Mathematical modeling of gene expression in space and time

Maria Knaiir Al Douaihy

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR
DE L'UNIVERSITÉ DE MONTPELLIER

En Mathématiques et modélisation

École doctorale I2S

Unité de recherche UMR5535-UMR5294

Mathematical modeling of gene expression in space and time

Présentée par Maria KNAIIR AL DOUAIHY

le 07 décembre 2023

Sous la direction de Ovidiu Radulescu
et Mounia Lagha

Devant le jury composé de

Ramon Grima, Professor, Chair of Mathematical Biology School of Biological Sciences University of Edinburgh, EH9 3BF, UK	**Rapporteur**
Michèle Thieullen, Maîtresse de conférences, Sorbonne Université, Paris, France	**Rapportrice**
Benoite De Saporta, Professor, Université de Montpellier, Montpellier, France	**Examinatrice**
Arnaud Debussche, Professor, ENS Rennes, Rennes, France	**Président du jury**

UNIVERSITÉ
DE MONTPELLIER

Résumé en Français

Au cours du développement embryonnaire, les cellules doivent adopter des programmes d'expression génique afin de se spécifier en plusieurs destins. L'activation de ces programmes, nécessite l'allumage de gêne du développement, leur transcription. Alors que la spécification cellulaire est précise dans l'espace et dans le temps, et pourrait sembler déterministe, la transcription reste un phénomène extrêmement stochastique.

Ainsi les techniques de visualisation de l'ARNm, soit dans des échantillons fixés à l'aide de smFISH (Trcek et al. (2016), Lyubimova et al. (2013)) ou dans des cellules vivantes (methode MS2/MCP Bertrand et al. (1998) (Figure 1 A)), ont révélé que la synthèse des ARN s'établissait de manière discontinue. La transcription s'établit par une alternance entre périodes actives, où plusieurs polymérases initient la transcription (bursts) et périodes inactives. Cette stochasticité dans la transcription (bursting) peut engendrer une hétérogénéité de transcription entre cellules voisines, phénomène connu sous le nom de "bruit biologique".

Au cours de ma thèse, je me suis intéressée aux sources de ce bruit et aux mécanismes permettant de le contrôler. J'ai appliqué des modèles mathématiques pour mieux appréhender des données biologiques, acquises au cours de l'embryogenèse précise de la drosophile. En effet l'embryon de drosophile est un système idéal pour étudier la stochasticité de la transcription parce que l'imagerie quantitative et les manipulations génétiques y sont aisées.

Je me suis particulièrement intéressée à la régulation d'un gène clé du développement, le gène *snail*, qui code pour un facteur de transcription, essentiel à la mise en place du mésoderme, aux transitions épithélio-mésenchymateuses (EMT) et à la gastrulation. Ce gène est conservé chez les vertébrés, et sa dérégulation est impliquée dans les EMT des cellules cancéreuses métastatiques chez l'Homme.

Je me suis concentrée sur le rôle des sequences cis-regulatrices de ce gène, son promoteur et ses deux enhancers (l'un proximal et l'un distal). La transcription de ce gène a été visualisée en temps réel, grâce à la méthode MS2/MCP (Figure 1 B)).

Mathématiquement, la transcription est modélisée comme une chaîne de Markov dans l'hypothèse où un nombre restreint d'étapes limites est modélisé comme une transition entre des états discrets. Nous classons ces états en trois catégories : les états ON productifs qui peuvent initier la transcription, les états OFF non productifs qui ne peuvent ni initier ni reprendre la transcription, et les états de pause dans lesquels la transcription initiée s'arrête et peut reprendre plus tard ou s'interrompre. Le promoteur ne démarre la transcription

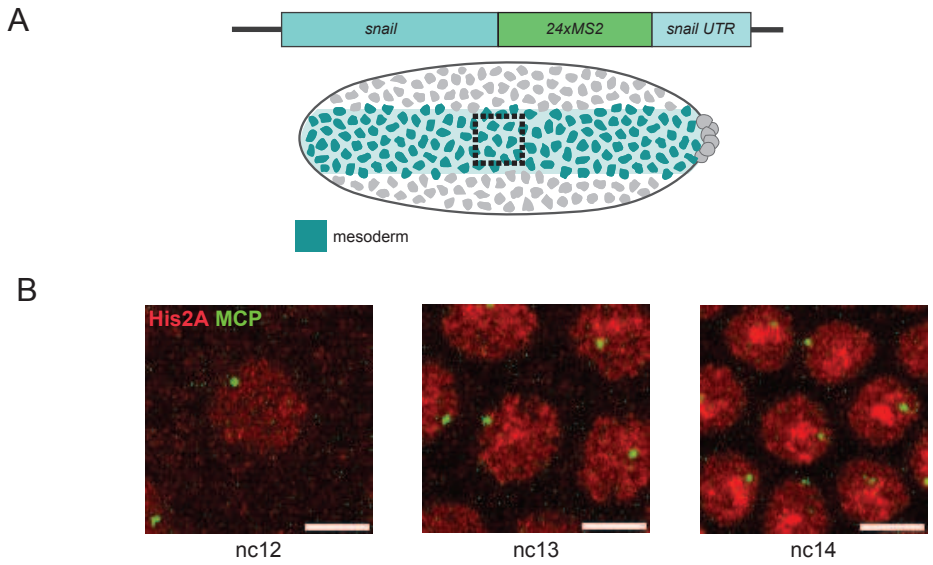


Figure 1: Imagerie de la transcription en temps réel (extrait de Pimmett et al in prep). A) Haut: Vue schématique des transgènes utilisés pour visualiser la transcription à l'aide du système MS2/GFP. Bas: Schéma de l'embryon de drosophile montrant la restriction spatiale de l'analyse au mésoderme présumé. B) Projection d'intensité maximale d'un Z-stack représentatif de l'embryon nc12, nc13, nc14 montrant les foci transcriptionnels liés à la MS2/MCP-GFP (GFP) et les noyaux (histone-RFP).

que dans l'état ON lorsqu'il peut déclencher plusieurs départs de molécules de RNAP le long de l'ADN. La RNAP peut finalement s'arrêter dans un état de pause ou s'engager dans une élongation irréversible que nous modélisons par l'état EL (Figure 2). Les techniques d'imagerie de la transcription en direct permettent de suivre la transcription en temps réel et pour chaque site de transcription.

Dans cette thèse, nous combinons des aspects théoriques, des techniques informatiques et des analyses d'imagerie en temps réel de cellules individuelles. Notre objectif princi-

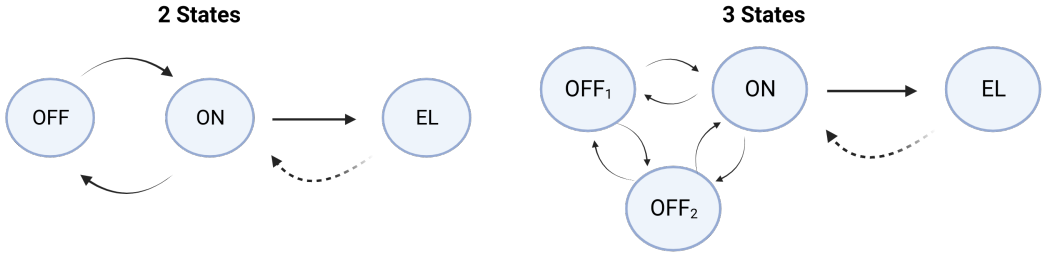


Figure 2: Illustration d'un processus de transcription modélisé comme un modèle de Markov. Sur le côté gauche, nous décrivons une représentation simplifiée avec deux modes d'états (modèle télégraphique) : l'état "OFF", "ON" et l'état "El". Sur la partie droite, une vue plus détaillée comprend deux états non obligatoires : 'OFF_1', 'OFF_2' et 'ON', suivis de l'état 'El'.

pal est d'élargir notre compréhension concernant la nature stochastique de l'expression des gènes, notamment au niveau transcriptionnel, et résoudre la contradiction apparente entre la stochasticité et la robustesse des mécanismes d'expression des gènes.

Nous utilisons Drosophila melanogaster comme organisme modèle pour trois raisons principales: le génome a été entièrement cartographié, ce qui permet une compréhension complète de ses enhancers et promoteurs. Le développement de techniques d'imagerie en direct nous permet de suivre la transcription et la traduction en temps réel et de manière fiable, de plus, le développement embryonnaire de Drosophilia melanogaster est un modèle reproductible.

Cependant, les différents paramètres des données transcriptionnelles qui aboutissent à la traduction nécessitent des approches de modélisation uniques. Nous distinguons deux critères principaux qui nous ont permis de connaître les limites de notre modèle dans l'extraction d'informations directes à partir des données :

1. Signal "homogène dans le temps" vs "homogène dans le temps".

2. Signal "homogène dans l'espace" vs "inhomogènes dans l'espace".

Par conséquent, nous divisons notre problématique en trois conditions principales : données homogènes dans le temps et dans l'espace, données inhomogènes dans le temps mais homogènes dans l'espace, et enfin données inhomogènes dans le temps et dans l'espace. Le terme inhomogène dans le temps (resp. l'espace) provient du fait que le taux de transition entre les différents états discrets du modèle markovien est dépendant du temps (resp. de l'espace) (Figure 3).

Dans un premier temps, nous avons examiné un cas restreint d'homogénéité temporelle et spatiale. Cette hypothèse simple sert de point de départ, nous permettant d'extraire directement des informations des données transcriptomiques. Cependant, lorsque nous introduisons des hypothèses plus complexes, nous observons un changement dans notre approche. À mesure que la complexité augmente, l'extraction de détails spécifiques des données est diminuée et nous commençons à extraire des conclusions plus théoriques. Cette transition reflète le compromis entre la richesse des informations expérimentales et la profondeur des connaissances théoriques au fur et à mesure que nous naviguons dans les complexités de la modélisation de l'expression génique.

Ce manuscrit est composé de quatre chapitres principaux. Les méthodes décrites dans les

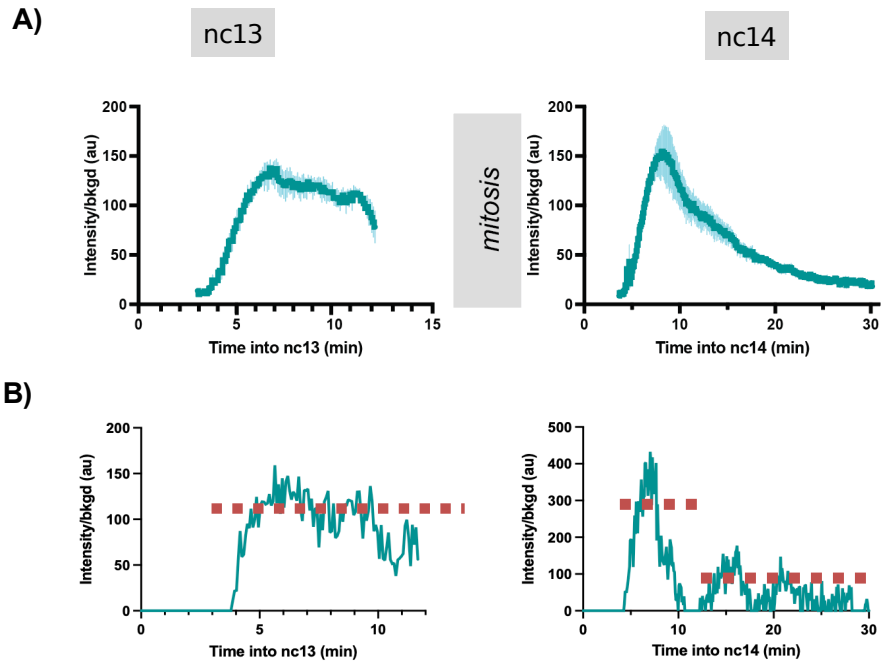


Figure 3: Homogénéité dans le temps (extrait de Pimmett et al in prep). A) Moyenne du signal d'intensité sur plusieurs noyaux de transcription dans nc13 (à gauche) et nc14 (à droite). B) Échantillon du signal d'intensité de nc13 mettant en évidence l'homogénéité du signal à gauche et l'inhomogénéité du signal dans nc14 (à droite).

chapitres 1 et 2 nécessitent des données homogènes dans le temps et dans l'espace. Une seconde partie de ma thèse vise à modéliser des données inhomogènes dans l'espace et dans le temps, présentées dans les chapitres 3 (inhomogènes dans le temps) et 4 (inhomogène dans le temps et l'espace).

Dans le Chapitre 1, nous présentons BurstDECONV, une méthode d'inférence statistique innovante conçue pour déconvoluer les traces de signaux en événements individuels d'initiation de la transcription. Nous appliquons ensuite la solution du problème inverse pour obtenir les paramètres de transition des données transcriptionnelles pour différents phénotypes. Une analyse comparative approfondie des paramètres de cette méthode d'inférence est présentée, ainsi qu'une comparaison avec d'autres méthodes utilisant des données synthétiques et réelles. Les données sont acquises par imagerie en direct d'une cellule unique, les signaux d'intensité étant calibrés à l'aide de smFISH Tantale et al. (2021); Pimmett et al. (2021).

Résultat: Le résultat de l'algorithme est triple : 1) Une carte temporelle détaillant les événements de transcription. Cette carte indique, pour chaque cellule, les moments précis où les différentes d'ARN polymérase initient la production d'ARNm. 2) Sélection du modèle : identification du nombre d'étapes limitant le taux de transcription. 3) Identification des paramètres de transition à partir de données d'imagerie en direct.

Dans le Chapitre 2, nous réalisons la résolution du problème inverse. La résolution du problème inverse consiste à obtenir les taux cinétiques du processus de Markov à partir des paramètres de la fonction de survie constituée par le temps d'attente entre les polymérases. En analysant la distribution des temps d'attente entre les événements successifs d'initiation de la polymérase, nous sommes en capacité de déduire des caractéristiques mécanistiques de la transcription, notamment le nombre d'étapes limitant la vitesse de la réaction et leur cinétique.

Résultat: Nous proposons une solution plus générale pour la résolution du problème inverse pour traiter des modèles plus compliqués que celui présenté dans le premier chapitre.

Dans le Chapitre 3: Nous souhaitons traiter des séries temporelles non homogènes. Un exemple d'inhomogénéité dans la transcription peut résulter de la présence d'un répresseur. Dans ce cas, les paramètres de commutation du modèle markovien dépendent de la concentration en répresseur.

Résultat: Nous avons développé une méthode supplémentaire pour analyser les données non homogènes. Cette méthode permet de diviser le signal temporel non homogène en signaux homogènes à l'aide d'une approche bayésienne. En simplifiant la complexité du problème, nous pouvons ensuite appliquer BurstDeconv à la partie du signal dont les paramètres de transition sont constants.

Dans le Chapitre 4: Nous avons effectué le traitement des signaux inhomogènes dans le temps et dans l'espace. Dans ce chapitre, nous avons étendu notre étude à des perspectives plus larges : les dimensions temporelles et spatiales. Comme ce problème est très complexe, nous ne pouvons pas obtenir d'informations directement à partir des données brutes. Par conséquent, dans ce chapitre, nous avons généré des méthodes numériques qui récapitulent les observations expérimentales (données transcriptionnelles) tout en étant performantes en termes de temps de calcul.

Résultat: Les résultats de ce chapitre sont au nombre de trois:

- Introduction d'une nouvelle méthode de simulation hybride qui implique des processus de Markov discrets et une approche déterministe (équations différentielles partielles EDP) pour modéliser l'expression des gènes avec une extension spatiale. Ce modèle est également capable de capturer la variabilité provenant de la stochasticité inhérente à la transcription.

- Comparaison de cette méthode avec d'autres techniques de simulation.

- Investigation des aspects critiques de la modélisation du blastoderme de la drosophile: autorégulation négative Coulier et al. (2021), mémoire transcriptionnelle Bellec et al. (2018); Dufourt et al. (2018).

Résumé en Anglais

During embryonic development, cells must adopt specific gene expression sequences in order to distinguish into several different fates. Activation of these genes requires their transcription. Whereas gene expression regulation is precise in space and time, and might seem deterministic, transcription remains an extremely stochastic phenomenon.

Single cell mRNA visualization techniques, either in smFISH-fixed samples (Trcek et al. (2016), Lyubimova et al. (2013)) or in living cells (MS2/MCP method Bertrand et al. (1998) (Figure 4 A)), have revealed that RNA synthesis is discontinuous. Transcription alternates between active periods, when several polymerases initiate transcription (bursts) and inactive periods. This stochasticity in transcription bursting can lead to transcriptional heterogeneity between neighboring cells, a phenomenon known as "biological noise".

In my thesis, I investigated the sources of this noise and the mechanisms by which it can be controlled. I applied mathematical models to better understand biological data acquired during the precise embryogenesis of Drosophila. Indeed, the Drosophila embryo is an ideal system for studying transcriptional stochasticity, because it is easy to image quantitatively and to manipulate genetically.

I was particularly interested in the regulation of a key developmental gene, *snail*, which encodes a transcription factor essential for mesoderm development, epithelial-mesenchymal transitions (EMT) and gastrulation. This gene is conserved in vertebrates, and its deregulation is implicated in the EMT of human metastatic cancer cells.

I focused on the role of the cis-regulatory sequences of this gene, its promoter and its two enhancers (one proximal and one distal). Transcription of this gene was visualized in real time, using the MS2/MCP method (Figure 4 B)).

Mathematically, transcription is modeled as a Markov Chain under the assumption that a small number of limiting steps are modeled as transition between discrete states. We classify these states in three categories: productive ON states that can initiate transcription, non-productive OFF states that can not initiate nor resume transcription, and paused states in which initiated transcription stops and can resume later or abort. The promoter starts transcription only in the state ON when it can trigger several departures of RNAP molecules along DNA. The RNAP can eventually stop in a paused state or commit to irreversible elongation that we model by the state EL (Figure 5). Live transcription imaging techniques allows the monitoring of transcription in real time and for each transcription site.

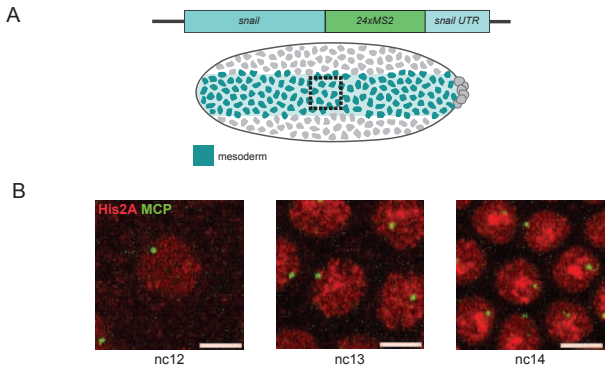


Figure 4: Real-time single cell transcription imaging (from Pimmett et al in prep). A) Top: Schematic view of the transgenes used to visualize transcription using the MS2/GFP system. Bottom: Schematic of Drosophila embryo showing spatial restriction of analysis to presumptive mesoderm. B) Maximum intensity projection of a representative Z-stack of the nc12, nc13, nc14 embryo showing transcriptional foci linked to MS2/MCP-GFP (GFP) and nuclei (histone-RFP).

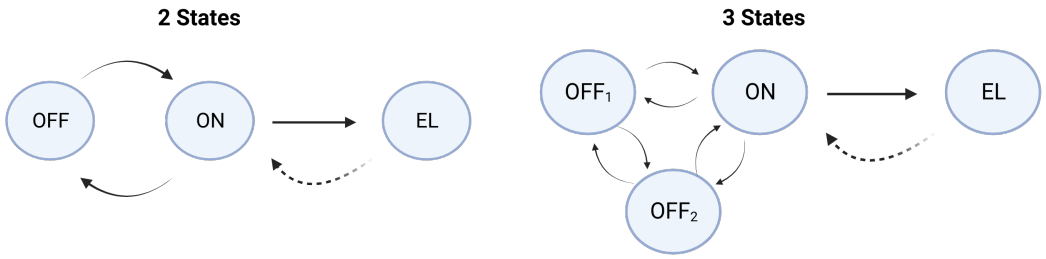


Figure 5: Illustration of a transcription process modeled as a Markov model. On the left side, we depict a simplified representation with two states mode (telegraph model): the 'OFF', 'ON' state and the 'Elongation' state. On the right side, a more detailed view includes two non-obligatory states: 'OFF_1', 'OFF_2' and 'ON,' followed by the 'Elongation' state.

In this thesis, we combine theoretical aspects, computational techniques and real-time imaging analysis of single cells. Our main aim is to broaden our understanding of the stochastic nature of gene expression, particularly at the transcriptional level, and to resolve the apparent contradiction between stochasticity and robustness of gene expression mechanisms.

We use Drosophila melanogaster as a model organism for three main reasons: the genome has been fully mapped, enabling a complete understanding of its enhancers and promoters. The development of live imaging techniques enables us to follow transcription and translation reliably in real time, and Drosophilia melanogaster embryonic development is a reproducible model.

However, the different parameters of transcriptional data that lead to translation require unique modeling approaches. We distinguish two main criteria that have enabled us to identify the limitations of our model in extracting direct information from the data:

1. 'Time homogeneous' vs 'time inhomogeneous' signal.

2. 'Space homogeneous' vs 'space inhomogeneous' signal.

Consequently, we divide our problem into three main conditions: data homogeneous in time and space, data inhomogeneous in time but homogeneous in space, and finally data inhomogeneous in time and space. The term inhomogeneous in time (resp. space) derives from the fact that the transition rate between the different discrete states of the Markov model is time (resp. space) dependent (Figure 6).

Initially, we examined a restricted case of temporal and spatial homogeneity. This simple assumption serves as a starting point, enabling us to extract information directly from the transcriptomic data. However, when we introduce more complex assumptions, we observe a shift in our approach. As complexity increases, the extraction of specific details from the data is diminished and we begin to extract more theoretical conclusions. This transition reflects the trade-off between the richness of experimental information and the depth of theoretical knowledge as we navigate the complexities of gene expression modeling.

This manuscript is composed of four main chapters. The methods described in Chapters 1 and 2 require spatially and temporally homogeneous data. A second part of my thesis aims to model data inhomogeneous in space and time, presented in chapters 3 (inhomogeneous in time) and 4 (inhomogeneous in time and space).

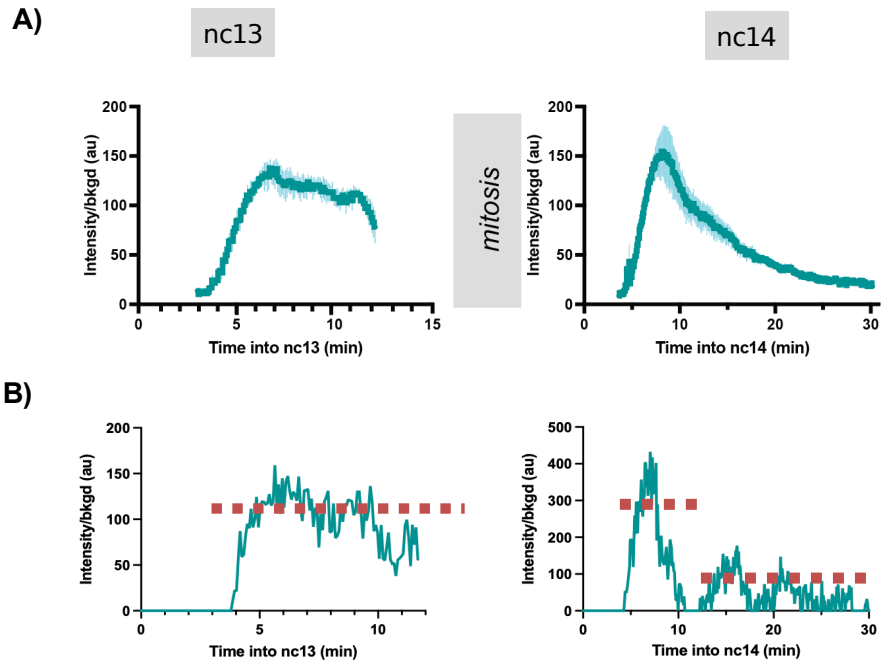


Figure 6: Homogeneity over time (extract from Pimmett et al in prep). A) Average of mean intensity signal across multiple transcriptional nuclei in nc13 (left) and nc14 (right). B) Sample of intensity signal of nc13 highlighting the homogenity of the signal on the left and the inhomogeneouty of the signal in nc14 (right)

In Chapter 1, we introduce BurstDECONV, an innovative statistical inference method designed to deconvolve signal traces into individual transcription initiation events. We then apply the inverse problem solution to obtain transition parameters from transcriptional data for different phenotypes. An in-depth comparative analysis of the parameters of this inference method is presented, as well as a comparison with other methods using synthetic and real data. Data are acquired by live single-cell imaging, with intensity signals calibrated using smFISH Tantale et al. (2021); Pimmett et al. (2021).

Result: The result of the algorithm is threefold: 1) A time map detailing transcription events. This map indicates, for each cell, the precise moments when different RNA polymerases initiate mRNA production. 2) Model selection: identification of the number of rate-limiting steps in transcription. 3) Identification of transition parameters from live imaging data.

In Chapter 2, we solve the inverse problem. Solving the inverse problem consists in obtaining the kinetic rates of the Markov process from the parameters of the survival function constituted by the waiting time between polymerases. By analyzing the distribution of waiting times between successive polymerase initiation events, we are able to deduce mechanistic features of transcription, notably the number of rate-limiting steps and their kinetics.

Result: We propose a more general solution for solving the inverse problem to deal with more complicated models than the one presented in the first chapter.

In Chapter 3: We want to deal with inhomogeneous time series. An example of inhomogeneity in transcription may result from the presence of a repressor. In this case, the switching parameters of the Markov model depend on the concentration of the repressor.

Result: We have developed an additional method for analyzing inhomogeneous data. This method divides the inhomogeneous temporal signal into homogeneous signals using a Bayesian approach. By simplifying the complexity of the problem, we can then apply BurstDeconv to the part of the signal whose transition parameters are constant.

Chapter 4: We have processed signals that are inhomogeneous in time and space. In this chapter, we have extended our study to broader perspectives: temporal and spatial dimensions. As this is a very complex problem, we cannot obtain information directly from the raw data. Therefore, in this chapter, we have generated numerical methods that summarize the experimental observations (transcriptional data) while being efficient in terms of computation time.

Result: There are three results in this chapter:

- Introduction of a new hybrid simulation method involving discrete Markov processes and a deterministic approach (partial differential equations PDEs) to model gene expression with spatial extension. This model is also capable of capturing the variability arising from the stochasticity inherent in transcription.

- Comparison of this method with other simulation techniques.

- Investigation of critical aspects of Drosophila blastoderm modeling: negative autoregulation Coulier et al. (2021) and transcriptional memory Bellec et al. (2018); Dufourt et al. (2018).

Acknowledgements

I would like to start by thanking Dr. Ramon Grima, Dr. Michèle Thieullen, Dr. Benoite De Saporta, and Dr. Arnaud Debussche for accepting to be members of the jury of my PhD thesis Defense. Their time reading and evaluating the manuscript is truly appreciated, and I feel honored to have you as my committee members. I look forward to the discussions during the defense.

I thank Dr. Antoine CLAESSENS, Dr. Benoîte DE SAPORTA, and Dr. John Reinitz for being part of my annual PhD committee (CSI). Thank you for all your time and guidance during the last three years.

I want to thank my two supervisors, Ovidiu Radulescu and Mounia Lagha, for giving me the chance to work on this interesting topic.
Ovidiu, your trust in me and your supervision-freedom balance have made me grow in my personal and professional life by learning to be more creative and to outgrow my hesitation in trusting my ideas and making decisions.
Mounia, without you, I would not have been here, and for that, I will always be grateful. I cannot express enough my appreciation for your support during the last three years. I would like to thank you, especially for all the advice regarding organization, communication skills, and the time spent talking science.

I would like to thank all members and past members of both of my teams, in LPHI and in IGMM. I would like to especially thank Inayat; you have been with me as a teammate and a friend since the beginning, and through it all, your presence in my life is a source of comfort, especially during the rainy days with a chai in your hand. Louise, not only have you supported me, especially in the beginning of my PhD with all of my basic biological questions, but the after-work we do in Broc Cafe will always remind me of you. Virginia, I never met anyone who can explain to me biology with food metaphors; not only was it useful, but it always made me hungry. Elea, Cas, Jason, Syrian, Quentin, Marie, each one has impacted my life during the last three years one way or another. Thank you for making life at LPHI and the after-work gatherings, either in movie nights, coffee outings, or parties, that much fun.

A special thank you to Dr. John Reinitz, with whom I have had multiple interesting discussions either through Zoom or when he welcomed me graciously in Chicago. Multiple of our discussions made me fascinated about the research field.

The most important people are always the hardest to acknowledge: my family.
Mom, you have always been an idol for me, either as a super-mum or a working woman. You never deprived us of any activity, spending hours driving and waiting for us for guitar lessons, gymnastics, football, libraries, AC-math classes, or whatever interested us. You were always creative in finding ways to make us motivated about school and our professional future. As a working woman, not only have you never stopped improving yourself with formation, but I will always be impressed with your determination in obtaining a second master with two teenagers, a toddler, and being pregnant.
Dad, half of my personality is impacted by you. You taught me by example how to always find a reason to understand others, help others, and love without judgment. Your commitment to helping the less fortunate, prisoners, refugees, the poor, people without anyone through various associations and NGOs has impacted me and many others profoundly.
To my sisters, Clara, Francesca, and Lea, with each passing day, I value your existence in my life more and more. I am so proud to see you grow and to watch all of the accomplishments you do each passing day. No matter how far or how much time has passed since I saw you, I always feel your love and support.

To my childhood friends, Vero, Rasha, Taty, Rita, Roodi, Saiid, we have been with each other through it all. The first people I called when I got the PhD to share my news with and the people I will probably be with at a nursing home at the age of 100. You are and always have been my chosen family. I am really thankful for how far you are willing to go for me and how you always made me feel special and loved.

To Pierre, you came into my life at my lowest, and with your tenderness and calmness, you made the last part of my PhD journey easier. I would like to take this chance to thank you for everything you have done and for all the small things you do that brighten up my mood. You are the most caring, honest, funny, and loving person; you are truly the "Pierre," for that I am grateful.

To the rest of my friends, especially Monica, Roberte, Carine, Maroun, Marcel, Roland, Carmel, Rose-Marie, Mike, thank you for your support and love. You have welcomed me and borne me whenever I was stressed, tired, and demotivated.

I have always been lucky to have found support and love wherever life takes me, and I have never taken that for granted. This PhD journey is not my work alone; it has been accomplished through the love, support, and trust in my capacities of every person I have encountered. So, thank you!

Contents

CHAPTER 1

General introduction

1.1.0 Gene expression

1.1.1 Into the world of gene expression

> Intelligence is the ability to adapt to change
>
> *Stephen Hawking*

The cell is the fundamental structural and functional biological unit of all known living beings. At the very core of its functioning are the genes, which constitute the building blocks of the genetic information. This genetic information is decoded to create functional components of the cells, the proteins. Decoding this information requires a tight control of gene expression, which involves two key steps transcription and translation (Figure 1.1). Gene expression must be tightly regulated to allow a cell to adapt to its changing environment and to adopt a specific cell fate (Lee and Young (2013)).

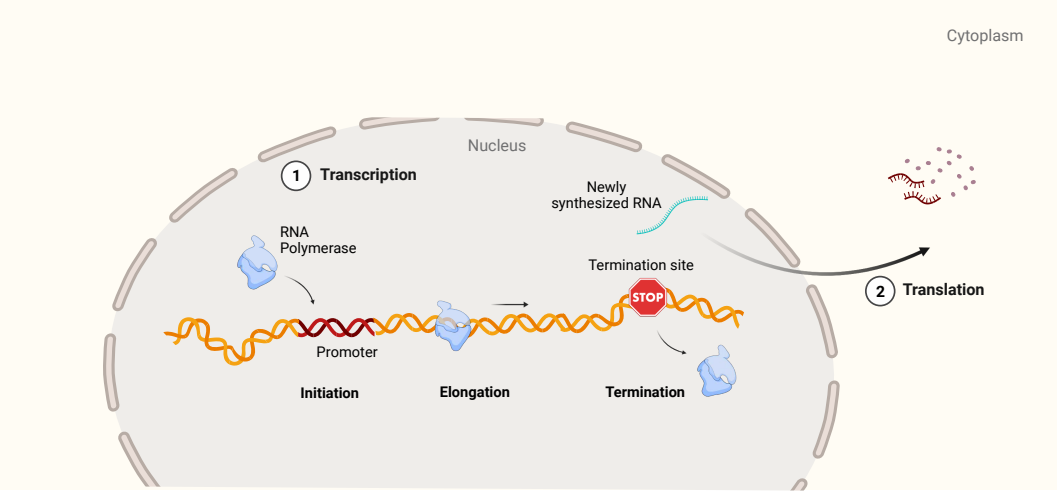


Figure 1.1: Overview of Gene Expression Process.

Transcription and Translation: two steps of the central dogma of molecular biology.

Stated by Francis Crick in 1957 Cobb (2017), the central dogma of gene expression states that genetic information flows from DNA to RNA (transcription) and then from RNA to protein (translation).

Transcription is the process by which the genetic information stored in the DNA sequence is converted into a messenger RNA (mRNA). This fundamental process requires the action of several transcription factors and the RNA Polymerase II (Pol II). Following transcription, the mRNA molecule carries the genetic information from the nucleus, where DNA resides, to the cytoplasm, where protein synthesis takes place.

Translation is the process by which the genetic instructions carried by mRNA, are converted into functional proteins. By coordinating the different phases of translation (Kasinath et al. (2006)) cells ensure the faithful translation of mRNA into proteins, enabling them to carry out essential biological processes and contribute to the overall complexity and diversity of cellular functions.

1.1.2 Transcription regulation

Rate limiting steps of transcription

Many aspects of gene regulation are universal, and the enzymatic machinery involved in DNA processing demonstrates significant adaptability. Notably, RNA Polymerase II emerges as a universal transcriptional engine, capable of transcribing a diverse array of protein-coding genes.

The transcription cycle, consisting of initiation, elongation, and termination, plays a piv-

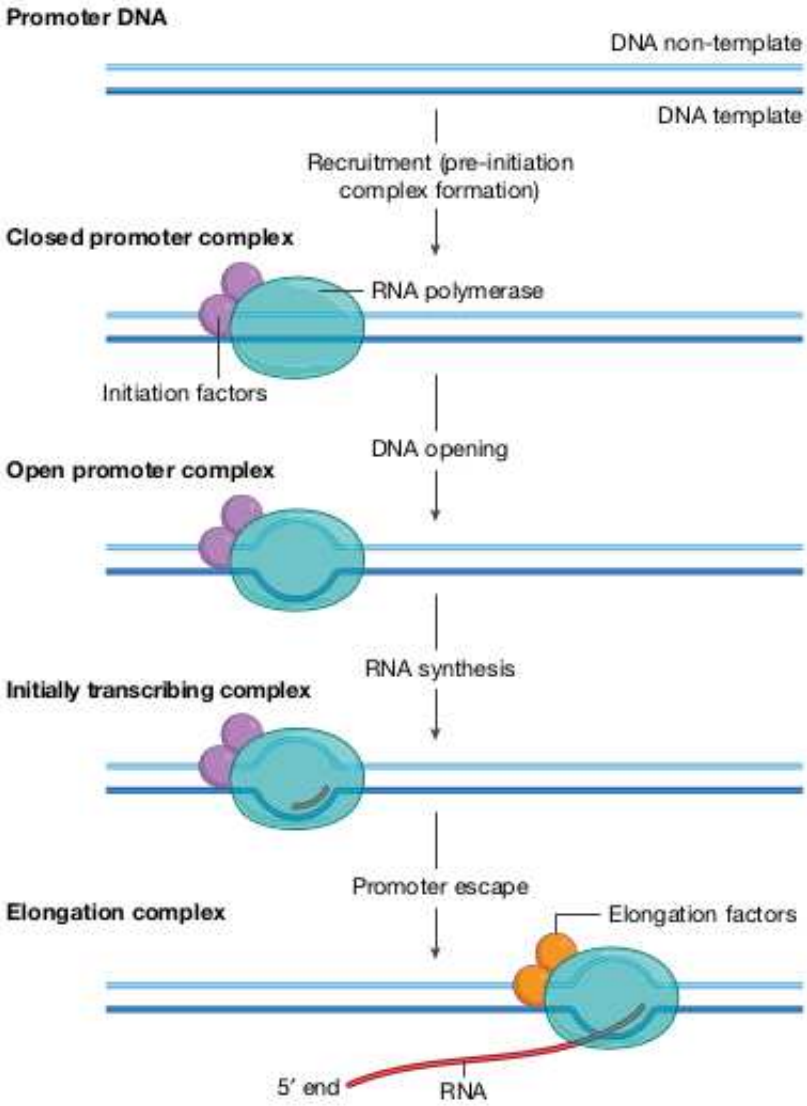


Figure 1.2: Key steps of gene transcription (Taken from Cramer (2019)).

otal role in this intricate regulatory process. Initiation involves the assembly of General Transcription Factors (GTFs) into the Pre-Initiation Complex (PIC) and the recruitment and modification of RNA Polymerase II (Figure 1.2). As transcription proceeds through elongation, it may encounter pausing and then transition to productive elongation before ultimately culminating in termination. Variability in transcription outcomes often traces back to the initiation phase, underscoring its significance in gene expression regulation.

At the core of transcription initiation is the chromatin accessibility. Chromatin, with its two distinct states of euchromatin and heterochromatin, assumes a central role in dictating the feasibility of transcription initiation. Euchromatin is characterized by its open and less compacted structure which fosters a favorable environment. In this environment, DNA readily interfaces with transcription factors and RNA polymerase. These interactions promote the initiation of transcription. Conversely, in heterochromatin, where chromatin adopts a highly

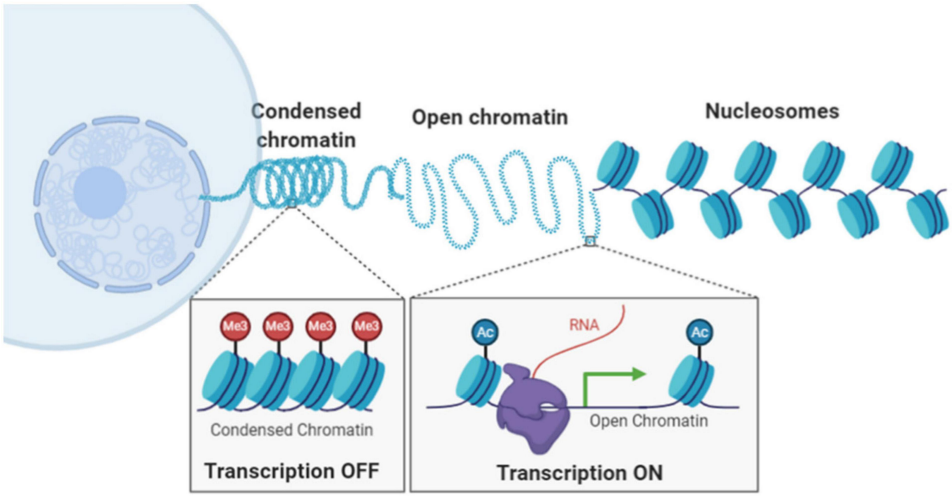


Figure 1.3: Visual representation depicting the hierarchical organization of chromatin and its varying compaction levels within the cellular nucleus (Taken from Xu and Liu (2021)).

condensed form, the initiation of transcription is impeded due to reduced DNA accessibility (Figure 1.3). The intricate relationship encompassing transcription initiation, chromatin dynamics (euchromatin and heterochromatin), and the pre-initiation complex underscores the profound impact of chromatin structure on the regulatory fate of a gene. By shaping the accessibility of the promoter region and guiding the formation of the pre-initiation complex, chromatin emerges as a pivotal determinant, either permitting active transcription or imposing repression upon a gene.

CIS regulatory elements

At the level of the DNA sequence, transcription is controlled by cis-regulatory elements, enhancers, promoters and insulators Levine (2010); Bulger and Groudine (2011) (Figure 1.4). These modules control where, when and at which levels transcription is activated. I will primarily describe promoters and enhancers as they are the main regulatory modules that I analysed during my PhD.

Cis-regulatory elements (CREs), are generally pieces of non-coding, containing binding sites for transcription factors (TFs) and/or other regulatory molecules Ong and Corces (2011). It is now widely recognized that mutations influencing the functioning of cis-regulatory sequences are the foremost contributors to phenotypic divergence, particularly in terms of morphology.

Promoter Promoters are DNA sequences that serve as the small stretch of DNA on which this Pre-Initiation Complex (PIC) or RNA Polymerase II (Pol II) is recruited. They are indispensable for eukaryotic transcription, yet they solely generate basal levels of mRNA. The majority of eukaryotic genes possess a single promoter containing the transcription start site.

Certain genes have alternative promoters, which initiate transcription at varying genomic positions, often occurring in specific conditions.

Enhancer Enhancers harbor binding sites for multiple transcription factors (TFs), featuring multiple sites for each TF. These enhancers usually position themselves upstream (5′), downstream (3′), or within gene introns. They can also manifest at considerable distances Kleinjan and van Heyningen (2004). They possess the capability to influence diverse genes in distinct contexts by orchestrating long-range chromatin loops, fostering proximity in a three-dimensional spatial framework Holwerda and De Laat (2012).

Multiple enhancers often govern gene expression, each exercising control over specific cell types or stages of development. The traditional notion was that each enhancer exclusively regulates a distinct segment of a gene's expression. Recent discoveries, however, have revealed pairs of enhancers which have substantial overlapping functions Hong et al. (2008a); Perry et al. (2010). These enhancers contribute to the stability of phenotypic traits. The functional independence of enhancers permits mutations in one enhancer to yield confined effects on gene expression facets regulated by other enhancers.

Enhancers can also function in synergy or antagonistically with other enhancers. In some instances, enhancers are active in the same cells, at the same time, and exert coordinated control over the same promoter. This phenomenon introduces the concept of redundant enhancers, wherein multiple enhancers collaborate to finely tune gene expression, enhancing robustness and adaptability in regulatory networks.

For example, the extensively studied enhancer governing *eve* stripe 2 orchestrates the second out of seven expression stripes of the even-skipped gene during the patterning of the Drosophila melanogaster embryo. Our particular emphasis centers on the genomic arrangement of enhancers. This arrangement includes their proximity or distance from the transcription site. We also focus on the intricate interplay among these enhancers.

1.2.0 Drosophila as optimal model to study gene expression

Drosophila, commonly known as fruit flies, have long been cherished by scientists for their key attributes. Firstly, their remarkably fast life cycle and the ease and cost-effectiveness with which it can be manipulated in laboratory experiments. Fruit flies are known for their rapid development from egg to adult, which takes only a matter of days.This short life cycle allows researchers to perform experiments and see many fruit fly generations quickly,

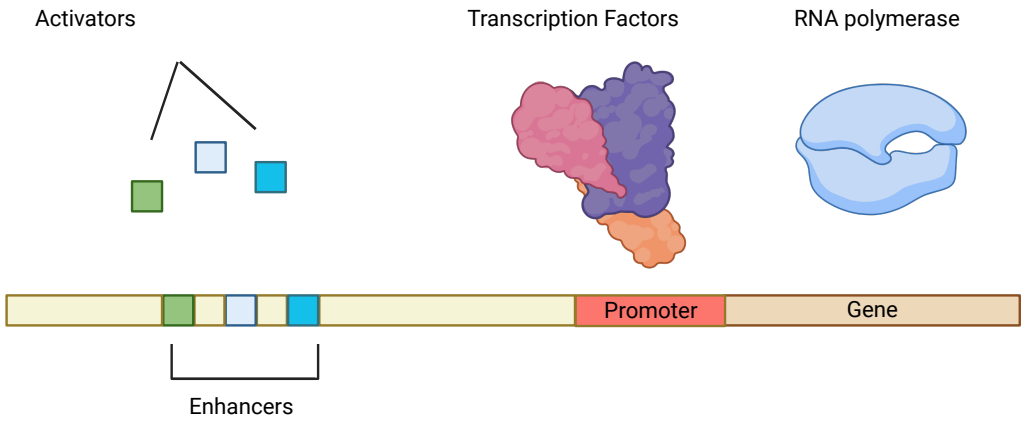


Figure 1.4: Different classes of regulatory elements

speeding up scientific progress.

Second of all, the fruit fly's genome has been fully mapped. Resulting in a comprehensive understanding of its enhancers and promoters. Notably, drosophila possess only four pairs of chromosomes, in sharp contrast to the 23 pairs in humans.

Thirdly, the affordability and simplicity associated with maintaining Drosophila colonies render them an ideal choice for genetic and experimental studies. Researchers can readily introduce genetic mutations, perform gene knockdowns or over-expressions, and manipulate various environmental factors to investigate specific biological processes.

The combination of these attributes has led to numerous groundbreaking discoveries in the realm of biology. Researchers have unraveled fundamental biological mechanisms related to development, genetics, and disease by studying fruit flies. Importantly, many of these discoveries have proven to be highly conserved in humans. This highlights the relevance of Drosophila research to our understanding of human biology. The conservation of these principles across species is significant underscoring the importance of Drosophila as a model organism. Drosophila is valuable for deciphering fundamental principles governing life processes and disease mechanisms.

These factors have resulted in a profound affinity between Nobel laureates and Drosophila, as notably evidenced by the following Nobel Prizes:

- In 1933, Thomas Hunt Morgan employed drosophila to unravel the role of chromosomes in heredity.

- Hermann Joseph Muller, in 1946, used X-ray irradiation to elevate mutation rates in fruit flies.

- In 1995, Edward B Lewis, Christiane Nüsslein-Volhard, and Eric F Wieschaus employed drosophila to understand the genetic control of embryonic development.

- Richard Axel's focus in 2004 revolved around odor receptors and the organization of the olfactory system.

- Jules A Hoffmann, awarded in 2011, made significant contributions to the understanding of innate immunity activation.

- Jeffrey C Hall, Michael Rosbash, and Michael W Young, who received the prize in 2017, unraveled the molecular mechanisms governing circadian rhythms.

1.2.1 Early Drosophila embryogenesis

1.2.2 Fast division and syncytium

Drosophila embryogenesis begins with a large egg. Within this egg, the maternal and paternal nuclei combine, and then they undergo rapid and synchronous division. These divisions include 13 successive stages. (Farrell and O'Farrell (2014). Notably, early nuclear cycles within this developmental process exhibit remarkable efficiency, with minimal intervals between replication and division events. This efficiency is exemplified by the nuclei alternating between division and S-phase (replication) in exceedingly brief time frames. To illustrate, the first 14 nuclear cycles occur in just an hour and a half, averaging about one division every 8.6 minutes Foe and Alberts (1983); Rabinowitz (1941) (see Figure 1.5).

Before cycle 14, which is the primary focus of our study, the nuclei divide without cytokinesis (divisions of membranes). Thus, nuclei shared a common cytoplasm, and this stage is referred to as the syncytial embryo or blastoderm embryo (Figure 1.5 top). During this phase most of the molecular diffusion remains unconstrained. However, during nc 14, which extends for $\sim$ 50 minutes, the plasma membrane invaginates from the apical side of the nuclei, progressing towards the basal side (embryo's interior). By the end of nc14, the embryo is fully cellularized and can be considered as a real multicellular organism. In contrast, at earlier stages, the cell membrane is only partially invaginated. Thus mRNA and protein products located in the apical cytoplasm will experience distinct diffusion constraints than those located in the basal cytoplasm. This specificity has some impact on the control of gene expression which justify our modeling assumption in Chapter 5.

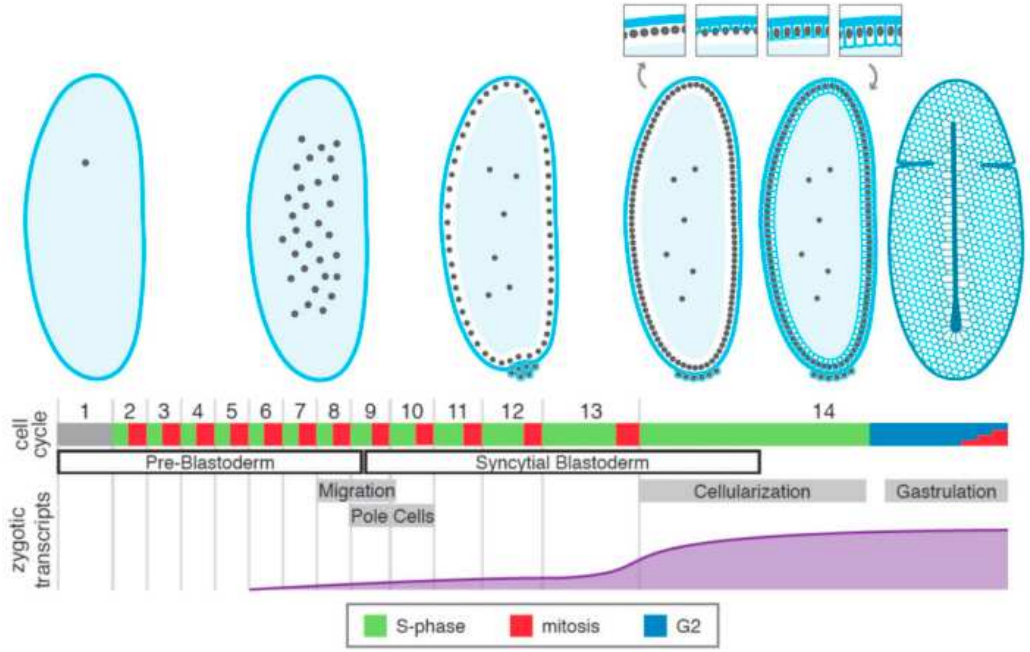


Figure 1.5: Early Development in Drosophila (Taken from Farrell and O'Farrell (2014)). A diagram of the first 14 cycles of Drosophila development with notable morphological stages illustrated at the top. Note that while most embryos are displayed as sections through the middle of the embryo with the ventral side to the right, the final illustration is a surface view, with the ventral side up. The process of cellularization is diagrammed in more detail in the insets. The duration of each phase of the cell cycle is below: S phase (green), mitosis (red), and G2 (blue). Mitosis 14 is represented as a series of small bars because the embryo is no longer synchronous at this time and individual groups of cells enter mitosis at different times according to a developmentally programmed schedule. The timing of notable morphological events are demarcated in grey boxes: the migration of the nuclei to the blastoderm, the insulation of the germline by cellularization of the pole cells, the cellularization of the blastoderm nuclei, and the onset of the first gastrulation movement—ventral furrow formation. Below this is diagrammed the approximate number of genes for which zygotic transcripts have been detected over time.

