jan. 2011. Available on: <https://doi.org/10.1103/physrevlett.106.038102>.

Bergen, V.; Lange, M.; Peidli, S.; Wolf, F. A.; Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. Cold Spring Harbor Laboratory, oct 2019.

Berkum, N. L. van et al. Hi-c: A method to study the three-dimensional architecture of genomes. *Journal of Visualized Experiments*, MyJove Corporation, n. 39, maio 2010. Available on: <https://doi.org/10.3791/1869>.

Björk, P.; Wieslander, L. Mechanisms of mRNA export. *Seminars in Cell & Developmental Biology*, Elsevier BV, v. 32, p. 47–54, ago. 2014. Available on: <https://doi.org/10.1016/j.semcdb.2014.04.027>.

Bompadre, O.; Andrey, G. Chromatin topology in development and disease. *Current Opinion in Genetics & Development*, Elsevier BV, v. 55, p. 32–38, apr 2019.

Breda, J.; Zavolan, M.; Nimwegen, E. van. Bayesian inference of the gene expression states of single cells from scRNA-seq data. Cold Spring Harbor Laboratory, dez. 2019. Available on: <https://doi.org/10.1101/2019.12.28.889956>.

Brivanlou, A. H.; Darnell, J. E. Signal Transduction and the Control of Gene Expression. *Science*, v. 295, n. 5556, p. 813–818, fev. 2002.

Brodsky, S. et al. Intrinsically disordered regions direct transcription factor in vivo binding specificity. *Molecular Cell*, Elsevier BV, v. 79, n. 3, p. 459–471.e4, aug 2020.

Brouwer, I.; Lenstra, T. L. Visualizing transcription: key to understanding gene expression dynamics. *Current Opinion in Chemical Biology*, Elsevier BV, v. 51, p. 122–129, aug 2019.

Bruneau, B. G. Epigenetic regulation of the cardiovascular system. *Circulation Research*, Ovid Technologies (Wolters Kluwer Health), v. 107, n. 3, p. 324–326, aug 2010.

Butler, J. E.; Kadonaga, J. T. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & Development*, Cold Spring Harbor Laboratory, v. 16, n. 20, p. 2583–2592, oct 2002.

Carmody, S. R.; Wente, S. R. mRNA nuclear export at a glance. *Journal of Cell Science*, The Company of Biologists, v. 122, n. 12, p. 1933–1937, jun 2009.

Chen, L.-F.; Greene, W. C. Shaping the nuclear action of NF-κb. *Nature Reviews Molecular Cell Biology*, Springer Science and Business Media LLC, v. 5, n. 5, p. 392–401, may 2004.

Chen, T.; He, H. L.; Church, G. M. MODELING GENE EXPRESSION WITH DIFFERENTIAL EQUATIONS. In: *Biocomputing '99*. : WORLD SCIENTIFIC, 1998.

Cherstvy, A. G.; Teif, V. B. Structure-driven homology pairing of chromatin fibers: the role of electrostatics and protein-induced bridging. *Journal of Biological Physics*, Springer Science and Business Media LLC, v. 39, n. 3, p. 363–385, jan 2013.

Chiu, C. et al. A two-scale mathematical model for DNA transcription. *Mathematical Biosciences*, Elsevier BV, v. 236, n. 2, p. 132–140, abr. 2012. Available on: <https://doi.org/10.1016/j.mbs.2011.12.006>.

Chou, I.-C. et al. Interleukin (IL)-1β, IL-1 receptor antagonist, IL-6, IL-8, IL-10, and tumor necrosis factor α gene polymorphisms in patients with febrile seizures. *Journal of Clinical Laboratory Analysis*, Wiley, v. 24, n. 3, p. 154–159, 2010.

Cloppet, F.; Boucher, A. Segmentation of complex nucleus configurations in biological images. *Pattern Recognition Letters*, Elsevier BV, v. 31, n. 8, p. 755–761, jun 2010.

Cobb, M. 60 years ago, francis crick changed the logic of biology. *PLOS Biology*, Public Library of Science (PLoS), v. 15, n. 9, p. e2003243, sep 2017.

Core, L.; Adelman, K. Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes & Development*, Cold Spring Harbor Laboratory, v. 33, n. 15-16, p. 960–982, maio 2019. Available on: <https://doi.org/10.1101/gad.325142.119>.

Coulon, A.; Chow, C. C.; Singer, R. H.; Larson, D. R. Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nature Reviews Genetics*, Springer Science and Business Media LLC, v. 14, n. 8, p. 572–584, jul 2013.

Courtois, G. et al. A hypermorphic iκbα mutation is associated with autosomal dominant anhidrotic ectodermal dysplasia and t cell immunodeficiency. *Journal of Clinical Investigation*, American Society for Clinical Investigation, v. 112, n. 7, p. 1108–1115, oct 2003.

Cremer, T. et al. The interchromatin compartment participates in the structural and functional organization of the cell nucleus. *BioEssays*, Wiley, v. 42, n. 2, p. 1900132, jan. 2020. Available on: <https://doi.org/10.1002/bies.201900132>.

Crick, F. H. On protein synthesis. *Symposia of the Society for Experimental Biology*, v. 12, p. 138–163, 1958. ISSN 0081-1386.

Dange, T.; Joseph, A.; Grünwald, D. A perspective of the dynamic structure of the nucleus explored at the single-molecule level. *Chromosome Research*, Springer Science and Business Media LLC, v. 19, n. 1, p. 117–129, set. 2010. Available on: <https://doi.org/10.1007/s10577-010-9156-5>.

Dergai, O.; Hernandez, N. How to recruit the correct RNA polymerase? lessons from snRNA genes. *Trends in Genetics*, Elsevier BV, v. 35, n. 6, p. 457–469, jun 2019.

Devos, D. P.; Gräf, R.; Field, M. C. Evolution of the nucleus. *Current Opinion in Cell Biology*, Elsevier BV, v. 28, p. 8–15, jun. 2014. Available on: <https://doi.org/10.1016/j.ceb.2014.01.004>.

Díaz, N. et al. Chromatin conformation analysis of primary patient tissue using a low input hi-c method. *Nature Communications*, Springer Science and Business Media LLC, v. 9, n. 1, nov. 2018. Available on: <https://doi.org/10.1038/s41467-018-06961-0>.

Dillencourt, M. B.; Samet, H.; Tamminen, M. A general approach to connected-component labeling for arbitrary image representations. *Journal of the ACM*, Association for Computing Machinery (ACM), v. 39, n. 2, p. 253–280, apr 1992.

Dion, V.; Gasser, S. Chromatin movement in the maintenance of genome stability. *Cell*, v. 152, n. 6, p. 1355–1364, 2013. ISSN 0092-8674. Available on: <https://www.sciencedirect.com/science/article/pii/S0092867413001992>.

Edelstein-Keshet, L. *Mathematical models in biology*: Originally published: 1st ed. new york : Random house, 1988, in series: The random house/birkhäuser mathematics series. with new pref. and slight corrections. Philadelphia, Pa.: Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2005. (Classics in applied mathematics, 46). System requirements: Adobe Acrobat Reader. Available on: <http://epubs.siam.org/ebooks/siam/classics_in_applied_mathematics/cl46>.

Elowitz, M. B. Stochastic gene expression in a single cell. *Science*, American Association for the Advancement of Science (AAAS), v. 297, n. 5584, p. 1183–1186, aug 2002.

Engeland, K. Cell cycle arrest through indirect transcriptional repression by p53: I have a DREAM. *Cell Death & Differentiation*, Springer Science and Business Media LLC, v. 25, n. 1, p. 114–132, nov 2017.

Fedorov, A. Comparison of intron-containing and intron-lacking human genes elucidates putative exonic splicing enhancers. *Nucleic Acids Research*, Oxford University Press (OUP), v. 29, n. 7, p. 1464–1469, apr 2001.

Femino, A. M.; Fay, F. S.; Fogarty, K.; Singer, R. H. Visualization of single RNA transcripts in situ. *Science*, American Association for the Advancement of Science (AAAS), v. 280, n. 5363, p. 585–590, apr 1998.

Findley, A. S. et al. Functional dynamic genetic effects on gene regulation are specific to particular cell types and environmental conditions. *eLife*, eLife Sciences Publications, Ltd, v. 10, maio 2021. Available on: <https://doi.org/10.7554/elife.67077>.

Fish, K. N. Total internal reflection fluorescence (TIRF) microscopy. *Current Protocols in Cytometry*, Wiley, v. 50, n. 1, oct 2009.

Fudenberg, G.; Abdennur, N.; Imakaev, M.; Goloborodko, A.; Mirny, L. Emerging evidence of chromosome folding by loop extrusion. *Cold Spring Harbor Symposia on Quantitative Biology*, v. 82, p. 034710, 05 2018.

Galitsyna, A. A.; Gelfand, M. S. Single-cell hi-c data analysis: safety in numbers. *Briefings in Bioinformatics*, Oxford University Press (OUP), v. 22, n. 6, ago. 2021. Available on: <https://doi.org/10.1093/bib/bbab316>.

Gao, P.; Honkela, A.; Rattray, M.; Lawrence, N. D. Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, Oxford University Press (OUP), v. 24, n. 16, p. i70–i75, ago. 2008. Available on: <https://doi.org/10.1093/bioinformatics/btn278>.

Garcia, D. A. et al. Power-law behaviour of transcription factor dynamics at the single-molecule level implies a continuum affinity model. Cold Spring Harbor Laboratory, may 2019.

Gibcus, J. H. et al. A pathway for mitotic chromosome formation. *Science*, American Association for the Advancement of Science, v. 359, n. 6376, 2018. ISSN 0036-8075. Available on: <https://science.sciencemag.org/content/359/6376/eaao6135>.

Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, Elsevier BV, v. 22, n. 4, p. 403–434, dec 1976.

Glauche, I.; Herberg, M.; Roeder, I. Nanog variability and pluripotency regulation of embryonic stem cells - insights from a mathematical model analysis. *PLoS ONE*, Public Library of Science (PLoS), v. 5, n. 6, p. e11238, jun. 2010. Available on: <https://doi.org/10.1371/journal.pone.0011238>.

Glisovic, T.; Bachorik, J. L.; Yong, J.; Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, Wiley, v. 582, n. 14, p. 1977–1986, mar. 2008. Available on: <https://doi.org/10.1016/j.febslet.2008.03.004>.

Gorin, G.; Fang, M.; Chari, T.; Pachter, L. RNA velocity unraveled. Cold Spring Harbor Laboratory, feb 2022.

Gorski, S. A.; Dundr, M.; Misteli, T. The road much traveled: trafficking in the cell nucleus. *Current Opinion in Cell Biology*, v. 18, n. 3, p. 284–290, 2006. ISSN 0955-0674. Nucleus and gene expression. Available on: <https://www.sciencedirect.com/science/article/pii/S0955067406000457>.

Grünwald, D.; Singer, R. H.; Rout, M. Nuclear export dynamics of RNA–protein complexes. *Nature*, Springer Science and Business Media LLC, v. 475, n. 7356, p. 333–341, jul 2011.

Guertin, M. J.; Lis, J. T. Mechanisms by which transcription factors gain access to target sequence elements in chromatin. *Current Opinion in Genetics & Development*, Elsevier BV, v. 23, n. 2, p. 116–123, apr 2013.

Guigas, G.; Weiss, M. Sampling the cell with anomalous diffusion—the discovery of slowness. *Biophysical Journal*, Elsevier BV, v. 94, n. 1, p. 90–94, jan. 2008. Available on: <https://doi.org/10.1529/biophysj.107.117044>.

Guiro, J.; Murphy, S. Regulation of expression of human RNA polymerase II-transcribed snRNA genes. *Open Biology*, The Royal Society, v. 7, n. 6, p. 170073, jun. 2017. Available on: <https://doi.org/10.1098/rsob.170073>.

Gupta, G. K.; Wallace, C. S. Some new multistep methods for solving ordinary differential equations. *Mathematics of Computation*, American Mathematical Society, v. 29, n. 130, p. 489–500, 1975. ISSN 00255718, 10886842. Available on: <http://www.jstor.org/stable/2005567>.

Haddad, R.; Akansu, A. A class of fast gaussian binomial filters for speech and image processing. *IEEE Transactions on Signal Processing*, Institute of Electrical and Electronics Engineers (IEEE), v. 39, n. 3, p. 723–727, mar 1991.

Hager, G. L.; McNally, J. G.; Misteli, T. Transcription dynamics. *Molecular Cell*, Elsevier BV, v. 35, n. 6, p. 741–753, sep 2009.

Hancock, R. Structures and functions in the crowded nucleus: new biophysical insights. *Frontiers in Physics*, Frontiers Media SA, v. 2, set. 2014. Available on: <https://doi.org/10.3389/fphy.2014.00053>.

He, Y.; Fang, J.; Taatjes, D. J.; Nogales, E. Structural visualization of key steps in human transcription initiation. *Nature*, Springer Science and Business Media LLC, v. 495, n. 7442, p. 481–486, fev. 2013. Available on: <https://doi.org/10.1038/nature11991>.

Heiberger, R. M.; Holland, B. *Statistical Analysis and Data Display An Intermediate Course with Examples in R*: An intermediate course with examples in r. : Springer London, Limited, 2015. ISBN 9781493921225.

Hettich, J.; Gebhardt, J. Transcription factor target site search and gene regulation in a background of unspecific binding sites. *Journal of Theoretical Biology*, Elsevier BV, v. 454, p. 91–101, oct 2018.

Hnisz, D.; Shrinivas, K.; Young, R. A.; Chakraborty, A. K.; Sharp, P. A. A phase separation model for transcriptional control. *Cell*, Elsevier BV, v. 169, n. 1, p. 13–23, mar 2017.

Hoboth, P.; Šebesta, O.; Hozák, P. How single-molecule localization microscopy expanded our mechanistic understanding of RNA polymerase II transcription. *International Journal of Molecular Sciences*, MDPI AG, v. 22, n. 13, p. 6694, jun. 2021. Available on: <https://doi.org/10.3390/ijms22136694>.

Hoffman, E. A.; Frey, B. L.; Smith, L. M.; Auble, D. T. Formaldehyde crosslinking: A tool for the study of chromatin complexes. *Journal of Biological Chemistry*, Elsevier BV, v. 290, n. 44, p. 26404–26411, out. 2015. Available on: <https://doi.org/10.1074/jbc.r115.651679>.