1.2.3 Maternal-to-Zygotic Transition

In the initial stages of embryonic development across various animal species, transcription does not take place. Instead, maternal RNAs and proteins controls the regulation of development. As development proceeds, control of this process transitions from maternally provided components to those produced by the developing zygote's genome. This shift is referred to as the Maternal-to-Zygotic Transition (MZT) Tadros and Lipshitz (2009). The MZT is characterized by two significant events: the degradation of maternal RNAs and the activation of transcription from the zygotic genome Kwasnieski et al. (2019); Schulz and Harrison (2019). As the MZT nears completion, there are notable changes. The cell division cycle slows down, and a gap phase is introduced, allowing cells to grow before the next division. These alterations prepare the embryo for gastrulation, a critical stage where cells begin to migrate and differentiate into the major germ layers of the organism Tadros and Lipshitz (2009).

Modern technologies have been developed to study the process of zygotic genome activation (ZGA), including methods like metabolic labeling, MS2-based reporters, and the use of RNA-targeted dead Cas9 (dCas9) fused with fluorescent proteins. These techniques enable the tracking of the activation of individual genes during ZGA. This tracking has led to the discovery of intricate transcriptional phenomena, such as mitotic memory Bellec et al. (2018); Ferraro et al. (2015) and transcriptional bursting Pimmett et al. (2021); Bothma et al. (2014); Senecal et al. (2014). Both discovery are of great importance in trancription process and I will further discuss it in Chapter 5.

In the case of Drosophila, a significant wave of ZGA occurs later, during a specific embryonic stage known as nc 14 Edgar and Schubiger (1986); De Renzis et al. (2007). This wave of ZGA leads to the cellularization introduced in the section above. Cellularization represents the first morphological event that depends on zygotic transcription.

1.2.4 Patterning

After fertilization, and before MZT, spatially varying gradients of maternal transcription factors that were initially placed within the egg during oogenesis come into play Jaeger et al. (2012) (Figure 1.7). These gradients play a role in setting up the embryo's dorsal-ventral (DV) and anterior-posterior (AP) orientation . This happens by triggering various signal pathways in the growing embryo. This is outlined in studies like Belvin and Anderson (1996); Kanodia et al. (2009); Levine and Davidson (2005). Consequently, these signaling pathways work together to partition the embryo into distinct tissue types along both the dorsal-ventral and anterior-posterior directions. Throughout the study we will focus more on the DV axis.

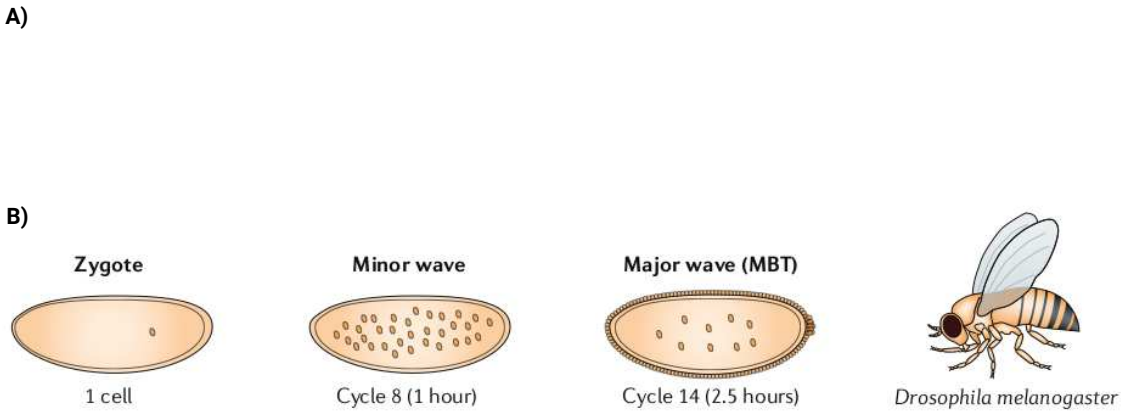


Figure 1.6: Zygotic genome activation (Adapted from Schulz and Harrison (2019). A) In the first hours of life, animals undergo a process called the maternal-to-zygotic transition (MZT) in which the clearance of maternal products is coordinated with the activation of zygotic transcription. A totipotent state (gray bar) is established during this transition. B) Key stages of zygotic genome activation are outlined . The absolute time (in hours post fertilization) is indicated below.

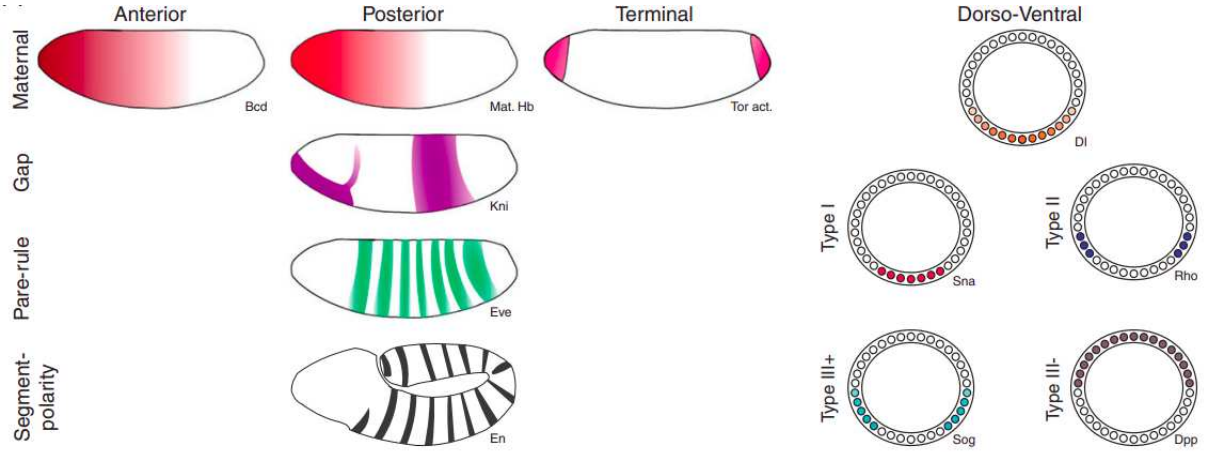


Figure 1.7: Pattern formation systems in the Drosophila blastoderm (Taken from Jaeger et al. (2012)). Maternal systems, both antero-posterior (A–P) and dorso-ventral (D–V), are illustrated in the top row with maternal morphogen gradients guiding their patterning. A–P patterns are presented as lateral views, while D–V patterns are shown in cross-section. Below, we depict representative expression patterns for each downstream gene class: gap, pair-rule, and segment-polarity for A–P, and types I, II, III+, and III- for D–V. Notably, *En* (*Engrailed*) expression is observed in an extended germ-band stage embryo. Key regulators include *Bcd* (*Bicoid*), *Hb* (*Hunchback*), *Kni* (*Knirps*), *Eve* (*Even-skipped*), *Dl* (*Dorsal*), *Sna* (*Snail*), *Rho* (*Rhomboid*), *Sog* (*Short-gastrulation*), and *Dpp* (*Decapentaplegic*).

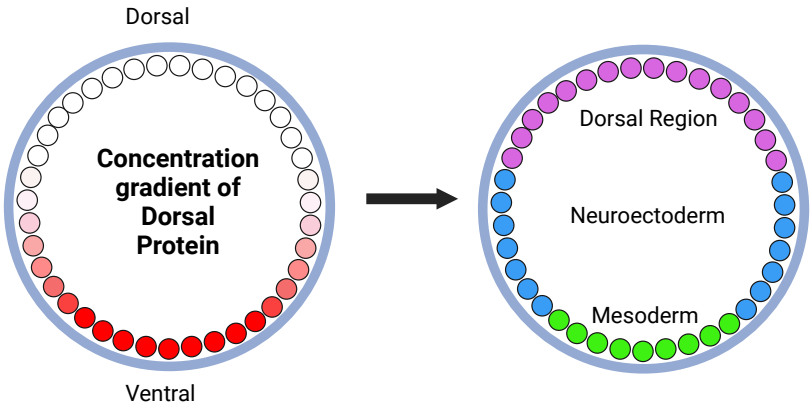


Figure 1.8: *Dorsal Gradient* and Germ Layer Specification. The *dorsal gradient* exhibits a bell-shaped curve, reaching its highest concentration in the embryo's ventral region. In proximity to the peak of this curve, the *dorsal gradient* displays a gradual decline, creating a wide area (containing 12-14 nuclei) with maximal *dorsal* activation, which subsequently specifies the Mesoderm. Along the lateral regions of the embryo, the *dorsal* concentration gradient defines the Neuroectoderm. Beyond the descending region of *dorsal*, the concentration decreases, and the gradient becomes shallow, allowing for the formation of the dorsal region

DV axis

In Drosophila, the establishment of dorsal-ventral (DV) polarity involves the maternal transcription factor *Dorsal*. *Dorsal* shares similarities with *NF-kB* Belvin and Anderson (1996). *Dorsal* forms a concentration gradient along the DV axis, reaching its highest levels in the ventral region of the embryo O'Connor et al. (2006); Umulis et al. (2010); Kanodia et al. (2009); Raser and O'Shea (2005); Hong et al. (2008b) (Figure 1.7 right). The graded distribution of *dorsal* is achieved through differential activation of the toll signaling pathway, originating from events preceding egg-laying Umulis et al. (2010); Kanodia et al. (2009). Intermediate *dorsal* levels penetrate nuclei in the embryo's lateral regions, while the dorsal-most region lacks *dorsal* presence O'Connor et al. (2006); Umulis et al. (2010). This graded *dorsal* distribution leads to distinct expression patterns in nearly 50 target genes involved in the DV system Kanodia et al. (2009).

Through a sophisticated threshold-response mechanism, different concentrations of *dorsal* activate distinct dorsal-target genes along the dorso-ventral axis. This creates the partitioning of the embryo into 3 tissue types: the mesoderm, the neurogenic ectoderm, and the dorsal ectoderm Kanodia et al. (2009) (Figure 1.8). Hence, we distinguish *dorsal* target genes that respond to high *dorsal* thresholds, such as *snail* and *twist*, that will specify the mesoderm in the most ventral part of the embryo. In contrast, intermediate levels responsive genes, such as *sog*/*brk* will specify the neurogenic ectoderm. At the most *dorsal* side of the embryo, the morphogen *dpp* counteracts the action of *dorsal*. There, nuclei are depleted of dorsal protein.

This patterning process occurs with extreme precision, reproducibility and within 2-3 hours. Therefore a question ask-itself: how come such speed and precision can be achieved?

1.2.5 Spatio-temporal precision in patterning

At the deterministic level, gene expression demonstrates remarkable stability and robustness, characterized by a highly stereotyped and predictable pattern. Cells adhere to a precise sequence of steps, faithfully executing genetic instructions without deviation. In optimal conditions, this well-coordinated process unfolds consistently. It exemplifies a widely prevalent and anticipated phenomenon (Dessalles (2017)).

However, when examining the molecular level, using single cell imaging techniques, phenotypic heterogeneity emerges, giving rise to the phenomenon known as biological "noise". This noise has been recognized to originate from various sources, as elucidated by (Raser and O'Shea (2005)). One notable contributor to this intricacy is the phenomenon of "transcriptional bursts". This dynamic behavior results in fluctuations in gene expression levels within an otherwise seemingly deterministic framework.

Transcriptional bursts

The earliest direct visual confirmation of this phenomenon dates back to the 1970s. During that time, Miller chromatin spreads from fruit fly embryos were examined using an electron microscope. These studies revealed an uneven distribution of nascent transcripts along gene sequences (as depicted in Figure 1.9) McKnight and Miller Jr (1979).

These active and inactive phases are described in terms of "promoter states" or levels of gene activity at which transcription can occur. Fluctuations between an 'active' state (also referred to as "ON") and an 'inactive' state (designated as "OFF") result in brief bursts of mRNA production interspersed with periods of transcriptional quiescence.

Transcriptional bursts are further characterized by two key parameters: burst frequency and burst size. The association between burst frequency and burst size can differ among genes and in various regulatory contexts. Understanding how these factors interact is crucial for deciphering the complexities of gene regulation and its impact on the diversity of cellular behaviors.

1.2.6 Methods to visualize transcription in live embryos

Transcriptional burst, chromatin accessibility and ZGA were revealed using live imaging techniques. These imaging techniques can be broadly categorized into two main groups:

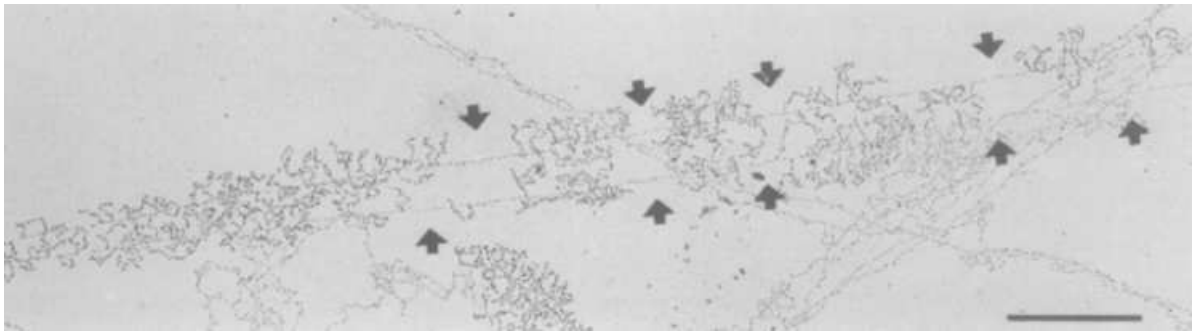


Figure 1.9: Chromatin spreads from Drosophila embryos (Taken from Tunnacliffe and Chubb (2020)). The image shows a pair of sister chromatids aligned in parallel, with inferred initiation sites marked by α and β. Note the increasing size of the fibres (transcripts) extending from the central axis of each chromatid with increasing distance from the initiation sites (scale bar $1\mu m$); also note the fibre-free gaps (marked by arrows).

static fluorescence microscopy, which captures stationary images, and live imaging, which involves recording dynamic movies of transcription processes in vivo. We will provide a comprehensive overview of the techniques employed to acquire the data utilized in this PhD thesis.

Labeling mRNA in fixed samples

Single Molecule Fluorescence in situ Hybridisation (smFISH) is a technique used to detect and visualize individual RNA molecules within fixed tissue at single-cell resolution (Trcek et al. (2016), Lyubimova et al. (2013)). It involves the use of fluorescently labeled probes that hybridize to specific RNA sequences of interest. By imaging and quantifying the fluorescent signals from the labeled RNA molecules, researchers can gain insights into the abundance, localization, and dynamics of transcriptional activity at the single-molecule level (Boettiger and Levine (2013), So et al. (2011)) (Figure 1.10).

Labeling mRNA in living cells

The MS2/MCP system is an amplification method consisting in two parts. First an array of MS2 sequences in inserted in the gene of interest (on transgenes or at the locus by crispr editing). Second, a fluorescent detector protein is provided as a free detector. Upon transcription, this array will form mRNA loops These loops exhibit a high affinity for the detector, which is the RNA-bound protein MCP which is used to a fluorescent detector, such as MCP-GFP. This process is typically visualized using a specific imaging setting. This binding event generates a fluorescent dot (spot) at the transcription site, whose fluorescence dynamically evolve with time (Figure 1.11). The fluctuations in MS/MCP-GFP signal at the transcription site can serve as a proxy for transcriptional activity (Gregor et al. (2014)). The MS2 system was initially developed in 1998 (Bertrand et al. (1998)) and further refined in 2004 (Golding and Cox (2004)). In the drosophila embryos, it was first used in 2013 by Lucas et al. (2013);

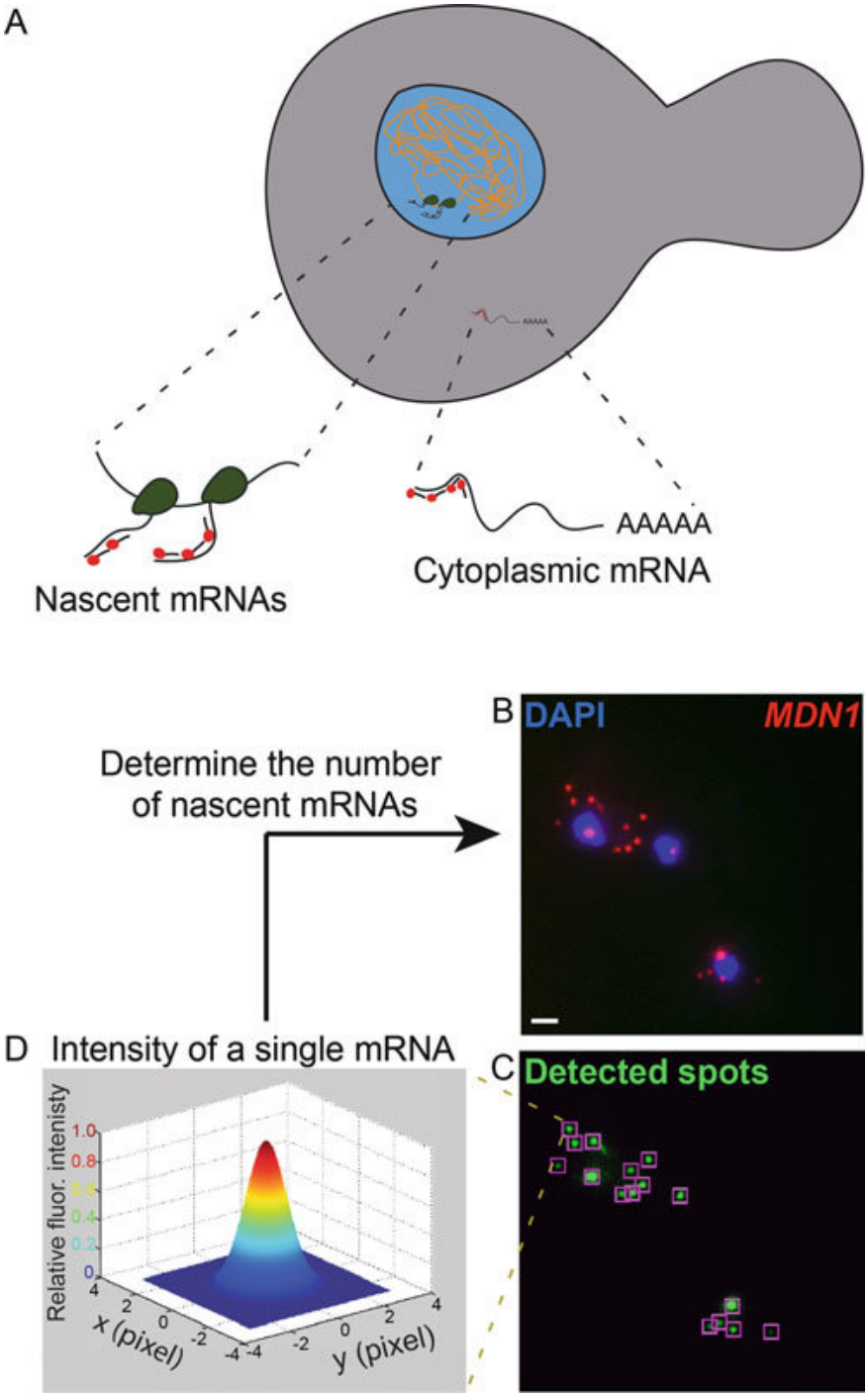


Figure 1.10: Transcription imaging in fixed embryos (Taken from Trcek et al. (2018)). A) A schematic of smFISH probe binding to nascent and mature mRNAs. B) Detection of cytoplasmic and nascent mRNAs (red) using smFISH. The nuclei are stained with a DAPI stain (blue). Note that transcription sites, located within the nuclei, are brighter than single cytoplasmic mRNAs indicating that several RNAP IIs are transcribing the *MDN1* gene concurrently. C) A spot-detection algorithm was applied to detect cytoplasmic and nascent *MDN1* mRNAs. Detected mRNAs are marked as green spots demarcated by purple squares. D) The intensity of a single cytoplasmic mRNA is determined. The total fluorescent intensity of a single spot in x, y, and z is determined. Once the average intensity of a single mRNA is known, it is used to calibrate the total intensity of transcription site to determine the absolute number of nascent chains associated with the active transcription sites detected in b and c. Scale bar: $1\mu m$.

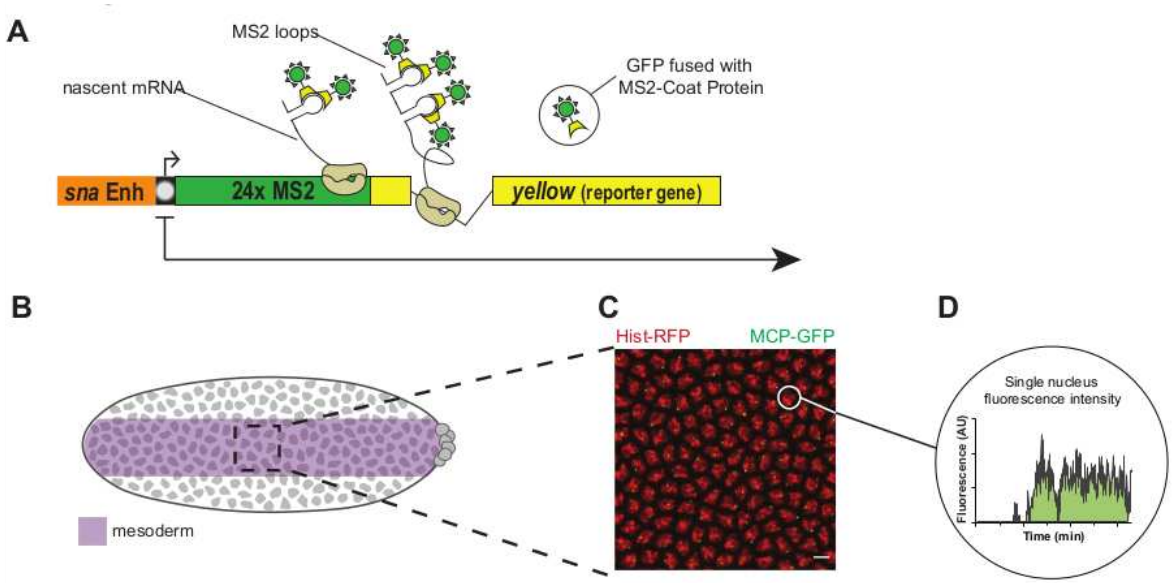


Figure 1.11: Transcription imaging in live embryos (adapted from Pimmett et al. (2021)). A) Schematic view of transgenes used to study transcriptional dynamics. A minimal enhancer was placed upstream of the core promoter followed by 24xMS2 repeats and a yellow reporter gene. B) Schematic of Drosophila embryo showing spatial restriction of analysis to presumptive mesoderm (purple). C) Maximum intensity projection of representative 15μm Z-stack of *snaE<snaPr<*24xMS2-y (*snaE<sna*) nc14 embryo showing MS2/MCP-GFP-bound transcriptional foci (GFP) and nuclei (histone-RFP). Scale bar is 5μm. D) Sample single nuclei trace showing GFP fluorescence during nc14. Surface of green region indicates trace integral amplitude.

Garcia et al. (2013). Since then, it has been extensively utilized in various biological systems Pimmett et al. (2021).

Image analysis

To better understand how intricate regulatory systems are interconnected, researchers would have to use mathematical models Fischer (2008); Joyce and Palsson (2006). Image analysis methods are essential for extracting quantitative information from microscopy data. These techniques involve the processing and analysis of images obtained from techniques such as MS2-GFP system and smFISH (Trullo et al. (2019)). Image analysis algorithms and software enable the quantification of parameters such as signal intensity, spatial distribution, co-localization, and temporal dynamics of transcriptional events. This facilitates the extraction of meaningful data from large datasets and provides quantitative insights into transcription processes.

Data calibration

Calibrating the fluorescent signal obtained from live imaging holds a critical role in the analysis of transcription. This process enables us to interpret data using precise counts of transcribing polymerases. It avoids the need to rely on arbitrary units. To achieve this calibration

for the MS2-GFP system, we employed single-molecule hybridization experiments as outlined in Garcia et al. (2013). By leveraging the fluorescence emitted by an individual mRNA molecule, we can derive an estimation of the average quantity of mRNA molecules. These molecules are situated at the transcription site (TS) within the nucleus. This estimation is made during a stable condition.

These techniques, when used individually or in combination, contribute to our understanding of transcriptional processes. They do so by providing valuable information about the localization, dynamics, and abundance of RNA molecules within cells. They are essential tools for studying gene expression and unraveling the complexities of transcriptional regulation.

1.2.7 Noise in gene expression

The question about how gene expression is tightly regulated despite the heterogeneity arising from the molecular level, or transcriptional bursts, is a fundamental question in molecular biology. This phenomenon plays a key role in many biological situations. It can be beneficial, allowing the exploration of multiple fate choices during development, such as during retinal photoreceptor specification (Urban and Johnston (2018)). But expression noise can also be harmful, as during HIV viral load ebbs (Weinberger et al. (2005)). Moreover, phenotypic heterogeneity has a key role in cancer growth and the emergence of therapeutic resistance (Blanco Calvo et al. (2009), Gupta et al. (2018a)).

We distinguish between two types of noise in gene expression: intrinsic noise and extrinsic noise (Figure 1.12).

Intrinsic noise

Intrinsic noise refers to the inherent stochasticity or randomness within individual cells. It arises from the discrete nature of molecular interactions and processes involved in gene expression. Various sources contribute to intrinsic noise, including the random arrival of transcription factors, the binding and unbinding of regulatory molecules, and the spontaneous nature of biochemical reactions within the cell. Intrinsic noise can result in cell-to-cell variability, causing fluctuations in gene expression levels even among genetically identical cells (highlighted by the different colors of the rectangles representing nuclei in Figure 1.12 below).

Extrinsic noise

Extrinsic noise, on the other hand, arises from external factors that influence gene expression across a population of cells. These external factors can include variations in environmental

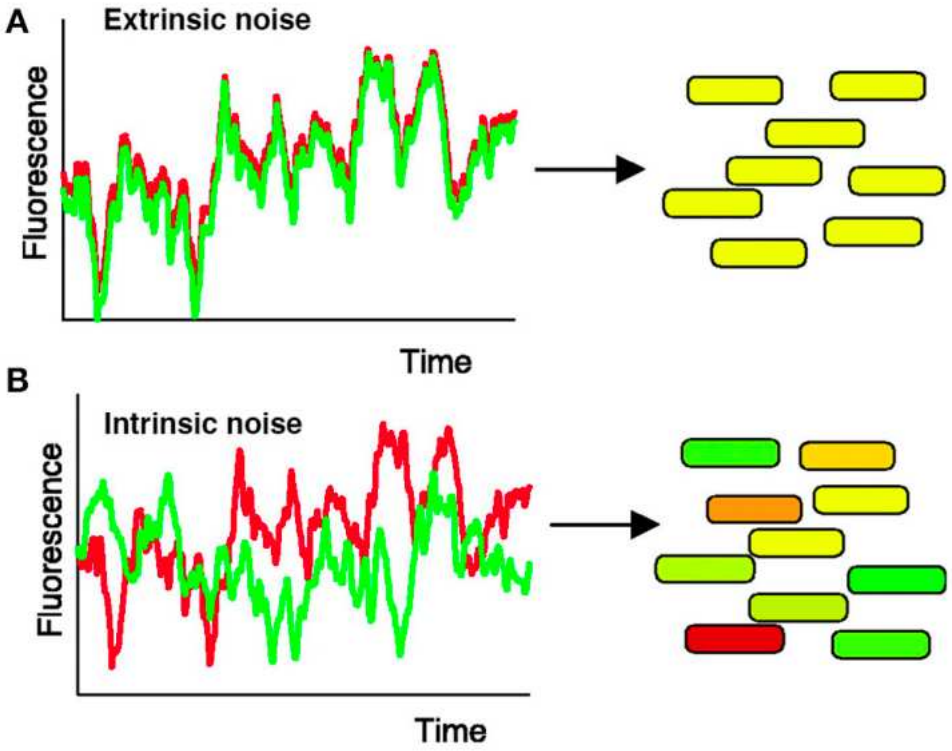


Figure 1.12: Overview of intrinsic and extrinsic noise in gene expression (taken from Meyer and Roeder (2014)).

conditions, fluctuations in available nutrients, or differences in cellular signaling and communication. Extrinsic noise leads to cell-to-cell variability. This variability is coordinated across the population, as highlighted by the same yellow color in Figure 1.12 above. It results in gene expression patterns that exhibit systematic deviations from cell-autonomous stochasticity.

Under this perspective, multiple mathematical frameworks have been developed (Swain et al. (2002); Paulsson (2004); Raser and O'Shea (2004); Meyer and Roeder (2014)). These frameworks are used to formally describe noise in biological systems. They address noise in general or dissect and quantify the contributions of each type of noise. In the simplest model of transcription, mRNA production is depicted as a Poissonian process, where transcripts are stochastically generated at a constant rate. However, this model proves inadequate for many genes. In such cases, the distribution of transcripts is "Super-Poissonian," meaning that the measured variance exceeds the mean (Nicolas et al. (2017); Lenstra et al. (2016)). However it was not until the discovery that transcription happens in burst with the developments of single cell live imaging technologies that major advances in the study of transcription has been made (Munsky et al. (2012), Gregor et al. (2014)).

Throughout this thesis we are interested in understanding and deciphering intrinsic noise through quantification of noise, manipulation of the biological context and prediction of

noise.

1.3.0 Introduction to mathematical modeling of stochastic gene expression

Mathematical modeling plays a crucial role in advancing our understanding of gene expression. Various frameworks have been developed to study distinct facets of gene expression, such as quantifying expression noise (section 1.2.7), deciphering the functions of cis-regulatory elements (section 1.1.2), and understanding the establishment of morphogen gradients (section 1.2.4). Consequently, in this section I will define the mathematical models and properties used throughout the PhD thesis to model the gene expression process from multiple point of view.

1.3.1 Stochastic chemical reaction networks as models of gene expression

A stochastic process is a mathematical model describing the evolution of a system over time, where the outcomes or states at each time point are subject to randomness or uncertainty. Stochastic processes are widely used in physics, engineering, economics, finance, and biology to model phenomena with inherent randomness or uncertainty. Common examples of stochastic processes include Poisson processes, Markov processes, and Brownian motion, among others. They provide valuable tools for analyzing and predicting probabilistic behavior of dynamical systems.

With the establishment of transcriptional "bursts" the need of a stochastic process to model transcription became trivial (section 1.2.5). Therefore we are interested in studying models of stochastic gene expression. These models rely on the more general theory of stochastic chemical reaction networks (CRN).

The stochastic CRN formalism was introduced by Delbrück (1940s), Rényi (1950s), and Bartholomay (1960s).

Delbrück's discussed one-species autocatalytic reaction models Delbrück (1940) and provided the solution of the master equations that describe the probabilistic dynamics of these models.

The Rényi formalism Rényi (1954), while similar to Delbruck's employs similar one species models, but proposes an alternative approach to solve the master equation. Rényi's approach uses probability generating functions to compute the distribution of the number of

copies of the chemical species resulting from the reactions. Like Delbrück, Rényi studies only univariate cases.

Building upon Dëlbruck's and Rényi's groundwork, the Bartholomay formalism Bartholomay (1957) introduces stochastic reaction networks in their full generality, going from the univariate to the multivariate case. His axiomatic description of enzymatic reaction introduces all the ingredients of the modern theory of stochastic chemical reaction networks (CRN) as continuous time jump Markov jump processes: propensity of reactions (probability per unit time that a reaction occurs), joint probabilities of species copy numbers, master equation, generating function for the multivariate case, differential equations for the moments.

More recently, in the 70s, Daniel T. Gillespie proposed an algorithm to simulate stochastic chemical reaction networks. This algorithm bridges the gap between theoretical frameworks and real-world simulations, offering a means to accurately model the evolution of stochastic chemical reaction networks Gillespie (1977). By simulating individual reaction events based on their propensity and stoichiometry, the Gillespie algorithm has become a cornerstone in the field of systems biology, enabling the detailed exploration of complex cellular dynamics influenced by stochastic elements.

One of the first models to use chemical reactions to model gene expression was established in 1997 McAdams and Arkin (1997). Rather generally, any intracellular process can be described at the molecular level in terms of biochemical reactions. In models of stochastic gene expression the synthesis of RNA and proteins has been modelled using stochastic CRNs.

Markov jump processes

As a matter of fact stochastic CRNs belong to a more general class of stochastic processes, the Markov jump processes, defined as follows.

For any continuous time Markov process $\boldsymbol{X}(t) \in E$, E being a metric space, we associate its natural filtration (its past and present) with $\mathcal{F}_t^{\boldsymbol{X}} = \sigma\{\boldsymbol{X}(s), s \leq t\}$ and a transition probability function $\mathbb{P} : \mathbb{R} \times E \times \mathbb{R} \times \mathcal{B}(E) \to [0,1]$ ($\mathcal{B}(E)$ is the Borel set),

$$\mathbb{P}(s,x,t,A) = \mathbb{P}\left[\boldsymbol{X}(t) \in A \mid \mathcal{F}_s^X\right] = \mathbb{P}[\boldsymbol{X}(t) \in A \mid \boldsymbol{X}(s) = x] \tag{1.1}$$

for $t \geq s$. The last equality being the Markov property (the future of a Markov process process depends on the past only through the present). $\mathbb{P}$ satisfies the following properties:

1. $\mathbb{P}[s,x,t,A]$ is measurable in x and represents a probability measure in A.

2. $\mathbb{P}[t, x, t, .] = \delta_x$ (Dirac mass in x).

3. $\mathbb{P}[t_1, x, t_3, A] = \int \mathbb{P}[t_1, x, t_2, dy]\, \mathbb{P}[t_2, y, t_3, A]$ (Chapman-Kolmogorov equation).

Markov Jump processes are defined by the existence of the following limit Gikhman and Skorokhod (1969):

$$\liminf_{t \to s} \frac{\mathbb{P}[s, x, t, A] - \delta_x(A)}{t - s} = \lambda(s, x)\mu(s, x, A) \tag{1.2}$$

for $t \in \mathbb{R}$, $x \in E$, $A \in \mathcal{B}(E)$.

The function $\lambda : \mathbb{R} \times E \to \mathbb{R}_+$ is the intensity (average number of jumps per unit of time). $\mu(s, x, A)$ is a probability measure in A known as the jump law. It represents the probability of choosing to jump from x to A. For homogeneous processes, the intensity and law of jumps do not depend on time s.

When E is separable, any Markov process with jumps is equivalent to a càd-làg process (right continuous with left limits) and there is almost certainly a sequence of random time intervals τ_i of exponential laws such that $lim_{n\to\infty} \sum_{i=1}^{n} \tau_i = \infty$ et $\boldsymbol{X}(t)$ is constant in $\left[\sum_{i=1}^{n} \tau_i, \sum_{i=1}^{n+1} \tau_i\right]$ (Gikhman and Skorokhod (1969)). Thus, a Markov process with jumps can be equivalently defined from a counting process $v(t)$ (of intensity λ) and a Markov chain X_n with transition kernel $\mu : \boldsymbol{X}(t) = \boldsymbol{X}_{\mu(t)}$.

Formalism of the chemical reaction and master equation

Stochastic CRNs are modelled as Markov jump processes (Rényi (1954); Bartholomay (1957)).

Let us consider a system of N chemical species $\{S_1, ..., S_N\}$ that interact within M chemical reactions $\{R_1, ..., R_M\}$. This system is assumed to be well-stirred, confined to a constant volume, and in thermal equilibrium at a constant temperature, but not necessarily in chemical equilibrium. Let $X_i(t)$ denote the copy number of molecules of species S_i in the system at time t. Our aim is to simulate the state vector $\boldsymbol{X}(t) = (X_1(t), ..., X_N(t))$ given the initial state $\boldsymbol{X}(t_0) = \boldsymbol{x}_0$ at some initial time t_0.

The changes in species populations are a direct result of the chemical reactions. Each reaction R_μ is characterized by two essential components. First, its state-change or stoichiometric vector $v_{\boldsymbol{\mu}} = (v_{1\mu}, ..., v_{N\mu})$ defines how one instance of reaction R_μ changes the molecular population of each species S_i. For instance, if the system is in state x, the occurrence of reaction R_μ causes an instantaneous transition to state $\boldsymbol{X} + v_{\boldsymbol{\mu}}$.

The second characteristic is the propensity function a_μ for reaction R_μ. This propensity function is designed such that when multiplied by τ, it gives the probability that a reaction R_μ will change the value of a system variable $\boldsymbol{X}$ within an infinitesimal time interval $(t, t+\tau)$. Rather generally, we can use the following relation to compute the propensity:

$$a_\mu(\boldsymbol{X}) = V h_\mu(\boldsymbol{X}/V) k_\mu \tag{1.3}$$

Here, V represents the volume of the well-stirred reactor, the arbitrary index μ refers to the M reaction, h_μ represents a function of the reactants concentrations $\boldsymbol{X}/V$, and k_μ is the kinetic rate constant of the reaction.

A stochastic CRN can be seen as a Markov jump process. The intensity of the Markov jump process is the total CRN propensity $a_{tot}(\boldsymbol{X}) = \sum_\mu a_\mu(\boldsymbol{X})$, whereas the jump measure is $\sum_\mu \frac{a_\mu(\boldsymbol{X})}{a_{tot}(\boldsymbol{X})} \delta_{\boldsymbol{X}+\nu_\mu}$.

The chemical master equation (CME) formally defines the equation that determines the probability of each species having a specific molecular population at a given time in the future. It is the forward Kolmogorov equation of the Markov jump process, resulting from the Chapman-Kolmogorov equation. For a stochastic CRN the master equation reads Gillespie (1992); McQuarrie (1967):

$$\frac{\partial P\left(\boldsymbol{x}, t \mid \boldsymbol{x}_0, t_0\right)}{\partial t} = \sum_{\mu=1}^{M} \left[a_\mu\left(\boldsymbol{x} - \boldsymbol{v}_\mu\right) P\left(\boldsymbol{x} - \boldsymbol{v}_\mu, t \mid \boldsymbol{x}_0, t_0\right) - a_\mu(\boldsymbol{x}) P\left(\boldsymbol{x}, t \mid \boldsymbol{x}_0, t_0\right)\right]. \tag{1.4}$$

Gillespie algorithm

Throughout this thesis, the Gillespie algorithm is used to simulate various stochastic models. Hence, it is useful to provide a formal introduction to this algorithm.

The Gillespie algorithm, often referred to as the Stochastic Simulation Algorithm (SSA) as well, serves as a means to numerically simulate the dynamic behavior outlined by the master equation. More generally it can be used to simulate continuous time Markov jump processes (last property in section 1.3.1). The mathematical underlying foundation behind this is that the time-to-the-next-jump in a Markov jump process is exponentially distributed and the probability of the next event is proportional to the rate. This was established by Feller (1940).

While there are situations where we can analytically solve the master equation, this approach may not be viable for intricate scenarios, such as when dealing with a multitude

of reactions. Examples of solvable master equations are: the autocatalytic reaction model of Delbrück (Delbrück (1940)), the two steps (transcription and translation) non-regulated, gene expression model (Ham et al. (2020)), the one step regulated gene expression model (Ramos et al. (2011)).

In such cases, the Gillespie algorithm, or some form of simulation, becomes essential.

The underlying concept of the Gillespie algorithm involves simulating a series of Markov processes. This is accomplished by sampling the probability distribution of two quantities: the time elapsed since the last reaction, denoted as τ, and the propensity function a_μ (see equation 1.3).
The intensity function $\lambda(s, x)$ of the formalism of a markov jump process (equation 1.2) corresponds to the propensity function $a_\mu(x)$ in the Gillespie algorithm for reaction μ when the system is in state x. Similarly the jump law $\mu(s, x, A)$ is related to the time until next event τ

In a nutshell, the algorithm is presented in algorithm 1.

Algorithm 1 Gillespie Algorithm

1. At some initial time $t(e.g. t = 0)$ select your initial state n and compute the propensities $a_\mu(\boldsymbol{x})$.
2. while $t < t_{max}$
3. Select two random numbers $r_1 \sim \mathcal{U}[0, 1]$ and $r_2 \sim \mathcal{U}[0, 1]$.
4. Compute τ using the formula
$$\tau = \frac{1}{a_0} ln(\frac{1}{r_1}) \tag{1.5}$$
where $a_0 = \sum_\mu a_\mu(\boldsymbol{x})$.
5. Find the smallest integer j that satisfies
$$\sum_{j'=1}^{j} a_{j'}(\boldsymbol{x}) > r_2 a_0(\boldsymbol{x}), \tag{1.6}$$
and set $j = \mu$.
6. Update the system according to $\boldsymbol{X}(t + \tau) = \boldsymbol{x}(t) + v_\mu$ and set $t = t + \tau$.
7. endwhile.

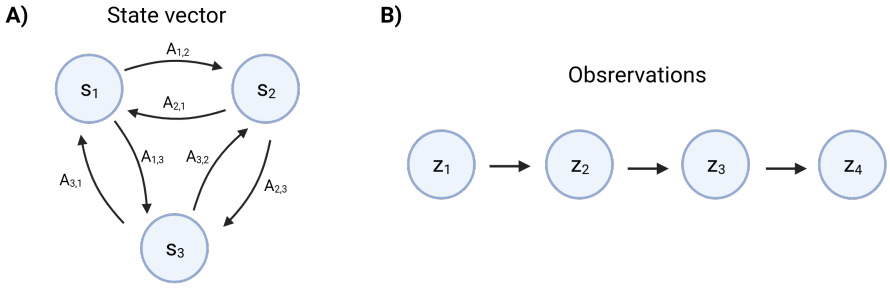


Figure 1.13: Markov model:
A) A simple first-order Markov Chain of state space $\boldsymbol{S}$ with N = 3. B) Observations z_t of the Markov model.

Markov Chain Models

Continuous time Markov chains find extensive utility as modeling tools in a diverse range of fields including biology, medicine, physics, chemistry, economics, and actuarial science Bharucha-Reid (1997). These chains serve as effective means of depicting complex systems characterized by various states and probabilistic transitions among them. Within the framework of Markov chain models, the evolution of states is memory-less, implying that the likelihood of transitions relies only on the present state and not on the past. Moreover, the waiting time to the next state is exponentially distributed.

A Markov chain model is simply defined by three core components: the state space, the transition rate matrix, and the initial state vector. The state vector $\boldsymbol{X}$ is defined as $\boldsymbol{X} = \{X_1, ..., X_N\}$ where N. The transition rate matrix, $\boldsymbol{A} \in \mathrm{R}^{(\mathrm{N}+1)\times(\mathrm{N}+1)}$, is an array of numbers describing the instantaneous rate at which a continuous-time Markov chain transitions between states. The value $A_{i,j} \geq 0$ and the diagonal elements $A_{i,i}$ are defined such that $A_{i,i} = -\sum_{j \neq i} A_{i,j}$ therefore the rows of the matrix sum to zero.

The heart of this modeling involves predicting the future state of a system using its current state and the rates of moving to different states Davis (1993).

Consider a sequence of observations denoted by $\mathrm{z}_1, ..., \mathrm{z}_\mathrm{T}$ providing the states of a system over a period of time. In this context, the notation $\mathrm{z(t)} = \mathrm{z_t}$ signifies that at a particular time instance, denoted as time step t, the system is found to be in a specific state (Figure 1.13 B)). This state is represented by the variable $\boldsymbol{z(t)}$ and can take on values corresponding to the predefined states of the Markov chain, which we've labeled as $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3$, and so on.

The memory-less or first-order Markov Chain property implies that

$$\mathbb{P}[\mathrm{z_t}|\mathrm{z}_1, ..., \mathrm{z}_{\mathrm{t}-1}] = \mathbb{P}[\mathrm{z_t}|\mathrm{z}_{\mathrm{t}-1}] \tag{1.7}$$

Therefore for a series of observations $z_1, ..., z_T$, the product rule can be used to express the joint distribution of the observations as

$$\begin{aligned}\mathbb{P}[z_1, ...z_T] &= \prod_{t=1}^{T} \mathbb{P}[z_t|z_1, ...z_{t-1}] \\ &= \prod_{t=2}^{T} \mathbb{P}[z_t|z_{t-1}]\, \mathbb{P}[z_1] \end{aligned} \tag{1.8}$$

While this approach involves making strong assumptions regarding the temporal relationships within the data, Markov Models have demonstrated effective utility in modeling time series data characterized by straightforward and short-term temporal connections. In the case where we are interested in integrating more extensive temporal relationships into the model, we can introduce conditional dependence of z_t on observations that extend further into the past. For instance, allowing z_t to be influenced by z_{t-1} and z_{t-2} leads to the creation of a Second-Order Markov Chain. In practical terms, this concept can be extended to what's known as an Mth order Markov Chain, where the conditional distribution of a given variable relies on the previous M variables.

However, with the introduction of longer-term temporal dependencies, the complexity of the model increases notably. As the value of M gets bigger, the number of parameters increases very quickly. This makes the method not work well for larger M values because it becomes too difficult to do the computations needed.

An example of a Markov model in the modeling of transcription process is the telegraph model described in the end of section 1.3.1 (also refer to Figure 1.14 left). The model can be extended to incorporate several ON and OFF states, depending on their time scale. Each ON/OFF state is represented as a Markovian state. An example of a transcription process occuring in a 3 state model where there are the promoter switches between two OFF states is shown in Figure 1.14 right). After modeling the transcription process as a Markov model we use the gillespie algorithm (Section 1.3.1) to simulate the time and the next state that the process goes to. Similarly throughout the thesis we will be modeling different transcription/translation steps as Markov chain states and then applying the gillespie algorithm to the Markov chain.

1.3.2 Spatial extension of gene expression modeling

Using stochastic (or deterministic) CRNs as models for gene expressions is possible under the assumption of well-stirred reactor, which means that molecular species are rapidly diffusing and uniformizing their concentration in space.

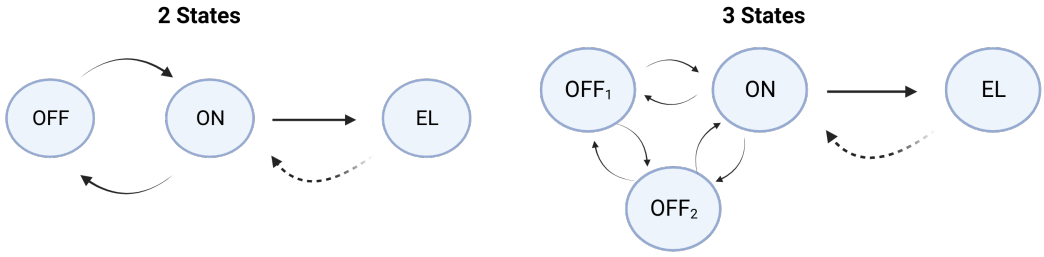


Figure 1.14: Illustration of a transcription process modeled as a Markov model. On the left side, we depict a simplified representation with two states mode (telegraph model): the 'OFF', 'ON' state and the 'Elongation' state. On the right side, a more detailed view includes two non-obligatory states: 'OFF_1', 'OFF_2' and 'ON,' followed by the 'Elongation' state.

This approximation is true for large diffusion coefficients and small compartments. In these thesis we deal with developmental biology models. In developping embryos diffusion can be limited by crowding and various sterical constraints (Section 1.2.2). Compartments are also large, as they can represent full embryos. In such cases, gene expression is no longer homogeneous and forms spatially heterogeneous patterns, that are used to organize the development of the embryo.

Furthermore, stochasticity of gene expression can lead to fluctuations in the spatial distribution of these patterns which could in principle lead to developmental defects. For this reason multiple studies have been dedicated to understand the mechanisms behind the reliable spatial control development in the embryo Dubuis et al. (2013); Tkačik and Gregor (2021); Teimouri and Kolomeisky (2022).

In developmental biology, two primary methodologies were used for investigating gene expression over spatial and temporal dimensions. The first one employs the concept of positional information that hypothesizes that the information needed for the embryo organization is provided by one or several spatially heterogeneous morphogenes. The second uses reaction-diffusion systems and explain patterning by instabilities of these systems that break translation symmetry (Figure 1.15).

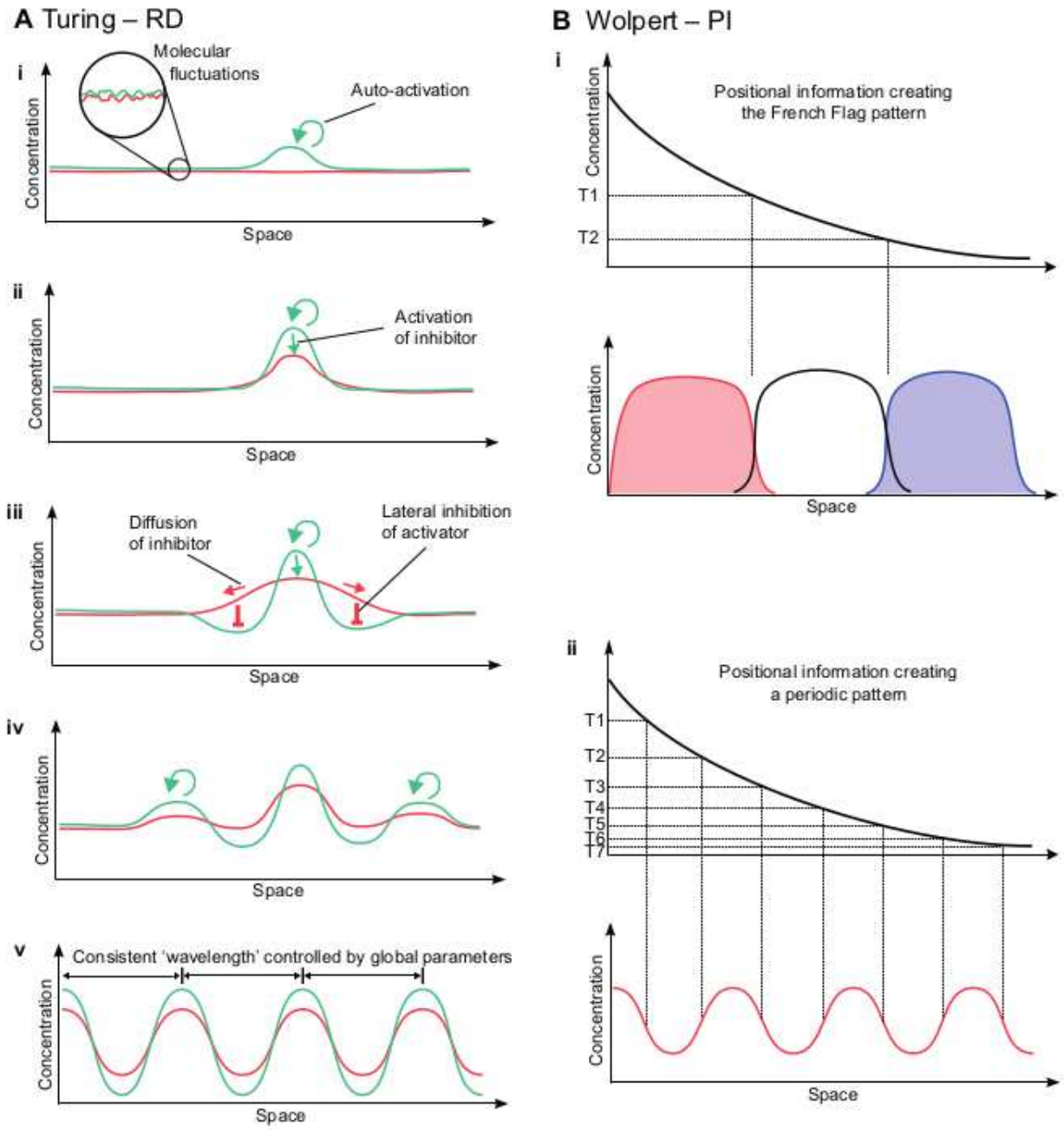


Figure 1.15: Principles of Reaction-Diffusion and Positional Information Systems: Self-Organization and Cell Fate Determination (taken from Green and Sharpe (2015)). (A) There are two broad categories of Turing RD systems: the activator-inhibitor model and the substrate-depletion model. In the former case, the two molecular concentrations make periodic patterns that are in phase with each other. In the latter case, the patterns are out of phase with each other. Here, we illustrate the general self-organising nature of RD systems by reference to the activator-inhibitor model. Even an apparently homogeneous distribution of molecules across space will display molecular fluctuations. Some cells with a slightly higher level of activator will thus auto-enhance these levels, pushing up the concentration (i). Since the activator also enhances production of the inhibitor, levels of inhibitor will also rise at that point (ii).