Hunter, D. J. Gene–environment interactions in human diseases. *Nature Reviews Genetics*, Springer Science and Business Media LLC, v. 6, n. 4, p. 287–298, abr. 2005. Available on: <https://doi.org/10.1038/nrg1578>.

Imbert, A. et al. FISH-quant v2: a scalable and modular analysis tool for smFISH image analysis. Cold Spring Harbor Laboratory, jul. 2021. Available on: <https://doi.org/10.1101/2021.07.20.453024>.

Ishihama, A. Prokaryotic genome regulation: A revolutionary paradigm. *Proceedings of the Japan Academy, Series B*, Japan Academy, v. 88, n. 9, p. 485–508, 2012. Available on: <https://doi.org/10.2183/pjab.88.485>.

Izeddin, I. et al. Single-molecule tracking in live cells reveals distinct target-search strategies of transcription factors in the nucleus. *eLife*, eLife Sciences Publications, Ltd, v. 3, jun 2014.

Jo, B.-S.; Choi, S. S. Introns: The functional benefits of introns in genomes. *Genomics & Informatics*, Korea Genome Organization, v. 13, n. 4, p. 112, 2015. Available on: <https://doi.org/10.5808/gi.2015.13.4.112>.

Johnnidis, J. B. et al. Chromosomal clustering of genes controlled by the aire transcription factor. *Proceedings of the National Academy of Sciences*, Proceedings of the National Academy of Sciences, v. 102, n. 20, p. 7233–7238, maio 2005. Available on: <https://doi.org/10.1073/pnas.0502670102>.

Johnstone, C. P.; Wang, N. B.; Sevier, S. A.; Galloway, K. E. Understanding and engineering chromatin as a dynamical system across length and timescales. *Cell Systems*, Elsevier BV, v. 11, n. 5, p. 424–448, nov 2020.

Jonge, W. J. de; Patel, H. P.; Meeussen, J. V.; Lenstra, T. L. Following the tracks: How transcription factor binding dynamics control transcription. *Biophysical Journal*, Elsevier BV, v. 121, n. 9, p. 1583–1592, maio 2022. Available on: <https://doi.org/10.1016/j.bpj.2022.03.026>.

Kadauke, S.; Blobel, G. A. Chromatin loops in gene regulation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, Elsevier BV, v. 1789, n. 1, p. 17–25, jan. 2009. Available on: <https://doi.org/10.1016/j.bbagrm.2008.07.002>.

Kærn, M.; Elston, T. C.; Blake, W. J.; Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, Springer Science and Business Media LLC, v. 6, n. 6, p. 451–464, may 2005.

Kalverda, B.; Röling, M. D.; Fornerod, M. Chromatin organization in relation to the nuclear periphery. *FEBS Letters*, v. 582, n. 14, p. 2017–2022, 2008. ISSN 0014-5793. Nuclear Dynamics and Cytoskeleton Signaling. Available on: <https://www.sciencedirect.com/science/article/pii/S0014579308003335>.

Kampmann, M. Facilitated diffusion in chromatin lattices: mechanistic diversity and regulatory potential. *Molecular Microbiology*, Wiley, v. 57, n. 4, p. 889–899, jun. 2005. Available on: <https://doi.org/10.1111/j.1365-2958.2005.04707.x>.

Klumpp, S. A superresolution census of RNA polymerase. *Biophysical Journal*, Elsevier BV, v. 105, n. 12, p. 2613–2614, dec 2013.

Kornberg, R. D. The molecular basis of eukaryotic transcription. *Proceedings of the National Academy of Sciences*, Proceedings of the National Academy of Sciences, v. 104, n. 32, p. 12955–12961, ago. 2007. Available on: <https://doi.org/10.1073/pnas.0704138104>.

Koşar, Z.; Erbaş, A. Can the concentration of a transcription factor affect gene expression? *Frontiers in Soft Matter*, Frontiers Media SA, v. 2, jul. 2022. Available on: <https://doi.org/10.3389/frsfm.2022.914494>.

Kowal, M.; Żejmo, M.; Skobel, M.; Korbicz, J.; Monczak, R. Cell nuclei segmentation in cytological images using convolutional neural network and seeded watershed algorithm. *Journal of Digital Imaging*, Springer Science and Business Media LLC, v. 33, n. 1, p. 231–242, jun 2019.

Krensky, A. M.; Ahn, Y.-T. Mechanisms of disease: regulation of RANTES (CCL5) in renal disease. *Nature Clinical Practice Nephrology*, Springer Science and Business Media LLC, v. 3, n. 3, p. 164–170, mar 2007.

Kuchler, O. et al. Single-molecule tracking (SMT) and localization of SRF and MRTF transcription factors during neuronal stimulation and differentiation. *Open Biology*, The Royal Society, v. 12, n. 5, maio 2022. Available on: <https://doi.org/10.1098/rsob.210383>.

Kuhn, T.; Hettich, J.; Davtyan, R.; Gebhardt, J. C. M. Single molecule tracking and analysis framework including theory-predicted parameter settings. *Scientific Reports*, Springer Science and Business Media LLC, v. 11, n. 1, may 2021.

Kyrchanova, O.; Georgiev, P. Mechanisms of enhancer-promoter interactions in higher eukaryotes. *International Journal of Molecular Sciences*, MDPI AG, v. 22, n. 2, p. 671, jan. 2021. Available on: <https://doi.org/10.3390/ijms22020671>.

Lafontaine, D. L.; Yang, L.; Dekker, J.; Gibcus, J. H. Hi-c 3.0: Improved protocol for genome-wide chromosome conformation capture. *Current Protocols*, Wiley, v. 1, n. 7, jul. 2021. Available on: <https://doi.org/10.1002/cpz1.198>.

Lee, E. S. et al. TPR is required for the efficient nuclear export of mRNAs and lncRNAs from short and intron-poor genes. *Nucleic Acids Research*, Oxford University Press (OUP), v. 48, n. 20, p. 11645–11663, oct 2020.

Lee, Y. et al. MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal*, Wiley, v. 23, n. 20, p. 4051–4060, set. 2004. Available on: <https://doi.org/10.1038/sj.emboj.7600385>.

Leijnse, N.; Jeon, J. H.; Loft, S.; Metzler, R.; Oddershede, L. B. Diffusion inside living human cells. *The European Physical Journal Special Topics*, Springer Science and Business Media LLC, v. 204, n. 1, p. 75–84, abr. 2012. Available on: <https://doi.org/10.1140/epjst/e2012-01553-y>.

Lettice, L. A. A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, Oxford University Press (OUP), v. 12, n. 14, p. 1725–1735, jul. 2003. Available on: <https://doi.org/10.1093/hmg/ddg180>.

Li, B.; Carey, M.; Workman, J. L. The role of chromatin during transcription. *Cell*, Elsevier BV, v. 128, n. 4, p. 707–719, feb 2007.

Li, G.-W.; Berg, O. G.; Elf, J. Effects of macromolecular crowding and DNA looping on gene regulation kinetics. *Nature Physics*, Springer Science and Business Media LLC, v. 5, n. 4, p. 294–297, mar. 2009. Available on: <https://doi.org/10.1038/nphys1222>.