Figure 1.15: However, the inhibitor can diffuse faster than the activator, which has two consequences: first, at the position of the peak, inhibitor levels fail to accumulate sufficiently to repress the activator, whose positive feedback is able to stabilise its own high levels; second, the increase in inhibitor levels in neighbouring cells prevents levels of the activator from growing, thus creating a zone on either side of the first peak where no new peaks can form (iii). However, beyond these regions of 'lateral inhibition' new peaks can form (iv), so the whole system dynamically changes until a regular array of peaks and valleys is formed across the whole field of cells (v). (B) Wolpert's concept of PI describes a very different process. A prior asymmetry results in a graded monotonic distribution of a variable (usually the concentration of a morphogen), and cells use this distribution to make fate choices. A popular illustration of this concept is the French Flag Problem (i), in which the field of cells must be divided into three equal regions of different cell fates (represented by red, white and blue). It is increasingly believed that small networks of cross-regulating genes constitute the mechanism of morphogen interpretation. However, irrespective of the molecular mechanism, the effective calculation is to define threshold levels of morphogen (T1, T2) and to associate prespecified fates to the different concentration ranges between these thresholds. In principle, any pattern can be defined in this way, including a periodic pattern similar to those produced by RD (ii). However, in this case a large number of different positional values (T1 to T7) would have to be accurately defined, even though they subsequently map to just two fate choices.

Reaction-diffusion system

In the middle of the 20th century, Alan Turing also known for his foundational work in computer science, introduced a model of morphogenesis: the initial symmetry in embryos can be broken by an intercellular diffusion reaction between two classes of molecules which are typically species of chemicals or biological agents: activator and inhibitor. The dynamics of the concentration of these morphogens, X_i, are dictated by the diffusion-reaction equations:

$$\frac{\delta X_i}{\delta t} = g_i(X_1, ..., X_M) + D_i \nabla^2 X_i \tag{1.9}$$

Where g_i are local functions of the full set of M morphogen concentrations and D_i is the diffusion coefficient for the i-th morphogen. The key point of these equations is that the interactions are fully local and posses translation symmetry. These equations describe how the concentration of each chemical species changes over time and space, taking into account the diffusion of the species through the system thus resulting in a pattern (Figure 1.15 A)).

Turing pattern formation arises from the spontaneous emergence of a spatial scale. This phenomenon is known as Turing instability. Turing instability occurs when the rate of diffusion of one species D_X is much greater than the rate of diffusion of another species D_Y. Furthermore, the more diffusive species is an inhibitor, whereas the less diffusive one is an activator. This mechanism, know as latteral inhibition, promotes the formation of spatially periodic patterns Turing (1952). This concept can be applied to a broad class of reaction dif-

fusion models Baurmann et al. (2007); K. Maini et al. (1997); Maini et al. (2006); Miura and Shiota (2000) .

Positional Information

The second most influential idea in morphogensis came from Lewis Wolpert between late 1960s and early 1970. Lewis Wolpert, introduced the idea of what is now know as "positional information" often represented by "The French Flag". The idea behind it is that the positional information is encoded in morphogen gradients, suggesting that there exists a predetermined initial symmetry breaking event for each cell depending on it's location. The cells are then able to "measure" their position within a morphogen gradient by comparing the concentration of the morphogen they receive to a threshold value. Cells can then activate different sets of genes depending on whether their morphogen concentration is above or below the threshold, leading to different cell fates Wolpert (1969, 1971) (Figure 1.15 B)).

These two ideas (reaction diffusion and positional information) are not mutually exclusive and are now used to model space dependent gene expression in developmental biology Green and Sharpe (2015); Gordon et al. (2020).

1.3.3 Modeling choice

To effectively accomplish the goals of modeling gene expression at multiple scales of time and space, it is essential to connect several models operating at various scales. Furthermore, it is crucial to use the appropriate model for each specific scale.

One of our aim is to extend CRN models from well-stirred reactors to spatially extended systems and use them for modelling gene expression across multiple scales.

The modeling choices needed for representing reactions in a spatially extended domain are determined by careful consideration of the following criteria:

1. global description vs local description.

    - Global description dictates spatially homogeneous, “well-stirred” case excluding diffusion.

    - Local description is the spatially heterogeneous, “spatial model” including diffusion.

2. deterministic description vs stochastic description.

    - Deterministic description usually concerns macroscopic scales and uses concentration variables.

    - Stochastic description is usually used for mesoscopic scale, taking into account fluctuations, at the molecular level.

3. one-scale vs multiple scales

    - one-scale description is characterized by a single, large population size scale, featuring solely high reaction rates and fast dynamics.

    - multiple scales which is distinguished by the presence of at least two population size scales, encompassing both high and low reaction rates, along with fast and slow dynamics.

The combination of the above criteria gives rise to six types of models:

(M1) Deterministic Homogeneous Model

(M2) Deterministic Spatial Model,

(M3) Stochastic Homogeneous Model,

(M4) Stochastic Spatial Model,

(M5) Multiscale Stochastic Homogeneous Model,

(M6) Multiscale Stochastic Spatial Model.

These models are not independent one from another.

(M3) , (M1): The relation between (M1) and (M3) has been extensively explored, particularly by Kurtz, as documented in various works including Kurtz (1970, 1971), and in collaboration with Ethier as shown in Ethier and Kurtz (1986). This investigation has yielded a **law of large numbers (LLN)** and a corresponding central **limit theorem (CTL)**, demonstrating the

convergence of (M3) toward (M1).

{(M1), (M2)}, {(M3), (M4)}: The **consistency** between (M1) and (M2), as well as between (M3) and (M4), has been established in Arnold (1981).

(M2), (M4): In their work, Arnold and Theodosopulu conducted a comparison of (M2) and (M4) using the L^2 **norm**, employing **a Law of Large Numbers (LLN)**. Subsequently, Blount, following in the footsteps of Kotelenez, extended this comparison extensively. They demonstrated various LLNs and associated **Central Limit Theorems (CLTs)** under progressively relaxed assumptions, including spaces of distributions. Blount's research even included a **LLN in the supremum norm**. Relevant references can be found in Kotelenez (1987, 1986, 1988), among others.

(M5), PDMP: In their research documented in Radulescu et al. (2007); Crudu et al. (2009, 2012), Crudu, Debussche, Muller, and Radulescu investigated (M5). In the latter publication, they adopted a modeling approach that incorporated the system's multiscale nature in the spatial homogeneous framework. They successfully demonstrated that the multiscale model exhibits **(weak) convergence** to a finite-dimensional Piecewise Deterministic Markov Process (PDMP). In finite dimensions, PDMPs are hybrid processes that follow ODE flows between consecutive jumps, with parameters that can undergo jumps. These processes have been thoroughly formalized and examined by Davis, as detailed in Davis (1993). Depending on the nature of interactions and scaling factors, Crudu et al. (2009) discerned various types of limiting PDMPs. It's worth noting that in hybrid simplification, stochasticity does not appear. However using multiscaling we can estimate asymptotically at the first order, the noise lost in the Law of Large Numbers.

(M6), (M2), (M1) In Debussche and Nguepedja Nankep (2019), they prove a new **law of large numbers** in the spatially heterogeneous framework, showing the convergence of (M6) to (M2) coupled with (M1), in the **supremum norm**.

Nonetheless, the resulting integrated models are frequently demanding in terms of computational resources and pose challenges in computing numerical solution. This is a common attribute of multi-scale problems, highlighting the need for appropriate multiscale algorithms. Examples of multi-scale numerical methods are demonstrated in (Dada and Mendes (2011); Smith and Yates (2018); Hepp et al. (2015)).

In Chapter 5, we compared numerically different modeling methods for gene expression, using both stochastic and deterministic approach taking into consideration local description

of the spatial scale. This comparison is done between Model (M2), (M4) and a third hybrid techniques that combine $\{(M3)and(M2)\}$

1.3.4 Non-Markovian models

When applied in real-world scenarios, it's important to note that not all states and transitions within a system are readily observable. This leads to the concept of an incompletely observed Markov chain, which can exhibit memory effects.

For instance, consider the context of transcription processes. These can be conceptualized as continuous-time Markov chains, where a promoter sequence stochastically triggers various non-productive states (OFF) before eventually transitioning to an active state (ON) capable of initiating transcription. However, our access to data is primarily limited to the transcribing state itself rather than the intricate transitions of the promoter molecule. As a result, the waiting times between observable states is **no longer exponentially** distributed. This deviation from the standard Markovian behavior necessitates a shift from using purely Markovian models to incorporating non-Markovian models.

Several strategies have been developed to tackle this complex phenomenon. Notable approaches involve the utilization of **hidden Markov processes** as demonstrated in works such as Lammers et al. (2020a); Bowles and Rattray (2021); Tantale et al. (2021); Douaihy et al. (2023).

Hidden Markov Models

The hidden Markov model (HMM) belongs to a category of doubly stochastic processes Rabiner and Juang (1986). These processes exhibit the Markov property and output independence. In a HMM, there is an underlying Markov process that remains hidden. This means that one cannot directly see the variable states, but can deduce them through a different set of stochastic processes. These processes become evident as a sequence of observed outputs. The Markov process generates the sequence of variable states. This is determined by the initial state probabilities and the probabilities of transitioning between states. On the other hand, the observation process produces measurable signals. These signals are determined by a state-dependent probability distribution. Essentially, this observation process can be thought of as a noisy version of a Markov process, providing insights into the hidden states.

Formally, an HMM constitutes a Markov model where we possess a sequence of observed outputs $\boldsymbol{X} = \{X_1, X_2, ..., X_T\}$ derived from an output alphabet $\boldsymbol{V} = \{v_1, v_2, ..., b_{N,}\}$, i.e. $X_t \in V$, $t = 1, ...T$. We also assume the presence of a sequence of states $\boldsymbol{z} = \{z_1, z_2, ..., z_T\}$ selected from a state alphabet $S = \{s_1, s_2, ...s_N\}$, $z_t \in S$, $t = 1, ..., T$. The vector state $\boldsymbol{z}$

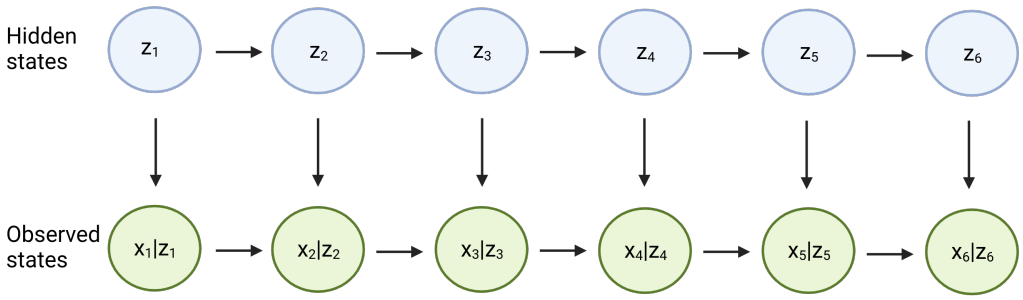


Figure 1.16: Illustration of Hidden markov Model.

follows a Markov model introduced in section 1.3.1, the actual values of these states remain unseen. The shift from one state, i, to another state, j, is once again denoted by the corresponding entry in our state transition matrix $A_{i,j}$ (Figure 1.16). We formulate the likelihood of generating an observed output by considering the influence of our hidden state. To achieve this, we operate under the assumption of output independence and define $\mathbb{P}\left[X_t = v_k | z_t = s_j\right] = \mathbb{P}[X_t = v_k | X_1, ..., X_T, z_1, ..., z_T] = B_{j,k}$. The matrix $\boldsymbol{B}$ contains the information about the likelihood that our hidden state produces the output v_k given that the state at the corresponding time was s_j.

Although in this thesis we don't use HMM explicitly but in Chapter 2 we compare a method that we have developed to extract information from transcription to another method based on HMM Lammers et al. (2020a); Bowles and Rattray (2021).

1.4.0 Thesis objectives

When examining the molecular level, using single cell imaging techniques, phenotypic heterogeneity emerges and gives rise to the phenomenon known as biological "noise". This noise has been recognized to originate from various sources Raser and O'Shea (2005). One contributor to this intricacy is the phenomenon of "transcriptional bursts". This dynamic behavior results in fluctuations in gene expression levels within an otherwise deterministic framework. Although the transcription cycle consists of 3 main phases: initiation, elongation and termination, variability in transcription outcomes often traces back to the initiation phase, underscoring its significance in gene expression regulation.

In this thesis we combine theoretical frameworks, computational techniques, and single-cell live imaging analyses. The primary goal is to expand our comprehension concerning the stochastic nature of gene expression, particularly at the transcriptional level, and to resolve the apparent contradiction between stochasticity and robustness of gene expression mecha-

nisms.

We aim to answer this question using the *Drosophila melanogaster* embryonic development. We have used this model organism for these three main reasons: genome has been fully mapped, resulting in a comprehensive understanding of its enhancers and promoters, the development of live imaging techniques that enables us to track transcription and translation in real time and the reliable, reproducible pattern.

However different criteria of transcriptional data that result in translation give rise to the need of unique modeling approaches. We distinguish two main criteria that helped us know our limitations in extracting direct information from the data:

1. 'Time homogeneous' vs 'time-inhomogeneous' signal.

2. 'Space homogeneous' vs space 'in-homogeneous'.

Therefore we divide our problem into three main conditions: time and space homogeneous data, time-inhomogeneous and space homogeneous data, and lastly time and space inhomogeneous. We start by examining a restricted case of time and space homogeneity. This simple assumption serves as a starting point, allowing us to extract information from transcription data directly. However, as we introduce more complex assumptions, we observe a shift in our approach. As complexity increases, we extract fewer specific details from the data itself and we start extracting more theoretical conclusions. This transition reflects the trade-off between the richness of experimental information and the depth of theoretical insights as we go through the complexities of gene expression modeling.

This PhD thesis is structured as follows.

In Chapter 2, we introduce BurstDECONV, an innovative statistical inference method designed to deconvolve signal traces into individual transcription initiation events which we then apply the solution of the inverse problem to obtain the switching parameters of the transcriptional data for different phenotypes. A thorough parameter benchmarking for this inference method is presented, along with a comparison against alternative methods using both synthetic and real-world data. The data is acquired through single-cell live imaging (see Section 1.2.6), wherein the intensity signals are calibrated using smFISH Tantale et al. (2021); Pimmett et al. (2021).

In 3, we developed the inverse problem used in 2 to deal with more complicated models.

Solving the inverse problem consisting of obtaining the kinetic rates of the markov Process from the parameters of the survival function consisting of the waiting time between polymerase. By analyzing the distribution of waiting times between successive polymerase initiation events and using the method described, we are able to deduce mechanistic characteristics of transcription, including the number of rate-limiting steps and their kinetics.

Both of these chapters require a time and space-homogeneous data.

When dealing with in-homogeneous time series we can no longer use the inverse problem. An example for in-homogeneity in transcription can result from the presence of a repressor Lagha et al. (2013). In this case the switching parameters of the Markovian model depend on the concentration of the repressor. Therefore, in Chapter 4 we have developed an additional method to analyze non-homogeneous data. This method can partition the in-homogeneous time signal into homogeneous signal using a Bayesian approach. By simplifying the complexity of our problem we are able then to apply BurstDeconv to the segment part of the signal that has constant parameters.

Finally in Chapter 5 we extended our investigation to include broader perspectives: temporal and spatial dimensions. As this problem is quite complex we can not obtain information directly from the data. Therefore, in this chapter, we constructed numerical methods that recapitulate experimental observations (transcriptional data) while it is also efficient in terms of computational time. This is achieved by employing a hybrid modeling framework that involves discrete Markov processes and deterministic approach (partial differential equations PDEs). Through a rigorous comparison of these models with real-world data, we were able to uncover critical bottlenecks in the expression of the *snail*, a crucial gene in the development of the drosophila embryo.

CHAPTER 2

Inference of bursting kinetics

2.1.0 Introduction

In this paper we employ machine learning techniques to extract features of single-cell transcription activity from MS2 data (Section 1.2.6) in the case of time and space homogeneous data. As mentioned in the introduction we assume that the key steps of transcription initiation process is modeled as Markov chain (Section 1.3.1). The assume of time homogeneous data is when the switching rate between these different states are constant.

The machine learning process yields three distinct outcomes.

Firstly, utilizing a deconvolution approach and high-resolution movies, we generate a time map detailing transcription events. This map indicates, for each cell, the precise moments when various RNA polymerase molecules initiate mRNA production. This direct transcriptional event readout within a cell population is a unique aspect of our methodology, distinguishing it from other approaches that directly fit specific transcription models to MS2 data, such as those relying on autocorrelation functions Coulon and Larson (2016); Ferguson and Larson (2013); Desponds et al. (2016), maximum likelihood estimation Corrigan et al. (2016), or Bayesian inference Rodriguez et al. (2019); Lammers et al. (2020a). This map enables di-

rect characterization of transcription features, including polymerase convoys and various inter-event time statistics.

Secondly, our approach yields a multiscale cumulative distribution function for the waiting times between consecutive transcription events (or its complement, the survival function). We provide both non-parametric Kaplan-Meier and parametric multi-exponential estimates of this multiscale distribution function. This distribution, derived from a combination of short, high-resolution and long, low-resolution movies, covers timescales ranging from seconds to 10 hours. For drosophila embryo we use short movies, however for human cells we use a combination of short plus long movies. The dynamic range of our method surpasses that of other existing methods, which often rely on smaller sampling rates and/or shorter movie duration. This waiting time distribution is model-independent but can be used to identify various transcription dynamics models.

Thirdly, our method delivers model parameter identification simultaneously for multiple transcription models that equally fit the data. Our primary focus is on discrete transcription models, which rely on Markovian transitions among hidden promoter states. These models vary in terms of state numbers and transition graph topologies. However, it it is important to note that this approach, described mathematically in the paper chapter, can be applied to very general models. In contrast to other methods, our approach stands out by streamlining the process. It doesn't demand separate fitting procedures for distinct models. Instead, it achieves the simultaneous identification of a diverse range of models that are all compatible with the data. Furthermore, it accomplishes this with a single parametric fit of the multiscale waiting time distribution function.

2.2.0 BurstDeconv

BurstDeconv is a tool for identification of transcription mechanism models from live transcription imaging data. The implementation and benchmarking of this tool are described in a paper published in Nucleic Acids Research (NAR) Douaihy et al. (2023).

This work is collaborative and interdisciplinary, encompassing mathematical modeling of the transcription process, live imaging techniques, CIS regulatory elements, image analysis, data calibration. Thanks to the collaborative environment fostered between the IGMM and IGH (for the biological aspect) and LPHI (for the mathematical modeling aspect), I had the opportunity to participate in various aspects of the project while also learning from my collaborators.

Upon joining Professor Radulescu's laboratory, a prototype Matlab pipeline had already been developed and applied to different biological contexts, ranging from the Drosophila Embryo (Pimmett et al. (2021)) to human cells (Tantale et al. (2021)). I implemented the version specifically tailored for short movie analysis in Python. Additionally, as these pipelines can always be improved, I worked on amelioring the local optimization method. However, the primary emphasis of the paper was on evaluating the robustness of BurstDeconv against errors and suboptimal experimental designs.

Initially, I tested the pipeline using artificial data generated through the Gillespie algorithm (Gillespie (1977)) for both 2-state and 3-state Markovian models, where the 3-state model included obligatory pause states and non-obligatory pause states. Subsequently, I assessed the pipeline's performance under changes in calibration (Section 1.2.6), highlighting the significance of accurate calibration. I also conducted benchmarking experiments related to time resolution. These experiments involved examining the trade-off between higher time resolution, which leads to longer movie duration (increasing error in most cases but decreasing error when inactive transcription states are prolonged), and lower time resolution, which results in shorter movie durations with the potential for reduced error. Crucially, I benchmarked the pipeline against methods based on Hidden Markov Models (HMM) Lammers et al. (2020a,b); Bowles and Rattray (2021), which necessitated an extensive understanding of those methods to implement the comparison.

Regarding figure construction, I designed and constructed figures 5, 6, and 7. Figure 1, 2, 9, and 10 were collaboratively worked on by Rachel Topno and myself.

https://doi.org/10.1093/nar/gkad629

BurstDECONV: a signal deconvolution method to uncover mechanisms of transcriptional bursting in live cells

Maria Douaihy[1,2,†], Rachel Topno[1,3,†], Mounia Lagha[2], Edouard Bertrand[3,*] and Ovidiu Radulescu[1,*]

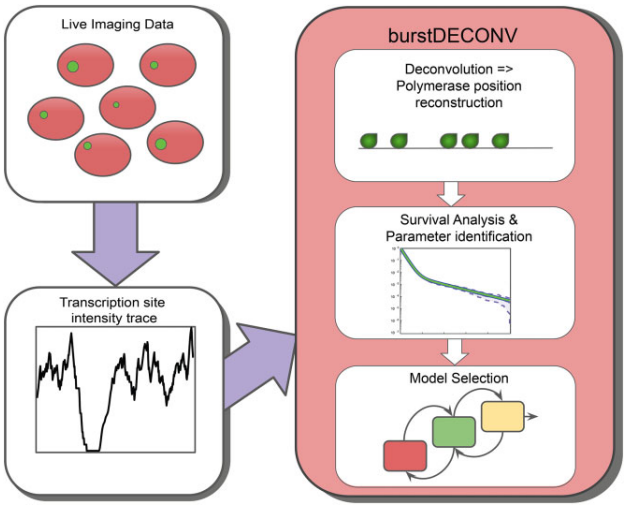
[1]LPHI, University of Montpellier and CNRS, Place Eugène Bataillon, Montpellier 34095, France, [2]IGMM, University of Montpellier and CNRS, 1919 Rte de Mende, Montpellier 34090, France and [3]IGH, University of Montpellier and CNRS, 141 Rue de la Cardonille, Montpellier 34094, France

Received January 25, 2023; Revised June 23, 2023; Editorial Decision July 08, 2023; Accepted July 17, 2023

ABSTRACT

Monitoring transcription in living cells gives access to the dynamics of this complex fundamental process. It reveals that transcription is discontinuous, whereby active periods (bursts) are separated by one or several types of inactive periods of distinct lifetimes. However, decoding temporal fluctuations arising from live imaging and inferring the distinct transcriptional steps eliciting them is a challenge. We present BurstDECONV, a novel statistical inference method that deconvolves signal traces into individual transcription initiation events. We use the distribution of waiting times between successive polymerase initiation events to identify mechanistic features of transcription such as the number of rate-limiting steps and their kinetics. Comparison of our method to alternative methods emphasizes its advantages in terms of precision and flexibility. Unique features such as the direct determination of the number of promoter states and the simultaneous analysis of several potential transcription models make BurstDECONV an ideal analytic framework for live cell transcription imaging experiments. Using simulated realistic data, we found that our method is robust with regards to noise or suboptimal experimental designs. To show its generality, we applied it to different biological contexts such as Drosophila embryos or human cells.

GRAPHICAL ABSTRACT



INTRODUCTION

The observation of transcription in live cells using methods such as MS2/MCP system (1,2) revealed that in most prokaryotic and eukaryotic cells, transcription is discontinuous and undergoes alternative periods of activity and inactivity, governed by stochastic laws. This phenomenon was called transcriptional bursting (3–8). The underlying mechanisms are complex because, even at the steady state, promoters can adopt multiple active and inactive states with distinct timescales and transition schemes, which modulate the variability of expression levels in single cells in non-trivial ways (9–12). Hence, it is necessary to infer these states and timescales from observations. The results of such inference are important as they provide insights into the

*To whom correspondence should be addressed. Tel: +33 4 6714 9221; Email: ovidiu.radulescu@umontpellier.fr
Correspondence may also be addressed to Edouard Bertrand. Tel: + 33 4 3435 9921; Email: edouard.bertrand@igh.cnrs.fr
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

molecular mechanisms underlying promoter dynamics and transcriptional regulation.

Transcriptional bursting with multiple active and inactive promoter states can be modeled using Finite-State Markov Models (FSMM) defined by a set of promoter states and by the rates of stochastic transitions between these states (13). The simplest FSMM, the random telegraph model, has two states and explains the alternation of transcriptionally active and inactive periods observed in transcriptional outputs (14). Commonly used to describe the bursting of prokaryote and less complex eukaryote promoters (6,15,16), this model fails to explain more complex eukaryote transcription mechanisms, as we and others have recently shown using live imaging data of human cell lines and *Drosophila* embryos (17–20). In this case, bursting models involving more than two states are required (21,22).

We must emphasize that the identification of models and rates describing the observed transcription dynamics is not merely a phenomenological description. Indeed, this gives direct access to key regulatory mechanisms at the molecular level. A variety of perturbation experiments have indicated that the states in the kinetic models correspond to well defined biochemical states of the promoter, and specific chromatin features and binding profiles of given transcription factors (e.g. assembled pre-initiation complex PIC, or TATA Binding Protein-bound, or nucleosome occupied promoter; (7,16,23–25)). Furthermore, recent advances in cryo-electron microscopy, as well as single molecule genomic methods (26) have revealed that promoters can be found in a multitude of molecular states as they undergo transcription initiation or early elongation (27–31). However, it is often difficult to figure out from molecular experiments which state is rate-limiting, and this is a key question as the rate-limiting steps are likely points of regulation. Live cell transcription imaging fills this gap, and robust methods to infer promoter dynamics from such data are thus essential for understanding the basic mechanisms of transcriptional control.

In order to decode single cell transcriptional traces, we developed BurstDECONV, a deconvolution based method for reconstructing FSMMs from live transcription imaging using RNA tagging. An overview of this method is presented in Figure 1. BurstDECONV first decomposes single cell MS2/MCP live imaging data into individual transcription initiation temporal events (Figure 1 C). This information is model agnostic and represents a comprehensive spatio-temporal map of transcription that can be used for multiple studies: identifying multiple temporal and spatial scales and kinetic parameters, testing the synchronicity or the correlation of transcription sites, detecting extrinsic noise events, and performing model selection and inference (19,20). In a second step, BurstDECONV computes the survival function characterizing the distribution of waiting times between successive polymerase initiation events (Figure 1D). Finally, multiexponential parametric survival models are inferred and mapped to FSMM kinetic promoter models. The number of exponentials required to fit the survival functions corresponds to the number of promoter states in the model, and this facilitates model comparison and selection (Figures 1D and 3). BurstDECONV has also been successfully applied to extracting transition rate parameters from real data (19,20) in human cells and *Drosophila* embryos. Importantly, this method revealed an alternative model of promoter pausing, described as facultative pausing, which could not be characterized by other live- or fixed-sample approaches.

We have performed a comparative benchmarking in which BurstDECONV was tested along with autocorrelation (32,33) and Hidden Markov Model (HMM) (34–36) methods, two other approaches previously employed for analysing transcriptional bursting data. Whenever comparison was possible, we found that the parameter reconstruction by BurstDECONV is significantly more accurate than by all other methods. Moreover, our method is precise for wide ranges of values of kinetic parameters of transcription processes. By combining short and long movies, we are able to quantify processes with timescales from seconds to days. This extremely wide dynamic range was not accessible with the previous quantitative live cell transcription imaging approaches.

Thus BurstDECONV proves to be a very effective tool for analysing live cell transcription, paving the way to exciting discoveries in the field of transcriptional control. For a wide usage, we provide Matlab and Python implementations of our method, and a user-friendly graphical interface that fits data to a variety of two and three state promoter models.

MATERIALS AND METHODS

Short, high resolution movie deconvolution

The MS2 signal from one transcription site is modeled as:

$$x(t) = \sum_{i=1}^{N_{\text{pol}}} x_{\text{pol}}(t - t_i), \tag{1}$$

where x_{pol} is the signal from one polymerase and t_i are the successive initiation times.

The initiation times are discretised $t_i = n_i\delta, n_i \in \mathbb{N}, 1 \leq i \leq N_{\text{pol}}$, where $\delta = D_{min}/V_{\text{pol}}$, with D_{min} a minimal inter-polymerase distance (in bp) and V_{pol} the polymerase speed (in bp/s) that we assume constant. The entire sequence of initiation times is then coded as a fixed size binary string $B = (b_1, \ldots, b_{N_{max}}), b_{n_i} = 1, b_{j\neq n_i} = 0, 1 \leq i \leq N_{\text{pol}}\}$, where $N_{max} = T/\delta$, T is the movie length.

If $x_{cal}(t)$ is the observed signal, calibrated in polymerase numbers, we find B and thus t_i by least-squares regression using a genetic algorithm GA and the objective function:

$$\mathcal{O}_1(B) = \sum_{k=1}^{N_{frames}} (x(k\Delta; B) - x_{cal}(k\Delta))^2, \tag{2}$$

where Δ is the movie time resolution and N_{frames} is the number of frames, $T = N_{frames}\Delta$.

The GA optimization follows four steps: estimating the amount of polymerases, generating an initial population, applying the genetic algorithm and the final local optimization. We estimate the number of polymerases N_{pol} as the ratio of integral intensities of the experimental signal and of the single polymerase signal. The resulting rough estimation is used to accelerate next steps. Then we prepare an

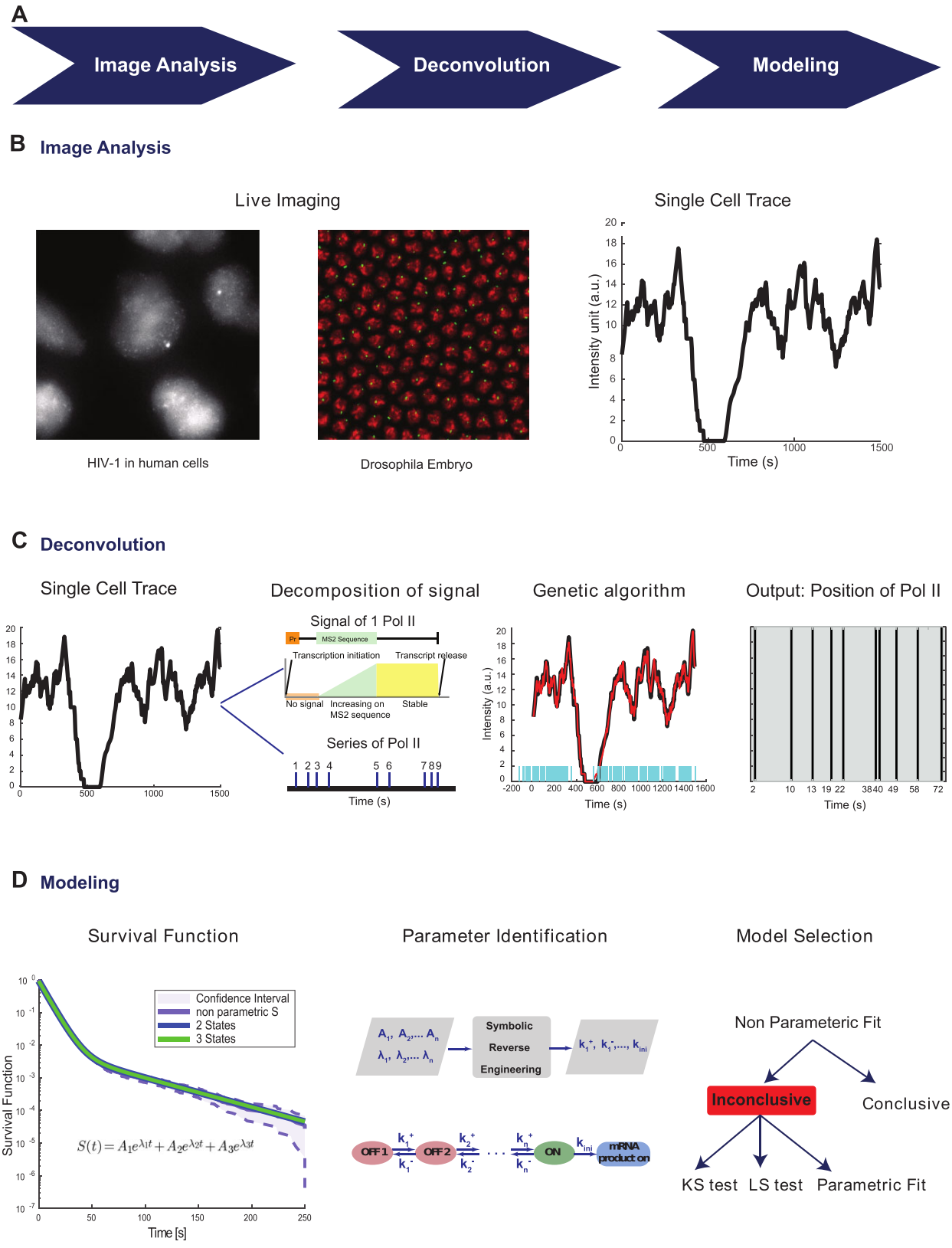


Figure 1. Overview of the live cell transcription imaging pipeline. (**A**) Workflow of the pipeline. (**B**) Movies are segmented to extract single cell signals. (**C**) For each single cell we compute the sequence of polymerase positions. (**D**) Single cell data is used to compute the survival function and identify parameters of transcriptional bursting models.

initial population of polymerase positions. Starting with a binary string B with N_{max} '0's, we randomly pick N_{pol} positions and change them into '1's. After the preparation of the initial population, we use the genetic algorithm (MATLAB built-in function ga or a modified Python function pygad.GA, depending on the implementation) to optimize the objective function. At each step, the genetic algorithm solver randomly selects a sub-population of parental individuals from which it produces the next generation by recombination, crossover and mutation. Over successive generations, the population keeps the best generated solutions and 'evolves' towards an optimal solution. The local optimization further decreases the objective function by displacing the polymerase positions a few steps to the right or to the left.

After optimization, the residuals $x(k\Delta;\ B_{optimal}) - x_{cal}(k\Delta)$ for all the transcription sites in the same movie are used for estimating the noise in the signal. We systematically find that noise is heteroscedastic with a variance depending non-linearly on the signal amplitude. We use cubic polynomial regression to approximate the dependence of the noise variance on the signal:

$$\sigma^2 = b_3 x^3 + b_2 x^2 + b_1 x + b_0. \tag{3}$$

The waiting times $\tau_i = t_{i+1} - t_i$, defined as intervals between successive initiation events coming from all the transcription sites in the movie, are considered as realizations of the same random variable τ. The survival function is defined as

$$S(t) = \mathbb{P}[\tau > t], \tag{4}$$

and estimated (non-parametrically) using the Kaplan-Meyer method (37) from the pooled series coming from all the transcription sites in the same movie.

Space dependent analysis can also be performed, by pooling the transcription sites region-wise (a prior spatial segmentation is needed).

We model the survival function using the multi-exponential family

$$S(t;\ \boldsymbol{A}, \boldsymbol{\lambda}) = \sum_{i=1}^{n_{\exp}} A_i \exp(\lambda_i t), \tag{5}$$

where $\sum_{i=1}^{n_{\exp}} A_i = 1$, $\lambda_i < 0$, $1 \leq i \leq n_{\exp}$.

The parametric estimate of the survival function is obtained by least square regression with an objective function that combines linear and logarithmic scales:

$$\mathcal{O}_2(\boldsymbol{A}, \boldsymbol{\lambda}) = \tfrac{\alpha}{n} \textstyle\sum_{i=1}^{n} (S(t_i;\ \boldsymbol{A}, \boldsymbol{\lambda}) - S_{KM}(t_i))^2 + \\ + \tfrac{1-\alpha}{n} \textstyle\sum_{i=1}^{n} (\log(S(t_i;\ \boldsymbol{A}, \boldsymbol{\lambda})) - \log(S_{KM}(t_i)))^2 \tag{6}$$

where $S(t)$ is defined by (4), $S_{KM}(t)$ is the non-parametric estimate of the survival function, $0 \leq \alpha \leq 1$ is a weight representing the relative importance of the linear scale compared to the logarithmic scale in the estimate of the survival function.

The use of linear and logarithmic scales was motivated by the fact that short timescales responsible of the large initial drop in the survival function are well captured by the linear scale, whereas longer timescales responsible for the smaller decrease in the tail of the survival function are well captured by the logarithmic scale.

The multi-exponential least-squares regression is performed for several values of the number of exponentials $n_{\exp}$. The selection of the number of exponentials is based on three criteria: the optimal value of $\mathcal{O}_2$, the Kolmogorov–Smirnov test using the optimal $S(t_i;\ \boldsymbol{A}, \boldsymbol{\lambda})$ as reference distribution and the uncertainty of the parameters $\boldsymbol{A}, \boldsymbol{\lambda}$ obtained by considering optimal and close to optimal solutions (Figure 5).

Combining two movies

The second version of the method uses two movies. The short high resolution movie is processed exactly as in the first method, resulting in the survival function $S_1(t)$. The transcription site signals from the long low resolution movie are thresholded. The sub-threshold intervals are used to estimate a survival function $S_2(t)$. Given that $S_1(t)$ misses waiting times longer than the short movie length T and that $S_2(t)$ misses waiting times shorter than T_{min} (estimated as the sum of the long movie resolution and the single polymerase signal duration), an interpretation of these two survival functions in terms of conditional probabilities is appropriate:

$$S_1(t) = \mathbb{P}[\tau > t \mid \tau < T], \\ S_2(t) = \mathbb{P}[\tau > t \mid \tau > T_{min}]. \tag{7}$$

Using the total probability theorem we obtain the multiple time scale survival function

$$S(t) = \begin{cases} (1 - p_s) S_1(t) + p_s, & t < T \\ p_l S_2(t), & t > T_{min} \end{cases} \tag{8}$$

where $p_s = \mathbb{P}[\tau < T]$ and $p_l = \mathbb{P}[\tau > T_{min}]$.

p_l is estimated using the formula (see (19)):

$$p_l = \frac{N_{inactive}}{N_{inactive} + N_{active}} = \\ = \frac{N_{inactive}}{N_{inactive} + \frac{P_{active}(1 - S(T_{min}))}{-T_{min} S(T_{min}) + \int_0^{T_{min}} S(u) du}}, \tag{9}$$

where $N_{inactive}$ is the number sub-threshold intervals (resolved and countable), N_{active} is the number of waiting times inside over-threshold intervals (not resolved), P_{active} is the probability to be over threshold (estimated as the time fraction from total that is over threshold) in the long movie signals; for this estimate we use $S(t) \approx S_1(t)$ for $t < T_{min}$.

p_s is optimised to minimize the gap between the short time and long time survival function branches in (8). The estimate of the gap uses interpolation and is possible only if there is an overlap between $S_1(t)$ and $S_2(t)$.

The multi-exponential parametric estimate of the survival function is now performed using the multiple time scale survival function (8).

Rate parameter identifiability

Both versions (short movie and short-long movie) of our method end with the identification of the FSMM rate parameters. This identification is possible symbolically,

using analytical formulas that relate the multi-exponential parameters $\boldsymbol{A}, \boldsymbol{\lambda}$ to the rate parameters.

For the sake of completeness we introduce, in the simplified case of the random telegraph model, the mathematical objects needed for solving this problem. Some solutions for FSMM with 2, 3, and 4 states can be found in (19). The algorithmic solution for an arbitrary number of states will be provided in a separate publication.

Let us denote by 1 and 2 the states OFF and ON of the random telegraph model, respectively. In order to study transcription initiation we add to the model a third state 3 representing the initiation event. The extended three states FSMM is defined by the transition rate matrix $\boldsymbol{Q}$ whose elements are the transition rates between the states of this model. For instance, the matrix element Q_{12} represents the transition rate from OFF to ON, which is k^+. Furthermore, we are interested in the waiting time to initiation, so we decide to stop the FSMM whenever we reach the state 3, which means that all the elements on the last row of $\boldsymbol{Q}$ are zero. The elements of the transition rate matrix sum to zero on any row, therefore

$$Q = \begin{pmatrix} -k^+ & k^+ & 0 \\ k^- & -(k^- + k_{ini}) & k_{ini} \\ 0 & 0 & 0 \end{pmatrix}$$

The vector

$$\boldsymbol{X} = \begin{pmatrix} \mathbb{P}[M(t) = 1 \mid M(0) = 2] \\ \mathbb{P}[M(t) = 2 \mid M(0) = 2] \\ \mathbb{P}[M(t) = 3 \mid M(0) = 2]) \end{pmatrix},$$

where $M(t)$ is the state of the FSMM at the time t satisfies the master equation:

$$\frac{d\boldsymbol{X}}{dt} = \boldsymbol{Q}^T \boldsymbol{X}, \ \boldsymbol{X}(0) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \tag{10}$$

where $\boldsymbol{Q}^T$ stands for the transpose of $\boldsymbol{Q}$.

Eq. (10) is equivalent to

$$\begin{pmatrix} \dot{X}_1 \\ \dot{X}_2 \end{pmatrix} = \tilde{\boldsymbol{Q}} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \tag{11}$$

$$\dot{X}_3 = k_{ini} X_2, \tag{12}$$

where

$$\tilde{\boldsymbol{Q}} = \begin{pmatrix} -k^+ & k^- \\ k^+ & -(k^- + k_{ini}) \end{pmatrix}.$$

The waiting time w between successive initiation events represents the first return time in the state 3 after starting in the state 3 (this is equivalent to starting in 2 because after initiation the promoter is immediately freed and gets to the ON state). The survival function is then $S(t) = \mathbb{P}[w > t] = 1 - \mathbb{P}[M(t) = 3 \mid M(0) = 2] = 1 - X_3(t)$, which shows that one can compute the survival function by solving the linear system of ODEs (11) with the initial conditions from (10). Interestingly, for constant parameters, the distribution of w does not change in time (it is the same during transient and steady state gene expression). In other words, the sequence of initiation events is a renewal process.

The solution of (11) reads

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = C_1 \begin{pmatrix} \alpha_1 \\ 1 \end{pmatrix} \exp(\lambda_1 t) + C_2 \begin{pmatrix} \alpha_2 \\ 1 \end{pmatrix} \exp(\lambda_2 t), \tag{13}$$

$$X_3 = A_1(1 - \exp(\lambda_1 t)) + A_2(1 - \exp(\lambda_2 t)), \tag{14}$$

where $\begin{pmatrix} \alpha_1 \\ 1 \end{pmatrix}$, $\begin{pmatrix} \alpha_2 \\ 1 \end{pmatrix}$ are eigenvectors and λ_1, λ_2 are eigenvalues of the matrix $\tilde{Q}$, and C_1, C_2 are the solutions of the system

$$\begin{aligned} C_1\alpha_1 + C_2\alpha_2 &= 0, \\ C_1 + C_2 &= 1. \end{aligned} \tag{15}$$

Furthermore,

$$S(t) = A_1 \exp(\lambda_1 t) + A_2 \exp(\lambda_2 t). \tag{16}$$

From (14) and (12)

$$A_1 = -k_{ini} C_1/\lambda_1, \ A_2 = -k_{ini} C_2/\lambda_2. \tag{17}$$

Eqs. (16), (15) and (17) provide the solution of the direct problem that consists in computing the survival function parameters given the transition rate parameters. The inverse problem consists in computing the rate parameters k^+, k^-, k_{ini} given the independent survival function parameters A_1, λ_1, λ_2. The rate parameters are identifiable if and only if the inverse problem is well posed, i.e. it has a unique solution.

The inverse problem for the random telegraph model corresponds to solving the system

$$\lambda_1 + \lambda_2 = -(k^+ + k^- + k_{ini}), \tag{18}$$

$$\lambda_1\lambda_2 = k_{ini}k^+, \tag{19}$$

$$A_1\lambda_1 + A_2\lambda_2 = -k_{ini}. \tag{20}$$

Eqs. (18) and (19) are the Vieta's formulas, resulting from the fact that λ_1, λ_2 are the solutions of the characteristic equation of the matrix $\tilde{Q}$. Eq. (20) follows from (15) and (17).

For the random telegraph model, the solution of the inverse problem is unique and the transition rate parameters are expressed in terms of symmetric rational functions in the variables $\lambda_1, \lambda_2, A_1, A_2$, i.e. ratios of polynomials invariant with respect to permutations of these variables. More precisely,

$$\begin{aligned} k_{ini} &= -S_1, \\ k^- &= (S_1 - L_1)S_1/L_2, \\ k^+ &= -L_2/S_1 \end{aligned} \tag{21}$$

where $S_1 = A_1\lambda_1 + A_2\lambda_2$, $L_1 = \lambda_1 + \lambda_2$, $L_2 = \lambda_1\lambda_2$, are symmetric polynomials.

More generally, one can show that whenever the inverse problem has a unique solution, this can be written in terms of symmetric polynomials. Of course, the inverse problem can also have no solutions, or have an infinity of solutions.

The question of model and parameter identifiability can be decomposed into two steps. First, the survival function parameters are uncertain because they are obtained from data. Second, the inverse problem, consisting in identifying

the model and its kinetic parameters for the survival function parameters can be not well posed and have infinitely many solutions. This source of uncertainty can be also addressed using symbolic methods (19). There are several situations of symbolic non-identifiability/uncertainty:

- Model non-identifibility/uncertainty. Model parameters are uniquely determined for each model, but different models give exactly the same survival function with different parameters (the case of models M_1, M_2, Figure 3C).
- Parameter non-identifiability/uncertainty. Model kinetic parameters leading to the same survival function form smooth manifolds, meaning that some of them are free. Concurrently, multi-exponential parameters of the survival function are constrained, meaning that there are less free parameters of the multi-exponential survival function (the case of the model M_3, Figure 3C).

In both cases of non-identifiability/uncertainty, more data is needed in order to directly identify one or several parameters. We have implemented this strategy in (19,20) where, using chromatin immunoprecipitation or genetic perturbations of pausing, the parameter k_2^+ was shown to correspond to exit from proximal pausing, indicating that the model M_2 should be preferred to M_1.

Determining the polymerase dwell time from the signal autocorrelation

The signal autocorrelation function is defined as $R(t, t') = \mathrm{Cov}(x(t), x(t'))$, where $x(t)$ is the single site MS2 signal. For a stationary MS2 signal, this function depends only on $\tau = t' - t$ and factorizes as:

$$R(\tau) = F(\tau; \mathbf{k})(H(\tau + d) - 2H(\tau) + H(\tau - d)), \quad (22)$$

where d is the dwell time, $\mathbf{k}$ contains all model parameters including the dwell time (for instance $\mathbf{k} = (k^+, k^-, k_{ini}, d)$ for the random telegraph model), $H(x) = -x\theta(-x)$, $\theta(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ if } x < 0 \end{cases}$ is the Heaviside function (see (32) for a derivation).

The determination of the dwell time results from fitting the theoretical model (22) to the empirical autocorrelation function resulting from data. The test of this method is illustrated in the Supplementary Table S1.

It turns out from (22) that the autocorrelation function R depends strongly on d and only weakly on the other parameters $\mathbf{k}$. For this reason d is precise, whereas $\mathbf{k}$ is uncertain when estimated from R.

RESULTS

Principles and workflow

The input data for our model are live imaging data of nascent transcription, with nascent RNAs labeled with a fluorescent tag. As test samples, we used MS2/MCP data collected from either cultured human cells or *Drosophila* early embryos. This labelling method is bipartite, with an RNA containing MS2 repeats of various lengths, detected by an RNA binding protein, here MCP, fused to a fluorescent protein (Figure 2D). After live imaging, MS2/MCP fluorescent signals of single transcription sites (temporal traces) are extracted through image analysis methods described in (18,20) that track each transcription site in 3D in order to extract the intensity of the MS2 signal over time. For each movie we produce an intensity matrix whose rows and columns represent transcription sites and time, respectively (Figure 2A). Because we wish to separate individual transcription initiation events we use movies with high temporal resolutions (typically 3–4 s) and the sequence of transcription initiation events is reconstructed independently for each transcription site (Figure 2B, C). The MS2/MCP fluorescent signals are calibrated to be expressed as polymerase numbers. In order to decompose the signal observed from multiple polymerases (Figure 2E) into initiation events, we first consider the signal expected from a single polymerase, schematized in Figure 2D. The single polymerase pattern is computed from n_{seq}, n_{post}, V_{pol} and t_{a}, representing the length in base pairs of the MS2 sequence, the remaining length after the MS2 sequence until the polyA site, the polymerase elongation speed and the 3'-end processing/polyadenylation time, respectively. In this notation, the polymerase dwell time on the DNA is $(n_{\mathrm{seq}} + n_{\mathrm{post}})/V_{\mathrm{pol}} + t_a$. In this model, we consider that a polymerase, once initiated, will continue transcription until it reaches the 3'-end. The estimated initiation times are obtained by least squares regression using a global genetic algorithm, followed by local optimization (Figure 2F). Multi-exponential parametric estimates of the survival function are then used to characterize the distribution of the waiting times between successive initiation events for the entire population of sites (see Figure 2G and Materials and Methods). The multi-exponential regression proposes one, two, or more exponentials. The number of exponentials corresponds to the number of states in the FSMM (Figure 3). Finally, comparison of the exponentials found by regression to the analytic solutions of the master equation satisfied by the survival function allows us to write explicit formulas for FSSM parameters in terms of the regression results (Figure 2H).