Li, Y.; Varala, K.; Coruzzi, G. M. From milliseconds to lifetimes: tracking the dynamic behavior of transcription factors in gene networks. *Trends in Genetics*, Elsevier BV, v. 31, n. 9, p. 509–515, set. 2015. Available on: <https://doi.org/10.1016/j.tig.2015.05.005>.

Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, American Association for the Advancement of Science (AAAS), v. 326, n. 5950, p. 289–293, oct 2009.

Liu, C.; Liu, Y.-L.; Perillo, E. P.; Dunn, A. K.; Yeh, H.-C. Single-molecule tracking and its application in biomolecular binding detection. *IEEE Journal of Selected Topics in Quantum Electronics*, Institute of Electrical and Electronics Engineers (IEEE), v. 22, n. 4, p. 64–76, jul 2016.

Liu, J. et al. Real-time single-cell characterization of the eukaryotic transcription cycle reveals correlations between RNA initiation, elongation, and cleavage. *PLOS Computational Biology*, Public Library of Science (PLoS), v. 17, n. 5, p. e1008999, maio 2021. Available on: <https://doi.org/10.1371/journal.pcbi.1008999>.

Liu, T.; Zhang, L.; Joo, D.; Sun, S.-C. NF-κb signaling in inflammation. *Signal Transduction and Targeted Therapy*, Springer Science and Business Media LLC, v. 2, n. 1, jul 2017.

Liu, X.; Kraus, W. L.; Bai, X. Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways. *Trends in Biochemical Sciences*, Elsevier BV, v. 40, n. 9, p. 516–525, set. 2015. Available on: <https://doi.org/10.1016/j.tibs.2015.07.003>.

Liu, Z. et al. 3d imaging of sox2 enhancer clusters in embryonic stem cells. *eLife*, eLife Sciences Publications, Ltd, v. 3, dec 2014.

Livingstone, M.; Atas, E.; Meller, A.; Sonenberg, N. Mechanisms governing the control of mrna translation. *Physical Biology*, v. 7, n. 2, p. 021001, may 2010. Available on: <https://dx.doi.org/10.1088/1478-3975/7/2/021001>.

Luo, H. et al. Cell identity bookmarking through heterogeneous chromatin landscape maintenance during the cell cycle. *Human Molecular Genetics*, Oxford University Press (OUP), v. 26, n. 21, p. 4231–4243, aug 2017.

Luse, D. S. The RNA polymerase II preinitiation complex. *Transcription*, Informa UK Limited, v. 5, n. 1, p. e27050, nov 2013.

Ma, J. Transcriptional activators and activation mechanisms. *Protein & Cell*, Springer Science and Business Media LLC, v. 2, n. 11, p. 879–888, nov. 2011. Available on: <https://doi.org/10.1007/s13238-011-1101-7>.

Maeshima, K.; Ide, S.; Babokhov, M. Dynamic chromatin organization without the 30-nm fiber. *Current Opinion in Cell Biology*, Elsevier BV, v. 58, p. 95–104, jun. 2019. Available on: <https://doi.org/10.1016/j.ceb.2019.02.003>.

Maeshima, K. et al. The physical size of transcription factors is key to transcriptional regulation in chromatin domains. *Journal of Physics: Condensed Matter*, IOP Publishing, v. 27, n. 6, p. 064116, jan 2015.

Magistris, P. D. The great escape: mRNA export through the nuclear pore complex. *International Journal of Molecular Sciences*, MDPI AG, v. 22, n. 21, p. 11767, oct 2021.

Malpica, N. et al. Applying watershed algorithms to the segmentation of clustered nuclei. *Cytometry*, Wiley, v. 28, n. 4, p. 289–297, dec 1998.

Manno, G. L. et al. RNA velocity of single cells. *Nature*, Springer Science and Business Media LLC, v. 560, n. 7719, p. 494–498, aug 2018.

Manzo, C.; Garcia-Parajo, M. F. A review of progress in single particle tracking: from methods to biophysical insights. *Reports on Progress in Physics*, IOP Publishing, v. 78, n. 12, p. 124601, out. 2015. Available on: <https://doi.org/10.1088/0034-4885/78/12/124601>.

Mao, X. et al. Regulation of translation initiation factor gene expression during human t cell activation. *Journal of Biological Chemistry*, Elsevier BV, v. 267, n. 28, p. 20444–20450, oct 1992.

Marini, M. et al. The structure of DNA by direct imaging. *Science Advances*, American Association for the Advancement of Science (AAAS), v. 1, n. 7, ago. 2015. Available on: <https://doi.org/10.1126/sciadv.1500734>.

Maston, G. A.; Evans, S. K.; Green, M. R. Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*, Annual Reviews, v. 7, n. 1, p. 29–59, set. 2006. Available on: <https://doi.org/10.1146/annurev.genom.7.080505.115623>.

Mazzocca, M.; Fillot, T.; Loffreda, A.; Gnani, D.; Mazza, D. The needle and the haystack: single molecule tracking to probe the transcription factor search in eukaryotes. *Biochemical Society Transactions*, Portland Press Ltd., v. 49, n. 3, p. 1121–1132, maio 2021. Available on: <https://doi.org/10.1042/bst20200709>.

McBride, K. M. Regulated nuclear import of the STAT1 transcription factor by direct binding of importin-alpha. *The EMBO Journal*, Wiley, v. 21, n. 7, p. 1754–1763, apr 2002.

McKinlay, A.; Podicheti, R.; Wendte, J. M.; Cocklin, R.; Rusch, D. B. RNA polymerases IV and v influence the 3′ boundaries of polymerase II transcription units in *Arabidopsis*. *RNA Biology*, Informa UK Limited, v. 15, n. 2, p. 269–279, dez. 2017. Available on: <https://doi.org/10.1080/15476286.2017.1409930>.

Meeussen, J. V. W. et al. Transcription factor clusters enable target search but do not contribute to target gene activation. Cold Spring Harbor Laboratory, dez. 2022. Available on: <https://doi.org/10.1101/2022.12.13.520200>.

Meier-Soelch, J. et al. Monitoring the levels of cellular NF-*kappa*b activation states. *Cancers*, MDPI AG, v. 13, n. 21, p. 5351, oct 2021.

Meldi, L.; Brickner, J. H. Compartmentalization of the nucleus. *Trends in Cell Biology*, Elsevier BV, v. 21, n. 12, p. 701–708, dez. 2011. Available on: <https://doi.org/10.1016/j.tcb.2011.08.001>.

Meyer, F. Topographic distance and watershed lines. *Signal Processing*, Elsevier BV, v. 38, n. 1, p. 113–125, jul 1994.

Meyer, T.; Vinkemeier, U. STAT nuclear translocation: potential for pharmacological intervention. *Expert Opinion on Therapeutic Targets*, Informa Healthcare, v. 11, n. 10, p. 1355–1365, oct 2007.

Miele, A.; Dekker, J. Mapping cis- and trans- chromatin interaction networks using chromosome conformation capture (3c). In: *The Nucleus*. Humana Press, 2008. p. 105–121. Available on: <https://doi.org/10.1007/978-1-60327-461-6_7>.

Mirny, L. et al. How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *Journal of Physics A: Mathematical and Theoretical*, IOP Publishing, v. 42, n. 43, p. 434013, oct 2009.

Mirny, L. A. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research*, Springer Science and Business Media LLC, v. 19, n. 1, p. 37–51, jan. 2011. Available on: <https://doi.org/10.1007/s10577-010-9177-0>.

Misteli, T. Physiological importance of RNA and protein mobility in the cell nucleus. *Histochemistry and Cell Biology*, Springer Science and Business Media LLC, v. 129, n. 1, p. 5–11, nov. 2007. Available on: <https://doi.org/10.1007/s00418-007-0355-x>.

Mitarai, N.; Semsey, S.; Sneppen, K. Dynamic competition between transcription initiation and repression: Role of nonequilibrium steps in cell-to-cell heterogeneity. *Physical Review E*, American Physical Society (APS), v. 92, n. 2, aug 2015.

Mizuno, K. et al. Novel multicellular prokaryote discovered next to an underground stream. *eLife*, eLife Sciences Publications, Ltd, v. 11, out. 2022. Available on: <https://doi.org/10.7554/elife.71920>.

Moerner, W. E. W. E. Single-molecule spectroscopy, imaging, and photocontrol: Foundations for super-resolution microscopy (nobel lecture). *Angewandte Chemie International Edition*, Wiley, v. 54, n. 28, p. 8067–8093, jun 2015.

Mor, A. et al. Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. *Nature Cell Biology*, Springer Science and Business Media LLC, v. 12, n. 6, p. 543–552, maio 2010. Available on: <https://doi.org/10.1038/ncb2056>.

Morrison, O.; Thakur, J. Molecular complexes at euchromatin, heterochromatin and centromeric chromatin. *International Journal of Molecular Sciences*, MDPI AG, v. 22, n. 13, p. 6922, jun. 2021. Available on: <https://doi.org/10.3390/ijms22136922>.

Moss, T.; Stefanovsky, V. Y. At the center of eukaryotic life. *Cell*, Elsevier BV, v. 109, n. 5, p. 545–548, maio 2002. Available on: <https://doi.org/10.1016/s0092-8674(02)00761-4>.

Murray, J. D. *Mathematical Biology*. Springer New York, 2007. Available on: <http://www.ebook.de/de/product/25193370/james_d_murray_mathematical_biology.html>.

Nagamine, N.; Kawada, Y.; Sakakibara, Y. Identifying cooperative transcriptional regulations using protein-protein interactions. *Nucleic Acids Research*, Oxford University Press (OUP), v. 33, n. 15, p. 4828–4837, aug 2005.

Nikolov, D.; Burley, S. RNA polymerase II transcription initiation: A structural view. *Proceedings of the National Academy of Sciences*, Proceedings of the National Academy of Sciences, v. 94, n. 1, p. 15–22, jan 1997.

Niu, L. et al. Amplification-free library preparation with SAFE hi-c uses ligation products for deep sequencing to improve traditional hi-c analysis. *Communications Biology*, Springer Science and Business Media LLC, v. 2, n. 1, jul. 2019. Available on: <https://doi.org/10.1038/s42003-019-0519-y>.

Nouaille, S. et al. The stability of an mRNA is influenced by its concentration: a potential physical mechanism to regulate gene expression. *Nucleic Acids Research*, Oxford University Press (OUP), v. 45, n. 20, p. 11711–11724, set. 2017. Available on: <https://doi.org/10.1093/nar/gkx781>.

Noursadeghi, M. et al. Quantitative imaging assay for nf-κb nuclear translocation in primary human macrophages. *Journal of immunological methods*, v. 329, p. 194–200, 02 2008.

Oluwadare, O.; Highsmith, M.; Cheng, J. An overview of methods for reconstructing 3-d chromosome and genome structures from hi-c data. *Biological Procedures Online*, Springer Science and Business Media LLC, v. 21, n. 1, abr. 2019. Available on: <https://doi.org/10.1186/s12575-019-0094-0>.

Pal, K.; Forcato, M.; Ferrari, F. Hi-c analysis: from data generation to integration. *Biophysical Reviews*, Springer Science and Business Media LLC, v. 11, n. 1, p. 67–78, dez. 2018. Available on: <https://doi.org/10.1007/s12551-018-0489-1>.

Palomo, A. B. A. et al. Plant hormones increase efficiency of reprogramming mouse somatic cells to induced pluripotent stem cells and reduce tumorigenicity. *Stem Cells and Development*, Mary Ann Liebert Inc, v. 23, n. 6, p. 586–593, mar 2014.

Pennacchio, L. A.; Bickmore, W.; Dean, A.; Nobrega, M. A.; Bejerano, G. Enhancers: five essential questions. *Nature Reviews Genetics*, Springer Science and Business Media LLC, v. 14, n. 4, p. 288–295, mar. 2013. Available on: <https://doi.org/10.1038/nrg3458>.

Peters, R. Translocation through the nuclear pore complex: Selectivity and speed by reduction-of-dimensionality. *Traffic*, Wiley, v. 6, n. 5, p. 421–427, mar 2005.

Peterson, Q. P. et al. A method for the generation of human stem cell-derived alpha cells. *Nature Communications*, Springer Science and Business Media LLC, v. 11, n. 1, maio 2020. Available on: <https://doi.org/10.1038/s41467-020-16049-3>.

Petrenko, N.; Jin, Y.; Dong, L.; Wong, K. H.; Struhl, K. Requirements for RNA polymerase II preinitiation complex formation in vivo. *eLife*, eLife Sciences Publications, Ltd, v. 8, jan 2019.

Petrenko, N.; Struhl, K. Comparison of transcriptional initiation by RNA polymerase II across eukaryotic species. *eLife*, eLife Sciences Publications, Ltd, v. 10, sep 2021.

Pharris, M. C. et al. An automated workflow for quantifying RNA transcripts in individual cells in large data-sets. *MethodsX*, Elsevier BV, v. 4, p. 279–288, 2017.

Piovesan, A. et al. Human protein-coding genes and gene feature statistics in 2019. *BMC Research Notes*, Springer Science and Business Media LLC, v. 12, n. 1, jun. 2019. Available on: <https://doi.org/10.1186/s13104-019-4343-8>.

Pokholok, D. K.; Hannett, N. M.; Young, R. A. Exchange of RNA polymerase II initiation and elongation factors during gene expression in vivo. *Molecular Cell*, Elsevier BV, v. 9, n. 4, p. 799–809, apr 2002.

Pollak, M.; Siegmund, D. A diffusion process and its applications to detecting a change in the drift of brownian motion. *Biometrika*, Oxford University Press (OUP), v. 72, n. 2, p. 267–280, 1985.

Popp, A. P.; Hettich, J.; Gebhardt, J. C. M. Transcription factor residence time dominates over concentration in transcription activation. Cold Spring Harbor Laboratory, nov 2020.

Pugh, B.; Tjian, R. Mechanism of transcriptional activation by sp1: Evidence for coactivators. *Cell*, Elsevier BV, v. 61, n. 7, p. 1187–1197, jun 1990.

Quintero-Cadena, P.; Lenstra, T. L.; Sternberg, P. W. Rna pol ii length and disorder enable cooperative scaling of transcriptional bursting. *Molecular Cell*, 2020.

Rao, S. S. et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, Elsevier BV, v. 159, n. 7, p. 1665–1680, dec 2014.