A few examples of FSMMs are represented in Figure 3. The random telegraph model (Figure 3A) contains two states: the ON state corresponding to active transcription, modeled by Poissonian initiation with constant initiation rate k_{ini}; and the inactive OFF state where no initiation events are observed. As the two state random telegraph model is generally too simplistic to fully describe the complexity of the transcription process (18,38,39), we also envisaged more complex models with three states, comprising two inactive OFF states (models M_1, M_2 and M_3). In the model M_1 (Figure 3C), an inactive promoter occupying the state OFF_2 can become active (state ON) or switch to a deeper inactive state OFF_1. Inactive states represent various molecular states of the promoter, such as chromatin states or assembly stages of the transcription pre-initiation complex (PIC). In the model M_2 (Figure 3C), the second inactive state was interpreted as proximal pausing. This interpretation is based on the experimental manipulation of pausing (in *cis* or in *trans*) that we performed with model

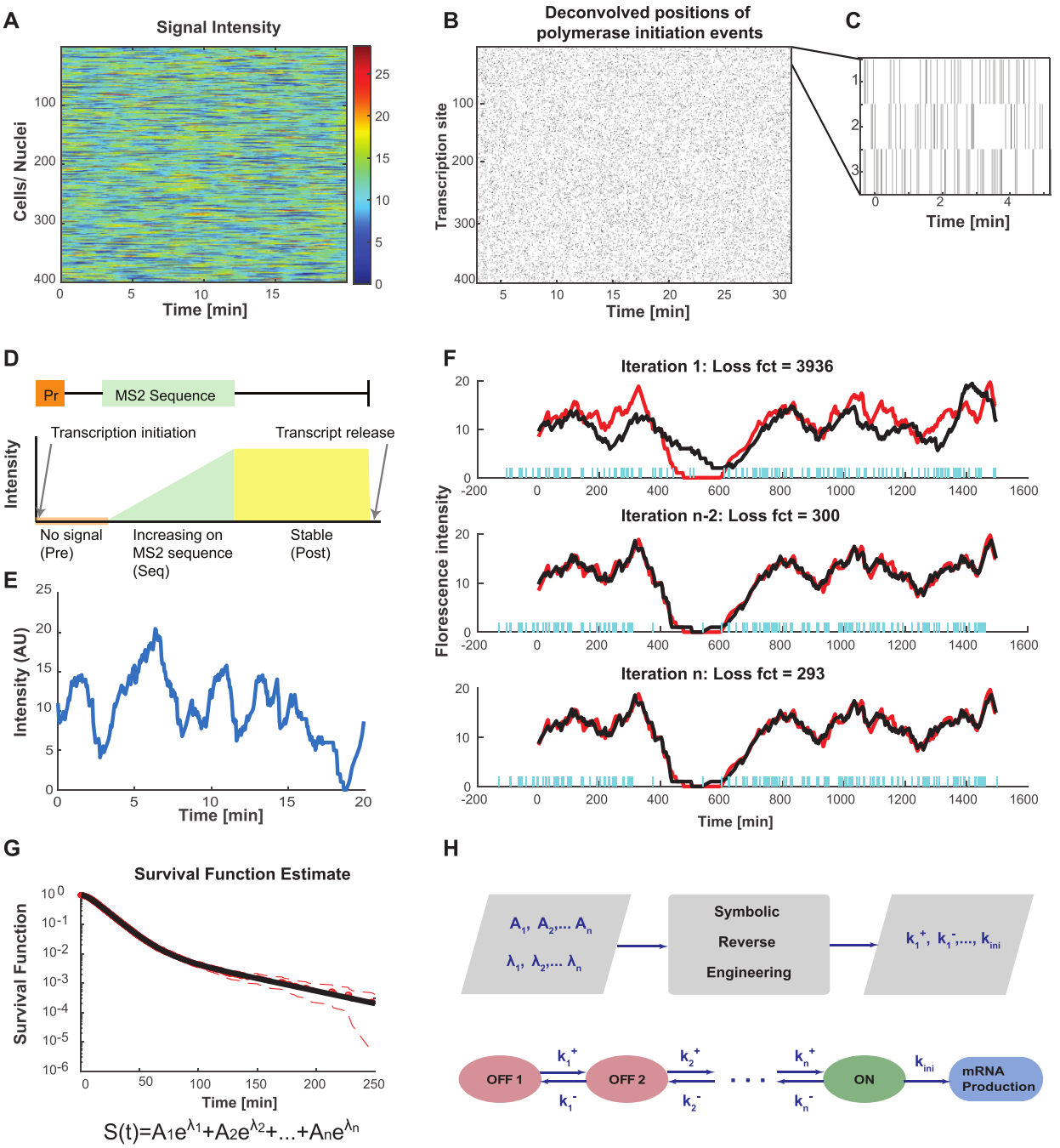


Figure 2. Overview of various steps of BurstDECONV. (**A**) Heatmap of the signal extracted from a high resolution movie. Each row represents a transcription site intensity in time (x-axis). The colour bar depicts the number of nascent RNA. (**B**) Timeline chart representing the transcription initiation events obtained for each corresponding transcription site intensity trace after performing deconvolution using the genetic algorithm. (**C**) Close up of the timeline chart. Each bar represents a single event; successive events are separated by waiting times. (**D**) The RNA tagging construct and its corresponding signal generated from a single polymerase. The orange box labeled Pr represents the promoter site where transcription initiation takes place. The MS2 sequence is located a few bases downstream of the promoter region. The signal profile is shown below the construct. (**E**) Intensity trace from one transcription site. (**F**) Example of polymerase positions reconstructing the transcription site intensity trace in the last three generations of the genetic algorithm. The red trace is the intensity trace that is to be reconstructed. The black trace is the reconstructed signal from the predicted polymerase positions (represented with blue bars). (**G**) Survival function estimated from the waiting times between the predicted polymerases. The dotted red curve represents the Greenwood confidence interval of the survival function. We obtain both non-parametric survival function (depicted with red circles) by Kaplan-Meier method, and the parametric survival function by least square regression (depicted with the black curve). The parametric survival function is a sum of N exponentials. (**H**) The various coefficients of the parametric survival function are used to obtain the model parameters (switching rates between different states) through symbolic reverse engineering.

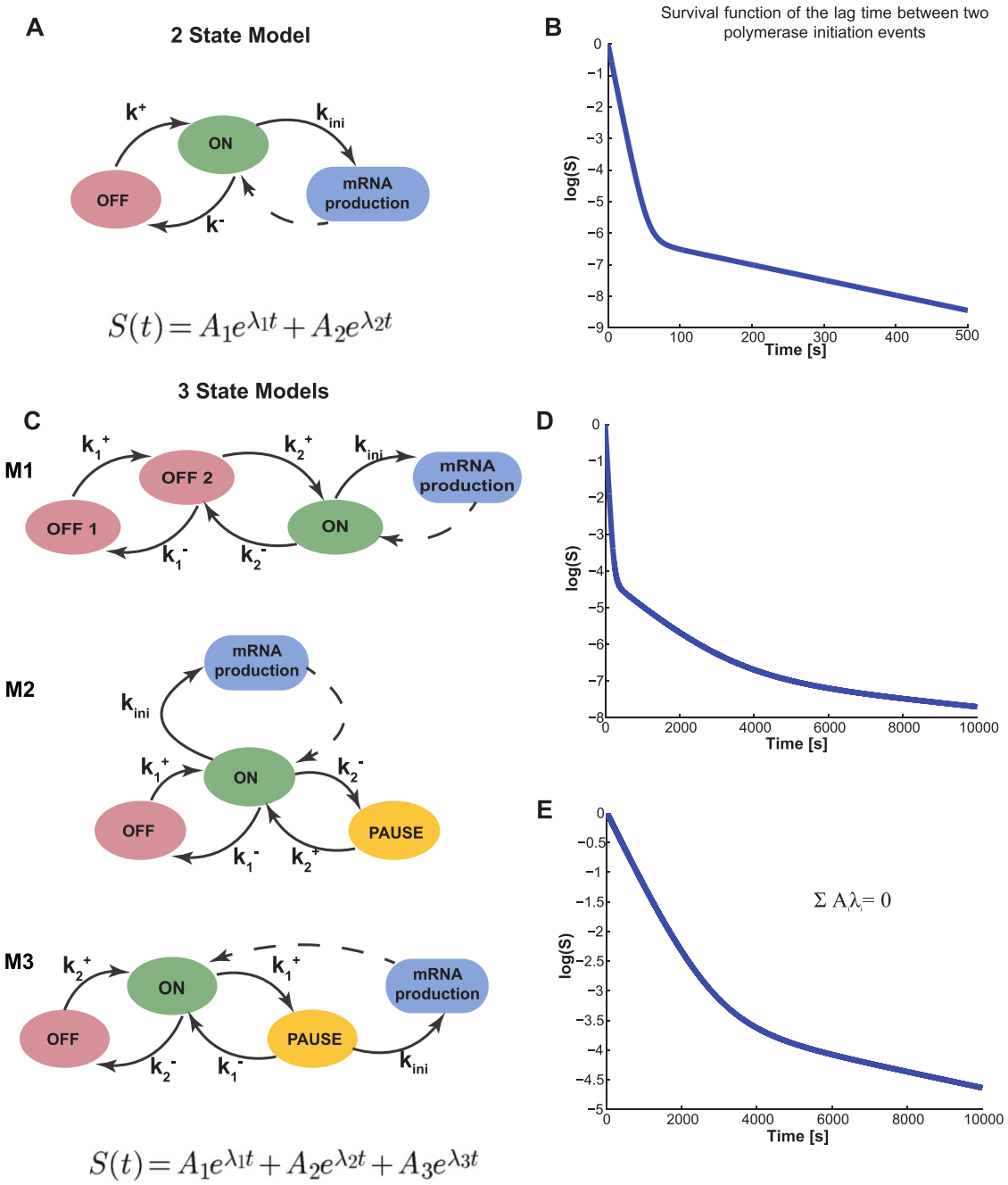


Figure 3. Finite-State Markov Models of transcription dynamics. (**A**) Model depicting two promoter states, ON and OFF with the respective transition rates. (**B**) The theoretical survival function corrresponding to the two-state exponential model has two timescales. The two separated timescales can be distinguished as two distinct slopes, piecewise in the semi-logarithmic representation. (**C**) Three-state models with different transition schemes. (**D**) The theoretical survival function of the three state models M_1, M_2 is a sum of three exponentials with no constraints on the amplitudes A_i. The three separated timescales can be distinguished as three distinct slopes, piecewise in the semi-logarithmic representation. These two models have the same type of survival function and can not be discriminated by BurstDECONV only. (**E**) The theoretical survival function of the three state model M_3 is a sum of three exponentials with constrains on the amplitudes A_i and exponents λ_i. Only two free timescales can be distinguished in the semi-logarithmic representation.

paused promoters such as HIV-1 in human cells or developmental core promoters in *Drosophila* embryos (19,20). In the model M_2 the transition from ON to PAUSE is stochastic, therefore pausing is facultative. This is at odds with the traditional obligatory pausing model M_3 (Figure 3C) in which the pausing occurs after initiation and systematically prevents elongation, as usually depicted in the literature (40). The model M_3 predicts a special type of survival function whose multi-exponential parameters are constrained by an additional relationship (see Figure 3E and (19)).

The inverse problem consisting in computing the model kinetic parameters from the survival function parameters (A_i, λ_i), $1 \leq i \leq n$, is well posed when it has a unique solution for all survival function parameters satisfying the constraints $\lambda_i < 0$, $1 \leq i \leq n$, $\sum_{i=1}^{n} A_i = 1$. This is the case for the random telegraph model, for the models M_1, M_2, for a family of models of arbitrary size discussed in (19), but not for the model M_3. For M_3, the survival function parameters are constrained by one bilinear equation in A_i and λ_i (see Materials and Methods and Figure 3); furthermore, in this case the inverse problem has infinitely many solutions, that depend on one free parameter.

Artificial data shows the robustness of BurstDECONV

In order to benchmark the method we use a collection of artificially generated datasets. These datasets consist of MS2 signals from N transcription sites. The models and corresponding parameter sets are given in Table 1, and they are chosen to mimic a variety of real biological situations. Indeed, the parameter sets simulate observations of wild type and mutated *snail* (D2,D3,D5,D7,D9) or *Kruppel* (D1,D4,D8) *Drosophila* promoters studied in (20), or from HIV-1 promoters inserted in Hela reporter cell line in various configurations (notably with and without the viral transactivator Tat; D12–14) studied in (19). We have added a few more parameter sets corresponding to the wild-type and mutant human EEF1A promoters inserted in human cell lines (D6;D10–11). These data cover a large range of expression levels, and correspond to promoters having two or three rate-limiting steps, and being mostly, or only episodically, active.

The artificial data was generated using the parameter estimates obtained with real data. Using the Gillespie algorithm we generated N independent trajectories of the FSMM that provide the initiation events over a time interval T corresponding to the movie length. Then, we use the single polymerase patterns to compute the MS2 signal. The single polymerase patterns correspond to 24xMS2 and 128xMS2 constructions in *Drosophila* and in human cell lines, respectively (see (18–20)). For more realism, we add noise to this signal. In analogy to real data, we use Gaussian heteroscedastic noise (see (19) and Material and Methods).

In order to evaluate the accuracy of the parameter reconstruction we use the logarithmic error defined as $|\log_{10}(k_r/k_{true})|$, where k_r, k_{true} are the reconstructed parameter and their true value, respectively. Errors were considered unacceptable if they correspond to one order of magnitude, i.e. if the logarithmic error is larger than one.

Table 1. FSMM parameters used to generate the artificial datasets. Furthermore, the MS2 sequence and elongation rate parameters were n_{seq} = 1292 bp (24×MS2), n_{post} = 4526 *bp*, V_{pol} = 45 bp × s^{-1} for D1–5 and D7–9,n_{seq} = 5800 bp (128×MS2), n_{post} = 8300 bp, V_{pol} = 67 bp × s^{-1} for D6 and D10–14. D1–5 and D7–9 parameters come from the study of *Drosophila* promoters in (20). D12–14 is based on estimates of HIV-1 transcription bursting in human cells studied in (19). D6 and D10–11 come from estimates of bursting from wild-type and mutated EEF1A promoters inserted in human cell lines

Dataset/Ref.	Parameters				
2 states	$k^+[s^{-1}]$		$k^-[s^{-1}]$		$k_{ini}[s^{-1}]$
D1 (20)	0.02036		0.00150		0.11432
D2 (20)	0.01117		0.00593		0.07637
D3 (20)	0.01189		0.01430		0.07745
D4 (20)	0.02439		0.00144		0.13397
D5 (20)	0.04169		0.00414		0.11277
D6	0.00484		0.00025		0.113
3 states M_2	$k_1^+[s^{-1}]$	$k_1^-[s^{-1}]$	$k_2^+[s^{-1}]$	$k_2^-[s^{-1}]$	$k_{ini}[s^{-1}]$
D7 (20)	0.01426	0.00339	0.06553	0.05751	0.17102
D8 (20)	0.00661	0.00013	0.05772	0.01054	0.13201
D9 (20)	0.00332	5.3×10^{-5}	0.05804	0.00586	0.13119
D10	0.0001	2.3×10^{-5}	0.00091	0.00024	0.019
D11	0.00023	3.2×10^{-5}	0.0011	0.00019	0.018
D12 (19)	0.0015	4.9×10^{-5}	0.01	0.0043	0.17
D13 (19)	0.00015	0.00031	0.0012	0.0028	0.1
D14 (19)	6×10^{-5}	0.00035	0.00089	0.003	0.063

BurstDECONV combines short high resolution with long low resolution movies to cover widely distributed timescales

Transcription bursting is a complex phenomenon involving processes with multiple timescales distributed over many orders of magnitudes (8,18). The movie length sets the upper bound of the timescales of processes that can be identified using live cell RNA imaging data. A short movie may fail to detect slow processes that involve long waiting times. In order to test this we have used models that have timescales ranging from 1s to 10^4s. Deconvolution of a short (20 min) signal (Figure 4 B) results in mediocre parameter reconstruction (Figure 4 F) and as expected, errors were larger for smaller kinetic parameters (large timescales). In order to illustrate this effect we have used the dataset D14 that includes very long waiting times (very small values of the parameters k_1^+ and k_2^+).

Due to bleaching of the signal, obtaining long movies (in the h scale) while imaging with a high temporal resolution of few seconds is extremely challenging.

Instead, we designed a version of BurstDECONV that combines short, high resolution movies, and long, low resolution movies. The first step consists in deconvolution of the short high-resolution movies and computation of their survival function. The second step processes the long movies, which last typically 10 h with a temporal resolution of 3 min. In this case, active and inactive periods are defined directly by considering the parts of the MS2 signal that are above and below a threshold, respectively, with the threshold corresponding to the brightness of 2–3 RNAs (Figure 4C). By thresholding, we miss the short waiting times between transcription events that occur during active periods, and we thus count only the long waiting times corresponding to inactive periods (i.e. waiting times greater than

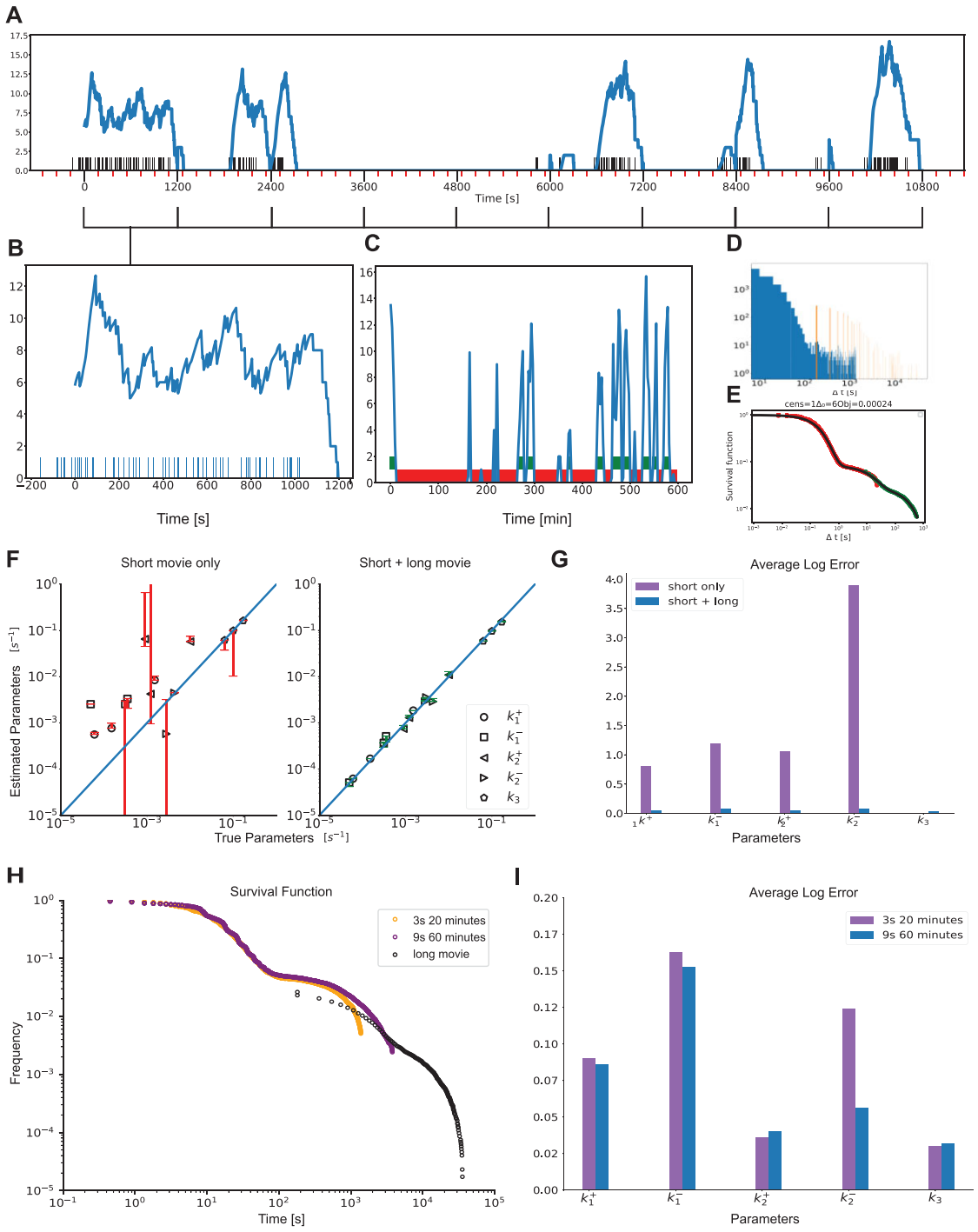


Figure 4. Combining short and long movies. (**A**) Simulated long duration signal using the three-state model M_2 and high temporal resolution, parameters corresponding to dataset D14. The timeline with black markers represent polymerase start time positions. The transcription site signal is represented in blue, using short movie high temporal resolution. The x-axis major tick marks in black (every 1200 seconds) represent the duration of a high resolution short movie (1 stack every 3 s for 20 min). The minor ticks in red represent 3 min marks, i.e. the resolution of a long duration movie (1 stack every 3 min). (**B**) Simulated low resolution short movie using the model M_2. Blue bars represent polymerase positions found by GA after deconvolution (blow-up start of the signal from **A**). (**C**) Low resolution long movie with thresholding to extract off periods or waiting times. (**D**) Histogram of length of waiting times obtained from short movies (in blue) and long movies (in orange). (**E**) Matched Survival function of the long (green) and short movie (red) with overlap in the middle. (**F**) Accuracy of parameter reconstruction of the model M_2 for the parameter sets D12–14 in Table 1 using only a short movie and a short + long movie. (**G**) Average logarithmic error of the parameter reconstruction of the model M_2 for the parameter sets D12–14 in Table 1 using a short movie (20 min every 3 s) combined or not with a long movie (10 h every 3 min). (**H**) Long and short movie survival functions and their overlap for different lengths and resolutions of the short movie. (**I**) Average logarithmic error of the parameter reconstruction of the model M_2 for the parameter sets D12–14 in Table 1 for different durations and resolutions of the short movie.

the dwell time of a polymerase, defined as the total time length of the signal generated by a single polymerase). The distribution of the long waiting times overlaps that of short waiting times obtained from the short movies (Figure 4D). This overlap permits us to reconstruct a multiscale survival function that covers many orders of magnitude in time (see Figure 4E and (19)). The multiscale survival function is then used for multi-exponential regression and provides the kinetic parameters of the model. The combination of the two movie types permits a good reconstruction of these parameters (Figure 4G).

The length of the short movie is critical for ensuring the overlap of the short and long timescale survival functions and an accurate parameter reconstruction. Interestingly, longer short movie with lower temporal resolution (60 min length with frames every 9 s, instead of every 3s for 20 min) can ensure a better reconstruction of the model parameters (Figures 4H, I), even with initiation rates (k_{ini}) in the range of 3 s. This provides useful guidelines for the experimental design and imaging conditions.

Thus, BurstDECONV allows to uncover processes with a remarkable distribution of timescales ranging from seconds to days.

BurstDeconv determines the number of states of the kinetic promoter model

A key question when analyzing live cell transcription imaging experiments, is the choice of the model used to fit the data. Instead of arbitrarily employing the simple random telegraph promoter model, our procedure uses the multi-exponential fit of the waiting time data to determine the number of promoter states that should be considered. We recall that, except for special cases when the spectrum of the matrix $\tilde{\boldsymbol{Q}}$ is degenerate, the number of states in the transcriptional bursting model is equal to the number of exponentials n_{exp} in the multi-exponential fit.

In order to compare models with different n_{exp} we use several indicators for the goodness of fit (Figure 5A). The experimental estimate of the survival function by the Kaplan–Meyer method provides a confidence interval based on Greenwood's formula (20). A first accuracy test consists in checking that the optimal parametric estimate of the survival function lies within this confidence interval. The Kolmogorov–Smirnov (KS) test and the optimal value of the objective function $\mathcal{O}_2$ (the mean squared deviation; see Material and Methods for a definition) provide alternative measures of the quantitative goodness of fit. The Greenwood's confidence interval and KS methods do not take into account errors resulting from the imperfect join of the short and long movie survival functions. Therefore, the only goodness of fit measure when one also uses long movies is the value of the objective function $\mathcal{O}_2$.

The goodness of fit systematically increases with n_{exp} and one would like to know when to stop. The stopping criteria can be based on parsimony (Figure 5 A): choose the less complex model (smallest n_{exp}) whose goodness of fit does not differ significantly from the next more complex one. An alternative strategy can use cross-validation. We illustrate cross-validation by splitting the artificial data in a training and a validation subset. Both training error and validation error decrease with n_{exp}. However, their difference (the validation gap) has a minimum. The optimal n_{exp} corresponds to training and validation errors that are as close as possible, i.e. corresponding to the minimal validation gap. A large validation gap indicates either underfitting (when both training and validation errors are large) or overfitting (Figure 5 B).

Cross-validation is usually difficult to set in practice when the number of available cells is not large enough. In this case we estimate overfitting by parametric uncertainty. Indeed, an overly complex model can fit data equally well for different values of its parameters. We use optimal and close to optimal solutions to define uncertainty intervals that contain the parameters leading to a close to optimal fit. We then gradually increase n_{exp} until the goodness of fit (training error) becomes sufficiently small while the uncertainty parametric intervals are not large (Figure 5A).

Alternative model selection procedures, based on hierarchical Bayesian learning have been proposed for obtaining the parametric survival function and the number of exponentials (chapter 5 of (41)). Their practical implementation will be tested in future work.

BurstDECONV is robust against error and suboptimal experimental designs

Robustness against changes in calibration. In this method and in any quantitative method based on live cell transcription imaging, the polymerase loading rate (k_{ini}) can be determined only if the signal intensity is expressed in units of full-length transcripts.

The calibration is performed by dividing the transcription site signal intensity by the calibration factor that is defined as the contribution to intensity of a single RNA molecule. This factor can be computed in different ways. In order to calibrate fluorescent signals from live *Drosophila* embryo imaging, we used single-molecule hybridization experiments (smFISH) as described in (20). In human cell lines, we collected right after the end of the movie one 3D stack—termed calibration stack—with increased laser intensity, which similarly allowed reliable detection and quantification of the brightness of individual RNA molecules (18,19).

We illustrate the importance of the calibration factor by testing the effect of altering it in artificial data (Figure 6). Decreasing the calibration factor corresponds to underestimating the contribution of one RNA to the signal and corresponds to more polymerases to model the same signal (Figure 6A). This also has an influence on the survival function, because more polymerases mean shorter waiting times between successive initiation events (Figure 6B). Increasing the calibration factor leads to decreased estimates of all kinetic parameters (Figure 6C). As expected, the polymerase initiation rate (parameter k_{ini} in the random telegraph model) scales like the inverse of the calibration factor (Figure 6C). The effect of the calibration factor on the switching parameters (k^+ and k^- in the random telegraph model) is asymmetric. It is weaker when the calibration factor is less than optimal and larger for calibration factor larger than optimal (Figure 6E). In other words overestimating the one polymerase signal leads to larger

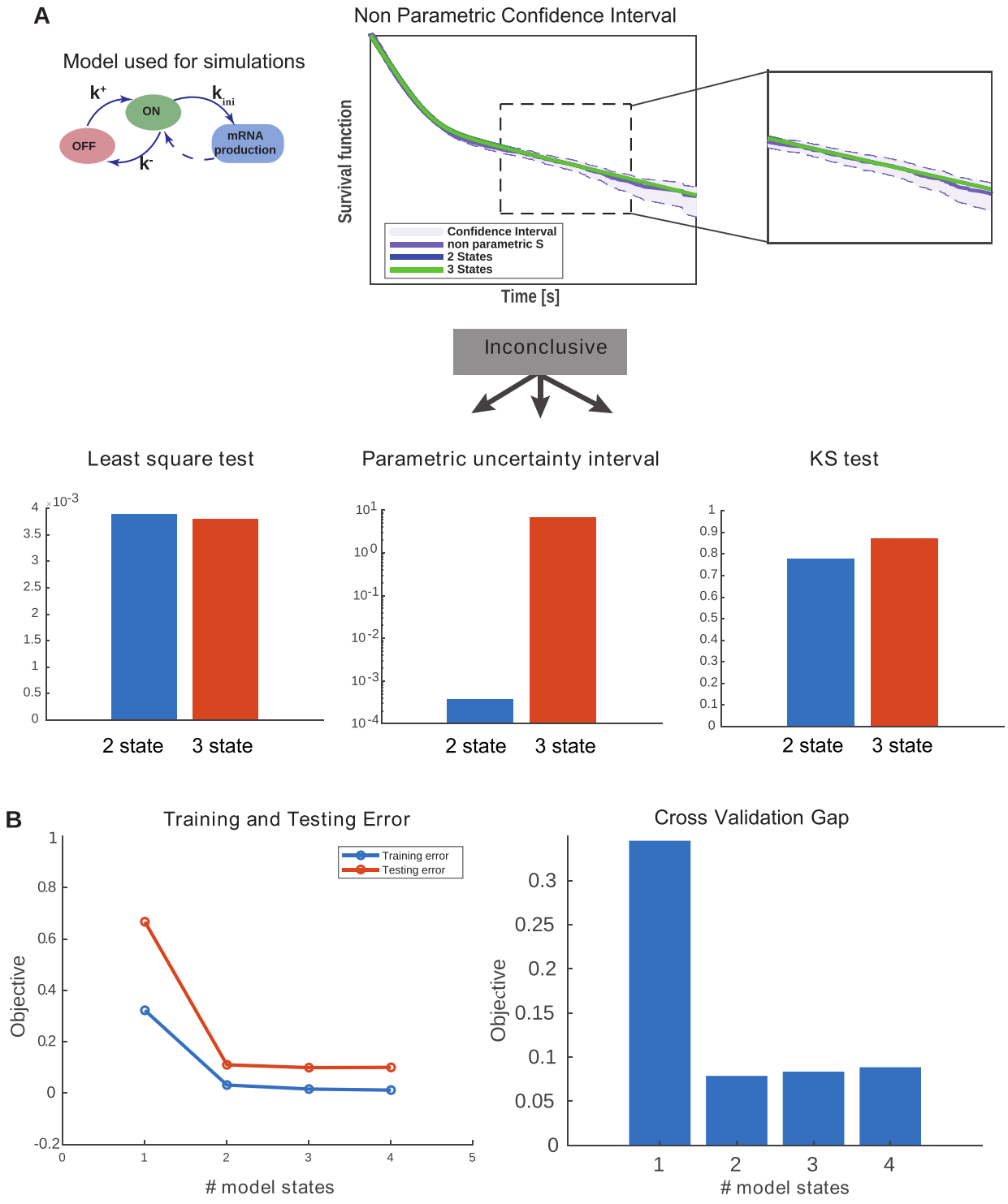


Figure 5. Selection of the number of exponentials in multi-exponential survival models. (**A**) The number of exponentials n_{exp} in the multi-exponential survival model is first selected using the Greenwood confidence interval for the Kaplan-Meyer non-parametric estimator. One verifies that the optimal parametric estimate is included in the confidence interval of the non-parametric estimate, for increasing n_{exp} starting with $n_{exp} = 1$. The selected n_{exp} value is the first one that satisfies this condition. If the result is inconclusive (borderline), we evaluate the training error by using the least-square error or the Kolmogorov–Smirnov test, and the overfitting by using the width of the parametric uncertainty intervals. The selected n_{exp} is the first one that has similar training error and lower parametric uncertainty than n_{exp} + 1. (**B**) Cross-validation. The dataset (set of nuclei) obtained from a two-state ground truth model (dataset D6) is split into a training and validation subsets. Then the model capacity is increased by increasing n_{exp}. Both training and testing errors decrease with n_{exp} but the difference between the two (the cross validation gap) has a minimum at the ground truth. The cross-validation can be used for selection when the number of samples (nuclei) is large enough.

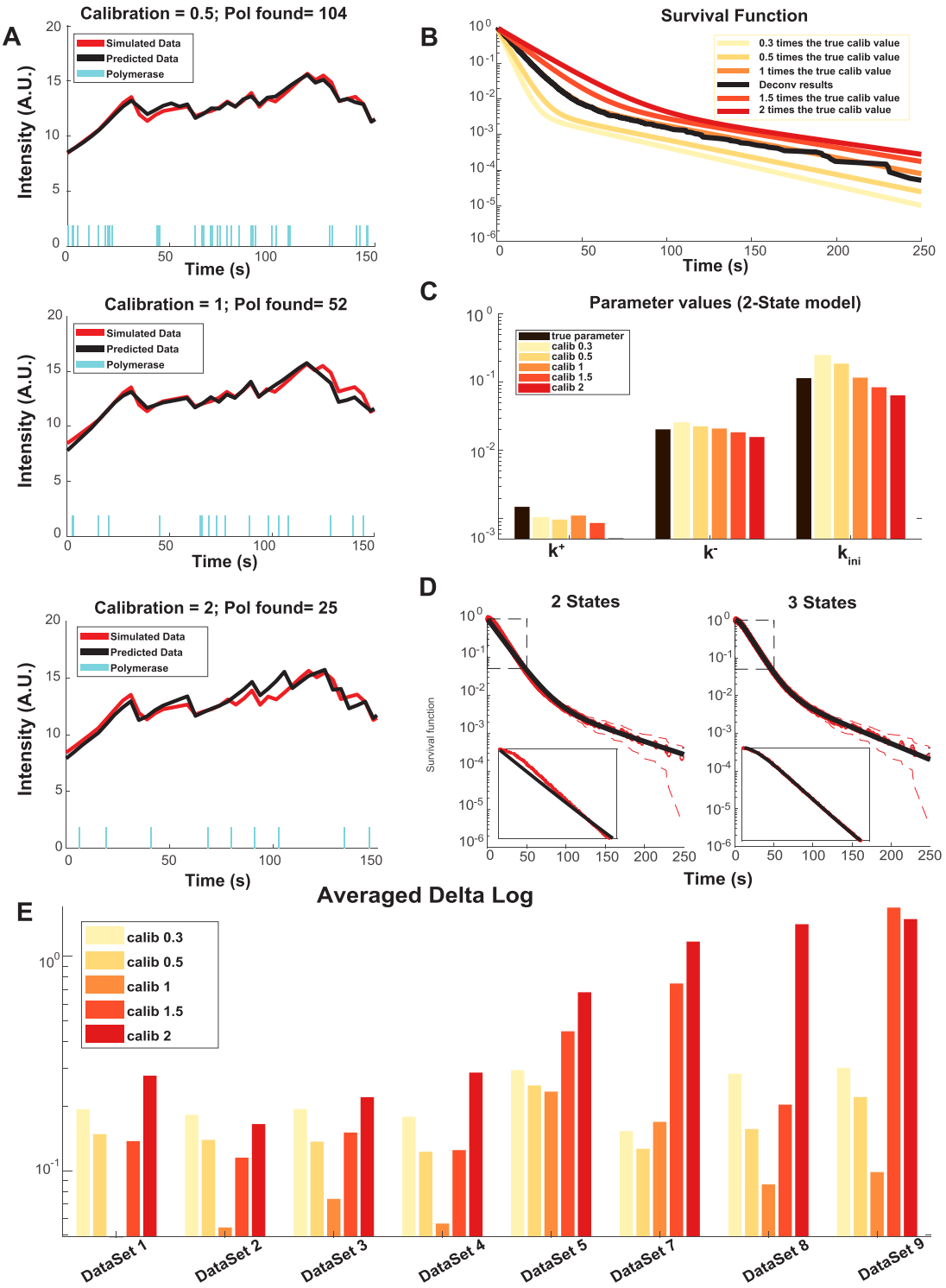


Figure 6. Testing the effect of a change in the calibration factor. (**A**) Simulated signal for a two-state model, dataset D1 in Table 1, for different values of the calibration factor. The cyan timeline bars indicate the start time positions. The simulated transcription site signal is represented in red. The reconstructed signal (after deconvolution) is represented in black. For the ground truth, the number of simulated polymerases is 52 and the calibration factor is one. (**B**) Survival functions reconstructed for different values of the calibration factor (two-state model, dataset D1 in Table 1). (**C**) Reconstructed parameter values for different calibration factors (two-state model, dataset D1 in Table 1). (**D**) Nonparametric survival function compared to parametric 2- and 3-exponential functions for a calibration factor = 2. Doubling the calibration factor with respect to the ground truth can mistakenly lead to a change in the model selection from two (ground truth, falsified by the confidence interval criterion) to three states. (**E**) Average logarithmic parameter reconstruction error for various datasets and calibration factors.

reconstruction errors than underestimating it. A possible explanation of this effect is that increasing the calibration factor reduces the apparent number of initiation events which renders the identification of the switching periods less reliable. In order to illustrate these effects we have used the dataset D1. This dataset (together with D4 that is very similar) proved to be the most sensitive to calibration, as in this case a twofold increase of the calibration factor with respect to the optimal value leads to selecting a three states instead of the ground truth two states model, see Figure 6D. This dataset corresponds to a highly active promoter as k^- and k^+ are small and large, respectively.

Robustness against changes in polymerase speed and dwell time. In our method, the polymerase speed is considered to be known. Changing this parameter is roughly equivalent to changing the polymerase dwell time and has effects on the number of polymerases (and loading rate parameter) opposed to changing the calibration.

Robustness against changes in time resolution. A low resolution movie provides poor representations of the MS2 intensity (Figure 7A). The deconvolution algorithm tends to interpret local drops in the MS2 signal as an OFF state. However, these drops and the corresponding OFF states may be missed for very low resolutions (such as 131.3 s in Figure 7A). Missing OFF states lead to a larger number of predicted polymerase positions (Figure 7 B), steeper survival functions (Figure 7C) and errors mostly in the ON to OFF transition rates (parameter k^- in Figure 7D). The shorter timescales, corresponding to the parameters k^+, k_{ini} are less affected. The critical resolution producing large errors in the number of polymerases, survival function and kinetic parameters is close to the polymerase dwell time. We compared the results obtained by our procedure on artificial datasets resampled with various temporal resolutions and found that the method is robust and tolerates resolutions (11–20 s) much lower than the ones currently employed (3–3.9 s). Thus, there is not significant gain when imaging every 3–4 s compared to imaging every 11–20 s. This again provides important guidelines to design optimal imaging conditions.

Robustness against noise in the data. In order to simulate a noise that resembles real experimental data, we analyzed the variance of the residuals resulting from the least-squares fitting. We have found that residuals are normally distributed with a variance increasing with the level of the predicted signal, which means that the experimental noise is heteroscedastic. A third order polynomial fitting was enough for approximating this dependence (see Materials and Methods). We thus have added Gaussian noise to the artificial data, whose variance has the same polynomial dependence on the mean as the experimental data. We have found that even for a noise amplitude multiplied by four with respect to the experimental values, BurstDECONV is able to reconstruct the parameter values used for the simulation (Figure 8). The accuracy is very good for experimental noise amplitudes. To some extent, the noise in the signal is averaged by the least-squares optimization step and therefore no noise subtraction or estimation is needed for the parametric model reconstruction in BurstDECONV. In order to illustrate these effects we have used the datasets D12–14 because they have multiple, well separated waiting times, which allow us to test the effect of noise on different timescales.

Robustness against the detection limit. It is very common in live cell imaging to have a background noise signal that sets a detection limit. To recreate this effect, we have added a supplementary component to the noise, which is independent of the MS2 signal. We tested different amplitudes of this basal noise corresponding to one, two or four molecules of RNA, respectively. The effect was tested on the datasets 12–14 as these include long waiting times.

The results are shown in the Supplementary Figure S1. The error induced by the background noise is small (smaller than one in base 10 logarithmic scale) for the parameters k_1^+, k_1^-, k_{ini} and for all the tested noise values. For the datasets 13, 14 the parameters k_2^+ and k_2^- are accurate for small noise, but can be inaccurate for a large background noise. However, reconstruction of parameters of the dataset 12 is particularly robust: the logarithmic error is smaller than one for all parameters and noise values.

It is reasonable to hypothesize that highly active promoters with high transcription site intensities are less affected by the RNA detection limit because they are above the detection threshold most of the time. This is indeed what we see as dataset 12 (known strong promoter) showed lower reconstruction errors as compared to the other datasets 13 and 14 (weak promoters).

Benchmark of BurstDECONV against state-of-the-art methods

We have compared BurstDECONV to the two main existing methods generally employed in quantitative transcriptional bursting, namely auto-correlation (33) and Hidden Markov Model (HMM) methods (34,36).

The auto-correlation method (32,33,42) uses the auto-correlation function of the single transcription site signal as a model-agnostic representation of the live cell transcription imaging data. Kinetic parameter inference can be performed by fitting theoretical auto-correlation functions to the empirical auto-correlation function obtained from the time series data. Theoretical auto-correlation function models are available for the random telegraph model (32,33) but also for a three state model (yet different from our models $M_{1\text{-}3}$) (33).

The HMM method (34) is based on a fixed choice of a mechanistic model. The model is inferred directly from data by the method of maximum likelihood. Like in our models, in the HMM model it is supposed that the promoter can be successively in one of the active or inactive states from a finite set of states. The transitions between states are modeled by a FSMM. Contrary to our models where the polymerase loading is a Poissonian process, in the HMM model the same process is modeled by a Gaussian process (34). This approximation is accurate for high polymerase loading rates, but may fail for lower rates of initiation. Moreover, in order to compute the likelihood function, the HMM method computes a sum over all the possible states of the

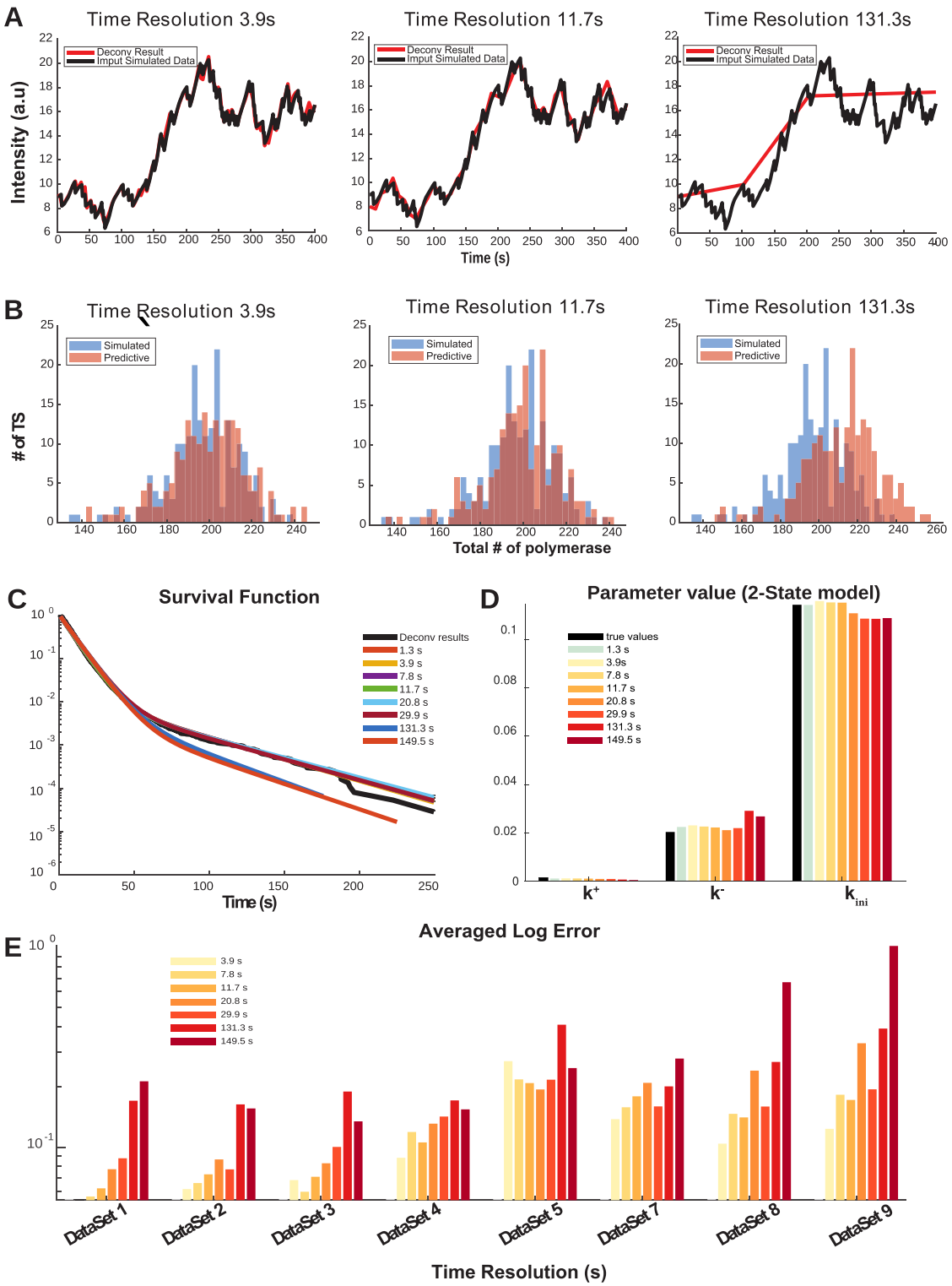


Figure 7. Testing the effect of a change in the movie time resolution. (**A**) Simulated and reconstructed signal for different time resolutions, for a two state model, dataset D1 in Table 1. The simulated transcription site signal is represented in black. The same signal is resampled with different rates and then reconstructed by deconvolution. The reconstructed signal (after deconvolution and with different sampling rates) is represented in red. (**B**) Histograms of the number of polymerases per analysed site for different time resolutions. (**C**) The reconstructed survival functions for different time resolutions. (**D**) Average logarithmic parametric reconstruction error for different time resolutions and kinetic parameters (dataset D1). (**E**) Average logarithmic error for different datasets and time resolutions.

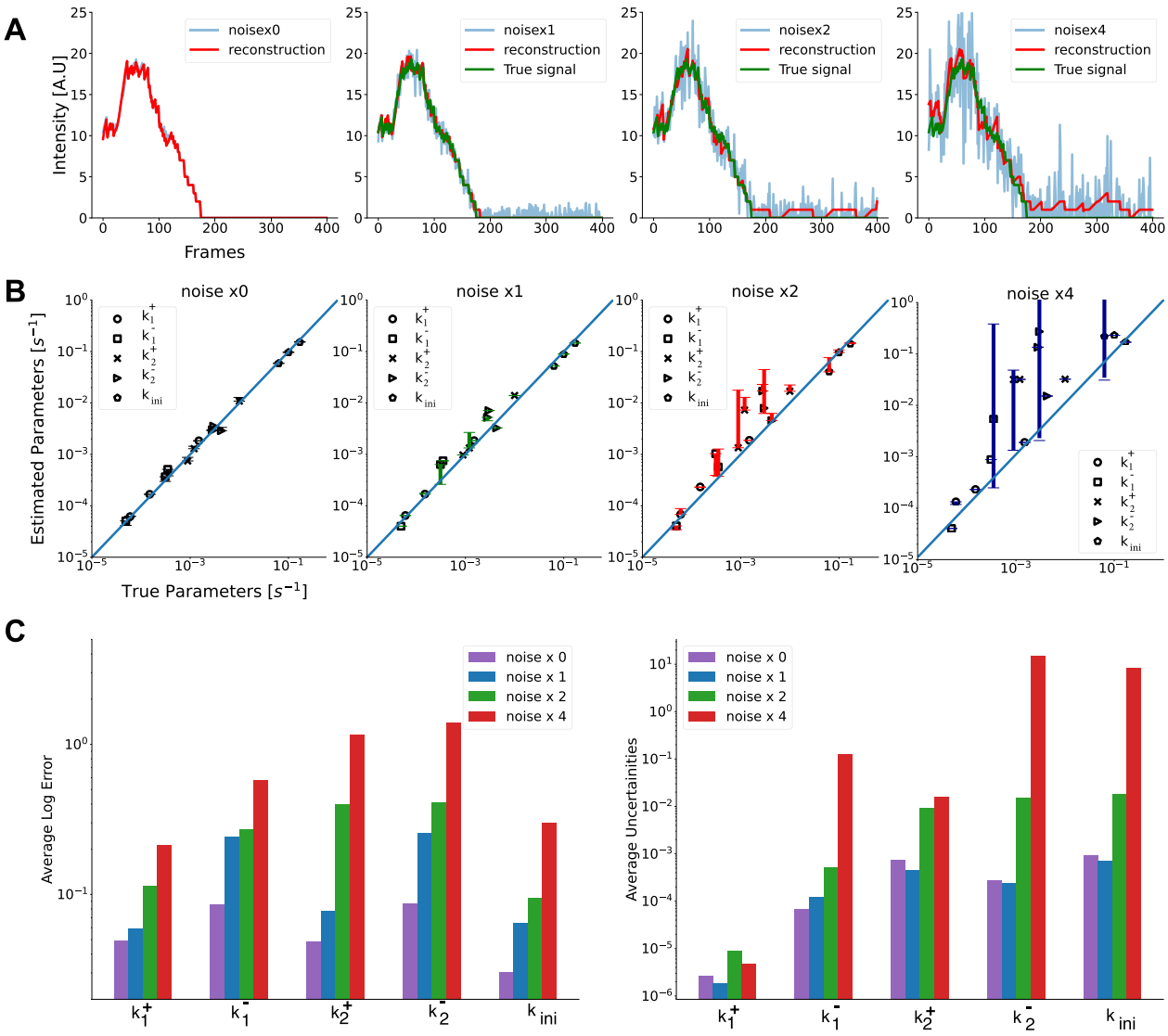


Figure 8. Robustness of the method against noise. All the simulations were performed using a 3-states model M_2 (Datasets D12-14 in Table1). (**A**) Reconstructed signal (shown in red) obtained using deconvolution from noisy artificial data (shown in blue). The artificial signal without noise is shown in green. Noise x0 represents the signal without any noise added to it and noise x1 is obtained by adding heteroscedastic noise to noise x0 equivalent to the one estimated from cultured human cells. noise x2 and x4 correspond to noise standard deviations two and four times larger, respectively. (**B**) Parameters reconstructed by the pipeline vs the true parameters used to simulate the artificial data for model M_2 datasets D12–14 in Table 1. (**C**) Logarithmic parametric reconstruction error D12–14 (left) Average uncertainty in the parameters (right).

promoter at several experimental time points spanning a memory interval equal to the dwell time. Thus, the computation time of this method increases exponentially with the dwell time and with the time resolution. An approximate version of the HMM method (36) trades accuracy for speed by considering only promoter states of large enough probability, for the computation of the likelihood function.

Given the difficulty of HMM in treating with high time resolutions, we have cross compared the kinetic parameter reconstruction error for several methods, using various artificial datasets and time resolutions. For the comparison we considered the two versions of BurstDECONV, the simple and the mixed one, using only short high resolution movies and both short and long movies, respectively. All the other methods were tested on short movies as they do not allow to combine movies of different time scales. All the artificial short movies last 26 min but their time resolution varies from 3.9 to 39 s. The HMM method was used in two versions: the 'exact' version implemented in (34) that explores the full state combinatorics of the promoter states and the 'burstInfer' version implemented in (36) that explores a reduced number of states. The auto-correlation method is represented by its implementation in (33). This implementation considers that the polymerase loading is deterministic with a known fixed rate (one polymerase every six seconds precisely, corresponding to our parameter $k_{ini} = 0.166\ s^{-1}$) and fits only the switching rates of the random telegraph model (corresponding to our parameters k^+ and k^-). We

have also tested inferring simultaneously all the parameters of the random telegraph model together with the polymerase dwell time using the auto-correlation method described in (32). In this case the parameters k^+, k^- and k_{ini} can not be reliably reconstructed, but interestingly, we obtain stable estimates of the polymerase dwell time (see Materials and Methods and Supplementary Table S1).

The results of the method comparison in terms of accuracy are shown in Figure 9. They show that BurstDECONV is robust and can be applied to all the datasets and time resolutions. Because of combinatorial issues described above, HMM method may fail in some cases (by memory overflow or execution timeout).

Whenever comparison was possible, for two-state datasets we found that the parameter reconstruction by BurstDECONV is significantly more accurate than by the other methods.

Some three-state datasets (D9,12–14) have very small switching rates (k^+ or k^- or both). In this case, the precision of BurstDECONV is limited by the length of the short movie. Then, the simple deconvolution method can generate large errors and the 'mixed' version of BurstDECONV, that combines short and long movies, is needed. Interestingly, the HMM method seems to be slightly less sensitive to the same phenomenon. Although large, the estimation errors of HMM are smaller than those of the simple BurstDECONV, for datasets D13 and D14 (Figure 9). However, in such difficult cases, the 'mixed' version of BurstDECONV significantly supersedes in accuracy all the other methods.

Testing BurstDECONV using an enriched collection of datasets

The datasets of Table 1 span a large parameter range but the parametric resolution is poor. In order to increase this resolution, we generated more parameter sets by latin hypercube sampling.

We generated 40 more short movie synthetic datasets corresponding to two states models. The parameter values where defined by latin hypercube sampling in linear (20 datasets) and logarithmic (20 datasets) scales.

We also set up 240 more datasets for three states models. These correspond to 60 parameter sets obtained by latin hypercube sampling in linear (30 datasets) and logarithmic (30 datasets) scales. We doubled the number of parameter sets by including both M2 and M1 three state models with parameters corresponding to the same theoretical survival function. Finally, the three states datasets were produced in two versions, simple (short movie) and mixed (short and long movies).

We have used BurstDECONV to reconstruct the parameters of these extra 280 synthetic datasets that add to those already presented in Table 1. The parameter values for these datasets can be found in the Supplementary Table S2. Figure 10 illustrates the result of these numerical experiments.

The initiation rate parameter k_{ini} is accurately reconstructed for all models and datasets (Figure 10A and C). Indeed, this parameter scales inversely with the signal amplitude and is robust with respect to signal sampling. The lack of robustness of k_{ini} against calibration was illustrated in Figure 6C.

In contrast to k_{ini}, the switching parameters can be inaccurately reconstructed using the simple version of BurstDECONV. We have identified two main sources of error. First, the reconstruction error is large when the lifetimes of the ON and OFF states are larger than the movie duration or, equivalently, when k^+ or k^- are smaller than the inverse of the short movie length. This effect is illustrated in Figure 10A–D. Second, when the lifetime of one of the OFF states becomes comparable to the interval between successive initiation events ($1/k_{ini}$) there is parametric uncertainty, as a model with less states fits equally well in this case. This effect, leading to large errors when k_1^+ or k_2^+ are large and close to k_{ini} is illustrated in Figure 10B, D, F.

As expected, the use of the mixed version of BurstDECONV (short and long movies) allows the reconstruction of very small switching parameters, corresponding to timescales larger than the length of the short movie (see Figure 10E). By using the mixed version, the error due to large lifetimes of ON and OFF states can be avoided.

DISCUSSION

While the development of imaging-based methods to monitor transcription in live cells and animals boomed over the last 20 years, the analytical frameworks extracting quantitative information from transcriptional bursting remained limited. Hence promoter switching was often modeled using ad hoc burst definitions, or using two states random telegraph model and rarely envisaging more complex models (18–20,39). Two main methods, namely the auto-correlation and the Hidden Markov Model (HMM) methods were employed in analysing transcriptional bursting data. However, there is no comparative benchmarking of the accuracy and robustness of these methods.

Here we provide BurstDECONV, a novel signal deconvolution method able to retrieve single polymerase initiation events from single cell transcription bursts and infer promoter states and their switching rates. We comparatively benchmark our method to the other two state of the art methods.

Our method is robust with respect to polymerase speed, signal calibration, time resolution, movie duration, and noise in the signal. The method is precise for wide values of kinetic parameters of the transcription regulation processes. By combining short and long movies, we are able to quantify processes with timescales from seconds to days. This extremely wide dynamic range was not accessible with the previous quantitative live cell transcription imaging approaches.

Thus, our method and tools are of interest in applications where a precise description of rate-limiting steps governing transcription dynamics is important: zygotic genome activation in model organisms, various aspects of gene expression regulation in human cells and tissues in health and disease, various studies of stochastic gene expression in prokaryotes and eukaryotes. Beyond transcription studies, they can be used for other applications where the signal can be deconvoluted into individual initiation events, for instance in studies of translation.

Another advantage of BurstDECONV resides in its ability to directly bridge agnostic representations of data to

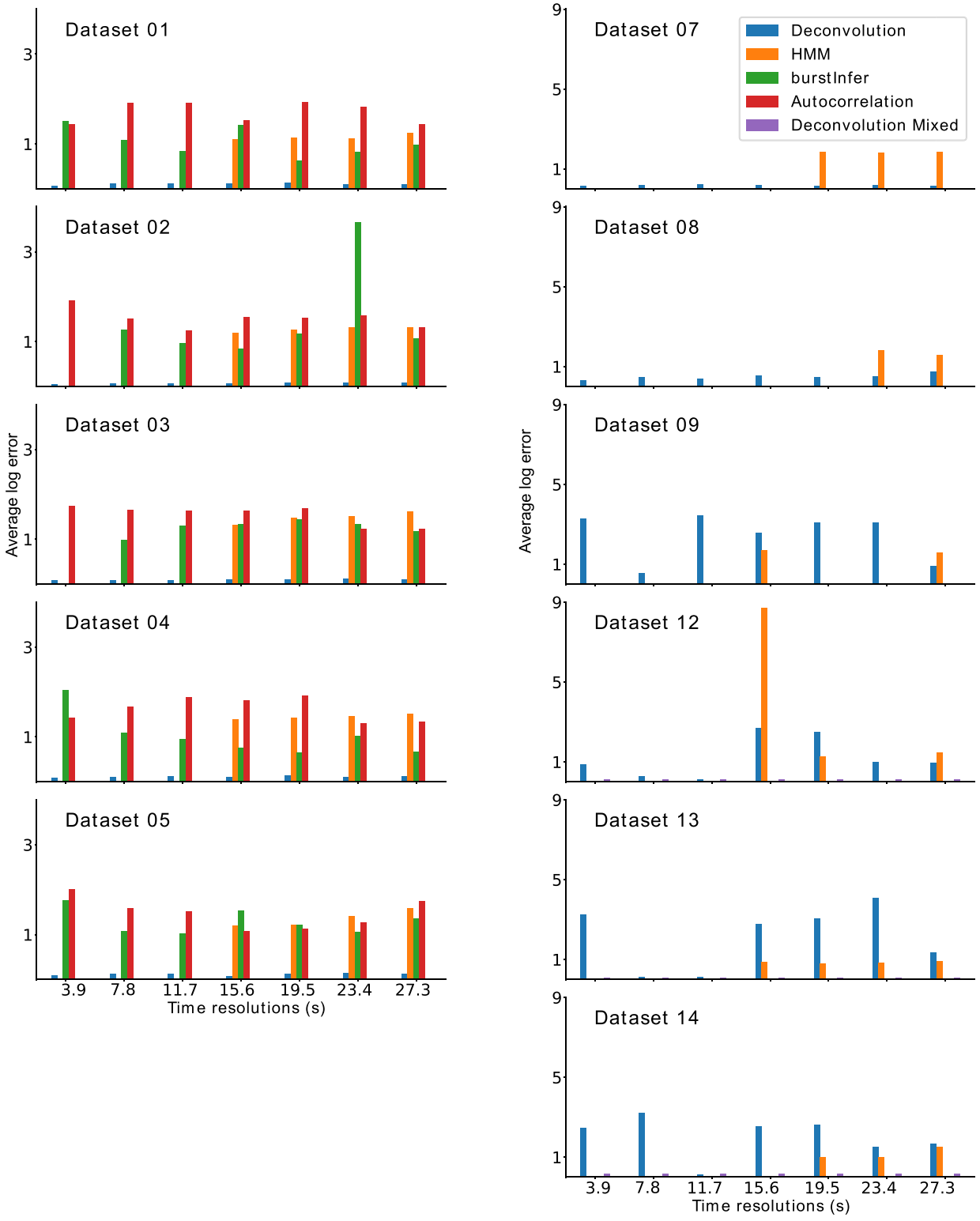


Figure 9. Parametric reconstruction accuracy of BurstDECONV vs. autocorrelation and HMM methods The bar plots on the left correspond to the parameter reconstruction errors for the four methods, BurstDECONV, cphmm, burstInfer (based on hmm) and Autocorrelation for datasets 1–5. These datasets correspond to a two state promoter model. The bar plots on the right depict the errors for BurstDECONV and for cphmm (datasets 7–9, 12–14 (3-state models)). BurstDECONV mixed refers to the deconvolution method combining high and low resolution movies. The x-axis for the plots have different time resolution of the short movies and the y-axis, the average log error (base 10). BurstDECONV mixed used short movies of resolution 3 s and long movies of resolution 3 min.

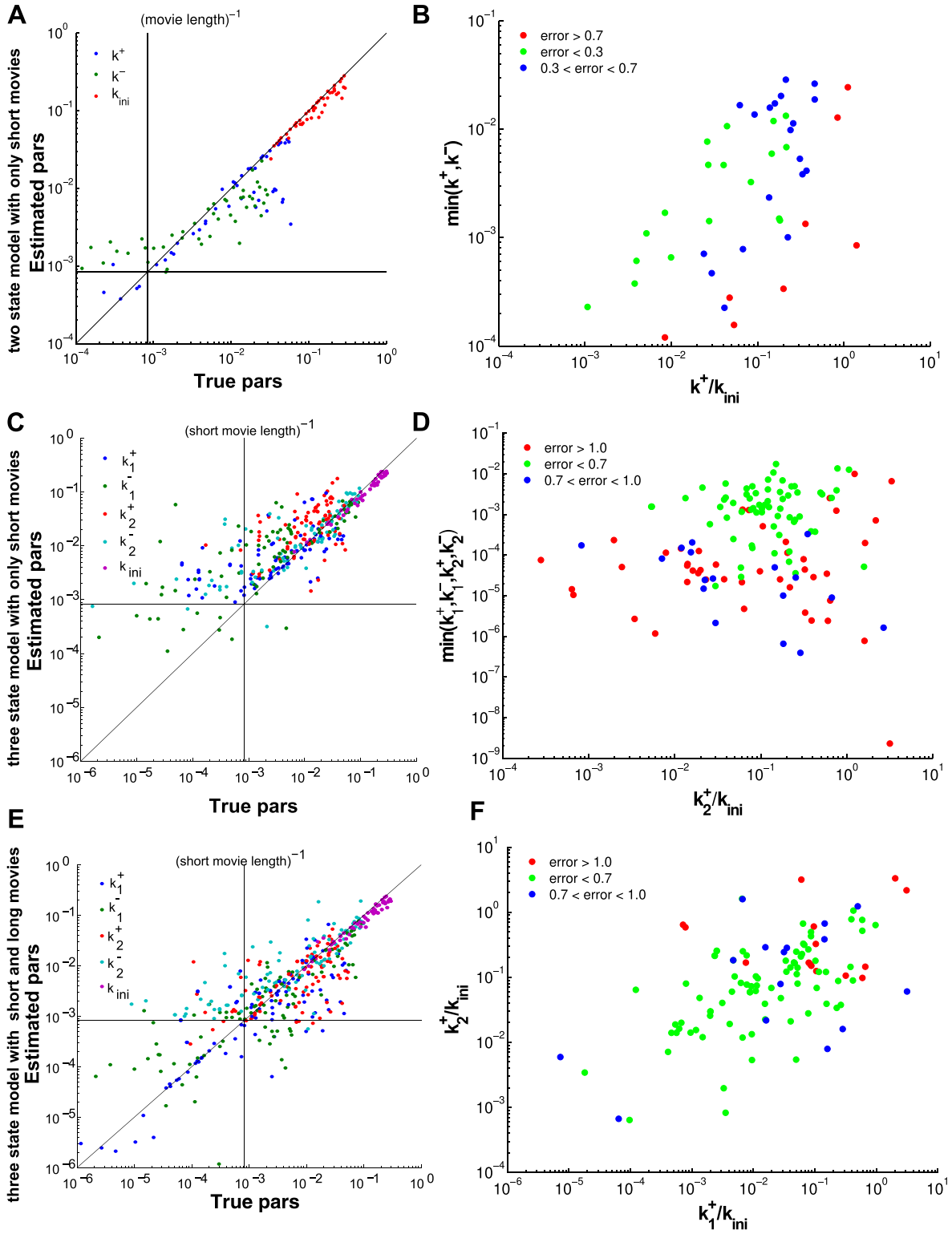


Figure 10. Parametric reconstruction accuracy for an enriched collection of 297 synthetic datasets. (**A**) Estimated vs. true parameters for 45 datasets generated with two states models. (**B**) Error versus parameters (all 2-state datasets); the mean logarithmic error is large when k^+ or k^- are small or when k^+ is close to k_{ini}. (**C**) Estimated versus true parameters for three states models using the simple version of BurstDECONV. Only datasets with error <1 are shown (81 out of 126). (**D**) Error versus parameters (all 3-state datasets); similarly to 2-state models, the error is large for small k_i^+ or k_i^-, $i \in \{1, 2\}$, or when k_1^+ or k_2^+ is close to k_{ini}. (**E**) Estimated versus true parameters for three states models using the mixed version of BurstDECONV. Only datasets with error <1 are shown (97 out of 126). (**F**) Error versus parameters (all 3-state datasets); the error is large when either k_1^+ or k_2^+ is close to k_{ini}.

kinetic parameters of discrete state models of transcription. This is not possible in the framework of the HMM method, where each model has to be fitted separately using a different likelihood function. The survival function used in BurstDECONV conveys different information than the auto-correlation function used in previous methods. This renders the two methods complementary. BurstDECONV can not determine the polymerase dwell time, but provides accurate estimates of the transition rate parameters. The autocorrelation method can estimate the dwell time, but is imprecise on the transition rates.

BurstDECONV can be extended to consider more complex transcriptional bursting models, with arbitrary number of states and transition schemes, or with multiple non-resolvable active sites resulting, for instance, from sister chromatids.

In its current setting our method considers that transcription sites are statistically equivalent. This assumption is valid when there is limited spatial and temporal heterogeneity. However, a segmentation step could be easily added to the image analysis in order to select statistically equivalent sites in the case of spatial or temporal heterogeneity. This is the case in a multicellular organism, where gene expression is submitted to positional information like Drosophila patterning instructed by gradients of morphogens.

The output of BurstDECONV is a set of promoter states and the transition rates between these states. The quantitative framework proposed in this study reveals the key bottlenecks responsible for the promoter switching dynamics. Moreover, by informing on the timescale of each rate-limiting step, BurstDECONV provides a hint on the nature of these rate limiting steps. We foresee that with the development of novel perturbation methods (as for example optogenetics), the molecular characterization of these steps will be more and more facilitated.

In addition, our stochastic models of transcription dynamics can be readily used to test mechanistic hypotheses. For example, by applying BurstDECONV to two biological systems, HIV-1 transcription in Hela cells and zygotic transcription in Drosophila embryos, we came to the conclusion that a classical view of polymerase pausing may not be accurate. Indeed, a scenario where all polymerase would experience a discernable paused state was not compatible with our data. This analysis led us to propose a new view of pausing, a non-obligatory pausing model, where only a subset of polymerase would experience stable pausing, whereas other initiated polymerases would not be kinetically limited by such long pauses (19,20). Thus, monitoring transcription in live cells and employing rigorous analytical framework, could in some cases affect our classical view of the transcription process, often raised from biochemical in vitro and static approaches.

DATA AVAILABILITY

The artificial data as well as the code used for benchmarking the pipeline are available on Zenodo at https://zenodo.org/record/7438759. BurstDECONV source code is available in both MATLAB and Python 3 versions under 3-clause BSD open license. BurstDECONV is also available as a Graphical User Interface. The source codes are available through Github at https://github.com/oradules/BurstDECONV. For increased portability, we have created a Docker container for the Python notebook. Instructions for using this container can be found in the same Github repository. The GUI and the user manual are available on Zenodo at https://zenodo.org/record/7443044. BurstDECONV does not include the image analysis part of the pipeline. This can be performed with MS2-Quant https://bitbucket.org/muellerflorian/ms2_quant/src/master/ for cell line movies, segment-track https://github.com/ant-trullo/SegmentTrack_v4.0 for Drosophila movies, or with any other equivalent software.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

RT is financed by the LabMuse PhD programme of the LABEX Epigenmed. MD is financed by the CNRS and University of Chicago PhD Joint Programme. ML is supported by ERC funding (SyncDev and LightRNA2Prot) and by the CNRS. OR and EB are financed by ANRS (project ANRS0068). We thank Dan Larson, Antoine Coulon, and Aleksandra Walczak for very useful discussions and for sharing with us their autocorrelation codes. We thank John Reinitz and Virginia Pimmett for careful reading of our manuscript and for their very helpful feedback. We thank Antonio Trullo for help with the implementation of the graphical user interface.

FUNDING

Université de Montpellier; Centre National de la Recherche Scientifique, CNRS; European Research Council, ERC; Agence Nationale de Recherches sur le Sida et les Hépatites Virales, ANRS. Funding for open access charge: Université de Montpellier.
Conflict of interest statement. None declared.

REFERENCES

1. Bertrand,E., Chartrand,P., Schaefer,M., Shenoy,S.M., Singer,R.H. and Long,R.M. (1998) Localization of ASH1 mRNA particles in living yeast. *Mol. Cell*, **2**, 437–445.
2. Chubb,J.R., Trcek,T., Shenoy,S.M. and Singer,R.H. (2006) Transcriptional pulsing of a developmental gene. *Curr. biol.*, **16**, 1018–1025.
3. Ozbudak,E.M., Thattai,M., Kurtser,I., Grossman,A.D. and Van Oudenaarden,A. (2002) Regulation of noise in the expression of a single gene. *Nat. Genet.*, **31**, 69–73.
4. Raser,J.M. and O'Shea,E.K. (2004) Control of stochasticity in eukaryotic gene expression. *Science*, **304**, 1811–1814.
5. Cai,L., Friedman,N. and Xie,X.S. (2006) Stochastic protein expression in individual cells at the single molecule level. *Nature*, **440**, 358–362.
6. Chong,S., Chen,C., Ge,H. and Xie,X.S. (2014) Mechanism of transcriptional bursting in bacteria. *Cell*, **158**, 314–326.
7. Nicolas,D., Phillips,N.E. and Naef,F. (2017) What shapes eukaryotic transcriptional bursting? *Mol. BioSyst.*, **13**, 1280–1290.
8. Tunnacliffe,E. and Chubb,J.R. (2020) What is a transcriptional burst? *Trends Genet.*, **36**, 288–297.

9. Sanchez,A., Garcia,H.G., Jones,D., Phillips,R. and Kondev,J. (2011) Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Comput. Biol.*, **7**, e1001100.
10. Sanchez,A. and Golding,I. (2013) Genetic determinants and cellular constraints in noisy gene expression. *Science*, **342**, 1188–1193.
11. Jones,D.L., Brewster,R.C. and Phillips,R. (2014) Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, **346**, 1533–1536.
12. Zoller,B., Little,S.C. and Gregor,T. (2018) Diverse spatial expression patterns emerge from unified kinetics of transcriptional bursting. *Cell*, **175**, 835–847.
13. Bharucha-Reid,A.T. (1960) Elements of the Theory of Markov Processes and their Applications, McGraw-Hill Inc., USA.
14. Peccoud,J. and Ycart,B. (1995) Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.*, **48**, 222–234.
15. Ferguson,M.L., Le Coq,D., Jules,M., Aymerich,S., Radulescu,O., Declerck,N. and Royer,C.A. (2012) Reconciling molecular regulatory mechanisms with noise patterns of bacterial metabolic promoters in induced and repressed states. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 155–160.
16. Lionnet,T. and Singer,R.H. (2012) Transcription goes digital. *EMBO Rep.*, **13**, 313–321.
17. Suter,D.M., Molina,N., Gatfield,D., Schneider,K., Schibler,U. and Naef,F. (2011) Mammalian genes are transcribed with widely different bursting kinetics. *Science*, **332**, 472–474.
18. Tantale,K., Mueller,F., Kozulic-Pirher,A., Lesne,A., Victor,J.-M., Robert,M.-C., Capozi,S., Chouaib,R., Bäcker,V., Mateos-Langerak,J. *et al.* (2016) A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting. *Nat. Commun.*, **7**, 12248.
19. Tantale,K., Garcia-Oliver,E., Robert,M.-C., L'Hostis,A., Yang,Y., Tsanov,N., Topno,R., Gostan,T., Kozulic-Pirher,A., Basu-Shrivastava,M. *et al.* (2021) Stochastic pausing at latent HIV-1 promoters generates transcriptional bursting. *Nat. Commun.*, **12**, 4503.
20. Pimmett,V.L., Dejean,M., Fernandez,C., Trullo,A., Bertrand,E., Radulescu,O. and Lagha,M. (2021) Quantitative imaging of transcription in living Drosophila embryos reveals the impact of core promoter motifs on promoter state dynamics. *Nat. Commun.*, **12**, 4504.
21. Sánchez,Á. and Kondev,J. (2008) Transcriptional control of noise in gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 5081–5086.
22. Innocentini,G. d. C.P., Forger,M., Ramos,A.F., Radulescu,O. and Hornos,J.E.M. (2013) Multimodality and flexibility of stochastic gene expression. *Bull. Math. Biol.*, **75**, 2600–2630.
23. Vos,S.M., Farnung,L., Urlaub,H. and Cramer,P. (2018) Structure of paused transcription complex Pol II–DSIF–NELF. *Nature*, **560**, 601–606.
24. Vos,S.M., Farnung,L., Boehning,M., Wigge,C., Linden,A., Urlaub,H. and Cramer,P. (2018) Structure of activated transcription complex Pol II–DSIF–PAF–SPT6. *Nature*, **560**, 607–612.
25. Rodriguez,J. and Larson,D.R. (2020) Transcription in living cells: molecular mechanisms of bursting. *Annu. Rev. Biochem.*, **89**, 189–212.
26. Krebs,A.R., Imanci,D., Hoerner,L., Gaidatzis,D., Burger,L. and Schübeler,D. (2017) Genome-wide single-molecule footprinting reveals high RNA polymerase II turnover at paused promoters. *Mol. Cell*, **67**, 411–422.
27. Roeder,R.G. (2019) 50+ years of eukaryotic transcription: an expanding universe of factors and mechanisms. *Nat. Struct. Mol. Biol.*, **26**, 783–791.
28. Osman,S. and Cramer,P. (2020) Structural biology of RNA polymerase II transcription: 20 years on. *Annu. Rev. Cell Dev. Biol.*, **36**, 1–34.
29. Patel,A.B., Greber,B.J. and Nogales,E. (2020) Recent insights into the structure of TFIID, its assembly, and its binding to core promoter. *Curr. Opin. Struct. Biol.*, **61**, 17–24.
30. Rengachari,S., Schilbach,S., Aibara,S., Dienemann,C. and Cramer,P. (2021) Structure of the human Mediator–RNA polymerase II pre-initiation complex. *Nature*, **594**, 129–133.
31. Fianu,I., Chen,Y., Dienemann,C., Dybkov,O., Linden,A., Urlaub,H. and Cramer,P. (2021) Structural basis of Integrator-mediated transcription regulation. *Science*, **374**, 883–887.
32. Coulon,A. and Larson,D.R. (2016) Fluctuation analysis: dissecting transcriptional kinetics with signal theory. In: *Methods in enzymology*. Elsevier, Vol. **572**, pp.159–191.
33. Desponds,J., Tran,H., Ferraro,T., Lucas,T., Romero,C.P., Guillou,A., Fradin,C., Coppey,M., Dostatni,N. and Walczak,A.M. (2016) Precision of readout at the hunchback gene: analyzing short transcription time traces in living fly embryos. *PLoS Comput. Biol.*, **12**. e1005256.
34. Lammers,N.C., Galstyan,V., Reimer,A., Medin,S.A., Wiggins,C.H. and Garcia,H.G. (2020) Multimodal transcriptional control of pattern formation in embryonic development. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 836–847.
35. Lammers,N.C., Kim,Y.J., Zhao,J. and Garcia,H.G. (2020) A matter of time: using dynamics and theory to uncover mechanisms of transcriptional bursting. *Curr. Opin. Cell Biol.*, **67**, 147–157.
36. Bowles,J.R., Hoppe, C., Ashe, H.L. and Rattray,M. (2021) Scalable inference of transcriptional kinetic parameters from MS2 time series data. *Bioinformatics (Oxford, England)*, **38**, 1030–1036.
37. Hougaard,P and . (2000) In: *Analysis of Multivariate Survival Data*. Springer, Vol. **564**.
38. Pichon,X., Lagha,M., Mueller,F. and Bertrand,E. (2018) A growing toolbox to image gene expression in single cells: sensitive approaches for demanding challenges. *Mol. Cell*, **71**, 468–480.
39. Corrigan,A.M., Tunnacliffe,E., Cannon,D. and Chubb,J.R. (2016) A continuum model of transcriptional bursting. *Elife*, **5**, e13051.
40. Wissink,E.M., Vihervaara,A., Tippens,N.D. and Lis,J.T. (2019) Nascent RNA analyses: tracking transcription and its regulation. *Nat. Rev. Genet.*, **20**, 705–723.
41. Liu,Y. (2022) In: *Non-Parametric Bayesian Inference with Application to System Biology*. PhD thesis, Department of Statistics, University of Chicago.
42. Ferguson,M.L. and Larson,D.R. (2013) Measuring transcription dynamics in living cells using fluctuation analysis. In: *Imaging Gene Expression*. Springer, pp. 47–60.

CHAPTER 3

Inferring stochastic gene expression bursting mechanisms from time-to-event data

3.1.0 Introduction

In Chapter 2 we have used phase-type distributions for inferring transcription regulation mechanisms from live transcription imaging data Tantale et al. (2021); Pimmett et al. (2021). However, in this Chapter we intend to elaborate the inverse problem method developed in the previous section in order to find solutions for more complex models.

Many problems in biology, medicine, physics, chemistry, economy, actuarial science can be modelled by using continuous time Markov chains Bharucha-Reid (1997). A common dataset produced by such models comprises a sequence of states that denote the incidence of specific events within the system. When all states within the model are observed, the time until the occurrence of the next event follows the memoryless exponential distribution. However, in many practical applications, not all the states of the model are observed. In this case the time to the next event is no longer exponentially distributed.

A particular case is when only one state of the model can be observed. In this case the time

to the next event distribution is a so-called *phase-type distribution*. Phase-type distributions were introduced by Neuts Neuts (1975) and are defined as the distribution of the time to absorbtion of a finite Markov chain when there is only a single absorbing state. For a single observed state, computing the time-to-event distribution is equivalent to computing a phase-type distribution (see Figure 3.1). Phase-type distributions have been used in various domains such as queueing theory Ramaswami and Neuts (1980); Ramaswami and Lucantoni (1985), drug kinetics Faddy (1993), public health problems Fackrell (2009); Liquet et al. (2012); Asanjarani et al. (2021); Stone et al. (2022). In spite of their importance for applications it is surprising how little is known about the general properties of phase-type distributions. Algorithmic calculations of phase-type distributions can be found in Asmussen and Bladt (1996). For some general results and open questions one could also see O'Cinneide (1999); Commault and Mocanu (2003); Bladt (2005).

In this paper we chose an inference perspective where we aim determining the Markov chain from the phase-type distribution. This inverse problem has already been addressed using methods based on maximum likelihood. In this approach, data likelihood is computed for a chosen model and its maximization provide the optimal model parameters Asanjarani et al. (2021); Stone et al. (2022). In contrast to this direct inference approach, here we decompose the inference problem into two parts. The first part consists in the regression of parametric multi-exponential representations of the phase-like distributions. This part has been developed in BurstDeconv and reviewed elsewhere Dufresne (2007) and will not be developed here. The second part consists in solving symbolically the inverse problem consisting in finding the Markov chain transition rate parameters from the parameters of the multi-exponential distributions. This approach has several advantages. It allows us to compute simultaneously all the Markov chain models, having different transition parameters and diagrams, but exactly the same likelihood. It is thus a precious tool for analysing parametric and structural uncertainty in inference.

Although we chose to formulate our problem in the context of transcriptional bursting, highly regarded in the gene expression research community, our approach can be applied to other applications that involve phase-type distributions: modeling failure of systems and components in reliability engineering, traffic flow characteristics in traffic engineering, optimization of queueing systems and telecommunication networks, claim inter-arrival times in insurance and actuarial science, survival times in epidemiology and medical research Aalen (1995); Fackrell (2009); Buchholz et al. (2014).

The structure of this chapter is the following. In Section 3.2 we introduce the class of models used for transcription bursting and show that for this class of models the phase-type distribution is multi-exponential. In Section 3.3 we introduce the symbolic inverse problem

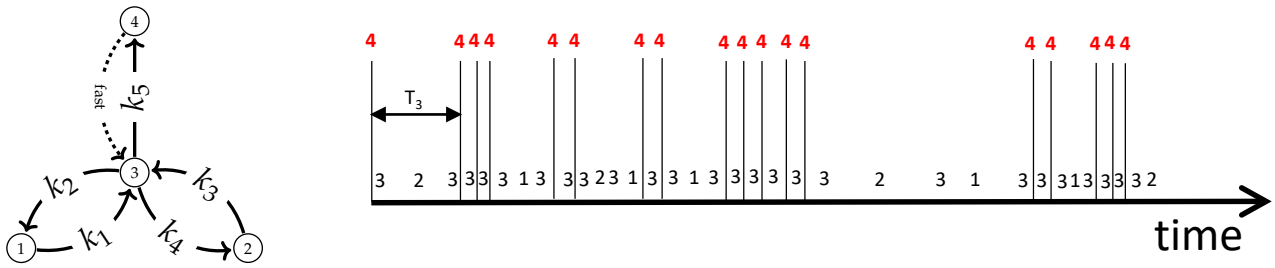


Figure 3.1: Phase-type distribution for a three state Markov chain model. 4 represents the unique observed state. A typical dataset is represented by the moments when the state 4 is observed. One does not know what's going on between two successive occurrences of 4 but can measure the time T_3 separating them. The probability distribution of this time is called phase-type distribution. The Markov model is represented by its transition diagram (arcs connecting states) and by the transition rate parameters. It is considered that from 4 the chain returns to 3 instantly. For this reason, T_3 is the first hitting time of 4 starting from 3. The timechart represents repeated observation of the state 4 for the same system. It could be equivalently seen as a succession of replicas of a system that stops when it reaches the state 4. The latter interpretation corresponds to the traditional definition of the phase-type distribution.

consisting in finding the transition rate parameters of the model from the parameters of the phase-type distribution. We show that the inverse problem consists in solving a system of polynomial equations, symmetric in the phase-type distribution parameters. We also discuss solvable models, i.e. for which the inverse problem has a unique solution. In section 3.4 we discuss an example of solvable model with arbitrary number of states. In section 3.5 we use Thomas decomposition from computational algebra to solve the inverse problem.

3.2.0 Modeling transcription bursting

3.2.1 Biological considerations

Transcription is a cellular process resulting in RNA synthesis. It involves a DNA sequence called promoter that binds the RNA polymerase (RNAP),the enzyme that synthesizes RNA. Transcription needs several preparatory steps such as chromatin opening and formation of transcription complexes needed for activation of the polymerase. Here we consider that there is a small number of limiting steps that can be modeled as transition between discrete states. We classify these states in three categories: productive ON states that can initiate transcription, non-productive OFF states that can not initiate nor resume transcription, and paused states PAUSE in which initiated transcription stops and can resume later or abort.

The promoter starts transcription in the productive state ON when it can trigger several departures of RNAP molecules along DNA. The RNAP can eventually stop in PAUSE or commit to irreversible elongation that we model by the state EL. Elongating polymerases

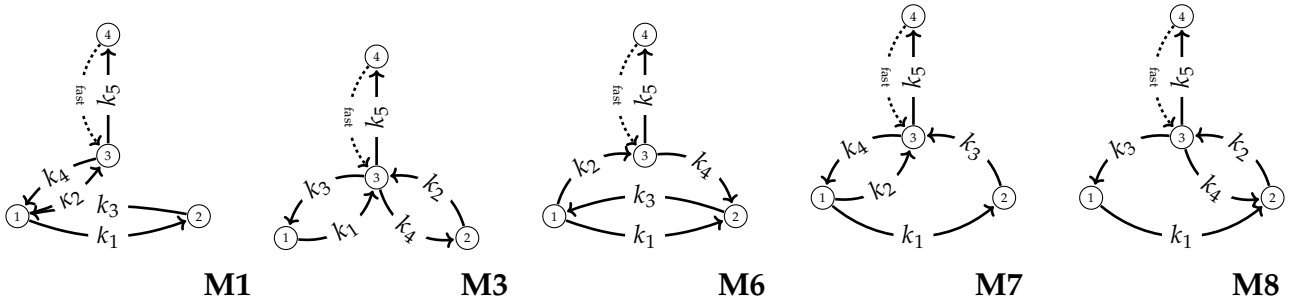


Figure 3.2: All strongly ergodic models with $N = 3$. We show only those models that are not related to one another by permutations of the states $1, \ldots, N$. All these models can generate exactly the same phase-type distribution. Model M3 is symmetric and generates the same phase-type distribution if its parameters are transformed by the permutation $k_1 \leftrightarrow k_2, k_3 \leftrightarrow k_4$ (the inverse problem has two solutions). Similarly, the inverse problem has two solutions for M7 and M8, whereas the solution is unique for M1. Models M_1, M_7, M_8, M_3 are solvable (the inverse problem has a unique or a finite number of solutions) and the model $M6$ is not solvable (the inverse problem has infinitely many solutions). Models generating the same phase-type distributions spend the same proportion of the total time in state 3. However, they could be discriminated by the proportion of the total time spent in the states 1 and 2 (any comparison between M7,M8,M3) or by the lifetimes of the states 1 and 2 or by both types of times (M1 compared to any other).

free the promoter that switches instantly to a productive or non-productive state.

Live transcription imaging allows the monitoring of transcription in real time and for each transcription site. Using the tool BurstDeconv we can deconvolve the live imaging signal and identify the time of each elongation event. This means that for each transcription site we observe the sequence of EL states. The other states of the promoter are not observed.

In this paper we consider that EL can be reached only from an unique state that can be either ON or PAUSE, and after reaching EL the promoter instantly switches to an unique return state S, that can be either productive or non-productive. This is consistent with models without pause or models where pausing is obligatory after initiation. When pausing is not obligatory it implies that pausing leads to systematic abortion. This important assumption simplifies the problem for two, closely related reasons. First, because elongation is instantly followed by the transition to the S state, therefore the observation of the elongation is equivalent to the observation of the S state. Second, the times of successive observations of elongations form a renewal process. For these reasons, the distribution of the waiting time between successive elongation events is a phase-type distribution.

Moreover, we demonstrate that the inverse problem for phase-type distributions is well-defined only when the productive and return states coincide, thereby ruling out the presence of multiple ON states. Due to this, our paper does not address scenarios in which transcription occurs simultaneously on duplicate DNA copies formed during recombination, a phe-

nomenon referred to as sister chromatids. This specific scenario, which could be effectively modeled using mixtures of phase-type distributions, will be addressed elsewhere.

3.2.2 Finite state continuous time Markov chain model

We model the promoter using a continuous time Markov chains with N+1 states $M(t) \in \{1, \ldots, N+1\}$. The first N states can be of the type ON, OFF, or PAUSE. A supplementary state $N+1$ designates EL, the event that we observe. For simplicity, we consider that $N+1$ can only be reached from N. After $N+1$ there is instantaneous transition to the state $s \in [1, N]$.

The chain is defined by its generator $\boldsymbol{Q}$, a $(N+1) \times (N+1)$ matrix such that $Q_{ij} > 0$ represents the probability per unit time to jump from the state A_i to the state A_j, $j \neq i$, and $Q_{ii} = -\sum_{j \neq i} Q_{ij}$ (zero row sum rule).

We are interested in the distribution of T_s, the first time to reach $N+1$ starting from the state $s, 1 \leq s \leq N$:

$$T_s = \inf\{t | t > 0, M(t) = N+1, M(0) = s\},$$

where $M(t)$ is the state of the chain at time t.

Let us consider two models. In the *model with return*, once in $N+1$, the Markov chain returns to s instantly, $Q_{N+1,s} = \infty$ and then the chain starts again. Because the reset state s is unique, the observation times in the model with return form a renewal process (the inter-event times are positive, independent, identically distributed random variables Feller (1966)). Because after each event, the model is instantaneously reset to s the inter-event time is always distributed as T_s.

In the *model with absorption*, $N+1$ is absorbing and $Q_{N+1,i} = 0$ for all $i \neq N+1$. In this case T_s is the time to absorption starting from s.

The distribution of T_s is the same for the two models, with return and with absorption.

If several return states are possible, then the model with return is no longer a renewal process. The distribution of the inter-event time depends on the return state and is no longer a phase-type distribution.

Although the model with return describes the biological situation, for technical reasons, we will use the model with absorption for computing the distribution of T_s.

For example, with the absorption assumption, the generator of the model M3 represented in the Figure 3.2 is

$$Q = \begin{pmatrix} -k_1 & 0 & k_1 & 0 \\ 0 & -k_3 & k_3 & 0 \\ k_2 & k_4 & -(k_2+k_4+k_5) & k_5 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.1}$$

For the calculation of the distribution of T_s it is convenient to introduce the state probabilities $X_i = \mathbb{P}[M(t) = i | M(0) = s], 1 \leq i \leq N+1$.

The variables $X_i(t)$ satisfy the following system of linear differential equations (the master equation) :

$$\frac{d\boldsymbol{X}}{dt} = \boldsymbol{Q}^T \boldsymbol{X}, \tag{3.2}$$

with the initial conditions $X_n(0) = \delta_{n,s}$, where δ is the Kronecker symbol and Q^T is the transpose of the generator matrix Q.

Because the last column of $\boldsymbol{Q}^T$ is zero ($N+1$ is absorbing), the variables $X_1, \ldots, X_N$ satisfy an autonomous ODE system. Indeed, let $\tilde{\boldsymbol{Q}}$ be the $N \times N$ matrix obtained by eliminating the last line and the last column of $\boldsymbol{Q}^T$. Then $\tilde{\boldsymbol{X}} = (X_1, \ldots, X_N)$ is the solution of

$$\frac{d\tilde{\boldsymbol{X}}}{dt} = \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{X}}, \tag{3.3}$$

with initial conditions $X_i(0) = \delta_{i,s}$.

The remaining variable X_{N+1} satisfies

$$\frac{dX_{N+1}}{dt} = Q_{N,N+1} X_N, \tag{3.4}$$

with the initial condition $X_{N+1}(0) = 0$.

Let us consider that the eigenvalues $\lambda_i, 1 \leq i \leq N$ of the matrix $\tilde{\boldsymbol{Q}}$ satisfy the following non-degeneracy condition

$$\lambda_i \neq \lambda_j, \text{ for all } 1 \leq i \neq j \leq N, \text{ and } \max_{1 \leq i \leq N}(\lambda_i) < 0. \tag{3.5}$$

Remark 1. *Using properties of solutions of* (3.3) *it can be shown that* $\max_{1 \leq i \leq N}(\lambda_i) \leq 0$ *is always satisfied. Indeed,* $X_i(t), 1 \leq i \leq N$ *are probabilities, therefore remain bounded for all* t*, which means*

that one can not have $\lambda_i > 0$ for some $1 \leq i \leq N$. If some $\lambda_i > 0$, then there are solutions of (3.3) *that increase without bound when t increases. We will show later that strong connectedness of the transition graph reduced to the vertices $1, \ldots, N$ implies that none of the eigenvalues λ_i can be zero. One should note that, unlike Q, $\tilde{Q}^T$ is not a continuous time Markov chain generator and does not satisfy the zero row sum rule.*

Considering that the condition (3.5) is satisfied, we have

$$\tilde{\boldsymbol{X}} = \sum_{i=1}^{N} C_i \boldsymbol{u}_i e^{\lambda_i t},\ X_{N+1} = Q_{N,N+1} \sum_{i=1}^{N} \frac{C_i u_{i,N}}{\lambda_i} (e^{\lambda_i t} - 1), \tag{3.6}$$

where λ_i and $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{i,N})^T$, $i \in [1, N]$ are eigenvalues and eigenvectors of $\tilde{\boldsymbol{Q}}$, respectively.

Because $N+1$ is absorbing, $X_{N+1} = \mathbb{P}[M(t) = N+1 | M(0) = s] = \mathbb{P}[T_s \leq t]$ is the cumulative distribution function of the time T_s. We also define the survival function of T_s as follows

$$S(t) = \mathbb{P}[T_s > t] = 1 - X_{N+1}(t) = \sum_{i=1}^{N} A_i exp(\lambda_i t), \tag{3.7}$$

where

$$A_i = -\frac{Q_{N,N+1} C_i u_{i,N}}{\lambda_i}. \tag{3.8}$$

The constants A_i are not independent. They satisfy

$$\sum_{i=1}^{N} A_i = 1, \tag{3.9}$$

which follows from $s(0) = 1$.

This shows that in the non-degenerate case the distribution of the first hitting time T_s is multi-exponentional with $2N - 1$ independent parameters $\lambda_1, \ldots, \lambda_N, A_1, \ldots, A_{N-1}$.

3.3.0 Inverse problem

The inverse problem consists in considering that the survival function and therefore the parameters $\lambda_1, \ldots, \lambda_N, A_1, \ldots, A_{N-1}$ are known. These can be inferred from data using least squares or maximum likelihood regression.

We want to find the transition rate parameters, which are the non-diagonal elements of the matrix Q. We are interested in models where this problem could have a unique solution, therefore the matrix Q has only $2N-1$ non-zero elements. We denote these transition rates by the $2N-1$ dimensional vector $\boldsymbol{k}$.

We show below that the inverse problem consists in solving $2N-1$ polynomial equations for the transition rates $\boldsymbol{k}$.

3.3.1 Vieta's formulas

The eigenvalues $\lambda_1, \ldots, \lambda_N$ are the roots of the the characteristic polynomial of $\tilde{\boldsymbol{Q}}$, defined as

$$P(\lambda) = det(\tilde{\boldsymbol{Q}} - \lambda \boldsymbol{I}) = (-1)^N \lambda^N + a_{N-1}(\boldsymbol{k})\lambda^{N-1} + \ldots + a_1(\boldsymbol{k})\lambda + a_0(\boldsymbol{k}). \tag{3.10}$$

The coefficients a_i of the characteristic polynomial are polynomials with integer coefficients on the transition rates $\boldsymbol{k}$.

A first set of equations relating eigenvalues to the kinetic equations results from the Vieta's formulas

$$\begin{aligned} L_1 &= \sum_{i=1}^{N} \lambda_i = (-1)^{N-1} a_{N-1}(\boldsymbol{k}), \\ L_2 &= \sum_{i<j} \lambda_i \lambda_j = (-1)^{N-2} a_{N-2}(\boldsymbol{k}), \\ &\vdots \\ L_N &= \lambda_1 \lambda_2 \ldots \lambda_N = a_0(\boldsymbol{k}). \end{aligned} \tag{3.11}$$

3.3.2 Eigenvector equations

The amplitude parameters of the survival function $A_1, A_2, \ldots, A_N$ occur in (3.8) together with eigenvector components $u_{i,N}$ and solution coefficients C_i. We need to relate the latter to transition rate parameters.

First, we need to solve the eigenvector equation

$$(\tilde{\boldsymbol{Q}} - \lambda \boldsymbol{I})\boldsymbol{u} = 0. \tag{3.12}$$

From (3.12) it follows that the eigenvector components are rational functions of λ and $\boldsymbol{k}$.

We consider that the eigenvectors can be chosen of the form $\boldsymbol{u}(\lambda, \boldsymbol{k}) = (u_1(\lambda, \boldsymbol{k}), \ldots, u_N(\lambda, \boldsymbol{k}))$ with $u_s(\lambda, \boldsymbol{k}) = 1$. In the subsection 3.3.4 we will see for which models this choice is possible. It follows that

$$u_n(\lambda, \boldsymbol{k}) = \frac{b_n(\lambda, \boldsymbol{k})}{b_s(\lambda, \boldsymbol{k})}, \quad n \in [1, N]$$

where $b_n(\lambda, \boldsymbol{k})$ are polynomials. We also choose $b_s(\lambda, \boldsymbol{k})$ prime relatively to $b_i(\lambda, \boldsymbol{k}), i \neq s$ in $\mathbb{Z}[\boldsymbol{k}, \lambda]$.

The initial conditions satisfied by the variables X_i provide a linear system of equations for the constants C_i:

$$\sum_{j=1}^{N} u_i(\lambda_j, \boldsymbol{k}) C_j = \delta_{i,s}, \quad i \in [1, N]. \tag{3.13}$$

Because of the non-degeneracy condition (3.5), if $u(\lambda, \boldsymbol{k})$ is not identically zero, then $u(\lambda_i, \boldsymbol{k}), 1 \leq i \leq N$ are independent. Therefore (3.13) has a unique solution $C_i(\boldsymbol{\lambda}, \boldsymbol{k})$, $i \in [1, N]$.

From (3.8) we obtain $N - 1$ equations for the transition rates $\boldsymbol{k}$:

$$k_{2N-1} u_N(\lambda_i, \boldsymbol{k}) C_i(\boldsymbol{\lambda}, \boldsymbol{k}) = -A_i \lambda_i, \quad i \in [1, N-1]. \tag{3.14}$$

The solution of the inverse problem is the solution of the system formed by (3.11) and (3.14). When this solution exists, the transition rates $\boldsymbol{k}$ can be expressed as functions of the survival function parameters $\lambda_i, A_i, 1 \leq i \leq N$.

3.3.3 Symmetrized system

There is a difference between (3.11) and (3.14). (3.11) is entirely expressed using elementary symmetric polynomials in λ_i, whereas there is no obvious symmetry in (3.14). We show here that the system formed by (3.11) and (3.14) is equivalent to a system symmetrized in both λ_i and A_i. The advantage of a symmetrized system over a non-symmetrized one is that it handles simpler formulas, decreasing the computational burden of the symbolic tools.

To this aim we use the following identities (due to Jacobi-Trudi Fulton and Harris (1991))

that are valid for any distinct N numbers $\lambda_i, 1 \leq i \leq N$

$$\sum_{i=1}^{N} \lambda_i^k \prod_{\substack{j=1 \\ j \neq i}}^{N} \frac{1}{\lambda_i - \lambda_j} = \begin{cases} 0, & \text{if } k < N-1 \\ 1, & \text{if } k = N-1 \\ h_{k-N+1}(\lambda_1, \ldots, \lambda_N), & \text{if } k > N-1 \end{cases}, \tag{3.15}$$

where h_{k-N+1} is the complete symmetric polynomial of degree $k - N + 1$ in N variables. More precisely

$$\begin{aligned} h_1 &= \sum_{i=1}^{N} \lambda_i, \\ &\vdots \\ h_m &= \sum_{1 \leq i_1 \leq \cdots \leq i_m \leq N} \lambda_{i_1} \lambda_{i_2} \ldots \lambda_{i_m}, \\ &\vdots \end{aligned} \tag{3.16}$$

In order to relate (3.14) and (3.15) we first related the coefficients C_i to eigenvalues. This can be done using Cramer's rule.

Denote by $D_i \in \mathbb{Z}[\lambda, \boldsymbol{k}], 1 \leq i \leq N$ the determinant of $(N-1) \times (N-1)$ submatrix of the matrix $\lambda \boldsymbol{I} - \tilde{\mathbf{Q}}$ obtained by deleting its s-th row and i-th column.

Lemma 2. *Assume that $\lambda_i \neq \lambda_j, 1 \leq i \neq j \leq N$, $D_s(\lambda_j) \neq 0$ and that $u_s(\lambda_j) = 1, 1 \leq j \leq N$. Then*

$$C_j = \frac{D_s(\lambda_j)}{\prod_{1 \leq l \neq j \leq N} (\lambda_j - \lambda_l)}, \; 1 \leq j \leq N, \tag{3.17}$$

where C_j form the unique solution of the system

$$\sum_{1 \leq j \leq N} u_s(\lambda_j) C_j = 1,$$

$$\sum_{1 \leq j \leq N} u_i(\lambda_j) C_j = 0, 1 \leq i \neq s \leq N.$$

Proof. If $D_s(\lambda_j) \neq 0$ then, using Cramer's rule $\boldsymbol{u}(\lambda_j) = (D_1(\lambda_j), \ldots, D_N(\lambda_j))/D_s(\lambda_j)$. Therefore, it holds

$$\sum_{1 \leq j \leq N} u_s(\lambda_j) C_j = \sum_{1 \leq j \leq N} \frac{D_s(\lambda_j)}{\prod_{1 \leq l \neq j \leq N} (\lambda_j - \lambda_l)} = 1,$$

while

$$\sum_{1\leq j\leq N} u_i(\lambda_j)C_j = \sum_{1\leq j\leq N} \frac{D_i(\lambda_j)}{\prod_{1\leq l\neq j\leq N}(\lambda_j-\lambda_l)} = 0,\ 1\leq i\neq s\leq N$$

(where $C_j, 1\leq j\leq N$ are taken from (3.17)) due to (3.15). □

Using (3.14) and (3.17) it follows

$$S_k = \sum_{i=1}^{N} A_i\lambda_i^k = -k_{2N-1}\sum_{i=1}^{N}\lambda_i^{k-1}\frac{D_N(\lambda_i)}{\prod_{1\leq j\neq i\leq N}(\lambda_i-\lambda_j)}. \tag{3.18}$$

For their definition, it follows that $deg_\lambda(D_s) = N-1$ and the (leading) coefficient of D_s at the monomial λ^{N-1} equals 1, while $deg_\lambda(D_i)\leq N-2, 1\leq i\neq s\leq N$. We therefore have:

$$D_s(\lambda) =: \lambda^{N-1} + c_{N-2}(\boldsymbol{k})\lambda^{N-2} + \ldots + c_1(\boldsymbol{k})\lambda + c_0(\boldsymbol{k}), \tag{3.19}$$

$$D_N(\lambda) =: d_{N-2}(\boldsymbol{k})\lambda^{N-2} + d_{N-3}(\boldsymbol{k})\lambda^{N-3} + \ldots + d_1(\boldsymbol{k})\lambda + d_0(\boldsymbol{k}),\ \text{if } N\neq s. \tag{3.20}$$

We can distinguish two cases:

$\underline{\mathbf{s}=N}$

In this case, after elongation the promoter returns to the same state. It is the case when the last state N is ON, that can initiate transcription without pausing (this may happen with pausing also, when pausing is facultative) and after initiation the promoter can recruit another polymerase immediately.