Ren, G. et al. CTCF-mediated enhancer-promoter interaction is a critical regulator of cell-to-cell variation of gene expression. *Molecular Cell*, Elsevier BV, v. 67, n. 6, p. 1049–1058.e6, set. 2017. Available on: <https://doi.org/10.1016/j.molcel.2017.08.026>.

Reuveni, E.; Getselter, D.; Oron, O.; Elliott, E. Differential contribution of cis and trans gene transcription regulatory mechanisms in amygdala and prefrontal cortex and modulation by social stress. *Scientific Reports*, Springer Science and Business Media LLC, v. 8, n. 1, abr. 2018. Available on: <https://doi.org/10.1038/s41598-018-24544-3>.

Robert, P. Mathematical models of gene expression. *Probability Surveys*, Institute of Mathematical Statistics, v. 16, n. none, jan. 2019. Available on: <https://doi.org/10.1214/19-ps332>.

Roeder, R. G. The role of general initiation factors in transcription by rna polymerase ii. *Trends in Biochemical Sciences*, v. 21, n. 9, p. 327–335, 1996. ISSN 0968-0004. Available on: <https://www.sciencedirect.com/science/article/pii/S0968000496100505>.

Roeder, R. G. 50+ years of eukaryotic transcription: an expanding universe of factors and mechanisms. *Nature Structural & Molecular Biology*, Springer Science and Business Media LLC, v. 26, n. 9, p. 783–791, aug 2019.

Rosenfeld, A.; Pfaltz, J. L. Sequential operations in digital picture processing. *Journal of the ACM*, Association for Computing Machinery (ACM), v. 13, n. 4, p. 471–494, oct 1966.

Roy, S. W.; Gilbert, W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Reviews Genetics*, Springer Science and Business Media LLC, v. 7, n. 3, p. 211–221, mar 2006.

Rzeszotarska, E. et al. IL-1β, IL-10 and TNF-α polymorphisms may affect systemic lupus erythematosus risk and phenotype. *Clinical and Experimental Rheumatology*, Clinical and Experimental Rheumatology, may 2021.

Saliba, A.-E.; Westermann, A. J.; Gorski, S. A.; Vogel, J. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Research*, Oxford University Press (OUP), v. 42, n. 14, p. 8845–8860, jul 2014.

Sanborn, A. L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, Proceedings of the National Academy of Sciences, v. 112, n. 47, out. 2015. Available on: <https://doi.org/10.1073/pnas.1518552112>.

Scharer, C. D.; Barwick, B. G.; Guo, M.; Bally, A. P. R.; Boss, J. M. Plasma cell differentiation is controlled by multiple cell division-coupled epigenetic programs. *Nature Communications*, Springer Science and Business Media LLC, v. 9, n. 1, abr. 2018. Available on: <https://doi.org/10.1038/s41467-018-04125-8>.

Schmitz, M. L.; Higgins, J. M. G.; Seibert, M. Priming chromatin for segregation: functional roles of mitotic histone modifications. *Cell Cycle*, Informa UK Limited, v. 19, n. 6, p. 625–641, jan 2020.

Schneider, N. et al. Liquid-liquid phase separation of light-inducible transcription factors increases transcription activation in mammalian cells and mice. *Science Advances*, American Association for the Advancement of Science (AAAS), v. 7, n. 1, p. eabd3568, jan 2021.

Schoenfelder, S.; Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, Springer Science and Business Media LLC, v. 20, n. 8, p. 437–455, may 2019.

Selvaraj, P.; Alagarasu, K.; Singh, B.; Afsal, K. CCL5 (RANTES) gene polymorphisms in pulmonary tuberculosis patients of south india. *International Journal of Immunogenetics*, Wiley, v. 38, n. 5, p. 397–402, jun 2011.

Shampine, L. F.; Reichelt, M. W. The MATLAB ODE suite. *SIAM Journal on Scientific Computing*, Society for Industrial & Applied Mathematics (SIAM), v. 18, n. 1, p. 1–22, jan 1997.

Shrinivas, K. et al. Enhancer features that drive formation of transcriptional condensates. 12 2018.

Slutsky, M.; Mirny, L. A. Kinetics of protein-DNA interaction: Facilitated target location in sequence-dependent potential. *Biophysical Journal*, Elsevier BV, v. 87, n. 6, p. 4021–4035, dec 2004.

Sokolik, C. et al. Transcription factor competition allows embryonic stem cells to distinguish authentic signals from noise. *Cell Systems*, Elsevier BV, v. 1, n. 2, p. 117–129, ago. 2015. Available on: <https://doi.org/10.1016/j.cels.2015.08.001>.

Soutoglou, E.; Misteli, T. Mobility and immobility of chromatin in transcription and genome stability. *Current Opinion in Genetics & Development*, Elsevier BV, v. 17, n. 5, p. 435–442, out. 2007. Available on: <https://doi.org/10.1016/j.gde.2007.08.004>.

Soutoglou, E.; Talianidis, I. Coordination of PIC assembly and chromatin remodeling during differentiation-induced gene activation. *Science*, American Association for the Advancement of Science (AAAS), v. 295, n. 5561, p. 1901–1904, mar. 2002. Available on: <https://doi.org/10.1126/science.1068356>.

Stark, R.; Grzelak, M.; Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics*, Springer Science and Business Media LLC, v. 20, n. 11, p. 631–656, jul 2019.

Sterne-Weiler, T. et al. Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Research*, Cold Spring Harbor Laboratory, v. 23, n. 10, p. 1615–1623, jun 2013.

Stewart, J. *Calculus.* : Cengage Learning, 2015. 1392 p. ISBN 9781285740621.

Strambio-De-Castillia, C.; Niepel, M.; Rout, M. P. The nuclear pore complex: bridging nuclear transport and gene regulation. *Nature Reviews Molecular Cell Biology*, Springer Science and Business Media LLC, v. 11, n. 7, p. 490–501, jul 2010.

Strogatz, S. *Nonlinear dynamics and chaos : with applications to physics, biology, chemistry, and engineering.* Boulder, CO: Westview Press, 2015. ISBN 9780429492563.

Struhl, K. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, Elsevier BV, v. 98, n. 1, p. 1–4, jul. 1999. Available on: <https://doi.org/10.1016/s0092-8674(00)80599-1>.

Sun, X.-M. et al. Size-dependent increase in rna polymerase ii initiation rates mediates gene expression scaling with cell size. *Current Biology*, v. 30, n. 7, p. 1217–1230.e7, 2020. ISSN 0960-9822. Available on: <https://www.sciencedirect.com/science/article/pii/S096098222030097X>.

Swift, J.; Coruzzi, G. M. A matter of time — how transient transcription factor interactions create dynamic gene regulatory networks. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, Elsevier BV, v. 1860, n. 1, p. 75–83, jan 2017.

Tang, G.-Q.; Roy, R.; Bandwar, R. P.; Ha, T.; Patel, S. S. Real-time observation of the transition from transcription initiation to elongation of the rna polymerase. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 106, n. 52, p. 22175–22180, 2009. ISSN 0027-8424. Available on: <https://www.pnas.org/content/106/52/22175>.

Tilgner, H. et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, Cold Spring Harbor Laboratory, v. 22, n. 9, p. 1616–1625, sep 2012.