Then, according to (3.20) D_N has degree $N-1$. Using (3.18) and (3.15) we find the following $N-1$ symmetrized equations:

$$\begin{aligned}
S_1 &= \sum_{i=1}^{N} A_i\lambda_i = -k_{2N-1},\\
S_2 &= \sum_{i=1}^{N} A_i\lambda_i^2 = -k_{2N-1}(h_1(\boldsymbol{\lambda}) + c_{N-2}(\boldsymbol{k})),\\
&\vdots\\
S_{N-1} &= \sum_{i=1}^{N} A_i\lambda_i^{N-1} = -k_{2N-1}(h_{N-2}(\boldsymbol{\lambda}) + c_{N-2}(\boldsymbol{k})h_{N-3}(\boldsymbol{\lambda}) + \ldots + c_2(\boldsymbol{k})h_1(\boldsymbol{\lambda}) + c_1(\boldsymbol{k})).
\end{aligned} \tag{3.21}$$

The symmetrized system is made of (3.11),(3.21) and is used to compute the transition rates $\boldsymbol{k}$ as functions of the symmetric polynomials $L_1, \ldots, L_N, S_1, \ldots, S_{N-1}$. We should note that the complete symmetric polynomials $h_1(\boldsymbol{\lambda}), \ldots, h_{N-2}(\boldsymbol{\lambda})$ can be expressed using the elementary symmetric polynomials $L_1, \ldots, L_{N-2}$: $h_1 = L_1, h_2 = L_1^2 - L_2, \ldots$.

$\underline{\mathbf{s} < N}$

In this case, after elongation the promoter goes to another state. It is the case when the last state N is PAUSE, that is followed by a ON state. Other situations are also compatible with this case.

Then, according to (3.20) D_N has degree at most $N-2$. Using (3.18) and (3.15) it follows

$$S_1 = \sum_{i=1}^{N} A_i \lambda_i = 0. \tag{3.22}$$

This means that the inverse problem is not well posed in this case. The $2N-1$ parameters $\lambda_1, \ldots, \lambda_N$, $A_1, \ldots, A_{N-1}$ are no longer independent and they do not determine $\boldsymbol{k}$ uniquely. Although we can no longer obtain $N-1$ independent symmetrized equations we can obtain a smaller number of equations

$$\begin{aligned}
S_2 &= -k_{2N-1} d_{N-2}(\boldsymbol{k}), \\
S_3 &= -k_{2N-1}(d_{N-2}(\boldsymbol{k}) h_1(\boldsymbol{\lambda}) + d_{N-3}(\boldsymbol{k})), \\
&\vdots \\
S_k &= -k_{2N-1}(d_{N-2}(\boldsymbol{k}) h_{k-2}(\boldsymbol{\lambda}) + d_{N-3}(\boldsymbol{k}) h_{k-3}(\boldsymbol{\lambda}) + \ldots + d_{N-k}), \\
&\vdots \\
S_{N-1} &= -k_{2N-1}(d_{N-2}(\boldsymbol{k}) h_{N-3}(\boldsymbol{\lambda}) + \ldots + d_2(\boldsymbol{k}) h_1(\boldsymbol{\lambda}) + d_1(\boldsymbol{k})).
\end{aligned} \tag{3.23}$$

In this case the symmetrized system made of (3.11),(3.23) can be used to compute the transition rates $\boldsymbol{k}$ as functions of the symmetric polynomials $L_1, \ldots, L_N, S_2, \ldots, S_{N-1}$ and of one or several indeterminate transition rates.

3.3.4 Solvable models

Different Markov chain models are distinguished by their transition graph defined as the directed graph $G = (V, A)$ with vertices $V = \{1, \ldots, N+1\}$ and arcs $A = \{(i,j), i \neq j, Q_{ij} \neq 0\}$.

This graph satisfies some conditions related to the underlying biological process:

C1. $N+1$ is reachable only from N and

C2. from $N+1$ there is only one outgoing arc, towards s, where $1 \leq s \leq N$.

Conditions C1 and C2 where used for setting the problem and writing the equations defining the inverse problem.

We would like to know for which models satisfying C1 and C2 the inverse problem has a unique solution, eventually up to transformation by discrete symmetries.

Definition 3. *We say that a model is solvable if the following solutions are satisfied:*

i) *The equation* $(\tilde{\boldsymbol{Q}} - \lambda \boldsymbol{I})\boldsymbol{u} = 0$ *has solutions* $\boldsymbol{u}(\lambda, \boldsymbol{k}) = (u_1(\lambda, \boldsymbol{k}), \ldots, u_N(\lambda, \boldsymbol{k}))$ *with* $u_s(\lambda, \boldsymbol{k})$ *not identically zero where s is the return state.*

ii) *The model has* $2N-1$ *transition rate parameters.*

iii) *The system made by the equations* (3.11) *and* (3.14) *or equivalently the system made by the equations* (3.11) *and* (3.21) *has unique solutions up to transformations by discrete symmetries, on a open domain of dimension* $2N-1$.

It is very difficult to obtain general sufficient conditions for solvability but we can state a number of necessary conditions.

The following proposition follows from the subsection 3.3.3.

Proposition 4. *A solvable model necessarily satisfies* $s = N$.

For a directed graph G and its vertices v, w we denote $v \preceq w$ if there is a path in G from v to w. We say that v, w are equivalent is $v \preceq w, w \preceq v$. If G consists of a single equivalence class then we call G *strongly connected*. A Markov chain model is called *strongly ergodic* if and only if its transition graph is strongly connected.

The following condition is another necessary condition of solvability.

Proposition 5. *A solvable model is strongly ergodic.*

Proof. We will prove that a model with a non strongly connected transition graph contradicts i).

Denote by G the graph of the Markov chain. Suppose the G is not ergodic. There exists its vertex v such that not for every vertex w of G it holds $w \preceq v$. Reorder the vertices of G in such a way that first are listed all r vertices w for which $w \preceq v$ holds, and v is listed as the last one among them (thus, v is numbered by r in the ordering). The rest of $N - r$ vertices are listed in an arbitrary order. The resulting matrix of the Markov chain we still denote by $\tilde{\mathbf{Q}}$ (slightly abusing the notations).

Making use of the notations from Lemma 2 we claim that the polynomial $D_{r+1} = 0$ vanishes identically. Indeed, in the matrix obtained from $\tilde{\mathbf{Q}}$ by means of deleting its r-th row and $r+1$-th column, all the entries in places (i, j) where $i \geq r+1, j \leq r$ vanish, which justifies the claim.

On the other hand $deg_\lambda(D_r) = N - 1$, and the coefficient of D_r at λ^{N-1} equals 1 (cf. the proof of Lemma 2). Hence $D_r(\lambda_t) \neq 0$ for a suitable $1 \leq t \leq N$. Therefore $\boldsymbol{u}_{r+1}(\lambda_t) = (D_{r+1}/D_r)(\lambda_t) = 0$ (see again the proof of Lemma 2). □

Reciprocally, we have

Proposition 6. *A strongly ergodic model satisfying the conditions C1, C2 and $s = N$ satisfies the condition i) of Definition 3, in other words the symmetrized equations* (3.21) *can be written down for such models.*

Proof. If $s = N$ then $D_N \in \mathbb{Z}[\lambda, \boldsymbol{k}]$ is the determinant of $(N-1) \times (N-1)$ submatrix of the matrix $\lambda \boldsymbol{I} - \tilde{\mathbf{Q}}$ obtained by deleting its N-th row and N-th column. This is a polynomial in λ whose leading term is λ^{N-1}, therefore not identically zero. In order to prove i) it is then enough to follow the proof of Lemma 2. □

Strongly ergodic models also satisfy the following property that has been used in condition (3.5) and is needed for writing the solution (3.6).

Proposition 7. *All strongly ergodic models satisfy* $\max_{1 \leq i \leq N} \lambda_i < 0$, *where λ_i are the eigenvalues of $\tilde{\boldsymbol{Q}}$.*

Proof. Suppose that some $\lambda_i = 0$. Then the corresponding eigenvector satisfies $\frac{d\boldsymbol{X}_i}{dt} =$

$\tilde{\boldsymbol{Q}}\boldsymbol{X}_i = 0$. Using the condition C1 we find that $\frac{d\left(\sum_{j=1}^{N} X_{ij}\right)}{dt} = -Q_{N,N+1}X_{iN}$. It follows that $X_{iN} = 0$. Also, $\boldsymbol{X}_i$ is a steady state distribution of the Markov chain having the generator $\tilde{\boldsymbol{Q}}^T$ in which $Q_{N,N+1}$ is set to zero. This chain is strongly ergodic. Or, any strongly ergodic Markov chain has a unique steady state distribution in which all states have non-zero probabilities, which contradicts $X_{iN} = 0$. □

Figure 3.2 shows all strongly ergodic models with $N = 3$ satisfying the conditions C1 and C2.

3.4.0 Solution of the inverse problem for the unbranched chain model

In this section we present an example of solvable model with an arbitrary number of states. For such model we propose a method to compute the solutions of the inverse problem.

Consider now a Markov chain with $N+1$ states located on a line and reversibly connected one to next such that

$$\tilde{Q}_{n+1,n} = k_n^+, \tilde{Q}_{n,n+1} = k_n^-, 1 \le n < N, \tilde{Q}_{n,n} = -k_n^+ - k_{n-1}^-, 1 \le n < N, \tilde{Q}_{N,N} = -k_N - k_{N-1}^-.$$

This model is a generalization, of arbitrary length N, of the model M1 represented in Figure 3.2 that has $N = 3$.

We have an algebraic map

$$f := f_N : \mathbb{C}^{2N-1} \to \mathbb{C}^{2N-1}, f(k_1^+, \dots, k_{N-1}^+, k_1^-, \dots, k_{N-1}^-, k_N) = (A_1, \dots, A_{N-1}, \lambda_1, \dots, \lambda_N). \tag{3.24}$$

The goal of this section is to prove that f is invertible and that its inverse is a rational function. In other words, we will show that the unbranched chain model is solvable for any N.

Note that (3.13), (3.14) imply that

$$C_1 + \dots + C_N = A_1 + \dots + A_N = 1, k_N = -\sum_{1 \le j \le N} A_j \lambda_j. \tag{3.25}$$

Denote a vector $C := (C_1, \dots, C_N)^T$.

For each eigenvalue λ of the matrix $\tilde{\mathbf{Q}}$ it holds

$$k^+_{N-1}u_{N-1}(\lambda) = k_N + k^-_{N-1} + \lambda,\ k^+_n u_n(\lambda) = (k^+_{n+1} + k^-_n + \lambda)u_{n+1}(\lambda) - k^-_{n+1}u_{n+2}(\lambda),\ 0 \le n \le N-2.$$

This provides by recursion on $N-n$ a rational function g_n with rational coefficients such that

$$k^+_n u_n(\lambda) = g_n(k^+_{n+1}, \ldots, k^+_{N-1}, k^-_n, \ldots, k^-_{N-1}, k_N, \lambda). \tag{3.26}$$

Moreover, the denominator of g_n equals $k^+_{n+1} \cdots k^+_{N-1}$.

Consider $N \times N$ matrix U with the columns $u(\lambda_1), \ldots, u(\lambda_N)$. Then for $s \ge 0$ it holds

$$\tilde{\mathbf{Q}}^s UC = U \cdot Diag(\lambda^s_1, \ldots, \lambda^s_N)C = \tilde{\mathbf{Q}}^s(0, \ldots, 0, 1)^T \tag{3.27}$$

where $Diag(\lambda^s_1, \ldots, \lambda^s_N)$ denotes a diagonal matrix.

The $(N-s)$-th coordinate of the middle vector in (3.27) equals

$$\sum_{1 \le j \le N} u_{N-s}(\lambda_j)\lambda^s_j(-A_j\lambda_j/k_N). \tag{3.28}$$

The same $(N-s)$-th coordinate of the right vector in (3.27) equals

$$k^-_{N-s} \cdots k^-_{N-1}. \tag{3.29}$$

Observe that the j-th coordinate of the right vector in (3.27) vanishes for $1 \le j < N-s$.

Multiplying the both sides of (3.26) for $n = N-s, \lambda = \lambda_j$ by $\lambda^s_j(-A_j\lambda_j/k_N)$ and summing them up over $1 \le j \le N$, we obtain that

$$k^+_{N-s}k^-_{N-s} \cdots k^-_{N-1} = G_{N-s,0}(A_1, \ldots, A_{N-1}, k^+_{N-s+1}, \ldots, k^+_{N-1}, k^-_{N-s}, \ldots, k^-_{N-1}, k_N, \lambda_1, \ldots, \lambda_N) \tag{3.30}$$

for a suitable rational function $G_{N-s,0}$ with rational coefficients and with a denominator $k^+_{N-s+1} \cdots k^+_{N-1}k_N$ taking into account the equality of (3.28) and of (3.29). Summarizing, we have established the following statement.

Lemma 8. *For a suitable rational function G_{N-s} with rational coefficients and with a denominator $k^+_{N-s+1} \cdots k^+_{N-1}k^-_{N-s} \cdots k^-_{N-1}k_N$ it holds*

$$k^+_{N-s} = G_{N-s}(A_1, \ldots, A_{N-1}, k^+_{N-s+1}, \ldots, k^+_{N-1}k^-_{N-s}, \ldots, k^-_{N-1}, k_N, \lambda_1, \ldots, \lambda_N).$$

The $(N-s+1)$-th coordinate of the middle vector in (3.27) equals

$$\sum_{1\le j\le N} u_{N-s+1}(\lambda_j)\lambda_j^s(-A_j\lambda_j/k_N). \tag{3.31}$$

The same $(N-s+1)$-th coordinate of the right vector in (3.27) equals

$$-k_{N-s}^-\cdots k_{N-1}^- + d_{N-s}(k_{N-s+1}^+,\dots,k_{N-1}^+,k_{N-s+1}^-,\dots,k_{N-1}^-,k_N) \tag{3.32}$$

for an appropriate polynomial h_{N-s} with integer coefficients.

Multiplying the both sides of (3.26) for $n=N-s+1,\lambda=\lambda_j$ by $\lambda_j^s(-A_j\lambda_j/k_N)$ and summing them up over $1\le j\le N$, we obtain that

$$-k_{N-s+1}^+k_{N-s}^-\cdots k_{N-1}^- + d_{N-s}(k_{N-s+1}^+,\dots,k_{N-1}^+,k_{N-s+1}^-,\dots,k_{N-1}^-,k_N) =$$

$$D_{N-s,0}(A_1,\dots,A_{N-1},k_{N-s+2}^+,\dots,k_{N-1}^+,k_{N-s+1}^-,\dots,k_{N-1}^-,k_N,\lambda_1,\dots,\lambda_N)$$

for a suitable rational function $D_{N-s,0}$ with rational coefficients and with a denominator $k_{N-s+2}^+\cdots k_{N-1}^+k_N$ taking into account the equality of (3.31) and of (3.32) (while the both latter multiplied by k_{N-s+1}^+). Summarizing, we have established the following statement.

Lemma 9. *For an appropriate rational function D_{N-s} with rational coefficients and with a denominator $k_{N-s+1}^+\cdots k_{N-1}^+k_{N-s+1}^-\cdots k_{N-1}^-k_N$ it holds*

$$k_{N-s}^- = D_{N-s}(A_1,\dots,A_{N-1},k_{N-s+1}^+,\dots,k_{N-1}^+,k_{N-s+1}^-,\dots,k_{N-1}^-,k_N,\lambda_1,\dots,\lambda_N).$$

Applying alternatingly Lemma 9 and Lemma 8 for $s=1,\dots,N-1$ consecutively, we conclude with the following main result of this section.

Theorem 10. *One can produce rational functions $F_n^+,F_n^-,1\le n\le N-1$ with rational coefficients such that*

$$k_n^+ = F_n^+(A_1,\dots,A_{N-1},\lambda_1,\dots,\lambda_N), k_n^- = F_n^-(A_1,\dots,A_{N-1},\lambda_1,\dots,\lambda_N).$$

Remark 11. *i) Together with (3.25) Theorem 10 assures the inverse rational map to f (see (3.24)) on the open dense subset of $\mathbb{C}^{2N-1}$ determined by conditions $k_1^+\cdots k_{N-1}^+\cdot k_1^-\cdots k_{N-1}^-\cdot k_N\neq 0$ and $\lambda_i\neq\lambda_j, 1\le i<j\le N$;*

ii) the proof of Theorem 10 provides an algorithm which produces explicitly rational functions F_n^+, F_n^- by recursion on $N - n$.

Remark 12. *In fact, for any model (not necessary, the unbranched chain model elaborated in this section) one can consider a similar to (3.24) rational map. Thus, $A_1, \dots, A_{N-1}, \lambda_1, \dots, \lambda_{N-1}$ are rational functions in $k_n^+, k_n^-, 1 \leq n < N, k_N$. Therefore, there is a fields extension*

$$\mathbb{C}(A_1, \dots, A_{N-1}, \lambda_1, \dots, \lambda_{N-1}) \subset \mathbb{C}(k_1^+, \dots, k_{n-1}^+, k_1^-, \dots, k_{n-1}^-, k_N).$$

It is known (see e.g. Shafarevich (1972), Ch. 1) that the degree of this extension equals the number of solutions in $\mathbb{C}^{2N-1}$ of the system of rational equations $A_1 = \alpha_1, \dots, A_{N-1} = \alpha_{N-1}, \lambda_1 = \beta_1, \dots, \lambda_N = \beta_N$ at a generic point $(\alpha_1, \dots, \alpha_{N-1}, \beta_1, \dots, \beta_N) \in \mathbb{C}^{2N-1}$. Recall that the degree is defined as the dimension of the vector space $\mathbb{C}(k_1^+, \dots, k_{n-1}^+, k_1^-, \dots, k_{n-1}^-, k_N)$ over the field $\mathbb{C}(A_1, \dots, A_{N-1}, \lambda_1, \dots, \lambda_{N-1})$. The degree can be infinite. Theorem 10 states that for the unbranched chain model the degree equals 1.

We conjecture that for all other models the degree is greater than 1.

3.5.0 Using the Thomas decomposition for solving the inverse problem

Thomas decomposition is a computational algebra algorithm allowing to decompose systems of polynomial equations and inequations into simple systems whose solutions can be more easily found by iteratively solving unvariate polynomial equations. We apply this technique to compute the solutions of the inverse problem for all the strongly ergodic models with $N = 3$ and $N = 4$.

3.5.1 The Thomas Decomposition of an Algebraic System

In this section we introduce briefly the notion of the algebraic Thomas decomposition (see the appendix of Lange-Hegermann et al. (2021) for a similar introduction). We will then apply the algebraic Thomas decomposition in the subsequent sections to our symmetrized systems to determine their solutions.

Let $\mathbb{C}[\boldsymbol{x}]$ be a polynomial ring in n variables $\boldsymbol{x} = (x_1, \dots, x_n)$ over the complex numbers $\mathbb{C}$. An algebraic system $\mathcal{S}$ is defined as a finite set of polynomial equations and inequations, that is as the set

$$\mathcal{S} = \{p_1(\boldsymbol{x}) = 0, \dots, p_r(\boldsymbol{x}) = 0, q_1(\boldsymbol{x}) \neq 0, \dots, q_s(\boldsymbol{x}) \neq 0\} \tag{3.33}$$

with polynomials $p_i(\boldsymbol{x})$, $q_j(\boldsymbol{x})$ in $\mathbb{C}[\boldsymbol{x}]$ and integers r, $s \in \mathbb{N}_0$. The solution set $\mathrm{Sol}(\mathcal{S})$ of the algebraic system (3.33) is defined as the set of all $\overline{\boldsymbol{x}} = (\overline{x}_1, \ldots, \overline{x}_n) \in \mathbb{C}^n$ satisfying the equations and inequations of $\mathcal{S}$, that is as

$$\mathrm{Sol}(\mathcal{S}) = \{\overline{\boldsymbol{x}} \in \mathbb{C}^n \mid p_i(\overline{\boldsymbol{x}}) = 0,\, q_j(\overline{\boldsymbol{x}}) \neq 0 \text{ for all } 1 \leq i \leq r \text{ and } 1 \leq j \leq s\}.$$

Geometrically, $\mathrm{Sol}(\mathcal{S})$ is the difference of the two varieties

$$\{\overline{\boldsymbol{x}} \in \mathbb{C}^n \mid p_1(\overline{\boldsymbol{x}}) = 0, \ldots, p_r(\overline{\boldsymbol{x}}) = 0\} \text{ and } \{\overline{\boldsymbol{x}} \in \mathbb{C}^n \mid q_1(\overline{\boldsymbol{x}}) \cdots q_s(\overline{\boldsymbol{x}}) = 0\}$$

and so it is a locally Zariski closed subset of $\mathbb{C}^n$.

In order to introduce the notion of an algebraic Thomas decomposition of a system $\mathcal{S}$ we make the following definitions. On the variables $\boldsymbol{x} = (x_1, \ldots, x_n)$ of our polynomial ring $\mathbb{C}[\boldsymbol{x}]$ we define a total ordering (sometimes also called a ranking) by setting $x_i < x_j$ for $i < j$. With respect to this ranking the leader $\mathrm{ld}(p(\boldsymbol{x}))$ of a non-constant polynomial $p(\boldsymbol{x})$ is defined as the greatest variable appearing in $p(\boldsymbol{x})$. In case $p(\boldsymbol{x}) \in \mathbb{C}$ is a constant polynomial, we set $\mathrm{ld}(p(\boldsymbol{x})) = 1$. If we consider every polynomial $p(\boldsymbol{x}) \in \mathbb{C}[\boldsymbol{x}]$ as a univariate polynomial in its leader, say $\mathrm{ld}(p(\boldsymbol{x})) = x_k$, then the coefficients of $p(\boldsymbol{x})$ as a polynomial in x_k are polynomials in $\mathbb{C}[x_1, \ldots, x_{k-1}]$. The coefficient of the highest power of $\mathrm{ld}(p(\boldsymbol{x}))$ in $p(\boldsymbol{x})$ is called the initial of $p(\boldsymbol{x})$ which we denote by $\mathrm{init}(p(\boldsymbol{x}))$. The separant $\mathrm{sep}(p(\boldsymbol{x}))$ of a polynomial $p(\boldsymbol{x})$ is defined as the partial derivative of $p(\boldsymbol{x})$ with respect to its leader.

Definition 13. *Let $\mathcal{S}$ be the algebraic system of* (3.33)*. Then $\mathcal{S}$ is called a simple algebraic system with respect to a ranking, if the following conditions are satisfied:*

1. *The leaders of all equations and inequations are pairwise different, i.e. we have*

$$\mathrm{card}\big(\{\mathrm{ld}(p_1(\boldsymbol{x})), \ldots, \mathrm{ld}(p_r(\boldsymbol{x})), \mathrm{ld}(q_1(\boldsymbol{x})), \ldots, \mathrm{ld}(q_s(\boldsymbol{x}))\} \setminus \{1\}\big) = r + s.$$

   This property is called triangularity.

2. *For every $p(\boldsymbol{x}) \in \{p_1(\boldsymbol{x}), \ldots, p_r(\boldsymbol{x}), q_1(\boldsymbol{x}), \ldots, q_s(\boldsymbol{x})\}$ the equation $\mathrm{init}(p(\boldsymbol{x})) = 0$ has no solution in $\mathrm{Sol}(\mathcal{S})$. We call this property non-vanishing initials.*

3. *For every $p(\boldsymbol{x}) \in \{p_1(\boldsymbol{x}), \ldots, p_r(\boldsymbol{x}), q_1(\boldsymbol{x}), \ldots, q_s(\boldsymbol{x})\}$ the equation $\mathrm{sep}(p(\boldsymbol{x})) = 0$ has no solution in $\mathrm{Sol}(\mathcal{S})$. This is called square-freeness.*

The advantage of a simple algebraic system $\mathcal{S}$ is that one can obtain its solution set by it-

eratively solving univariate polynomials. This is a consequence of the triangularity of a simple algebraic system. Indeed, the triangularity implies that there is at most one equation $p(\boldsymbol{x}) = 0$ with leader x_1 or at most one inequation $q(\boldsymbol{x}) \neq 0$ with leader x_1. Note that $p(\boldsymbol{x})$ or respectively $q(\boldsymbol{x})$ is a univariate polynomial in x_1. The square-freeness implies that the number of zeros in $\mathbb{C}$ of $p(\boldsymbol{x})$ (respectively $q(\boldsymbol{x})$) is equal the degree of $p(\boldsymbol{x})$ (respectively $q(\boldsymbol{x})$). In case of the equation $p(\boldsymbol{x}) = 0$, any root $\overline{x}_1 \in \mathbb{C}$ of $p(\boldsymbol{x})$ can be chosen as the first coordinate of a solution $\overline{\boldsymbol{x}}$ of $\mathcal{S}$. In case of the inequation $q(\boldsymbol{x}) \neq 0$, all elements of $\mathbb{C}$ except for the roots of $q(\boldsymbol{x})$ can here be chosen as the first coordinate of a solution. If there is no equation or inequation with leader x_1, then the first coordinate is free, that is $\overline{x}_1$ can be chosen arbitrary in $\mathbb{C}$. Now we make the first iteration step. Again by triangularity there is at most one equation or inequation with leader x_2. If there is an equation or inequation with leader x_2, then we substitute $\overline{x}_1$ for x_1 in this equation or inequation and obtain so an univariate polynomial in x_2. Condition 2 of Definition 13 guarantees that the degree of the so obtained polynomial is independent of the choice of $\overline{x}_1$ and the square-freeness implies that the number of roots of this polynomial is equal to its degree. According to the three possible cases, that is there is an equation, inequation or neither of them, we determine as described above $\overline{x}_2 \in \mathbb{C}$ for the second coordinate of a solution of $\mathcal{S}$. An iteration of this process yields successively a solution $\overline{\boldsymbol{x}} = (\overline{x}_1, \ldots, \overline{x}_n) \in \mathbb{C}^n$ of $\mathcal{S}$. Moreover, any solution of the simple algebraic system $\mathcal{S}$ can be obtained by this process.

Definition 14. *A Thomas decomposition of an algebraic system $\mathcal{S}$ as in* (3.33) *consists of finitely many simple algebraic systems $\mathcal{S}_1, \ldots, \mathcal{S}_m$ such that* $\mathrm{Sol}(\mathcal{S})$ *is the disjoint union of* $\mathrm{Sol}(\mathcal{S}_1), \ldots, \mathrm{Sol}(\mathcal{S}_m)$.

It was proved by Thomas in Thomas (1962, 1937) that any algebraic system has a Thomas decomposition which is in general not unique. A Thomas decomposition can be determined algorithmically (see Bächler et al. (2012)) and there is an implementation in MAPLE. A description of the implementation can be found in Bächler and Lange-Hegermann (2008-2012); Gerdt et al. (2019).

We will apply in the subsequent sections the Thomas decomposition to the symmetrized systems. In other words we will use the MAPLE implementation to compute a Thomas decomposition for them. To this end we need to define a ranking on the polynomial ring

$$\mathbb{C}[k_1, k_2, k_3, k_4, k_5, L_1, L_2, L_3, S_1, S_2].$$

To simplify notation we collect the variables into $\boldsymbol{k} = (k_1, k_2, k_3, k_4, k_5)$ and $\boldsymbol{v} = (L_1, L_2, L_3, S_1, S_2)$. Since we want to solve the symmetrized systems for the variables $\boldsymbol{k}$, we rank them always higher than the collection of variables $\boldsymbol{v}$. Among the variables $\boldsymbol{k}$ we some-

times change the ranking for the different symmetrized systems. This is done to minimize the output of the Thomas decomposition and has no deeper meeting. The ranking of the variables $\boldsymbol{k} > \boldsymbol{v}$ implies that each simple algebraic system returned by the Thomas decomposition yields solutions for the variables $\boldsymbol{k}$ which are only valid for solutions of the variables $\boldsymbol{v}$ satisfying equation and inequation conditions in $\boldsymbol{v}$ over $\mathbb{C}$.

3.5.2 A Thomas Decomposition for Model M1

The model studied here is a particular case of the one from Section 3.4 and Theorem 10. We are going to determine the solutions of the algebraic system of equations

$$\begin{aligned}\mathcal{S} \;=\; \big\{ & -S_1 = k_5,\ -S_2 = k_5\,(L_1 + k_1 + k_2 + k_3),\ L_1 = -k_1 - k_2 - k_3 - k_4 - k_5,\\ & L_2 = k_1\,k_4 + k_1\,k_5 + k_2\,k_3 + k_2\,k_5 + k_3\,k_4 + k_3\,k_5,\ L_3 = -k_2\,k_3\,k_5\big\}.\end{aligned}$$

To this end we compute with MAPLE an algebraic Thomas decomposition of $\mathcal{S}$ with respect to the ranking

$$k_1 > k_3 > k_2 > k_4 > k_5 > \boldsymbol{v}.$$

MAPLE returns eleven simple systems $\mathcal{S}_1, \ldots, \mathcal{S}_{11}$. We will only compute here the solutions in $\boldsymbol{k}$ for the first simple systems, since it the most generic one, meaning that the solutions are valid for specializations of the variables $\boldsymbol{v}$ satisfying only inequations, i.e the solutions for $\boldsymbol{k}$ are valid over an Zariski open subset of $\mathbb{C}^5$. The first simple system consists of the equations and inequations

$$\begin{aligned}\mathcal{S}_1 \;=\; \big\{ & L_1^2\,S_1^3\,S_2 + L_2^2\,S_1^3 + S_1\,S_2^3 + L_3^2\,S_1 + (-S_1^2\,S_2^2 + S_2^3 + (S_1^3\,S_2 - S_1\,S_2^2)\,L_1\\ & + (-S_1^4 + S_1^2\,S_2)\,L_2 + (S_1^3 - S_1\,S_2)\,L_3)\,k_1 + (-2\,S_1^2\,S_2^2 + (-S_1^4 - S_1^2\,S_2)\,L_2\\ & + (S_1^3 + S_1\,S_2)\,L_3)\,L_1 + (S_1^3\,S_2 - 2\,L_3\,S_1^2 + S_1\,S_2^2)\,L_2 + (S_1^4 - 3\,S_1^2\,S_2)\,L_3 = 0,\\ & (L_1\,S_1\,S_2 - L_2\,S_1^2 + L_3\,S_1 - S_2^2)\,k_3 + (-S_1^2 + S_2)\,L_3 = 0,\\ & -L_1\,S_1\,S_2 + L_2\,S_1^2 - L_3\,S_1 + S_2^2 + (S_1^3 - S_1\,S_2)\,k_2 = 0,\ k_4\,S_1 - S_1^2 + S_2 = 0,\\ & k_5 + S_1 = 0,\ L_1\,S_1\,S_2 - L_2\,S_1^2 + L_3\,S_1 - S_2^2 \neq 0,\ S_1^3 - S_1\,S_2 \neq 0,\ S_2 \neq 0\big\}.\end{aligned}$$

One easily checks that the last three inequations do not involve any variable $\boldsymbol{k}$. Thus the inequations

$$L_1\,S_1\,S_2 - L_2\,S_1^2 + L_3\,S_1 - S_2^2 \neq 0, \quad S_1^3 - S_1\,S_2 \neq 0, \quad S_2 \neq 0 \tag{3.34}$$

define the above described Zarisk open subset. Solving now successively the remaining equations for the variables $\boldsymbol{k}$ (they are all linear in their respective leaders), we obtain the

solutions

$$
\begin{aligned}
k_1 &= \frac{-1}{(-S_1^2\,S_2^2+S_2^3+(S_1^3\,S_2-S_1\,S_2^2)\,L_1+(-S_1^4+S_1^2\,S_2)\,L_2+(S_1^3-S_1\,S_2)\,L_3)} \\
&\quad (L_1^2\,S_1^3\,S_2+L_2^2\,S_1^3+S_1\,S_2^3+L_3^2\,S_1+(-2\,S_1^2\,S_2^2+(-S_1^4-S_1^2\,S_2)\,L_2+(S_1^3+S_1\,S_2)\,L_3)\,L_1 \\
&\quad +(S_1^3\,S_2-2\,L_3\,S_1^2+S_1\,S_2^2)\,L_2+(S_1^4-3\,S_1^2\,S_2)\,L_3), \\
k_3 &= \frac{(S_1^2-S_2)\,L_3}{(L_1\,S_1\,S_2-L_2\,S_1^2+L_3\,S_1-S_2^2)}, \\
k_2 &= \frac{L_1\,S_1\,S_2-L_2\,S_1^2+L_3\,S_1-S_2^2}{S_1^3-S_1\,S_2}, \\
k_4 &= \frac{S_1^2-S_2}{S_1}, \\
k_5 &= -S_1,
\end{aligned}
$$

which are only valid for those $\overline{v} \in \mathbb{C}^5$ satisfying the inequations in (3.34). If one is now interested in a solutions for $\boldsymbol{k}$ with respect to a $\overline{v} \in \mathbb{C}^5$ which does not satisfy the inequations of (3.34), then there is a simple system among $\mathcal{S}_2, \ldots, \mathcal{S}_{11}$, which one can solve successively for $\boldsymbol{k}$ as described in the previous subsection. The remaining simple systems $\mathcal{S}_2, \ldots, \mathcal{S}_{11}$ can be found in subsection A.1 of the appendix.

The conditions (3.34) do not contain equalities, therefore the inverse problem has solutions on an open domain of dimension 5. According to the Section 3.3.4 this means that the model M1 is solvable.

3.5.3 A Thomas Decomposition for Model M6

The symmetrized algebraic system for model M6 is

$$
\begin{aligned}
\mathcal{S} \;=\; & \{-S_1 = k_5,\ -S_2 = k_5\,(L_1+k_1+k_2+k_3),\ L_1 = -k_1-k_2-k_3-k_4-k_5, \\
& L_3 = -k_2\,k_3\,k_5,\ L_2 = k_1\,k_4+k_1\,k_5+k_2\,k_3+k_2\,k_4+k_2\,k_5+k_3\,k_4+k_3\,k_5\}
\end{aligned}
$$

and we compute an algebraic Thomas decomposition for it with respect to the ranking

$$k_1 > k_2 > k_3 > k_4 > k_5 > \boldsymbol{v}.$$

MAPLE returns 10 simple systems $\mathcal{S}_1, \ldots, \mathcal{S}_{10}$. One easily checks by comparing the number of equations only involving the variables $\boldsymbol{v}$ (see appendix A.2 for the remaining simple systems $\mathcal{S}_2, \ldots, \mathcal{S}_{10}$), that the first simple sytem

$$
\begin{aligned}
\mathcal{S}_1 \;=\; & \{k_1\,k_3\,S_1\,S_2+k_3^2\,S_1\,S_2+L_3\,S_2+(L_2\,S_1^2-L_3\,S_1)\,k_3=0,\ k_2\,k_3\,S_1-L_3=0,\ k_3\neq 0, \\
& k_4\,S_1-S_1^2+S_2=0,\ k_5+S_1=0,\ L_1\,S_1\,S_2-L_2\,S_1^2+L_3\,S_1-S_2^2=0,\ S_1\neq 0,\ S_2\neq 0\}
\end{aligned}
$$

is the most generic one. Here, even in the most generic case, the solutions for $\boldsymbol{k}$ are only valid for $\overline{v} \in \mathbb{C}^5$ lying on a Zariski open subset of a hypersurface defined by

$$L_1\,S_1\,S_2 - L_2\,S_1^2 + L_3\,S_1 - S_2^2 = 0,\ S_1 \neq 0,\ S_2 \neq 0. \tag{3.35}$$

Solving again the remaining equations successively for k_5, k_4, k_3, k_2, k_1, we obtain the solutions

$$\begin{aligned}
k_5 &= -S_1\\
k_4 &= \frac{S_1^2 - S_2}{S_1}\\
k_3 &\quad \text{arbitrary in}\,\mathbb{C}\setminus\{0\},\\
k_2 &= \frac{L_3}{k_3\,S_1},\\
k_1 &= \frac{-1}{k_3\,S_1\,S_2}(k_3^2\,S_1\,S_2 + L_3\,S_2 + (L_2\,S_1^2 - L_3\,S_1)\,k_3),
\end{aligned}$$

where the expression for k_2 and k_1 depend on the choice made for k_3.

The conditions (3.35) contain one equality, therefore the inverse problem has solutions on an open domain of dimension 4. Furthermore, the set of solutions is infinite. According to the Section 3.3.4 this means that the model M6 is not solvable.

3.5.4 A Thomas Decomposition for Model M7

In case of model M7 the symmetrized algebraic system is

$$\begin{aligned}
\mathcal{S} \;=\; &\big\{ -S_1 = k_5,\ -S_2 = k_5\,(k_1 + k_2 + k_3 + L_1),\ L_1 = -k_1 - k_2 - k_3 - k_4 - k_5,\\
&L_2 = k_1\,k_3 + k_1\,k_4 + k_1\,k_5 + k_2\,k_3 + k_2\,k_5 + k_3\,k_4 + k_3\,k_5,\ L_3 = -k_1\,k_3\,k_5 - k_2\,k_3\,k_5\big\}.
\end{aligned}$$

Using the ranking

$$k_1 > k_2 > k_3 > k_4 > k_5 > \boldsymbol{v}$$

the MAPLE implementation of algebraic Thomas decomposition returns 21 simple systems (see subsection A.3 of the appendix for the systems $\mathcal{S}_2, \ldots, \mathcal{S}_{21}$). The first simple system

$$\begin{aligned}
\mathcal{S}_1 \;=\; &\big\{L_1\,S_1^2 - L_2\,S_1 - S_1\,S_2 + (S_1^2 - S_2)\,k_1 + (S_1^2 - S_2)\,k_3 + L_3 = 0,\\
&L_2\,S_1^2 - L_1\,S_1\,S_2 - L_3\,S_1 + S_2^2 + (S_1^3 - S_1\,S_2)\,k_2 = 0,\ k_3^2\,S_1 + (L_1\,S_1 - S_2)\,k_3 + L_3 = 0,\\
&k_4\,S_1 - S_1^2 + S_2 = 0,\ k_5 + S_1 = 0,\ L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2 \neq 0,\ L_3 \neq 0,\\
&S_1^3 - S_1\,S_2 \neq 0,\ S_2 \neq 0\big\}.
\end{aligned}$$

is the most generic one. Indeed, all inequations appearing in $\mathcal{S}_1$ involve only the variables $\boldsymbol{v}$, namely

$$L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2 \neq 0,\ L_3 \neq 0,\ S_1^3 - S_1\,S_2 \neq 0,\ S_2 \neq 0,$$

and so the solutions for $\boldsymbol{k}$ are valid for all $\overline{v} \in \mathbb{C}^5$ of the Zariski open subset defined by these inequations. We can now successively solve the remaining equations in $\mathcal{S}_1$ for the variables k_5, k_4, k_3, k_2 and k_1, where the equations for k_5, k_4, k_2, k_1 are all linear in their respective leaders except for the equation with leader k_3,

$$k_3^2\,S_1 + (L_1\,S_1 - S_2)\,k_3 + L_3 = 0,$$

which is quadratic. Since we want to consider only real roots for k_3, we require that the discriminant is positive, that is we change the inequation $L_1^2S_1^2 - 2L_1S_1S_2 - 4L_3S_1 + S_2^2 \neq 0$ into $L_1^2S_1^2 - 2L_1S_1S_2 - 4L_3S_1 + S_2^2 > 0$. Solving successively the equations with leader k_5, k_4, k_3, k_2, k_1 using for k_3 the quadratic formula, we obtain the solutions

$$\begin{aligned}
k_5 &= -S_1,\\
k_4 &= \frac{S_1^2 - S_2}{S_1},\\
k_3^{1,2} &= -\frac{L_1\,S_1 - S_2 \pm \sqrt{L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2}}{2\,S_1},\\
k_2 &= \frac{L_1\,S_1\,S_2 - L_2\,S_1^2 + L_3\,S_1 - S_2^2}{S_1^3 - S_1\,S_2},\\
k_1 &= \frac{-L_1\,S_1^2 + L_2\,S_1 + S_1\,S_2 - (S_1^2 - S_2)\,k_3^{1,2} - L_3}{S_1^2 - S_2}
\end{aligned}$$

subject to the conditions

$$L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2 > 0,\ L_3 \neq 0,\ S_1^3 - S_1\,S_2 \neq 0,\ S_2 \neq 0. \tag{3.36}$$

Note that the solution for k_1 depends on the choice of the root $k_3^{1,2}$.

The generic simple system provides a finite set of solutions of the inverse problem on the open domain of dimension 5 defined by (3.36). According to the Section 3.3.4 this means that the model M7 is solvable.

3.5.5 A Thomas Decomposition for Model M8

In case of model M8 the symmetrized algebraic system is

$$\begin{aligned}\mathcal{S} \;=\; \big\{ &-S_1 = k_5,\; -S_2 = k_5\,(k_1+k_2+L_1),\; L_1 = -k_1-k_2-k_3-k_4-k_5,\\ &L_3 = -k_1\,k_2\,k_5,\; L_2 = k_1\,k_2+k_1\,k_3+k_1\,k_4+k_1\,k_5+k_2\,k_3+k_2\,k_5\big\}.\end{aligned}$$

We compute the algebraic Thomas decomposition of $\mathcal{S}$ with respect to the ranking

$$k_1 > k_3 > k_4 > k_2 > k_5 > \boldsymbol{v}$$

and obtain from MAPLE twelve simple systems $\mathcal{S}_1, \ldots, \mathcal{S}_{12}$. One easily checks by comparing the number of equations which have leader one of the variables of $\boldsymbol{v}$, that the first simple system

$$\begin{aligned}\mathcal{S}_1 \;=\; \big\{ &S_1\,k_1\,k_2 - L_3 = 0,\; k_3\,k_2\,S_1^2 + L_1\,S_1\,S_2 - L_2\,S_1^2 + L_3\,S_1 - S_2^2 + (-S_1^3 + S_1\,S_2)\,k_2 = 0,\\ &S_1^2\,k_2\,k_4 - L_1\,S_1\,S_2 + L_2\,S_1^2 - L_3\,S_1 + S_2^2 = 0,\; k_2^2\,S_1 + (L_1\,S_1 - S_2)\,k_2 + L_3 = 0,\\ &k_5 + S_1 = 0,\; L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2 \neq 0,\; L_3 \neq 0,\; S_1 \neq 0\big\}\end{aligned}$$

is the most generic one. Analogously as in case of model M7 one determines the solutions

$$\begin{aligned}k_5 &= -S_1,\\ k_2^{1,2} &= -\frac{L_1\,S_1 - S_2 \pm \sqrt{L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2}}{2\,S_1},\\ k_4 &= \frac{L_1\,S_1\,S_2 - L_2\,S_1^2 + L_3\,S_1 - S_2^2}{k_2^{1,2}\,S_1^2},\\ k_3 &= -\frac{-S_1^3\,k_2^{1,2} + L_1\,S_1\,S_2 - L_2\,S_1^2 + S_1\,S_2\,k_2^{1,2} + L_3\,S_1 - S_2^2}{k_2^{1,2}\,S_1^2},\\ k_1 &= \frac{L_3}{S_1\,k_2^{1,2}}\end{aligned}$$

of $\mathcal{S}_1$ subject to the conditions

$$L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2 > 0,\; L_3 \neq 0,\; S_1 \neq 0. \tag{3.37}$$

Note that here the solution for k_4, k_3 and k_1 depend on the choice of the root $k_2^{1,2}$ and that we changed the inequation $L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2 \neq 0$ for the discriminant into $L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2 > 0$ to guarantee that the roots are real. The remaining simple systems are presented in subsection A.4 of the appendix.

The generic simple system provides a finite set of solutions of the inverse problem on the

open domain of dimension 5 defined by (3.37). According to the Section 3.3.4 this means that the model M8 is solvable.

3.5.6 A Thomas Decomposition for Model M3

We compute an algebraic Thomas decomposition of the algebraic system

$$\begin{aligned} rcl\mathcal{S} &= \{ -S_1 = k_5,\ -S_2 = k_5\,(k_1+k_2+L_1),\ L_1 = -k_1-k_2-k_3-k_4-k_5, \\ &\quad L_3 = -k_1k_2k_5,\ L_2 = k_1\,k_2 + k_1\,k_4 + k_1\,k_5 + k_2\,k_3 + k_2\,k_5\} \end{aligned} \tag{3.38}$$

for model M3 with respect to the ranking

$$k_3 > k_1 > k_4 > k_2 > k_5 > \boldsymbol{v}.$$

The MAPLE implementation returns 15 simple systems $\mathcal{S}_1, \ldots, \mathcal{S}_{15}$, where the most general simple system is

$$\begin{aligned} \mathcal{S}_1 &= \{L_1\,S_1\,S_2 - L_2\,S_1^2 + L_3\,S_1 - S_2^2 + (2\,k_2\,S_1^2 + L_1\,S_1^2 - S_1\,S_2)\,k_3 + (-S_1^3 + S_1\,S_2)\,k_2 = 0, \\ &\quad -L_1\,S_1^2 + L_2\,S_1 + S_1\,S_2 + (2\,k_2\,S_1 + L_1\,S_1 - S_2)\,k_4 + (-S_1^2 + S_2)\,k_2 - L_3 = 0, \\ &\quad k_1\,S_1 + k_2\,S_1 + L_1\,S_1 - S_2 = 0,\ k_2^2\,S_1 + (L_1\,S_1 - S_2)\,k_2 + L_3 = 0, \\ &\quad k_5 + S_1 = 0,\ L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2 \neq 0,\ L_3 \neq 0,\ S_1 \neq 0\}. \end{aligned}$$

Similar as for model M7 and M8 one determines the solutions

$$\begin{aligned} k_5 &= -S_1, \\ k_2^{1,2} &= -\frac{L_1\,S_1 - S_2 \pm \sqrt{L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2}}{2\,S_1}, \\ k_4 &= \frac{L_1\,S_1^2 + S_1^2\,k_2^{1,2} - L_2\,S_1 - S_1\,S_2 - S_2\,k_2^{1,2} + L_3}{2\,k_2^{1,2}\,S_1 + L_1\,S_1 - S_2}, \\ k_1 &= -\frac{k_2^{1,2}\,S_1 + L_1\,S_1 - S_2}{S_1}, \\ k_3 &= -\frac{-S_1^3\,k_2^{1,2} + L_1\,S_1\,S_2 - L_2\,S_1^2 + S_1\,S_2\,k_2^{1,2} + L_3\,S_1 - S_2^2}{S_1\,(2\,k_2^{1,2}\,S_1 + L_1\,S_1 - S_2)} \end{aligned} \tag{3.39}$$

subject to the conditions

$$L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2 > 0,\ L_3 \neq 0,\ S_1 \neq 0. \tag{3.40}$$

Note that the solutions for k_4, k_1, k_3 depend on the choice of the root $k_2^{1,2}$ and to guarantee real roots we changed the inequation representing the discriminant $L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2 \neq 0$ into $L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2 > 0$. The remaining 14 simple algebraic systems can be found in subsection A.5 of the appendix.

The generic simple system provides a finite set of solutions of the inverse problem on the open domain of dimension 5 defined by (3.40). According to the Section 3.3.4 this means that the model M3 is solvable.

3.6.0 Conclusion

Mapping the parameters of the phase-type distribution to kinetic model parameters allows to identify transcription bursting mechanisms from live transcription imaging. This has practical applications and led to a tool, BurstDeconv Douaihy et al. (2023).

In contrast to statistical inference that treats one model at a time, using symbolic solutions for the inverse problem allowed us to infer simultaneously all solvable models.

Thomas decomposition can be used to solve the inverse problem for models with a small number of states (we provide results for $N = 3$ but we tested the method also for $N = 4$). This allowed us to extend the list of models that can be analysed with BurstDeconv (the current implementation of this tool covers only the two state random-telegraph and the three states M1 and M3 models).

The algebraic structure of the phase-type distribution inverse problem lead us a classification of different bursting mechanisms models. Some models are solvable and others are not. Among the three state solvable models we found relations resulting from similar symmetry properties of the inverse problem. Its generalization possibly asks for different mathematical tools and we leave it for future work. Moreover, we have found examples of solvable models with an arbitrary number of states and necessary solvability conditions but we did no find sufficient solvability conditions. The task of finding sufficient solvability conditions, as well as classifying all solvable models, can be very challenging.

Phase-type distributions arise when the sequence of transcription events resulting from transcription imaging data is a renewal process. However, we can consider other situations when this property is no longer valid, for instance in the presence of sister chromatids, i.e. homologous DNA sequences resulting from replication and capable of undergoing transcription, when the transcription site should be considered as having multiple ON states. This case has to be treated using different methods and the corresponding problem will be addressed elsewhere.

CHAPTER 4

Time-inhomogeneous signal

4.1.0 Introduction

In Section 2, I've delved into an algorithm able to extract bursting kinetics from a time-homogeneous signal. Importantly, the process of deconvolving the signal is entirely independent of the underlying Markovian model governing it. However, when we venture into addressing the issue associated with solving the inverse problem for the deconvolved data, as outlined in Section 3, we encounter a critical problem. This problem revolves around the assumption that we are operating within a time homogeneous regime where the kinetic parameters remain constant over time.

In the context of time-series data originating from transcription processes, it's important to acknowledge the presence of regulatory checkpoints occurring at various timings and operating at various timescale. These checkpoints, such as changes in transcription factor (TF) concentration or the presence of feedback loops, can lead to moments when the relationships between various components of the system undergo alterations in response to external stimuli.

The dataset discussed in Pimmett et al. (2021) relates to transgenic lines where the cis-

regulatory element, was deliberately constrained and kept identical between various experimental conditions. Indeed, instead of considering the full regulatory repertoire of the gene *snail*, a synthetic construct was considered, expressing only one truncated enhancer, the *sna* core part of the distal *sna* enhancer. The data was purposefully acquired in a region where the activator *dorsal* was at peak levels, in order to solely focus on the impact of core promoter motifs. The idea was to start with a controlled, simple synthetic construct. However, when our objective is to truly comprehend the dynamics of cis-regulatory elements, one needs to consider transcription controlled by various enhancers and one also needs to consider the impact of TF regulators. Thus we can not avoid dealing with time-series data in which all binding sites are present and therefore which does not lead to time-homogeneous signal.

In such cases, we need to develop an alternative approach, suited to analyse time inhomogeneous data This necessity remains irrespective of whether the inhomogeneous nature of the signal arises from the complexities of a feedback loop or the dynamic alterations in transcription factors binding to the gene of interest.

We model the inhomogeneous time signals in this case as piecewise Markov process. A piecewise Markov process exhibits Markovian behavior within distinct time intervals or segments. When this process begins within a particular initial state at the start of a segment, it proceeds in a Markovian fashion until the segment ends at a random time point (a jump). On a segment the state has a distribution determined by the initial state. The concluding state of the segment then dictates the starting state for the subsequent segment Kurtz (1970, 1971); Kuczura (1973).