Timney, B. L. et al. Simple rules for passive diffusion through the nuclear pore complex. *Journal of Cell Biology*, Rockefeller University Press, v. 215, n. 1, p. 57–76, out. 2016. Available on: <https://doi.org/10.1083/jcb.201601004>.

Tokunaga, M.; Imamoto, N.; Sakata-Sogawa, K. Highly inclined thin illumination enables clear single-molecule imaging in cells. *Nature Methods*, Springer Science and Business Media LLC, v. 5, n. 2, p. 159–161, jan 2008.

Trask, J. O. J. Nuclear factor kappa b (nf-κb) translocation assay development and validation for high content screening. 2012. Available on: <https://www.semanticscholar.org/paper/7aae48f1b599bcb4e3614169e44e5c7e6a4fd003>.

Tsai, F. T. Structural basis of preinitiation complex assembly on human pol II promoters. *The EMBO Journal*, Wiley, v. 19, n. 1, p. 25–36, jan. 2000. Available on: <https://doi.org/10.1093/emboj/19.1.25>.

Tsanov, N. et al. smiFISH and FISH-quant – a flexible single RNA detection approach with super-resolution capability. *Nucleic Acids Research*, Oxford University Press (OUP), v. 44, n. 22, p. e165–e165, set. 2016. Available on: <https://doi.org/10.1093/nar/gkw784>.

Turner, B. *Chromatin and Gene Regulation*: Mechanisms in epigenetics. : Blackwell Publishing Limited, 2002. 284 p. ISBN 9780865427433.

Vargas, D. Y.; Raj, A.; Marras, S. A. E.; Kramer, F. R.; Tyagi, S. Mechanism of mRNA transport in the nucleus. *Proceedings of the National Academy of Sciences*, Proceedings of the National Academy of Sciences, v. 102, n. 47, p. 17008–17013, nov 2005.

Veloso, A. et al. Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Research*, Cold Spring Harbor Laboratory, v. 24, n. 6, p. 896–905, apr 2014.

Vestergaard, C. L.; Génois, M. Temporal gillespie algorithm: Fast simulation of contagion processes on time-varying networks. *PLOS Computational Biology*, Public Library of Science, v. 11, n. 10, p. 1–28, 10 2015. Available on: <https://doi.org/10.1371/journal.pcbi.1004579>.

Viertl, R. *Probability and Bayesian Statistics*. : Springer US, 1987. 510 p. ISBN 9781461318859.

Volanakis, A. et al. Spliceosome-mediated decay (SMD) regulates expression of nonintronic genes in budding yeast. *Genes & Development*, Cold Spring Harbor Laboratory, v. 27, n. 18, p. 2025–2038, sep 2013.

Vrljic, M.; Nishimura, S. Y.; Moerner, W. E. Single-molecule tracking. In: *Methods in Molecular Biology*. : Humana Press, 2007. p. 193–219.

Wade, J. T.; Struhl, K. The transition from transcriptional initiation to elongation. *Current Opinion in Genetics & Development*, Elsevier BV, v. 18, n. 2, p. 130–136, apr 2008.

Wang, Y.; Stumph, W. E. RNA polymerase II/III transcription specificity determined by TATA box orientation. *Proceedings of the National Academy of Sciences*, Proceedings of the National Academy of Sciences, v. 92, n. 19, p. 8606–8610, set. 1995. Available on: <https://doi.org/10.1073/pnas.92.19.8606>.

Watson, J. D.; Crick, F. H. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, v. 171, p. 737–738, 1953.

Weiss, M. Crowding, diffusion, and biochemical reactions. In: *International Review of Cell and Molecular Biology*. : Elsevier, 2014. p. 383–417.

Whalen, S.; Truty, R. M.; Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics*, Springer Science and Business Media LLC, v. 48, n. 5, p. 488–496, abr. 2016. Available on: <https://doi.org/10.1038/ng.3539>.

Wickramasinghe, V. O.; Laskey, R. A. Control of mammalian gene expression by selective mRNA export. *Nature Reviews Molecular Cell Biology*, Springer Science and Business Media LLC, v. 16, n. 7, p. 431–442, jun 2015.

Willis, I. M. RNA polymerase III. genes, factors and transcriptional specificity. *European Journal of Biochemistry*, Wiley, v. 212, n. 1, p. 1–11, fev. 1993. Available on: <https://doi.org/10.1111/j.1432-1033.1993.tb17626.x>.

Wit, E. de; Laat, W. de. A decade of 3c technologies: insights into nuclear organization. *Genes & Development*, Cold Spring Harbor Laboratory, v. 26, n. 1, p. 11–24, jan. 2012. Available on: <https://doi.org/10.1101/gad.179804.111>.

Wollman, A. J. et al. Transcription factor clusters regulate genes in eukaryotic cells. *eLife*, eLife Sciences Publications, Ltd, v. 6, ago. 2017. Available on: <https://doi.org/10.7554/elife.27451>.

Woringer, M.; Darzacq, X. Protein motion in the nucleus: from anomalous diffusion to weak interactions. *Biochemical Society Transactions*, Portland Press Ltd., v. 46, n. 4, p. 945–956, jul. 2018. Available on: <https://doi.org/10.1042/bst20170310>.

Woringer, M.; Darzacq, X.; Izeddin, I. Geometry of the nucleus: a perspective on gene expression regulation. *Current Opinion in Chemical Biology*, Elsevier BV, v. 20, p. 112–119, jun 2014.

Wunderlich, Z.; Mirny, L. A. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in Genetics*, Elsevier BV, v. 25, n. 10, p. 434–440, oct 2009.

Xia, Y.; Shen, S.; Verma, I. M. NF-κb, an active player in human cancers. *Cancer Immunology Research*, American Association for Cancer Research (AACR), v. 2, n. 9, p. 823–830, sep 2014.

Xiao, J.; Hafner, A.; Boettiger, A. N. How subtle changes in 3d structure can create large changes in transcription. Cold Spring Harbor Laboratory, oct 2020.

Xie, Y.; Ren, Y. Mechanisms of nuclear mRNA export: A structural perspective. *Traffic*, Wiley, v. 20, n. 11, p. 829–840, set. 2019. Available on: <https://doi.org/10.1111/tra.12691>.

Yang, Y. et al. HIF-1 interacts with TRIM28 and DNA-PK to release paused RNA polymerase II and activate target gene transcription in response to hypoxia. *Nature Communications*, Springer Science and Business Media LLC, v. 13, n. 1, jan. 2022. Available on: <https://doi.org/10.1038/s41467-021-27944-8>.

Zabet, N. R.; Adryan, B. A comprehensive computational model of facilitated diffusion in prokaryotes. *Bioinformatics*, Oxford University Press (OUP), v. 28, n. 11, p. 1517–1524, abr. 2012. Available on: <https://doi.org/10.1093/bioinformatics/bts178>.

Zabet, N. R.; Adryan, B. The effects of transcription factor competition on gene regulation. *Frontiers in Genetics*, Frontiers Media SA, v. 4, 2013.

Zambrano, S. et al. First responders shape a prompt and sharp nf-κb-mediated transcriptional response to tnf-α. *iScience*, v. 23, n. 9, p. 101529, 2020. ISSN 2589-0042. Available on: <https://www.sciencedirect.com/science/article/pii/S2589004220307215>.