However, estimating the parameters of this process through an inverse problem can be quite intricate. Consequently, I have applied an existing method, as described in Adams and MacKay (2007), to facilitate the determination of the transition time points between these Markov models.

In this chapter, I will first introduce the Bayesian Change Point Detection (BOCPD) method developed by Adams and MacKay (2007), which is used to determine the onset of repression. Subsequently, I will apply this method to artificial data generated using the Gillespie algorithm in the presence of a feedback loop (see Section 1.3.1). I will then investigate the process of distinguishing between homogeneous and inhomogeneous time signals on real data. This discrimination will be achieved through a sliding time window approach, coupled with an evaluation of the stability of the product $P_{ON} \times k_r$ where P_{ON} represents the probability that the promoter is in an ON state (transcribing state), and k_r denotes the transcription initiation rate. Finally, I will put the pipeline to the test with real data and proceed to compare the obtained kinetics results with those that can be directly extracted from the

datasets.

4.2.0 Bayesian change point detection

Changepoints denote sudden change in the underlying parameters that generate the data. In our context, these changepoints manifest as the jumps that occurs between segments within the piecewise Markov process. Such transitions result in a modification of the switching rate of the Markov process governing transcription. Multiple Research papers on Bayesian changepoint detection exists, as cited in Aminikhanghahi and Cook (2017); Van den Burg and Williams (2020). In Van den Burg and Williams (2020) they compared different Change point detection algorithms and concluded using the $F1$-score (Van Rijsbergen (1979)) that the Bayesian online changepoint detection is the most optimal to deal with univariate and multivariate time series.

Bayesian online changepoint detection, as introduced by Adams and MacKay (2007), has undergone several extensions in subsequent works. These extensions include online hyperparameter optimization, as discussed by Turner et al. (2009), and the incorporation of Gaussian Process segment models, as explored in works by Garnett et al. (2009); Saatçi et al. (2010). Recent research by Knoblauch and Damoulas (2018) has expanded this framework to encompass model selection and spatiotemporal models, along with robust detection techniques using β-divergences, as detailed in Knoblauch et al. (2018). Its purpose is to accurately predict the distribution of the next unseen data point in the sequence (in our case the next transcriptional intensity signal), based solely on the data observed up to that point.

4.2.1 Method

In these two sections I will present the algorithm introduced by Adams and MacKay (2007), so that I can apply it to our case.

Let $\mathrm{x_t} \in \mathbb{R}_d$, in our case $d = 1$, represent the t-th observation within a sequence of data, and let $\boldsymbol{x}_{s:t}$ represent the sequence $\mathrm{x_s}, \mathrm{x_{s+1}}, ..., \mathrm{x_{t-1}}, \mathrm{x_t}$ for $s \leq t$ where s is the time of the last change in the distribution. We make an assumption that our dataset consisting of T data points, denoted as $\boldsymbol{x}_{1:T}$, can be divided into $\boldsymbol{\eta}^{(p)}$ partitions where the data within each partition are independent and identically distributed (i.i.d.) samples originating from a common distribution. This concept aligns with what is known as the product partition model Barry and Hartigan (1992). In our context, $\boldsymbol{x}$ is the transcription intensity.

To be more precise, let $p = 1, 2, ..., \boldsymbol{\eta}^{(P)}$, represent the generative parameters pertaining to partition P, and let η_0 stand for the hyperprior. Consequently, $\boldsymbol{\eta}^P \sim \boldsymbol{\eta}_0$, and these parameters

are independent and identically distributed across partitions, where p ranges from 1 to P.

Bayesian online changepoint detection operates through the representation of the time elapsed since the previous changepoint, which is referred to as the "run length". The run length at time t is symbolized as r_t. r_t increases by 1 when the next time step is part of the distribution which still didn't see a change in its parameters.

Let $x_t^{(r)}$ denotes the set of observations in the run r_t, since the last changepoint. The objective of the BOCPD is to compute the distribution of probability of the time to the next changepoint given the observed data $\mathbb{P}[\mathrm{r}_t|\boldsymbol{x}_{1:t}]$. This distribution can be computed by using

$$\mathbb{P}[r_t|x_{1:t}] = \frac{\mathbb{P}[r_t, x_{1:t}]}{\mathbb{P}[x_{1:t}]} \tag{4.1}$$

where the joint distribution $\mathbb{P}[r_t, \boldsymbol{x}_{1:t}]$ can be computed recursively using the following equation:

$$\mathbb{P}[r_t, \boldsymbol{x}_{1:t}] = \sum_{r_{t-1}} \overbrace{\mathbb{P}\left[\mathrm{x}_t|r_{t-1}, \mathrm{x}_t^{(r)}\right]}^{\text{predictive distribution}} \overbrace{\mathbb{P}[r_t \mid r_{t-1}]}^{\text{changepoint prior}} \overbrace{\mathbb{P}[r_{t-1}, \mathbf{x}_{1:t-1}]}^{\text{Message}}. \tag{4.2}$$

Note that, based on the independence of r_t and x_t, we use

$$\mathbb{P}[r_t, x_t|r_{t-1}, x_{1:t-1}] = \mathbb{P}\left[\mathrm{x}_t|r_{t-1}, \mathrm{x}_t^{(r)}\right] \mathbb{P}[r_t \mid r_{t-1}] \tag{4.3}$$

One can notice that the algorithm is recursive since the "message" is the same as the joint distribution at $t-1$. Therefore after we compute the joint distribution $\mathbb{P}[r_{t-1}, \mathbf{x}_{1:t-1}]$, we can forward message-pass this distribution to calculate $\mathbb{P}[r_t, \boldsymbol{x}_{1:t}]$ once we know $\mathbb{P}[r_1, \boldsymbol{x}_1]$

It's worth noting that the predictive distribution, denoted as $\mathbb{P}\left[\mathrm{x}_t|r_{t-1}, \mathrm{x}_t^{(r)}\right]$, relies only on the most recent data $\mathrm{x}_t^{(r)}$, where r_t indicates that a changepoint occurred r_t time steps ago. Consequently, we can establish a recursive message-passing algorithm for the joint distribution over the current run length and the data. This algorithm relies on two fundamental distributions: 1) the changepoint prior distribution $\mathbb{P}[r_t \mid r_{t-1}]$, and 2) the predictive distribution $\mathbb{P}\left[\mathrm{x}_t|r_{t-1}, \mathrm{x}_t^{(r)}\right]$.

Predictive distribution

To compute the predictive distribution, we utilize Conjugate-exponential models. We assume that the predictive distribution belongs to the exponential family, a class of probability distributions Pitman (1936); Koopman (1936). One important property about the exponential family is that it has a conjugate priors. In this case the the prior distribution possesses

the same mathematical structure as the posterior distribution. Therefore we avoid the need for integration in the posterior distribution.

Let $\mathrm{x}_t^{(r)}$ represent the most recent data, x_t denote a new observation, $\boldsymbol{\theta}$ represent our model parameters, and $\boldsymbol{\alpha}$ stand for the hyperparameters of the conjugate prior. The conjugate prior gives that:

$$p(\boldsymbol{\theta} \mid \mathrm{x}_t^{(r)}, \boldsymbol{\alpha}) = p\left(\boldsymbol{\theta} \mid \boldsymbol{\alpha}'\right) \tag{4.4}$$

for different hyperparameters $\boldsymbol{\alpha}'$ related to the conjugate prior.

If the predictive distribution belongs to the exponential family then the predictive distribution $\mathbb{P}\left[\mathrm{x}_t | r_{t-1}, \mathrm{x}_t^{(r)}\right]$ of equation 4.2 as a conjugate prior and therefore:

$$\begin{aligned} p(x_t \mid \mathrm{x}_t^{(r)}, r_{t-1}, \boldsymbol{\alpha}) &= \int p(x_t \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathrm{x}_t^{(r)}, r_{t-1}, \boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\theta} \\ &= \int p(x_t \mid \boldsymbol{\theta}) p\left(\boldsymbol{\theta} \mid r_{t-1}, \boldsymbol{\alpha}'\right) \mathrm{d}\boldsymbol{\theta} \\ &= p\left(x_t \mid r_{t-1}, \boldsymbol{\alpha}'\right). \end{aligned} \tag{4.5}$$

As a consequence, the posterior predictive distribution is identical to the prior predictive distribution, with the only distinction being the utilization of hyperparameter $\boldsymbol{\alpha}'$ instead of $\boldsymbol{\alpha}$. This signifies that if we can determine the values of $\boldsymbol{\alpha}'$, in function of $\boldsymbol{\alpha}'$ we can calculate the posterior predictive distribution without the need for integration.

In our case, I assume that the transcriptional signal follows a normal distribution with an unknown mean μ and variance σ, where $\boldsymbol{\theta} = \{\mu, \sigma\}$. Since the Gaussian distribution is a member of the exponential family we can benefit from the conjugate prior.
In this case we have the following

$$\begin{gathered} x \sim \mathcal{N}(\mu_x,\, \sigma_x^2) \\ (\mu_x, \sigma_x \mid \mu_0, \kappa_0, \alpha_0, \beta_0) \sim \mathcal{N}\left(\mu \mid \mu_0, (\kappa_0\, \sigma_x)^{-1}\right) Ga\left(\sigma_x \mid \alpha_0,\, \text{rate } = \beta_0\right) \end{gathered} \tag{4.6}$$

This is given in more details in Murphy (2007) and is needed to compute the predictive distribution.

Due to the fact that the conjugate prior of a Gaussian distribution with unknown mean and variance is normal-Gamma distribution Murphy (2007). μ_x, σ_x changes according to a changepoint prior. The set of parameters $\boldsymbol{\alpha_0} = \{\mu_0, \kappa_0, \alpha_0, \beta_0\}$ are parameters of the prior also known as hyperparameters.

The posterior predictive distribution for this specific model, after seeing n data points, is given by

$$p(\mathrm{x}_t|r_{t-1}, \mathrm{x}_t^{(r)}) = (\pi)^{-1/2}\frac{\Gamma\left(\left(2\alpha_n+1\right)/2\right)}{\Gamma\left(\left(2\alpha_n\right)/2\right)}\left(\frac{\Lambda}{2\alpha_n}\right)^{\frac{1}{2}}\left(1+\frac{\Lambda\left(x-\mu_n\right)^2}{2\alpha_n}\right)^{-(2\alpha_n+1)/2} \tag{4.7}$$

which is a T-distribution with center at μ_n , precision $\Lambda = \frac{\alpha_n\kappa_n}{\beta_n(\kappa_n+1)}$ and degree of freedom $2\alpha_n$.

The hyperparameters are updated according to the following:

$$\begin{aligned}
\alpha_{n+1} &= \alpha_n + 1/2 \\
\kappa_{n+1} &= \kappa_n + 1 \\
\beta_{n+1} &= \beta_n + \frac{\kappa_n(x-\mu_n)^2}{2(\kappa_n+1)} \\
\mu_{n+1} &= \mu_n\frac{\kappa_n}{\kappa_{n+1}} + \frac{1}{\kappa_{n+1}}
\end{aligned} \tag{4.8}$$

For the full details about the conjugate of a Gaussian distribution and the equations 4.8 please refer to Murphy (2007).

To clarify, the subscripts in the expressions above represent the number of data points linked to a specific hypothesis about the run length. For instance, μ_2 denotes the posterior predictive mean when we consider that the most recent changepoint took place two observations ago.

Changepoint prior

As for the second calculation, $\mathbb{P}[r_t \mid r_{t-1}]$, required for equation 4.2, let $H(\tau)$ be the hazard function defined by

$$H(\tau) = \frac{f(\tau)}{S(\tau)} \tag{4.9}$$

where $f(\tau)$ denotes the probability that the current run length is τ and $S(\tau)$ is the survival function associated to it i.e. to τ. The hazard function provides a measure of the response to the following question: "If a changepoint has not taken place by the time we reach a run length of τ, what is the likelihood that it will indeed happen at τ?" The hazard function is a concept often utilized in survival analysis and reliability engineering.

Our modeling assumption is that our changepoint prior is

$$p(r_t|r_{t-1}) = \begin{cases} H(r_{t-1}+1) & \text{if } r_t = 0 \\ 1 - H(r_{t-1}+1) & \text{if } r_t = r_{t-1}+1 \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

In our case, we make the assumption that τ follows a geometric distribution with a success probability of p. Consequently, the hazard function $H(\tau)$ is equivalent to p, as stated in Forbes et al. (2011). This type of process is referred to as "memoryless" because the hazard function remains constant and does not vary with time.

4.2.2 Algorithm

I have implemented the algorithm of BOCPD proposed in Turner et al. (2009) on equations 4.7, 4.8, 4.10. We recall it in Algorithm 2, where the input is the prior hyperparameters of the conjugate prior of the Gaussian distribution, and the transcriptional data. The algorithm will output a $T \times T$ triangular matrix r, such that $r_{i,j} = 0$ for $i > j$ and

$$r_{i,j} = \mathbb{P}\left[r_j = i | x_{1:j}\right] \tag{4.11}$$

where r_j is the run length at time j (see Figure 4.1 for a visual explanation).

4.2.3 Application on simulated data

In order to test the algorithm, I have generated artificial data using Gillespie algorithm (see section 1.3.1) with the following chemical reactions corresponding to the random-telegraph model of auto-regulated gene expression, with two stages, RNA and protein (P)

$$\begin{aligned} OFF &\xrightarrow{k_{ON}(P)} ON \\ ON &\xrightarrow{k_{OFF}(P)} OFF \\ ON &\xrightarrow{k_r} ON + \mathrm{RNA_{TS}} \\ \mathrm{RNA_{TS}} &\xrightarrow{dwell} \phi \\ ON &\xrightarrow{k_r} ON + \mathrm{RNA} \\ \mathrm{RNA} &\xrightarrow{\lambda_r} \phi \\ \mathrm{RNA} &\xrightarrow{k_p} \mathrm{RNA} + \mathrm{P} \\ \mathrm{P} &\xrightarrow{\lambda_p} \phi \end{aligned} \tag{4.16}$$

Algorithm 2 BOCPD Algorithm

(A1) **Set the priors $\boldsymbol{\alpha}_0$ and initial conditions.**

$$p\left(r_0\right) = 1 \quad \text{Since we consider there is a changepoint time } t = 0. \tag{4.12}$$

Choose the priors $\boldsymbol{\alpha_0} = \{\mu_0, \kappa_0, \alpha_0, \beta_0\}$ that best suits the data

(A2) **Observe new data point** x_t.

(A3) **Compute predictive probabilities.** This calculation $\pi_{t-1}^{(r)} = p(\mathrm{x}_t | r_{t-1}, \mathrm{x}_t^{(r)})$ is for each possible run length value r, according to equation 4.7
At time $t-1$, there exist a total of t potential run lengths. We proceed by forwarding the accumulated information associated with these values up through the trellis structure. This enables us to compute the predictive distribution for x_t

(A4) **Compute growth probabilities.** The growth probabilities are the probabilities $p\left(r_t = r_{t-1} + 1, \mathbf{x}_{1:t}\right)$ for each possible run length value. To compute the probability for a specific value $r_t = r$, the growth probability equation is given by

$$p\left(r_t = r, \mathbf{x}_{1:t}\right) = p\left(r_{t-1}, \mathbf{x}_{1:t-1}\right) \pi_{t-1}^{(r)} \left(1 - H\left(r_{t-1}\right)\right). \tag{4.13}$$

where $\left(1 - H\left(r_{t-1}\right)\right)$ is given in 4.10. This is the solution of equation 4.2
It's important to note that there is no summation considered over r_{t-1} because, within the context of a specific run length, the only applicable "growth value" corresponds to $r_{t-1} = r - 1$.

(A5) **Compute changepoint probability.** The changepoint probability is the probability that the run length drops to 0 . Again using equation 4.13 , we see

$$p\left(r_t = 0, \mathbf{x}_{1:t}\right) = \sum_{r_{t-1}} p\left(r_{t-1}, \mathbf{x}_{1:t-1}\right) \pi_{t-1}^{(r)} H\left(r_{t-1}\right). \tag{4.14}$$

In this case, a summation over r_{t-1} is necessary because r_{t-1} can assume any value within the range from 0 to t.

(A6) **Compute the evidence.** This is just the normalization:

$$p(r_t | \boldsymbol{x}_{1:t}) = \frac{p(r_t, \boldsymbol{x}_{1:t})}{\sum_{r_{t\prime}} p(r_{t\prime}, \boldsymbol{x}_{1:t})} \tag{4.15}$$

(A7) **Compute the posterior.** Equation 4.15.

(A8) **Update sufficient statistics** according to equation 4.8

(A9) **Set** $t = t + 1$. Return to Step (A2) .

Where ON, OFF represents the promoter state (active or not), the promoter switching rates, $k_{ON}(P)$, $k_{OFF}(P)$ are functions of the protein P introducing a negative feed-back. $k_{ON}(P)$, $k_{OFF}(P)$ are decreasing and increasing functions of P, respectively. In order to take into account cooperativity of the repression I have used the Hill equation to model these functions:

$$\begin{aligned} k_{OFF}(P) &= k^m_{min} + (k^m_{max} - k^m_{min})\frac{th^{nh_m}}{th_m^{nh_m} + P^{nh_m}} \\ k_{ON}(P) &= k^p_{max} + (k^p_{min} - k^p_{max})\frac{P^{nh_p}}{th_p^{nh_p} + P^{nh_p}} \end{aligned} \tag{4.17}$$

nh_m (resp. nh_p) is the Hill coefficient, which represents the cooperativity of binding, th_m (resp. th_p) represents the concentration of protein at which half-maximal response of the switching rates to protein is achieved, k^m_{min}, k^m_{max} (resp k^p_{min}, k^p_{max}) represents the minimal/maximal switching rate from $OFF \longrightarrow ON$ (resp. $ON \longrightarrow OFF$).

In this model I expect a changepoint because the switching parameters change from (k^m_{min}, k^p_{max}) to (k^m_{max}, k^p_{min}) when P increases from zero to a value larger than the thresholds.

Once I determined the initiation time of RNA_{TS}, I generated the transcription signal. This signal was constructed using experimentally measured elongation and 3'-end processing rates Tantale et al. (2016). Subsequently, I passed this signal through the algorithm to assess the stability of the switching rates with respect to the identified changepoint. I apply the algorithm by optimising it to data of a normal distribution with unknown mean and variance.

I ran a sample of 500 simulation for a duration of 40min (approximate duration of nuclear cycle 14). The switching parameters used are found in table 4.1 as for the values of the signal construction I used the ones in Pimmett et al. (2021).

In figure 4.2 A), I plotted a sample of the simulations output, mRNA, TS RNA, protein (upper pannel) and the constructed signal (middle pannel).

To test the results of the simulation of BOCPD I plotted the point where the change was detected by changing the color of the constructed signal (middle pannel). By comparing the changepoint found with the switching parameters, k_{ON}, k_{OFF}, of the same nuclei I noticed that indeed these switching parameters are rather constant after the change point (figure 4.2 A).

Finally in order to compute the stability of the parameters after the cut point I computed the

Name	Symbol	Formula
Number of simulations	N	500
minimal Promoter switching to OFF	k^m_{min}	0.004 s^{-1}
maximal Promoter switching to OFF	k^m_{max}	$k^m_{min} \times 20$ s^{-1}
maximal Promoter switching to ON	k^p_{max}	0.042 s^{-1}
minimal Promoter switching to ON	k^p_{min}	$k^p_{max}/20$ s^{-1}
Transcription rate	k_r	0.113 s^{-1}
mRNA half-life	$T^m_{1/2}$	13×60 s^{-1}
Translation rate	k_p	$\lambda_r \times 4$ s^{-1}
Protein half-life	$T^p_{1/2}$	26×60 s^{-1}
Hill coefficient for $k_O N$	nh_p	2
Hill coefficient for $k_O FF$	nh_m	2
Threshold on protein for inhibition for k_{ON}	th_p	40
Threshold on protein for inhibition for k_{OFF}	th_m	40
dwell time	$dwell$	2*60 s^{-1}

Table 4.1: Parameter Values for BOCPD

variance of k_{ON}, k_{OFF} for each nuclei in the "repressed part" (figure 4.2 B). The stability of the parameters is needed to be able to run and extract the switching parameters from real data at least in the "repressed part".

4.3.0 Application to real data

To emphasize the importance of our methodology and clarify the essential stages leading to Bayesian Online Change Point Detection (BOCPD), I will conduct a comparative analysis involving two mutations of *snail* gene within Drosophila melanogaster: "transgenic snail" and "CRISPR snail".

The first mutation, introduced in the introduction of this chapter, is what I refer to as the "transgenic snail." This is essentially a synthetic platform designed to eliminate looping effects in transcription. In this platform, only the Core promoters were inserted immediately downstream of the *snail* distal minimal enhancer (snaE). The platform was then cloned into a minigene and integrated into the Drosophila genome at the same genomic location. To enable the tracking of transcription, we incorporated 24 MS2 stem loops in the 5′ UTR downstream of the promoter, followed by the insertion of the yellow reporter gene 1.2.6.

On the other hand, the second mutation, which we refer to as the "CRISPR *snail*", represents the endogenous form of the *snail* gene, unaltered except for the insertion of the 24 MS2 stem loops. The native *snail* gene is known to auto-regulate its own transcription, resulting in an inherently inhomogeneous time series Lagha et al. (2013).

Figure 4.3 A) presents a heatmap featuring the *transgenic snail* (on the left side), where it becomes evident that the density of Pol II, as determined through signal deconvolving (see section 2), remains relatively constant over time following the variable time to activation experienced by different nuclei. In contrast, on the right side of Figure 4.3, the heatmap showcasing the CRISPR-edited *snail* clearly reveals a decline in Pol II density.

4.3.1 Identification of inhomogeneous data

To discern between homogeneous and inhomogeneous temporal data, I adopt an approach where I evaluate the product of $\mathbb{P}[ON]$ and k_r over time. Instead of computing each component separately, I calculate this product directly since applying them separately using BurstDeconv might not be appropriate when the signal's homogeneity is uncertain.

The first method consists of considering that the initiation rate remains at 0 during "OFF" periods and takes on the value k_r during "ON" periods. As a result, the mean initiation rate τ_r can be expressed as $\tau_r = \mathbb{P}[ON] \times k_r$. Now since the mean waiting time is the inverse of the mean initiation rate then to obtain τ_r I calculate the mean waiting time τ_w for Pol II within a sliding time window of length w.

The signal's homogeneity is then assessed based on the constancy of τ_r over time, which is determined by the mean waiting time. The results of τ_r are visualized in Figure 4.3 B) for both phenotypes, utilizing a time window comprising 8 frames (around 30s).

The second method for confirmation involves applying a sliding time window technique to the signal. Here, I select a fixed time window denoted as TI and then employ the complete BurstDeconv algorithm within each of these time windows. This approach operates under the assumption that within sufficiently small time windows, there isn't a significant alteration in the data's underlying distribution. I then assume the constancy of the kinetic parameters within this time window.

The results of this sliding time window analysis on both phenotypes, using an 8-minute time window, are presented in Figure 4.3. It's important to note that I excluded the output from the initial time window of the movie since during this time, transcription is shut down in most of the nuclei.

By utilizing both of these methods, I can confirm that the "transgenic *snail*" indeed exhibits temporal homogeneity. Consequently, I can directly apply the BurstDeconv algorithm to extract valuable information. However, in the case of the "CRISPR *snail*," transcription exhibits temporal heterogeneity, necessitating the application of BOCPD before employing BurstDeconv.

4.3.2 BOCPD results

Once I am able to identify the inhomogeneous data and before I run the BOCPD, I needed to set the proper priors for the conjugate of the Gaussian distribution $\boldsymbol{\alpha_0} = \{\mu_0, \kappa_0, \alpha_0, \beta_0\}$

Let $\mathfrak{D}_i$ denotes the transcription data for nuclei i between 5min and 8 min after mitosis. I assume that during this time window the data is rather homogeneous since I avoid the time into activation, known as post-mitotic reactivation time, and therefore I can extract the prior $\boldsymbol{\alpha_0}$ from it.

According to equation 4.6 the variance of the data σ_x follows a gamma distribution $Ga\left(\sigma_x \mid \alpha_0, \text{ rate } = \beta_0\right)$ Therefore the parameters α_0, β_0 were chosen according to the maximum likelihood estimation of a gamma distribution for the data consisting of the $\{variance(\mathfrak{D}_i)\}_{i \in N}$ where N denotes the total number of nuclei.

As for the mean μ_0 it was set to be the mean of the dataset $\mathfrak{D}_i$. And finally the parameter κ_0 was chosen by benchmarking against artificial data by changing the hyperparameter κ_0 in both the simulations of the artificial data and the input in BOCPD and choosing the least sensitive value of input of BOCPD regardless of the value of κ_0 initially used in the simulations.

I employed BOCPD with these priors to determine an independent changepoint, denoted as T_0, for each nucleus. A representative subset of the results is visualized in Figure 4.4 A), where I illustrated the transcription intensity, the complete BOCPD output. At each time point the probability of the BOCPD output is plotted by the gradient color of the black lines: the darker the black lines are, the higher probability. Whenever a changepoint emerges within the distribution, I can see a discontinuity in the increasing line of r_t. I also plotted the output of BOCPD when the probability exceeds 0.8. It's important to note that BOCPD does not aim to overestimate changes in the distribution. Consequently, there may be nuclei where BOCPD does not identify a changepoint, as illustrated in Figure 4.4 D).

Once I have established the changepoint for each nucleus, I can determine the time $T_{1/2}$ at which 50% of the repression process was built, as demonstrated in Figure 4.4 B).

Now that I possess the stable, repressed segment of the signal, I proceed to apply BurstDeconv on the reoressed oart of the signal. This yields $\mathbb{P}[ON] \times k_r$, which I then compare with τ_r as an additional validation of our methodology, as presented in Figure 4.4 C).

4.4.0 Conclusion

In our current biological and mathematical landscape, we face inherent limitations that challenge our ability to extract meaningful information from time-inhomogeneous data that are due to the presence of feedback loop or other biological limitations in transcription. One significant constraint lies in our capacity to simultaneously capture both protein and nascent mRNA intensity signals through live imaging techniques. Indeed, if we were equipped with the capability to conduct live imaging for both protein and mRNA, it would unlock the potential for comprehensive Hill function fitting. Additionally, when it comes to mathematical modeling, obtaining non-constant switching parameters within a Markovian framework, whether utilizing a hidden Markov approach or inverse problem solving, remains a formidable challenge.

To address these limitations, I have used online changepoint detection techniques. This approach serves a dual purpose: it simplifies the problem at hand and facilitates the transition from a piecewise deterministic Markov model to a Markov model with constant switching rates. By doing so, I hope to obtain more manageable and interpretable analyses of dynamic transcription processes.
Through our optimized online changepoint detection method, we've achieved the remarkable ability to pinpoint the exact moment when repression commences for each individual nucleus. This represents a substantial advancement in our capacity to discern key temporal events.

Nonetheless, challenges persist, particularly in extracting information from the active phase of transcription. The switching parameters governing these dynamic processes do not have sufficient time to stabilize and attain constancy before the transition from the time of nucleus activation to the onset of repression. This part of the signal requires further exploration and innovative solutions to fully elucidate the intricacies of transcription dynamics.

In summary, while our current limitations impose formidable constraints, our approach represents a significant step forward in unraveling the complexities of time-inhomogeneous data. It not only simplifies the problem but also enables us to capture critical moments in biological processes.

Figure 4.1: Diagram of the message-passing algorithm. Each node has associated mass. For example, the probability of $p(r_6 = 5|x_{1:6}) = p(r_6 = 5|x_{2:5}) = 0$ if changepoint occured 7 else is associated with the node indexed by $t = 5$ and $r_t = 5$. In other words, at this node, the run length can either increase by one, hence $r_t = 6$ or we have a changepoint and r_t will go back to 0.

Figure 4.2: Efficiency of BOCPD on Simulated Data with Feedback Loop Regulation.

A) In each subplot, the upper panel depicts the Gillespie output, emphasizing how mRNA levels decrease as protein concentration increases due to the feedback loop. In the middle panel, the Bayesian Online Change Point Detection (BOCPD) output is displayed by transitioning the mRNA signal color from blue (before the changepoint) to orange (after the changepoint is detected). The changepoint is detected when the output of BOCPD r, given by equation 4.11), is such that $r_{i,j} \leq 0.2$ given that $r_{i-1,j-1} \geq 0.8$ for $j < i$. The green bar represents the initiation time of transcription. The lower panel shows the changes in the values of k_{ON} and k_{OFF} based on the protein concentration obtained from each simulation using the hill function (refer to Equation 4.17).

B) Histogram displaying the variance of the kinetic parameters k_{ON} and k_{OFF} after identifying the changepoint for each dataset. The variance highlights the stability of these parameters in the repressed portion of the signal. The dashed black line (resp. green line) is the mean of the variance (resp. median).

Figure 4.3: Evaluating the stability of time sereis using different criteria by comparing *snail* transgene to *snail* CRISPR.

A) Heatmap displaying the number of polymerases for the transgenic *snail* (left) and CRISPR-edited *snail* (right) as a function of time. Each row corresponds to a nucleus, and the color of each time bin represents the count of Pol II initiation events per 30-second interval.

B) Probability of transcription initiation (τ_r) which is computed as the inverse of the mean waiting time τ_w for Pol II within a sliding time window of length w for each nucleus (black) and the average value across all nuclei (green) for the *snail* transgene (left) and the CRISPR-edited *snail* (right).

C) Each subplot represents one of the kinetic parameters estimated using a two-state model with a sliding time window approach over an 8-minute window for the *snail*

transgene (left) and the CRISPR-edited *snail* (right).

Figure 4.4: Evaluating BOCPD output and results on CRISPR *snail* data.

A) Three instances showcasing the application of BOCPD to distinct nuclei, each represented by a column. Top row: of each column, you'll find the transcription signal displayed in yellow, along with the time that the changepoints occured indicated by the red line.
Middle row: comprehensive output of r_t wit. The output of BOCPD r is plotted such that $r_{i,j}$ is the value found at the coordinate $\{i, j\}$. The value of $r_{i,j}$ is shown by the gradient color of the black lines: the darker the black lines are, the higher probability. Whenever a changepoint emerges within the distribution, we can see a discontinuity in the increasing line of r_t.
Last row: represents r solely when the probability of a changepoint surpasses 0.8, i.e. $r_{i,j} \geq 0.8$

B) Cumulative distribution function of the identified changepoints for each nucleus. The black line here represents $T_{1/2}$, which marks the point in time when 50% of the nuclei have transitioned through the repression phase.

C) Probability of transcription initiation (τ_r) which is computed as the inverse of the mean waiting time τ_w for Pol II within a sliding time window of length w for each nucleus (**black**) and the average value across all nuclei (**green**). In **blue**, resp, **yellow**, resp. **red** I plotted the value $\mathbb{P}[ON] \times k_r$ coming from each output of BurstDeconv for the models two states, 3states M1 and 3states M2 (see figure 3 in appendix 2.2 for the reference of the models).

D) Example of a nucleus that did not experience repression. This illustration serves to emphasize that BOCPD does not overestimate the occurrence of changepoints. The subplot is the same as subplot A).

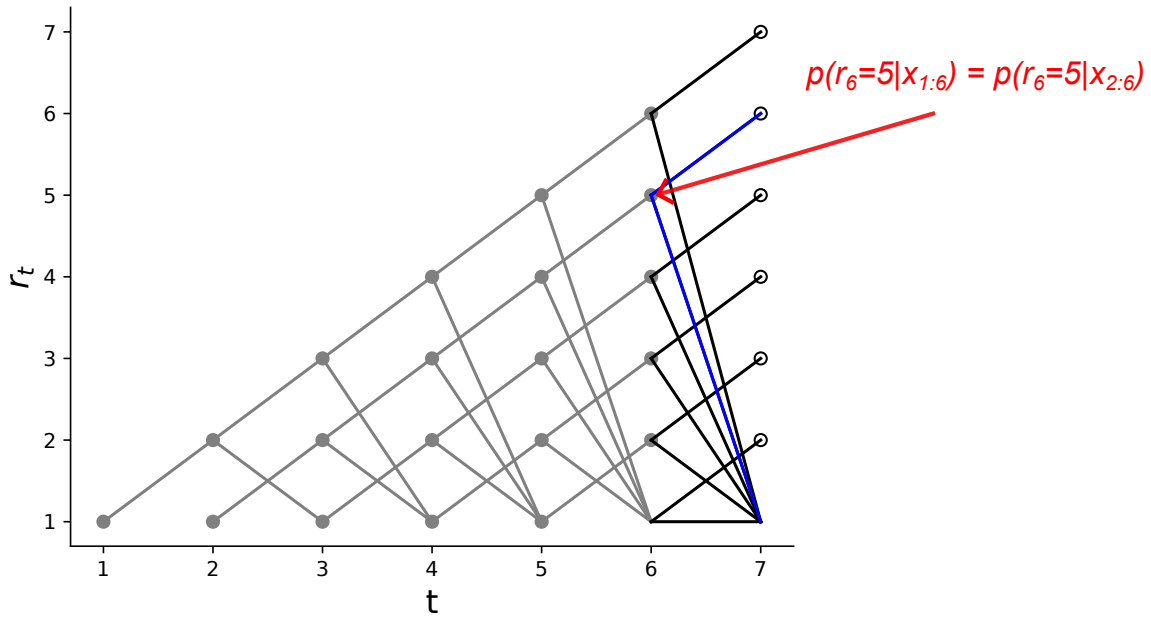


Figure 4.1: Diagram of the message-passing algorithm.

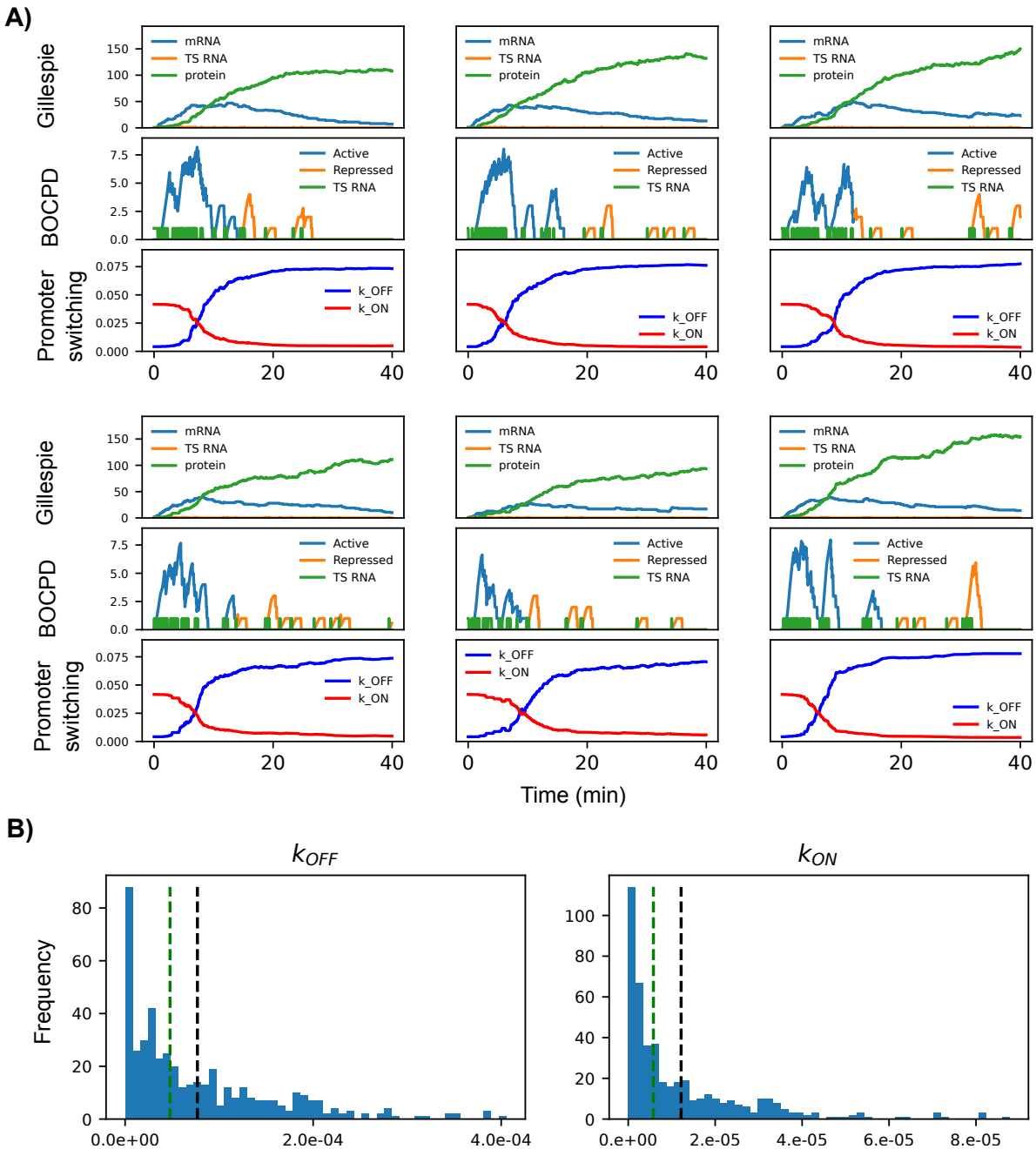


Figure 4.2: Efficiency of BOCPD on Simulated Data with Feedback Loop Regulation.

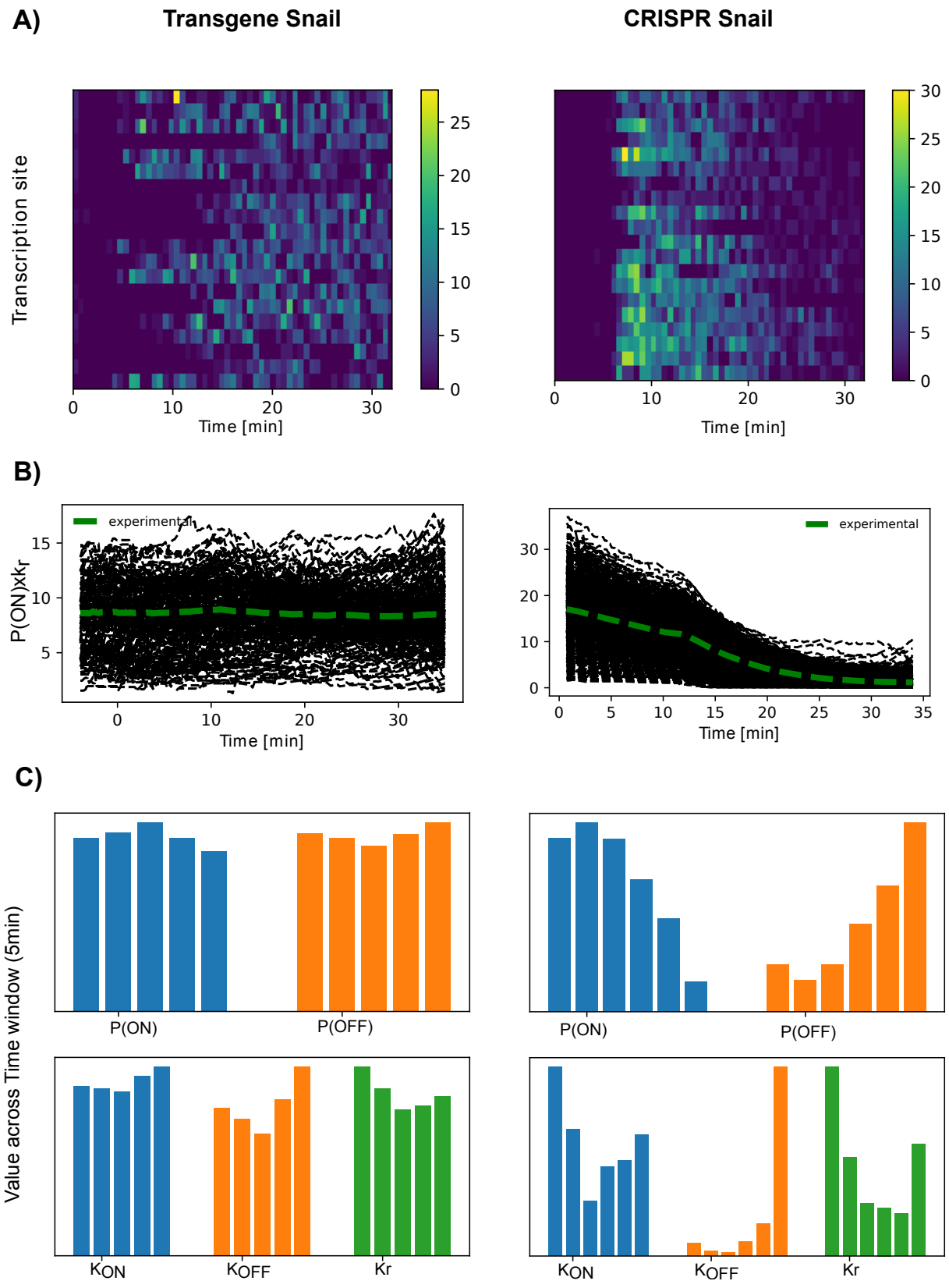


Figure 4.3: Evaluating the stability of time sereis using different criteria by comparing *snail* transgene to *snail* CRISPR.

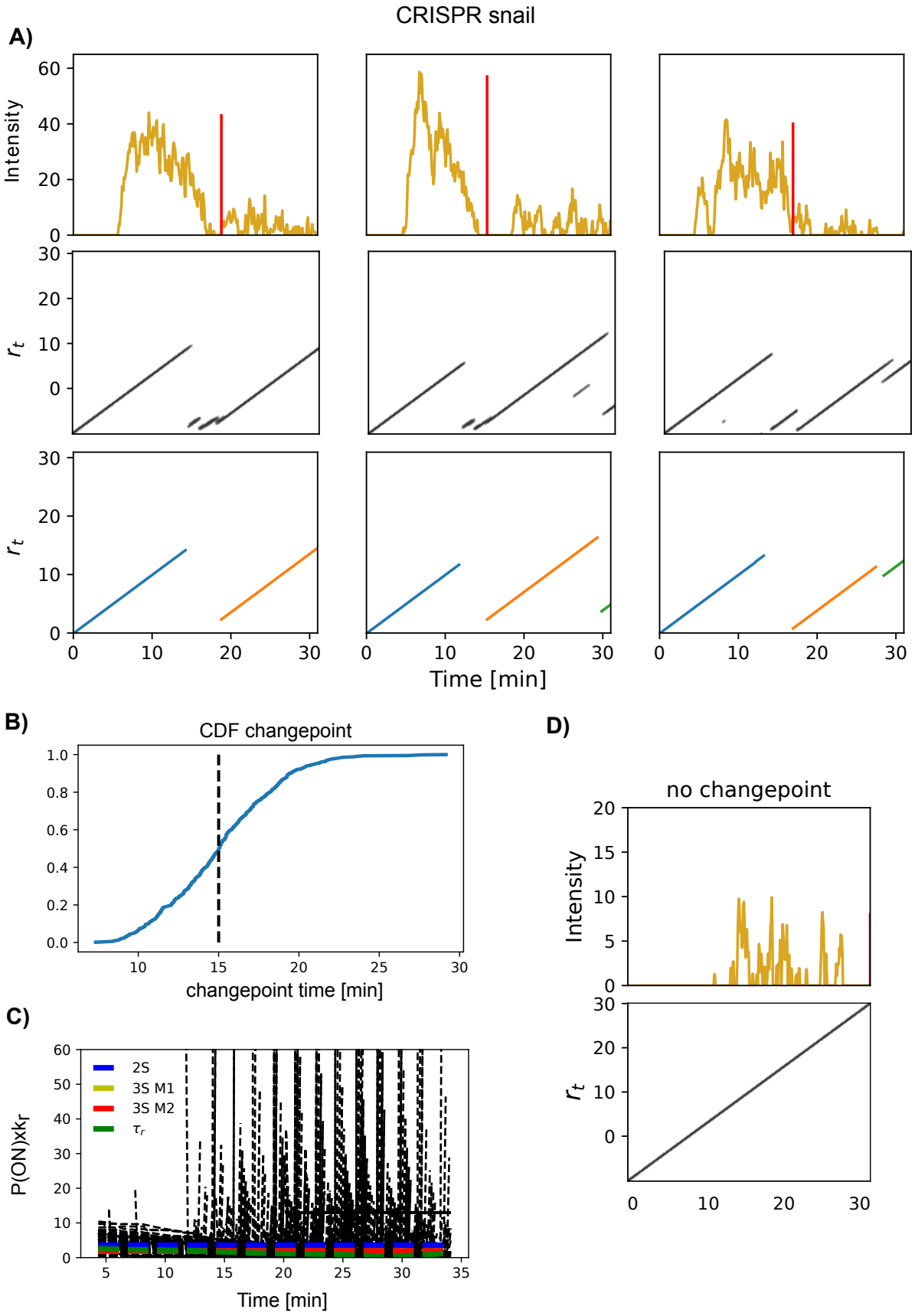


Figure 4.4: Evaluating BOCPD output and results on CRISPR *snail* data.

CHAPTER 5

Space and time modeling of gene expression

5.1.0 Introduction

Zygotic gene expression during early stages of embryogenesis of Drosophila is submitted to complex regulation resulting in spatio-temporal patterns. For instance, the protein snail required to coordinate the mesoderm invagination during gastrulation, is expressed within a spatial domain sharply delimited along the dorso-ventral embryo axis (Figure 5.1 A). Microscopy studies using single-molecule fluorescence in situ hybridization, smFISH, (section 1.2.6) in fixed embryos and MS2-MCP dual mRNA synthesis reporter in live embryos (section 1.2.6) showed that *snail* mRNA levels undergo stochastic fluctuations in populations of nuclei and in single nuclei time series (Figure 5.1 F). It was proposed that these fluctuations result from transcriptional bursting, a stochastic phenomenon consisting in alternating transcriptionally active and inactive promoter states (Pimmett et al. (2021); Tantale et al. (2021)). Mathematically, incorporating both spatial and stochastic effects, even within uncomplicated geometries and homogeneous environments, presents a formidable challenge Gardiner et al. (1985).

Deterministic patterning of zygotic genes, *snail* in particular, was intensively modeled using thermodynamic or differential equations models (Bieler et al. (2011); Jaeger et al. (2012);

Kanodia et al. (2012)). Much less is known about stochastic patterning in a transcriptional bursting regime. In order to approach this regime we use models of stochastic biochemical networks with spatial extension as well as hybrid models consisting in coupling stochastic biochemical networks and differential equations.

Stochastic reaction-diffusion models (SRDM) were first introduced in non-equilibrium statistical physics by van Kampen (Van Kampen (1976)), Nicolis and Prigogine (Nicolis and Prigogine (1977)), and Haken (Haken (1978)). A one dimensional SRDM is illustrated in Figure 5.1 D. In SRDMs space is divided into compartments in which reactions take place and neighboring compartments are coupled by diffusion reactions. These (quasi-)compartments are needed for replacing continuous with discrete diffusion (that is more convenient for simulation and analysis purposes) and do not necessarily have physical or biological significance. The choice of compartments size should adhere to Kuramoto's length, denoted as l_k, such that it satisfies the condition

$$l_k = \sqrt{DT^p_{1/2}} \geq h \tag{5.1}$$

Here, D represents the diffusion rate, $T^p_{1/2}$ signifies the average lifespan of the protein, and h corresponds to the size of the cell. This criterion ensures that within each compartment, any concentration fluctuations at a specific location within the compartment swiftly propagate throughout the entire compartment. Consequently, homogeneity is consistently maintained, allowing us to disregard diffusion for enter-compartment reactions Cottrell et al. (2012); Grima and Schnell (2008); Van Kampen (1992).

SRDMs are mathematically described as Markov processes (continuous time Markov chains). Therefore, the chemical master equation approach as well as the Gillespie simulation algorithm known for well-stirred reactors apply to SRDMs as well. In the macroscopic limit corresponding to large number of particles in each cell, the Markov jump processes can be approximated by deterministic reaction-diffusion partial differential equations, PDE, (Arnold and Theodosopulu (1980); Debussche and Nguepedja Nankep (2019)) or by stochastic partial differential equations (Van Kampen (1976); Blount (1992, 1993); Asllani et al. (2013)).

SRDMs have already been used to model fluctuations of gene expression in tissues and developing embryos (Pfaffelhuber and Popovic (2015); Smith and Grima (2018); McFann et al. (2021)).

One should note that a generalisation of SRDM is the Reaction-Diffusion Master Equation (RDME) (Gardiner et al. (1976)) which uses voxels for dividing space instead of a regular

grid. The RDME can be simulated using specialized algorithms like the Next Subvolume Method (NSM) (Fange et al. (2010)), which optimize the stochastic simulation for reaction-diffusion scenarios. The RDME has been extended to handle unstructured meshes, enabling simulations in complex geometries Engblom et al. (2009). However, it presents additional numerical challenges, particularly when the mesh size approaches zero, which require careful consideration for reliable results (Fange et al. (2010), Hellander et al. (2012), Isaacson (2009)).

In this study, our primary objective is to employ a hybrid approach. This approach combines both SRDM (Stochastic Reaction-Diffusion Modeling) and differential equations models. The goal is to effectively represent gene expression with spatial considerations. This model should be capable of capturing the variability coming from the inherent stochasticity of transcription. The decision regarding which approach to employ is not solely dependent on accuracy of the model but also on the delicate balance between efficiency, in terms of obtaining results quickly, and accuracy, in terms of having results aligned with experimental knowledge.
Moreover, our investigation centers on critical aspects when modeling the drosophila's blastoderm, with a specific emphasis on the significance of two regulatory factors contributing to spatial stochastic fluctuations: negative self-regulation Coulier et al. (2021) and transcriptional memory Bellec et al. (2018); Dufourt et al. (2018).

The structure of this chapter is as follows: Firstly, in section 5.2, we conduct a comprehensive review of existing modeling methods found in the literature. Subsequently, in section 5.3, we introduce our own model. Following that, in section 5.4, we present a range of simulation methods employed throughout the chapter, taking into account both temporal and spatial scales (refer to Figure 5.1 E, F). Finally, in section 5.5.2, we apply our model within a biological context, with a specific focus on the drosophila's blastoderm. We utilize *snail* as a foundational model

5.2.0 State of the art numerical methods for gene expression modeling

Beside SRDM, there are various simulations methods that have been developed to study gene expression ranging from simple models with deterministic transcription initiation models without diffusion to models that incorporate diffusion and promoter switching. Here are some of the main simulation methods used to model gene expression.