Zhang, H. et al. Chromatin structure dynamics during the mitosis-to-g1 phase transition. *Nature*, Springer Science and Business Media LLC, v. 576, n. 7785, p. 158–162, nov 2019.

Zhang, J.-M.; An, J. Cytokines, inflammation, and pain. *International Anesthesiology Clinics*, Ovid Technologies (Wolters Kluwer Health), v. 45, n. 2, p. 27–37, 2007.

Zhao, Q.; Liu, H.; Yao, C.; Shuai, J.; Sun, X. Effect of dynamic interaction between microRNA and transcription factor on gene expression. *BioMed Research International*, Hindawi Limited, v. 2016, p. 1–10, 2016. Available on: <https://doi.org/10.1155/2016/2676282>.

Zidovska, A. The rich inner life of the cell nucleus: dynamic organization, active flows, and emergent rheology. *Biophysical Reviews*, Springer Science and Business Media LLC, v. 12, n. 5, p. 1093–1106, out. 2020. Available on: <https://doi.org/10.1007/s12551-020-00761-x>.

Zon, J. S. van; Morelli, M. J.; Tănase-Nicola, S.; Wolde, P. R. ten. Diffusion of transcription factors can drastically enhance the noise in gene expression. *Biophysical Journal*, Elsevier BV, v. 91, n. 12, p. 4350–4367, dez. 2006. Available on: <https://doi.org/10.1529/biophysj.106.086157>.

APPENDIX A – Appendix

The values of the residence times (τ) and number of connections (d) from specific regions in Figs. 2.15, 2.19 and 2.24 are available in Table 1.

Parameter	Value	Unit	Name
τ_{162}	0.8454	s	residence time for region 162
d_{162}	9.4694	-	number of connections for region 162
τ_{590}	2.9863	s	residence time for region 590
d_{590}	5.0000	-	number of connections for region 590
τ_{850}	4.0076	s	residence time for region 850
d_{850}	9.7143	-	number of connections for region 850

Table 1 – Values of τ and d for the regions in Figs. 2.15, 2.19 and 2.24.

We present the parameters for the model from Eq. (3.3) in Table 2.

Parameter	Value	Unit	Name
k_{3D}^{T}	240/7	s^{-1}	TF effective diffusing rate
α	4.8452×10^{-2}	-	TF effectiveness binding
k_{3D}^{P}	19.20/7	s^{-1}	RNAP effective diffusing rate
q	0.27	$(\text{molecules} \times s)^{-1}$	RNAP success binding rate
k_I	1/10	s^{-1}	initiation rate
k_ε	4/500	s^{-1}	elongation rate

Table 2 – Parameters for our model from Eq. (3.3).

The parameters for the flux function in Eq. (3.11) are available in Table 3

Parameter	Value	Unit	Name
k_{im}	1.5797×10^{-3}	s^{-1}	TF import rate
T_{total}	269.9929	molecules	TF total concentration
μ	5.3394×10^{-3}	s^{-1}	exporter production rate
δ	2.3075×10^{-5}	s^{-1}	exporter degradation rate

Table 3 – Parameters for the flux function in Eq. (3.11).

Parameters for the mRNA exportation Eq. (4.1) are available in Table 4.

Parameter	Value	Unit	Name
k_{3D}^{r}	19.20/14	s^{-1}	mRNA effective diffusive rate
γ	1/480	s^{-1}	mRNA degradation rate

Table 4 – Parameters for the mRNA exportation function in Eq. (4.1).

The random Parameters for the RNA Velocity model (4.2) are available in Table 5.

Parameter	Value	Unit	Name
α	0.9572	s^{-1}	mRNA transcription rate
β	1/3	s^{-1}	mRNA splicing rate
k	0.1429	s^{-1}	mRNA export rate
γ	1/15	s^{-1}	mRNA degradation rate

Table 5 – Parameters for the RNA Velocity in Eq. (4.2).

Karen AMARAL DE OLIVEIRA

Modelling Molecular Diffusion in the Nucleus and its role in gene regulation

Résumé

La diffusion des facteurs de transcription (FTs) dans le noyau joue un rôle crucial dans la régulation transcriptionnelle. La recherche par les TF d'une séquence d'ADN spécifique est l'un des principaux facteurs de l'expression des gènes. Ainsi, les interactions entre deux FTs dues à de faibles interactions protéine-protéine (IPPs) forment des clusters de FTs, influençant leur occupation à un site cible particulier. Pour comprendre comment la structure 3D de la chromatine affecte l'agglutination des FTs, nous avons proposé un modèle pour traduire la présence des IPPs dans le noyau et vérifié comment l'agglutination affecte l'allocation des FTs, en considérant la diffusion 3D/1D comme notre mécanisme de recherche. Ensuite, une fois qu'un FT est lié à une région, il recrute l'ARN polymérase (ARNP). En outre, les FTs inductibles restent dans le cytoplasme et translocation dans le noyau par le biais du complexe du pore nucléaire (CPN) après une signalisation appropriée. Afin d'intégrer ces mécanismes, nous avons proposé un autre modèle pour comprendre la dynamique de recherche des TF et de recrutement de l'ARNP. Nous avons obtenu des solutions déterministes et stochastiques vérifiant comment la transcription est renforcée à la périphérie du CPN et confirmée par l'analyse d'imagerie de gènes spécifiques. Enfin, nous avons incorporé le processus d'exportation de l'ARNm pour vérifier les différentes concentrations de transcrits cytoplasmiques, prouvant ainsi que le volume d'ARNm disponible dépend également du CPN. Par conséquent, notre travail montre des liens pertinents entre la structure de la chromatine, l'allocation des ressources transcriptionnelles et la stochasticité de la régulation des gènes.

Recherche par les FTs, Interactions Protéine-Protéine, Recrutement de l'ARNP, exportation d'ARNm, modèle mathématique, structure de la chromatine

Résumé en anglais

The diffusion of transcription factors (TFs) within the nucleus plays a crucial role in transcriptional regulation. The TF search for a specific DNA sequence is one of the main factors in gene expression. Thus, the interactions between two TFs due to weak protein-protein interactions (PPIs) form TF clusters, influencing their occupancy at a particular target site. To understand how the 3D structure of the chromatin affects the TF agglutination, we proposed a model to convey the presence of PPIs in the nucleus and verified how the clustering affects the TF allocation, considering the 3D/1D diffusion as our search mechanism. Then, once a TF is bound to a region, it recruits RNA Polymerase (RNAP). Besides, inducible TFs remain in the cytoplasm and translocate into the nucleus through the nuclear pore complex (NPC) upon proper signalling. To incorporate these mechanisms, we proposed another model to understand the TF search and RNAP recruitment dynamics. We obtained deterministic and stochastic solutions verifying how transcription is enhanced at the NPC periphery and confirmed through imaging analysis of specific genes. Finally, we incorporated the mRNA export process to verify the different cytoplasmic transcripts concentrations proving how the volume of available mRNA is also NPC-dependent. Therefore, our work shows relevant connections between the chromatin structure, the allocation of transcriptional resources and the stochasticity in gene regulation.

TF search, Protein-protein Interactions, RNAP recruitment, mRNA export, mathematical model, chromatin structure