5.2.1 Green Function Reaction Dynamics (GFRD):

GFRD is a method used to model diffusion-limited reactions in continuous space and time (van Zon and ten Wolde (2005); van Zon et al. (2006); Belousov et al. (2018)). GFRD simulates particle-based reactions by considering the interactions between particles and their surrounding environment. It uses the analytical solution of the diffusion reaction using Green's functions to combine in one step the propagation of the particles in space while also taking into consideration the reactions between them. GFRD offers an efficient approach to study reaction dynamics but may face challenges when dealing with complex geometries or when analytical solutions are not readily available. In addition to GFRD, Enhanced Green Function Reaction Dynamics (eGFRD) has been developed in order to incorporate additional enhancements, such as protective domains, to optimize computational efficiency (Sokolowski et al. (2019)).

5.2.2 Mcell and Smoldyn:

Mcell (Gupta et al. (2018b)) and Smoldyn (Andrews and Bray (2004), Andrews et al. (2010), Coulier et al. (2021)) are particle-based simulation methods used in spatial stochastic modeling. They focus on tracking the positions of relevant molecules within a continuous space. Unlike some other simulation methods, Mcell and Smoldyn avoid the use of a mesh, which simplifies their computational approach. Instead of discretizing space, these methods discretize time. This unique approach enables them to effectively simulate various molecular processes, including diffusion, membrane interactions, and reactions of individual molecules. These methods simulate the movement of particles by determining their new random positions based on Smoluchowski's dynamics at each time step. Smoluchowski derives the steady-state reaction rate for diffusion-limited bimolecular reactions, expressing it in terms of the molecular radii and the diffusion coefficients of the reactant species Smoluchowski (1917). Mcell and Smoldyn offer advantages such as efficient handling of complex geometries but introduce a discretization error due to the discretization of time. However they are computationally expensive Coulier et al. (2021).

5.2.3 Hybrid Models:

Multiple hybrid models that combine different simulations method such as particle based models, ODEs, stochastic simulations, have been developed.

For example in Intep et al. (2009), they compare different hybrid methods while also proving uniqueness and convergence of numerical solutions for models transcription/translation without diffusion of molecules.

In Smith and Yates (2018) they give a review of different Spatially coupled hybrid methods in three steps. They give the appropriate modeling technique at each scale. These hybrid models are restricted to cases where different regions of space represent different scales and therefore are modeled using distinct modelling paradigms. The models in these distinct regions of space are typically coupled together through an interface or overlap.

Utilizing the RDME framework to handle spatial simulations, a number of algorithms employing tau-leaping assumptions for species with sufficiently large copy numbers were developed (Marquez-Lago and Burrage (2007), Ferm et al. (2010)).

Additionally, in the work by Coulier et al. (2021), a combination of compartment-based models and the RDME framework is utilized to model various biological processes, including negative-feedback loops, promoter switching, and diffusion of proteins and mRNA while taking into consideration the different abundance of mRNA and protein molecules by deriving transition rates using first-exit times. They compare the different hybrid methods to Smoldyn models as these are the most accurate representations of not well stirred systems.

5.3.0 Model introduction

Stochastic reaction-diffusion models are powerful mathematical frameworks employed to investigate the behavior of dynamic systems in which both random fluctuations and spatial variations play significant roles. At the core of stochastic reaction-diffusion models is the recognition that many natural processes involve not only the diffusion of particles through space but also random encounters and interactions between these particles. To explore the behavior of these systems computationally, we will be using the stochastic algorithm 1.3.1 in order to capture the essence of gene expression in the Drosophila blastoderm.

5.3.1 Example of model

We are interested in modeling gene expression in Drosophila blastoderm which is composed of two main steps: transcription and translation. The inter-compartments reactions that are of interest for us are summarized in Figure 5.1 C and consists of the following reactions

- Promoter switching in a two state telegraph process where the ON and OFF states represent the transcriptionally active and inactive promoter, respectively.

- mRNA production in the nucleus.

- Active transport of mRNA to the cytoplasm.
- Translation of the cytoplasmic mRNA to cytoplasmic protein.
- Active transport of cytoplasmic protein back to the nuclei (useful to when we want to add the auto feedback-loop).

The ON state mentioned in the first bullet point means that the transcription pre-initiation complex is formed and the polymerase is ready to initiate transcription. The OFF state is any other state where these conditions are not fulfilled.
It is of course possible to have several OFF states. Degradation reactions are ubiquitous, as molecular species can be degraded anywhere in space.

We focus on modeling the expression of one gene in 1D. We divide our space domain (which will represent the DV axis of the embryo) $[0, L]$ into N compartments of equal length $h = L/N$ where h needs to justify the Kuramoto's length 5.1.

We denote by $\mathrm{P}_{\mathrm{TS}i}$, $\mathrm{P}_{\mathrm{cyt}_i}$, $\mathrm{RNA}_{\mathrm{TS}i}$, $\mathrm{RNA}_{\mathrm{cyt}_i}$, s_i the concentration of proteins in each compartment, proteins in each compartment with active source, mRNA in compartment, mRNA in compartment with active source and the gene state in the i-th compartment $[(i-1)h,\ ih]$, $i = 1, \ldots, N$, respectively. The active source represent the presence of at least one nuclei.
In each compartment we can have 0, 1, or more than 1 nuclei. In the case of 0 nuclei we consider that $\mathrm{P}_{\mathrm{TS}i} = \mathrm{RNA}_{\mathrm{TS}i} = 0$. This case, when the distance between nuclei is larger than the compartment size h, and is discussed in section 5.3.5. If we have more than 1 nuclei in each compartment than we consider that $\mathrm{P}_{\mathrm{TS}i}$ (resp. $\mathrm{RNA}_{\mathrm{TS}i}$ is the result of averaging of several nuclei that are present in the compartment.

The gene state s_i specifies the state of the promoter, that can be one of several OFF and ON states. For simplicity, we will consider that we have only one OFF and one ON state (telegraph model). In this case $s_i \in \{0, 1\}$ where 0 is the OFF state and 1 is the ON state.

We consider that the diffusivity of the mRNA is limited and therefore they can not move between compartments. On the other hand the cytoplasmic protein can move between compartments by diffusion and this diffusitivity is the between-compartment reaction of our model. This constraint, can be easily lifted by considering that both the protein and the mRNA diffuse with various diffusivities (Figure 5.1 B).

All of the transitions (inter-compartment and between-compartments) are considered to be

Markovian and the resulting model is a Markov jump process described by the following set of chemical reactions:

$$\mathrm{P}_{\mathrm{cyt}_1} \underset{d_h}{\overset{d_h}{\rightleftarrows}} \mathrm{P}_{\mathrm{cyt}_2} \underset{d_h}{\overset{d_h}{\rightleftarrows}} \mathrm{P}_{\mathrm{cyt}_3} \underset{d_h}{\overset{d_h}{\leftrightarrows}} \dots \underset{d_h}{\overset{d_h}{\rightleftarrows}} \mathrm{P}_{\mathrm{cyt}_N} \tag{5.2}$$

$$\mathrm{RNA}_{\mathrm{cyt}_i} \xrightarrow{k_p} \mathrm{RNA}_{\mathrm{cyt}_i} + \mathrm{P}_{\mathrm{cyt}_i} \tag{5.3}$$

$$\mathrm{P}_{\mathrm{cyt}_i} \xrightarrow{k_{entry}} \mathrm{P}_{\mathrm{TS}i} \tag{5.4}$$

$$\mathrm{P}_{\mathrm{cyt}_i} \xrightarrow{\lambda_p} \phi \tag{5.5}$$

$$\mathrm{P}_{\mathrm{TS}i} \xrightarrow{\lambda_p} \phi \tag{5.6}$$

$$s_i \xrightarrow{k_{r,h}} s_i + \mathrm{RNA}_{\mathrm{TS}i} \tag{5.7}$$

$$\mathrm{RNA}_{\mathrm{TS}i} \xrightarrow{k_{exit}} \mathrm{RNA}_{\mathrm{cyt}_i} \tag{5.8}$$

$$\mathrm{RNA}_{\mathrm{cyt}_i} \xrightarrow{\lambda_r} \phi \tag{5.9}$$

$$1 - s_i \xrightarrow{k_{ON_i}} s_i \tag{5.10}$$

$$s_i \xrightarrow{k_{OFF}} 1 - s_i \tag{5.11}$$

where $i = 1, \dots, N$.

Equation 5.2 to 5.11 represent consecutively: protein diffusion, translation, import of protein to the nuclei, protein degradation in cytoplasm, protein degradation in nuclei, transcription, mRNA export to cytoplasm, mRNA degradation in cytoplasm, promoter switching to ON and finally promoter switching to OFF. We disregard the degradation of nuclei mRNA within the nucleus, similar to the approach taken in Bahar Halpern et al. (2015) and Battich et al. (2015). To maintain simplicity, we do not make assumptions about mRNA import into the nucleus or protein export from it.

Our model distinguishes itself from the one outlined in Coulier et al. (2021) through its explicit integration of interactions between neighboring nuclei (by protein diffusion) and its focus on revealing the impact of the morphogene gradient on gene expression within our framework.

5.3.2 Modeling the morphogen gradient

To incorporate the breaking of spatial translation symmetry, we rely on modeling the morphogen gradient. The morphogen gradient refers to the spatial distribution of signaling molecules, characterized by varying concentration levels that form a pattern across space,

influencing transcription regulation and development during embryogenesis. In the case of *snail* , it is well-established that the primary regulator is a transcription factor known as *dorsal* (dl) Ip et al. (1992). Dl directly activates sna expression in the ventral region of the early embryo. We model this morphogen gradient by assuming that the switching rate from the OFF to ON state for the different gene states s depends on the concentration levels of dl, which takes the form of a sum of sigmoidal functions (see Figure 5.1 E). While this concept could be applied to all switching rates, for biological simplicity, we simplify it to a single switching rate, k_{ON}, which is modulated by the sum of sigmoidal functions. Consequently, k_{ON} is given by

$$k_{ON_i} = k_{ON} \times k(i, x_0, x_1, h) = k_{ON} \left(\frac{1}{1 + e^{-(i-x_0)\chi}} + \frac{1}{1 + e^{-(i-x_1)\chi}} - 1 \right) \tag{5.12}$$

where χ represents the steepness of the gradient, for $i \in [0, N]$, $x_0 < x_1$ are the positions where dorsal gradient is maximal (in absolute value).

5.3.3 Modeling the transcription memory

All nuclear cycles start with a period of transcription inactivity due to the structural constraints imposed by the important nuclear re-organization operating during mitosis. This period of inactivity changes stochastically from one nucleous to another, but can depend on the history of activation of ancestor nuclei, a phenomenon called transcriptional memory (Dufourt et al. (2018)). We model trancriptional memory using a finite state Markov chain model proposed in (Bellec et al. (2018); Dufourt et al. (2018)). In the transcriptional memory model, the transcriptionally active ON state is reached by a chain of irreversible transitions made of OFF states denoted by I_j where $1 \leq j \leq M$ where M is the number of OFF state, different from the one used in the telegraph model to describe bursting (figure 5.4 B for $M = 2$). The transcriptional memory is modeled by the following chain of reactions

$$\mathrm{I}^i_M \xrightarrow{k_{I_M}} \mathrm{I}^i_{M-1} \tag{5.13}$$

$$\mathrm{I}^i_{M-1} \xrightarrow{k_{I_{M-1}}} \mathrm{I}^i_{M-2} \tag{5.14}$$

$$\vdots \tag{5.15}$$

$$\mathrm{I}^i_1 \xrightarrow{k_{I_1}} ON_i \tag{5.16}$$

which we incorporate the set of chemical reactions described in 5.3.1 with the minor modification of replacing equations 5.10, 5.11 by

$$\begin{aligned} OFF_i &\xrightarrow{k_{ON_i}} ON_i \\ ON_i &\xrightarrow{k_{OFF}} OFF_i \end{aligned} \tag{5.17}$$

and equation 5.7 by

$$ON_i \xrightarrow{k_{r,h}} ON_i + \mathrm{RNA}_{\mathrm{TS}i} \tag{5.18}$$

5.3.4 Modeling the negative feedback loop

Based on the hypothesis that the *snail* gene represses itself Boettiger and Levine (2013) we introduce a negative feedback loop in the model (see figure 5.4-C. We consider that the protein concentration modulates the promoter switching rates k_{ON} and k_{OFF} (figure 5.4 C). The switching parameters in this case are given by Hill functions

$$k_{OFF_i} = k_{OFF}^{min} + (k_{OFF}^{max} - k_{OFF}^{min}) \times \frac{{\mathrm{P}_{\mathrm{TS}}}_i^{n_{OFF}}}{T_{OFF}^{n_{OFF}} + {\mathrm{P}_{\mathrm{TS}}}_i^{n_{OFF}}} \tag{5.19}$$

$$k_{ON_i} = k_{ON}^{max} + (k_{ON}^{min} - k_{ON}^{max}) \times \frac{{\mathrm{P}_{\mathrm{TS}}}_i^{n_{ON}}}{T_{ON}^{n_{ON}} + {\mathrm{P}_{\mathrm{TS}}}_i^{n_{ON}}} \tag{5.20}$$

where T_{OFF}, T_{ON} represents the concentration threshold of $\mathrm{P_{TS}}$ that results in half-maximal repression, n_{OFF}, n_{ON} represents the Hill coefficient, which describes the steepness of the the repression curve for each switching parameter respectively. These modeling choices are standard (see also Hooshangi and Weiss (2006); To and Maheshri (2010)) and take into account the possibility of cooperative repression corresponding to $n_{OFF} > 1$ or $n_{ON} > 1$.

5.3.5 Modeling distant transcription site

In cases where the distance between nuclei, denoted as d, exceeds the Kuramoto's length, it is better to chose a compartment size h such that $h \leq d$.

In this scenario, we effectively create a delta Dirac source, for RNA production, where not every mesh grid point serves as a source. To account for the finite internuclear distance, we introduce sources only within every $\lfloor d/h \rfloor$ compartments, rather than in every compartment. This approach helps incorporate the influence of the internuclear distance into the simulation (figure 5.4 A). This boils down to writing reactions (5.3), (5.4), (5.7), (5.8), (5.10), (5.11),

only for $i = 1, 1 + r, 1 + 2r, \ldots$, where $r = \lfloor d/h \rfloor$.

5.4.0 Approximations and numerical schemes

We simulated our model with 4 different algorithms, using a full deterministic approach which we call *Det*, then one using a fully stochastic approach *SRDM* with the assumption

of a well stirred model and two hybrid approaches called $H1$ and $H2$ resp. We first apply the different simulation methods to the simplest model, i.e. without feedback loop and memory. Then we use different type of statistical test to uncover the optimal simulation method.

5.4.1 The deterministic limit

Let us consider that the size of each compartment is small but remains large enough to contain a large number of molecules. This situation corresponds to the deterministic limit, that can be obtained from a re-scaling of the model's variables and parameters. More precisely, molecular species copy numbers are replaced by concentrations and promoter state variables by occupation probabilities. We also suppose that

$$\begin{aligned} d_h &= \frac{D}{h^2} = \frac{DN^2}{L^2} \\ k_{r,h} &= h\mu K_r \end{aligned} \tag{5.21}$$

Then $h\mu$ sets the scale of the number of particles in each compartment.

Considering that $N \to \infty$, $\mu \to \infty$ such that $\log N/(h\mu) \to 0$, classical results Blount (1992, 1993); Debussche and Nguepedja Nankep (2019) show that the SRDM converges in probability for the supremum norm to the solution of the following system of differential equations:

$$\frac{\partial p_{ON}}{\partial t} = k_{ON}(x)p_{OFF} - k_{OFF}p_{ON}, \tag{5.22}$$

$$\frac{\partial p_{OFF}}{\partial t} = -k_{ON}(x)p_{OFF} + k_{OFF}p_{ON}, \tag{5.23}$$

$$\frac{\partial [\mathrm{RNA_{TS}}]}{\partial t} = K_r p_{ON} - K_{exit}[\mathrm{RNA_{TS}}], \tag{5.24}$$

$$\frac{\partial [\mathrm{RNA_{cyt}}]}{\partial t} = K_{exit}[\mathrm{RNA_{TS}}] - \lambda_r[\mathrm{RNA_{cyt}}], \tag{5.25}$$

$$\frac{\partial [\mathrm{P_{cyt}}]}{\partial t} = D\frac{\partial^2 [\mathrm{P_{cyt}}]}{\partial x^2} + k_p[\mathrm{RNA_{cyt}}] - \lambda_p[\mathrm{P_{cyt}}] - K_{entry}[\mathrm{P_{cyt}}], \tag{5.26}$$

$$\frac{\partial [\mathrm{P_{TS}}]}{\partial t} = K_{entry}[\mathrm{P_{cyt}}] - \lambda_p[\mathrm{P_{TS}}], \tag{5.27}$$

with Neumann, no flux boundary conditions

$$\frac{\partial [\mathrm{P_{cyt}}]}{\partial x}(x,t) = 0 \text{ for } x \in \partial([0,L]) \quad \forall t > 0, \tag{5.28}$$

where $[\mathrm{RNA_{TS}}] = \mathrm{RNA_{TS}}/(\mu h)$, $[\mathrm{RNA_{cyt}}] = \mathrm{RNA_{cyt}}/(\mu h)$, $[\mathrm{P_{TS}}] = \mathrm{P_{TS}}/(\mu h)$, $[\mathrm{P_{cyt}}] = \mathrm{P_{cyt}}/(\mu h)$ and p_{ON}, p_{OFF} are the probabilities of promoters being ON and OFF, respectively.

We have opted to utilize the finite differences method and more precisely the Forward Euler

Scheme to numerically solve Equation 5.26 and, consequently, Equation 5.27. Although there are other established numerical techniques available for solving such linear systems, as demonstrated in references Smith et al. (1985) and Brenner and Carstensen (2004), we have chosen the finite differences method due to its efficiency, in terms of speed, in obtaining numerical solutions for these partial differential equations.

In this approach we consider that all molecules are deterministic. The Euler Scheme involves dividing the spatial and temporal domains into a mesh, where the solution of the PDE is estimated by taking into consideration the stability condition given by

$$\Delta_t \leq \frac{1}{2}\frac{h^2}{D} \tag{5.29}$$

where h and Δ_t are the space and time mesh sizes, respectively.

5.4.2 Stochastic approach

We are going to use Gillespie algorithm for the simulations of the SRDM (Gillespie (1977)). To cite Gillespie, "the probability, given $X(t) = x$, that the system's next reaction will occur in the infinitesimal time interval $[t, t + \tau]$, and will be of stoichiometry corresponding to the jth reaction" where τ is the time until next event given by

$$\tau = \frac{1}{\alpha_0} ln \left(\frac{1}{r_1}\right) \tag{5.30}$$

with α_0 is the combined rate of all possible reactions also known as the sum of propensities.

The algorithm for our SRDM problem is the classical gillespie algorithm given in section 1.3.1 by taking into account the multiple compartments separated by Kuramoto length. When there is no diffusion and no feedback-loop, the sites function can be simulated independently.

5.4.3 Hybrid methods

Gillespie simulation of the SRDM is too costly in terms of execution time. The finite differences integration of deterministic PDE approximation does not render the stochastic fluctuations generated by the SRDM. A compromise can be obtained by hybrid modeling where some variables are considered discrete and are simulated using the Gillespie algorithm and other variables are considered continuous and follow PDEs. For our model, two discrete/continuous splittings are natural, leading to two hybrid models: H1, where the promoter states are discrete whereas the mRNA and protein concentrations are continuous, and H2, where the promoter states and the mRNA are discrete and the protein is continuous.

More precisely, in H1 and H2 the variables s_i resp. $(s_i, \mathrm{RNA_{TS}}_i, \mathrm{RNA_{cyt}}_i)$ are considered discrete and the rest of the variables are considered continuous. The continuous variables follow PDEs as follows:

$$\begin{aligned}
\frac{\partial[\mathrm{RNA_{TS}}]}{\partial t} &= K_r \textstyle\sum_{i=1}^{N_{TS}} s_i \delta(x - x_i) - K_{exit}[\mathrm{RNA_{TS}}] \sum_{i=1}^{N_{TS}} \delta(x - x_i), \\
\frac{\partial[\mathrm{RNA_{cyt}}]}{\partial t} &= K_{exit}[\mathrm{RNA_{TS}}] - \lambda_r[\mathrm{RNA_{cyt}}] \textstyle\sum_{i=1}^{N_{TS}} \delta(x - x_i), \\
\frac{\partial[\mathrm{P_{cyt}}]}{\partial t} &= D\frac{\partial^2[\mathrm{P_{cyt}}]}{\partial x^2} + k_p[\mathrm{RNA_{cyt}}] - \lambda_p[\mathrm{P_{cyt}}] - K_{entry}[\mathrm{P_{cyt}}] \textstyle\sum_{i=1}^{N_{TS}} \delta(x - x_i), \\
\frac{\partial[\mathrm{P_{TS}}]}{\partial t} &= K_{entry}[\mathrm{P_{cyt}}] \textstyle\sum_{i=1}^{N_{TS}} \delta(x - x_i) - \lambda_p[\mathrm{P_{TS}}],
\end{aligned} \tag{5.31}$$

for H1 and

$$\begin{aligned}
\frac{\partial[\mathrm{P_{cyt}}]}{\partial t} &= D\frac{\partial^2[\mathrm{P_{cyt}}]}{\partial x^2} + \textstyle\sum_{i=1}^{N_{TS}} (k_p[\mathrm{RNA_{cyt}}] - K_{entry}[\mathrm{P_{cyt}}])\delta(x - x_i) - \lambda_p[\mathrm{P_{cyt}}], \\
\frac{\partial[\mathrm{P_{TS}}]}{\partial t} &= K_{entry} \textstyle\sum_{i=1}^{N_{TS}} [\mathrm{P_{cyt}}]\delta(x - x_i) - \lambda_p[\mathrm{P_{TS}}]
\end{aligned} \tag{5.32}$$

for H2, where N_{TS}, x_i are the number and positions of the nuclei and δ is the Dirac delta distribution.

The sums of Dirac delta distributions in (5.31) and (5.32) are spatially discretized in the same way as for modeling the finite distance between nuclei in the section 2.

H1

The algorithm in this case is going to calculates which reaction occurs next (between the different promoter state s) and when. The next reaction is obtained based on the propensity function and the time until next event follows an exponential distribution in the case where we don't have a feedback loop. The adaptation of the algorithm to the presence of a feedback loop is discussed at the end of this section. Thus the algorithm realisation of the H1 is provided in Algorithm 3.

H2

The algorithm for H2 (see Algorithm 4) proceeds by firstly generating N independent time until the next event for each compartment (a diffusive jump between mRNA production, degradation and switching of the states) according to the Gillespie algorithm Gillespie (1977)

Algorithm 3 H1 algorithm

(B1) Set the initial condition for the stochastic variables (promoter state).

(B2) Set the initial condition for the deterministic ${RNA_{TS}}_i^0$, ${RNA_{cyt}}_i^0$ ${P_{TS}}_i^0$, ${P_{cyt}}_i^0$ for each $i \in [0, LN \times h]$.

(B3) Initiate the timing for the simulations t.

(B4) Discretize the space into mesh points and specify the PDE-update time step Δ_t satisfying condition 5.29.

(B5) Generate $2N_a$ random variables uniformly distributed in $(0, 1)$, where N_a is the number of transcriptionally active compartment. The first N_a will choose the time for the next event, and the other N_a will be responsible for the next reaction.

(B6) Compute propensity functions of the transcription model $\alpha_j^i(t)$ at time t for each position $i \in [0, N_a]$.

(B7) Calculate the sum of the propensity function $\alpha_0^i = \sum_{j=1}^{R} \alpha_j^i(t)$ where R is the number of reactions per position.

(B8) Determine the time for the next 'stochastic simulation' event, $t^i = t^i + \tau^i$, where τ^i is given 5.30 for each i and keep $\tau = min(t_i)_{i \in [1, N_a]}$.

(B9) Find the next reaction $j = \{1, ..., S\}$ according to

$$\sum_{r=0}^{r=i-1} \alpha_r < r_2 \alpha_0 \leq \sum_{r=0}^{r=i} \alpha_r \tag{5.33}$$

for each i.

(B10) Update the stochastic variable (s_i) at $t + \tau$.

(B11) Update the deterministic variables
- (a) solve the analytical equation of equations 5.31, 5.31 between t^p and $t + \tau$ for each $i \in [0, N \times h]$.
- (b) solve equations 5.31, 5.31 according to the finite difference method between t and $t + \tau$ with initial condition ${P_{cyt}}_0$, ${P_{TS}}_0$.
- (c) Update the PDE simulator time $t = t + \tau$.
- (d) Set ${RNA_{cyt}}_i^0 = {RNA_{cyt}}_i^{t_1^p}$, ${RNA_{TS}}_i^0 = {RNA_{TS}}_i^{t_1^p}$
- (e) Update the initial condition for the next PDE simulator

$$ {P_{cyt}}_0 = {P_{cyt}}^t, \quad {P_{TS}}_0 = {P_{TS}}^t. \tag{5.34}$$

which will represent the transcriptional model. Then it calculates the deterministic variables using finite-difference method. The algorithm is given in Algorithm 4.

Algorithm 4 H2 algorithm

(C1) Set the initial condition for the stochastic variable (promoter state, mRNA) in each compartment.

(C2) Set the initial condition for the deterministic variable ($P_{cyt} = P_{cyt_0}$), ($P_{TS} = P_{TS0}$)

(C3) Initiate the timing for the simulations t.

(C4) Discretize the space into mesh points and specify the PDE-update time step Δ_t satisfying condition 5.29.

(C5) Generate $2N_a$ random numbers uniformly distributed in (0, 1) (the first N_a will chose the time into the next event and the other N_a will be responsible for the next reaction), where N_a is the number of "active position"

(C6) Compute propensity functions of the transcription model $\alpha_j^i(t)$ at time t for each position $i \in [0, N_a]$

(C7) Calculate the sum of the propensity function $\alpha_0^i = \sum_{j=1}^{R} \alpha_j^i(t)$ where R is the number of reactions per position.

(C8) Determine the time for the next 'stochastic simulation' event, $t^i = t^i + \tau^i$, where τ^i is given by equation 5.30 for each i and keep $\tau = min(t_i)_{i\in[1,N_a]}$.

(C9) Find the next reaction $j = \{1, ..., S\}$ according to 5.33 for each i.

(C10) update the stochastic variables at $t + \tau$

(C11) Update the deterministic variables:
- (a) solve equations 5.32, 5.32 according to the finite difference method between t and $t + \tau$ with initial condition P_{cyt_0}, P_{TS0}.
- (b) update the PDE simulator time $t = t + \tau$.
- (c) update the initial condition for the next PDE simulator

$$P_{TS0} = {P_{TS}}^t, \quad P_{cyt_0} = {P_{cyt}}^t \tag{5.35}$$

Adaptation of the hybrid models to feedback loop

In the case of a feedback loop, Algorithm 3 and Algorithm 4 need to be modified where step (B8) in Algorithm 3 (resp. (C8) in Algorithm 4) is replaced by

While $\Delta_t < \tau_i$ for all $i \in [1, N]$

1. Update the deterministic variables according to step (B11) in Algorithm 3 (resp. (C11) in Algorithm 4)
2. Update the propensity function to take into consideration the change in P_{TS} concentration
3. Calculate τ_i according to 5.30

EndWhile

where Δ_t is the time mesh size used for Euler Scheme (see equation 5.29.

5.5.0 Numerical results

In this section, we assess the accuracy and limitations of our proposed simulation techniques by subjecting them to a series of carefully selected statistical tests. These tests aim to measure the precision of our methods in reproducing the behavior described by the Stochastic Reaction-Diffusion Model (SRDM). We conduct these evaluations for both mRNA and protein levels to comprehensively validate our approach. Then, we demonstrate a practical application of the simulations to biological data.

5.5.1 Approximation quality

For each algorithm, we generated a 41-compartment grid for a simulation duration of 60 minutes. We repeated the simulations 500 times. The chosen parameters for these simulations were based on information found in the bibliography given below.

In the study by Pimmett et al. (2021), the switching rates K_{ON}, K_{OFF}, and K_{ini} of *snail* were computed in the central region of the pattern, where the gene is assumed to be at its maximum transcription capacity. This was achieved through live imaging and deconvolution of the signal.
Additionally, in Boettiger and Levine (2013), the half-life of mRNA ($T_{1/2}$) was computed, which relates to the degradation rate (γ_r) by $\gamma_r = \frac{ln(2)}{T_{1/2}}$. Furthermore, Boettiger et al. Boettiger and Levine (2013) computed the half-life of mRNA ($T_{1/2}$), which is related to the degradation rate (γ_r) by $\gamma_r = \frac{ln(2)}{T_{1/2}}$.
Regarding the diffusion rate, it was required to be sufficiently fast so that a well-mixed model could effectively capture the crucial aspects of the spatial dynamics that we are interested in therefore we chose a diffusion rate of $D = 0.0005 \mu m^2 s^{-1}$
As for the translation rate, import rate, export rate and the minimum value for K_{ON} there is

limited information available.

The summary of the parameters used in these simulations are provided in the table 5.1.

Name	Symbol	Formula
Number of compartments	N	41
Compartment size	h	0.0128 μm
Length of the space	L	1 μm
Diffusion rate	D	5×10^{-4} $\mu m^2 s^{-1}$
Promoter switching to OFF	k_{OFF}	0.004 s^{-1}
Promoter switching to ON	$k_{ON}(x)$	0.042 s^{-1}
Translation rate	k_p	$\lambda_r \times 4$ s^{-1}
Protein half-life	$T^p_{1/2}$	26×60 s^{-1}
Transcription rate	k_r	0.113 s^{-1}
mRNA half-life	$T^m_{1/2}$	13×60 s^{-1}
Lower bound of k_{ON}	k^{min}_{ON}	$k_{ON} \times 10^{-4}$ s^{-1}
mRNA export rate	K_{entry}	0.0417 s^{-1}
Protein import rate	K_{exit}	0.005 s^{-1}

Table 5.1: Parameter Values

A sample of each algorithm was chose randomly for visualisation the protein concentration in the cytoplasm in figure 5.2-A. On the side of each method a time and space point was chosen, to plot the cytoplasmic mRNA concentration that contributed to generate the protein concentration shown in the heatmap. A comparison between the different simulation methods was done using the Kolmogorov distance metrics and by comparing the distance between a set of summary statistics.

Kolmogorov distance

To measure the disparity between the data distributions of the various simulation methods used, we employed the Kolmogorov distance. This metric is frequently employed in the comparison of simulation methods Cao and Petzold (2006) Coulier et al. (2021).

The Kolmogorov distance (also known as the Kolmogorov-Smirnov distance) is the maximum difference between two cumulative distribution functions (CDFs) given by

$$D_{KS} = \sup_a |F_X(a) - F_Y(a)| \tag{5.36}$$

where D_{KS} represents the Kolmogorov distance, $F_X(a) = \mathbb{P}[X \leq a]$, (resp. $F_Y(a) = \mathbb{P}[Y \leq a]$) is the the cumulative distribution function (CDF) of the distribution X and Y respectively.

We conducted comparisons of the Kolmogorov distance between the hybrid methods and SRDM, examining every combination of time and space coordinates denoted as (t, x). For each specific (t, x) point, and for each simulation method, we have a dataset consisting of 500 values coming from the number of simulations we generated. This approach verifies that the datasets for all simulation methods originate from the same underlying distribution. However, the deterministic approach, denoted as *Det*, lacks variability. Consequently, at each (t, x) point, we have only one value, making it impossible to apply the Kolmogorov distance. The outcomes of these analyses are illustrated on the heatmap presented in Figure 5.3 A.

Summary of statistics

Summary statistics, such as moments, play a crucial role in understanding data as they provide insights into a dataset's shape and characteristics without relying on knowledge of the underlying distribution. Selecting appropriate summary statistics can be challenging in real-world scenarios.

For the sake of simplicity, we have chosen a set of three commonly used summary statistics: the mean value, the standard deviation, and the Fano factor.

The Fano factor represents the variance-to-mean ratio, a significant parameter for quantifying the departure from Poisson statistics. It is commonly employed to describe the variability arising from gene expression. These summary statistics offer valuable information about the central tendency, variability, and distribution characteristics of the the different simulation methods.

Mean error and standard deviation We calculated the average values from multiple simulations (n=500) and determined the relative error w.r.t. the mean in SRDM. This resulted in a 2D error map. Subsequently, we generated plots illustrating the mean and standard deviation of these errors across space (Figure 5.3-B). To emphasize the influence of the morphogen gradient, represented by the variation in k_{ON}, we also plotted a spatially normalized visualization of k_{ON} in the same plot.

Fano Factor An essential criterion for gene expression modeling is the capacity to capture the noise inherent to transcription, stemming from its stochastic nature. While cells might leverage gene expression noise for fitness gains in fluctuating environments (Acar et al.,

2008), noise is generally detrimental. Precise internal regulation of biochemical reactions is vital for cell growth and survival. In the context of the gene *snail*, expression variability could disrupt gastrulation, yielding incomplete ventral furrows or even halting gastulation.

A crucial factor in mathematical model selection is the model's ability to replicate inherent transcription noise, particularly at the protein level (Perry et al., 2010; McFann et al., 2021). The Fano Factor (FF) assesses this variability. To compute the FF with spatial extension, our domain was divided into three sections based on spatial switching parameter percentages relative to the highest value, as shown in Figure 5.3-C. Figure 5.3-E displays FF results over time for mRNA and protein.

Our analysis of the simulations results using the statistical tests mentioned above revealed that the hybrid model $H2$ outperformed the others in predicting gene expression in space and time in function of the balance between accuracy and timing (Figure 5.3 D). While yielding comparable statistical test results to the other models, its distinction lies in its capacity to accurately reproduce the inherent noise. It has approximately the same noise output as SRDM at the level of mRNA and is still able to capture the noise at the level of protein. However, it is important to note that our simulation experiments had some limitations, such as the simplifications and assumptions made in the models, and the specific parameters used in the simulations. Nonetheless, our results provide strong evidence that the hybrid model $H2$ is the most effective approach for modeling gene expression in space and time and can be used to gain deeper insights into the complex regulatory mechanisms underlying gene expression in living organisms.

5.5.2 Application to biology

We modeled each regulatory mechanisms presented in Section 5.3 in conjunction with an appropriate propensity function in the Hybrid method. The simulation results for all these mechanisms (transcriptional memory, negative feedback, finite distance between nuclei) are illustrated in Figure 5.4-D. Each subplot is presented in the same format as Figure 5.2-A.

Relative contribution of different layers of spatial stochastic fluctuations

To assess the relative contributions of the aforementioned mechanisms to the spatial stochastic fluctuations of the Drosophila blastoderm, we compared the mRNA output from real data generated in Lagha's lab Virginia_Paper to simulation results at the nuclei mRNA level, using three criteria: heatmaps illustrating the distribution of transcription activity in time (x-axis) and space (y-axis) (figure 5.5-A), the Fano Factor of transcription as a function of time (figure 5.5-b), the percentage of active nuclei over time (figure 5.5-C). The mRNA intensity signal was obtained using the MS2-MCP system from the central region of the *snail* expres-

sion domain, and was calibrated to provide the number of mRNA molecules present in each nucleus at every time step.

5.6.0 Conclusion

In conclusion, our study underscores the critical role of spatial dynamics in shaping the behavior of biological systems. When dealing with large-scale modeling tasks, such as model exploration and parameter inference, it becomes imperative to identify the most cost-effective simulation method capable of accurately capturing the dynamics of interest.

Throughout our investigation, we conducted a comprehensive comparison between various simulation methods, including classic Gillespie algorithm and alternative approaches, both hybrid and deterministic. Our findings unequivocally demonstrate that the newly proposed hybrid approximation, specifically $H2$, significantly broadens the applicability of the SRDM model. This computational innovation offers a remarkable advantage by accurately replicating the the behavior of interest: noise originating from the promoter level to the level of protein process. Additionally, this model provides valuable insights into the key regulatory elements governing *snail* protein dynamics, with a particular focus on feedback loops and memory effects.

The choice of noise quantification as a metric for selecting the simulation method arises from the increasing relevance of cell-to-cell variability in contemporary research. This source of variation has accumulated significant attention due to its potential to influence various statistical properties of biochemical reactions and, consequently, the probability of specific phenotypic outcomes. However one needs to note that, despite the presence of stochastic fluctuations, many phenotypic traits exhibit a degree of robustness.

Moving forward, our model serves as a solid foundation for further exploration and benchmarking. Future efforts will investigate into an exhaustive analysis of the free parameters within our simulations, seeking to answer crucial questions about the optimal parameter ranges for auto-feedback mechanisms in noise reduction. This research direction promises to enhance our understanding of the intricate interplay between spatial dynamics, cellular variability, and gene expression, ultimately advancing our knowledge of complex biological systems.

Figure 5.1: Model introduction for spatial gene expression.

A) On the left we have a Drosophila embryo showing the gradient distribution of the

snail mRNA. On the right, we have a closer look at the embryo, highlighting how the diffusion can differ between the location of transcription site (apically or basally) before invagination.

B) A simplified model showing bursty transcription leading to protein production.

C) A model of simple diffusion. In each pseudo-compartment containing a nucleus there is stochastic mRNA and protein production. Both products can diffuse locally with varied diffusion lengths.

D) *Dorsal gradient* affecting the switching rate $k_O N(s)$ of transcription in space.

E) Stochastic signal example of nascent mRNA in time for the same space point.

Figure 5.2: Simulation results.

A) Simulations output of the different methods. For each method we have a heatmap of the protein concentration in time (xaxis) and space (yaxis). On top of the heatmap is the output of the mRNA in time of the respected method at the middle point of space. On the right of the heatmap is the output of mRNA in space at the final time point (60 mn).

B) Different variables plotted together. Above slow $K_{ON} = 0.00013$ rate. Below: Fast $k_{ON} = 0.00024$ rate. Blue denotes RNA_{TS}, orange denotes RNA_{cyt}, green denotes P_{cyt}

Figure 5.3: Simulation performance.

A) KS distance output. Each column represent a variable and each row represent the output of the different hybrid method w.r.t. SRDM in time (x-axis) and space (y-axis).

B) Error of the mean and std of the different methods over the different simulation method and over time. The output is then plotted in function of space. The dashed line are a normalized k_{ON} value w.r.t. the maximum error in order to visualized the error in space in function of the morphogene gradient.

C) Sections of the space dimension that was used to compute FF in different spatial conditions. The choice of the division was taken w.r.t. the switching parameters are ,40%,

between 40 and 80% and above 80% of the maximal intensity respectively.

D) Table of execution time of the different methods. The timing is done for 1 simulations with 40 mesh points and 60 min duration.

E) FF results over the different sections at the level of TS mRNA, cytoplasmic mRNA and cytoplasmic protein.

Figure 5.4: Additional biological layers.

A) Representation of delta Dirac source point. We divide the space into a mesh grid with h step (represented in squares). We then consider that not all of the compartments are active sources highlighting the active compartment with navy blue according to a delta Dirac source .

B) Transcriptional memory: To incorporate the time that each gene is active with respect to mitosis, n inactive states are added to the Markov model of transcription. We initiate each nuclei randomly between one of the n inactive states.

C) Another important property in gene expression is the presence of feedback loop. In this case we consider that we have a negative feedback loop affecting the switching between promoter state ON, OFF.

D) Simulations results of the hybrid methods with different added properties.

Figure 5.5: Visual comparison of the simulations output w.r.t. real data

A) HeatMap of mRNA output in section III of the spatial space. First plot is real data of *snail* transgenic line NC14 followed by simulations output.

B) on the left, F.F. of the mRNA output of the different properties added to the hybrid method. On right, FF output of different transgenic lines of *snail*.

C) Percentage of non zero nuclei in function of time with simulated output on the left and real data on the right.

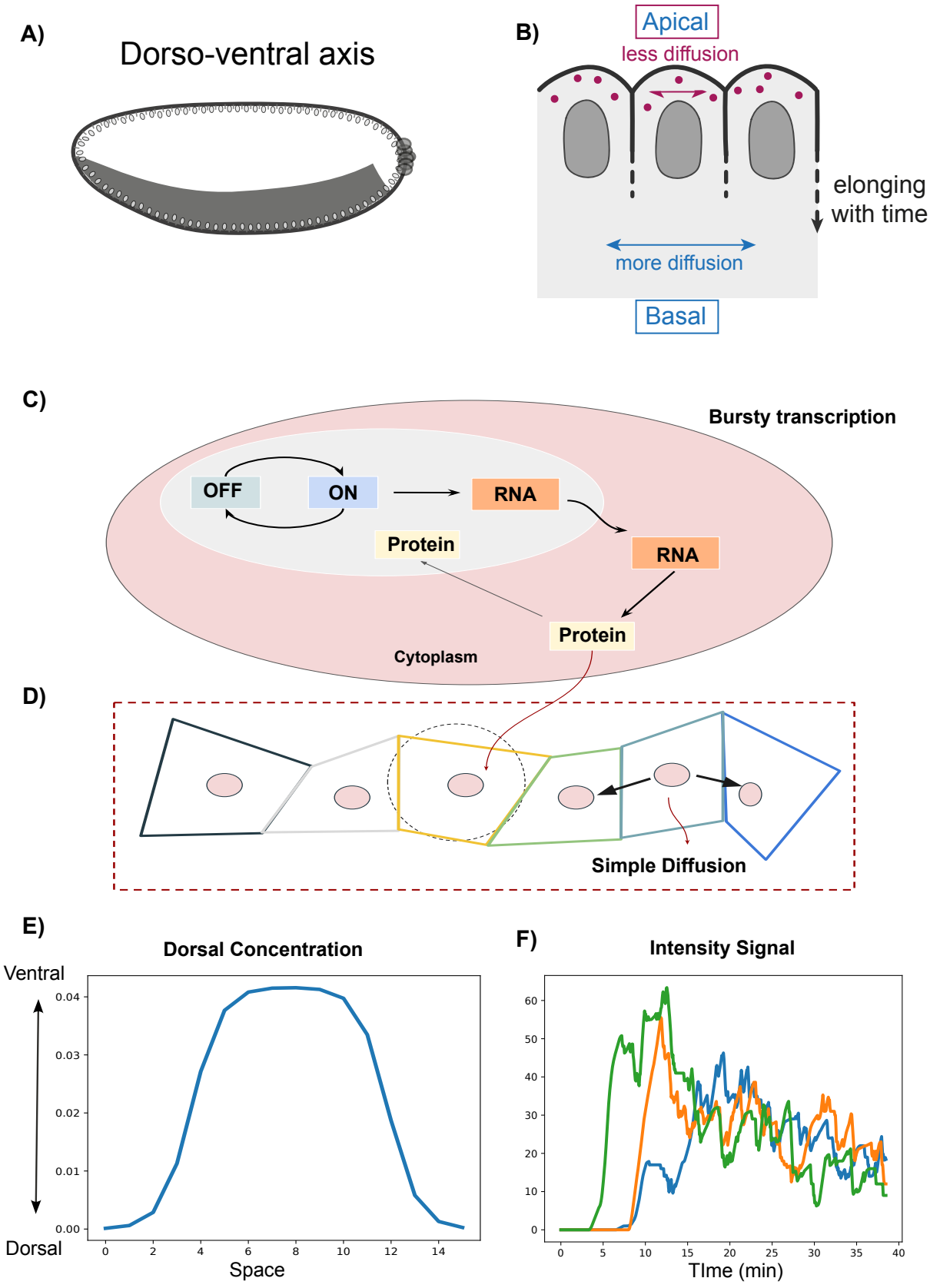


Figure 5.1: Model introduction for spatial gene expression.

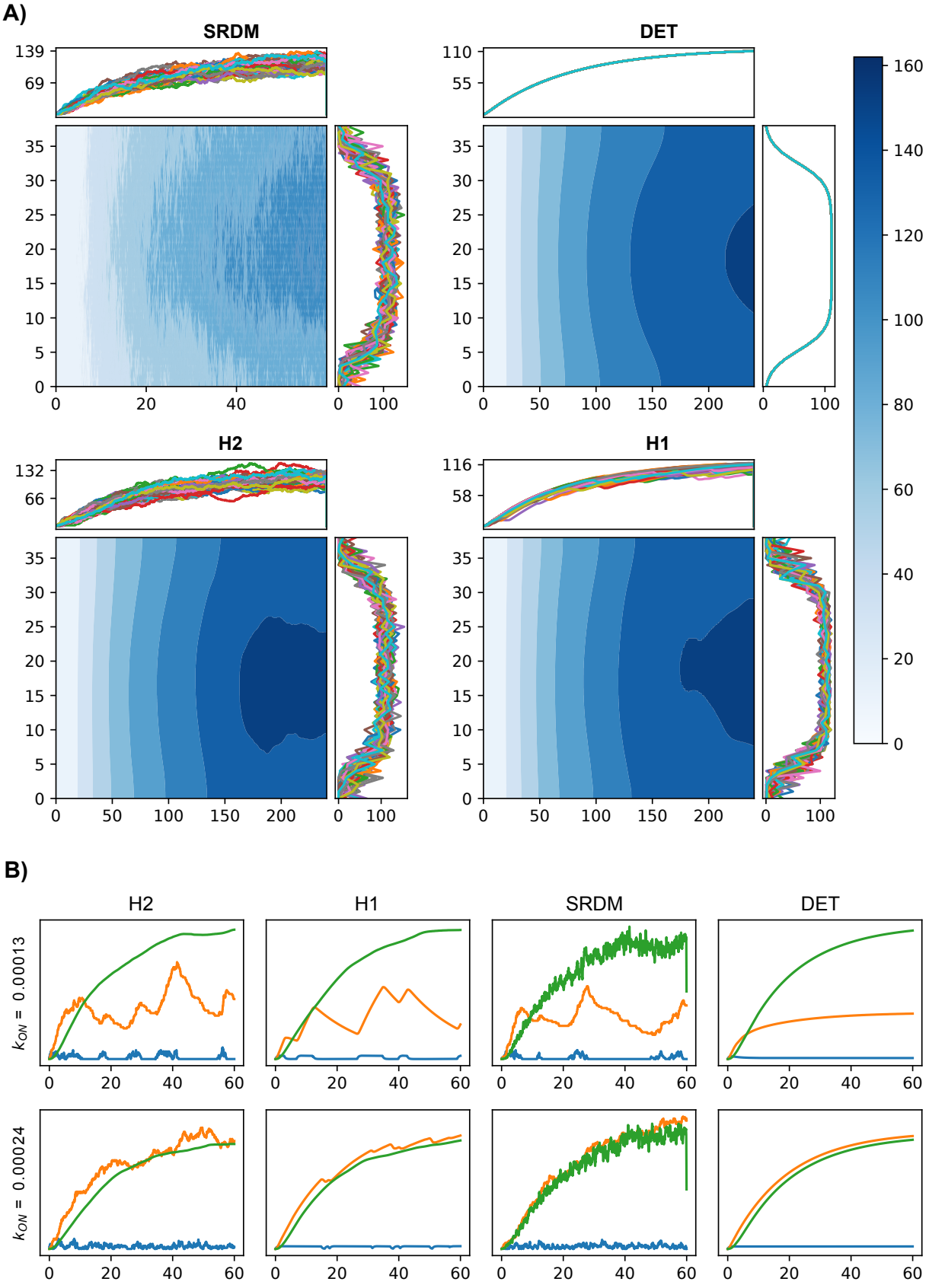


Figure 5.2: Simulation results.

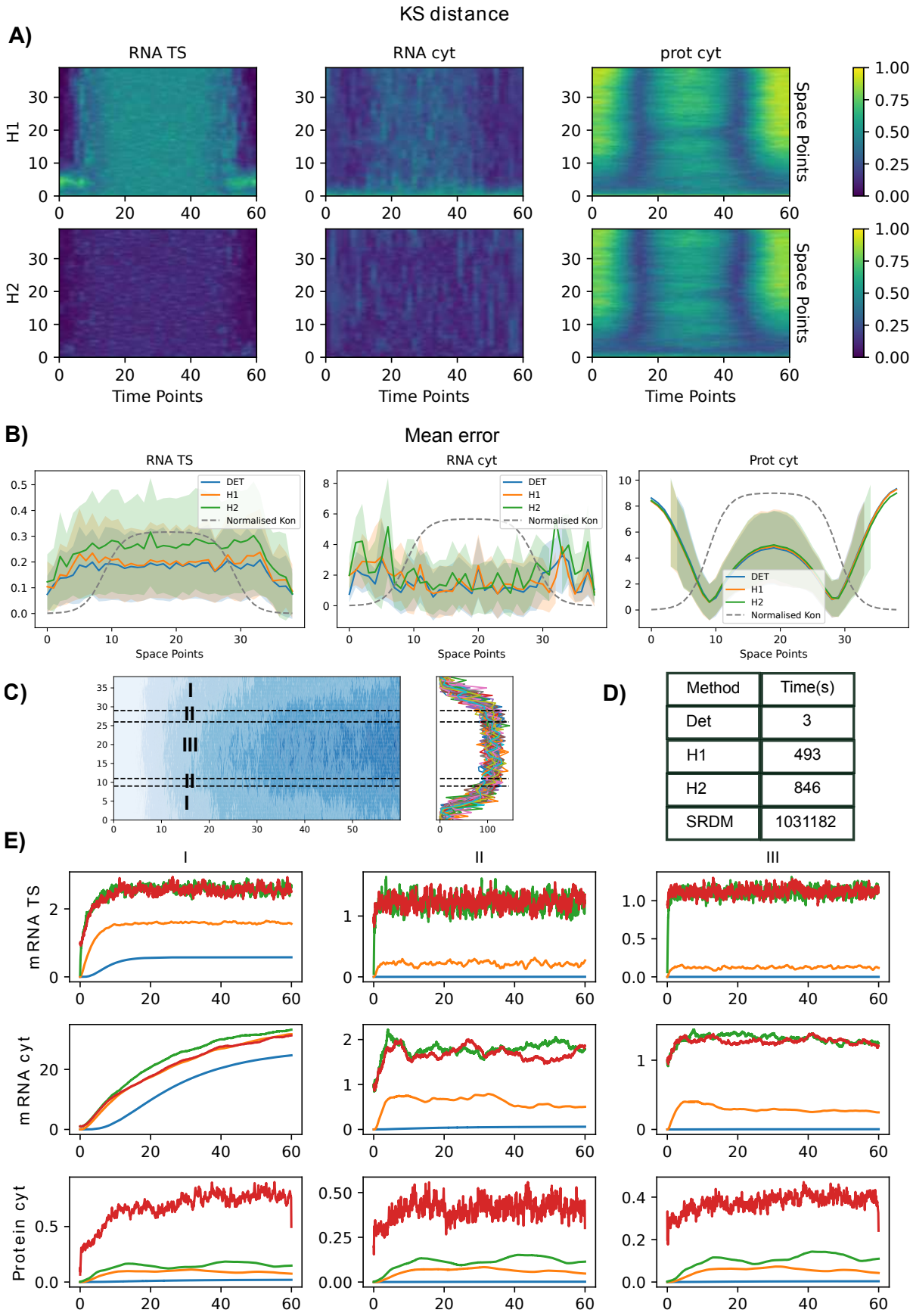


Figure 5.3: Simulation performance.

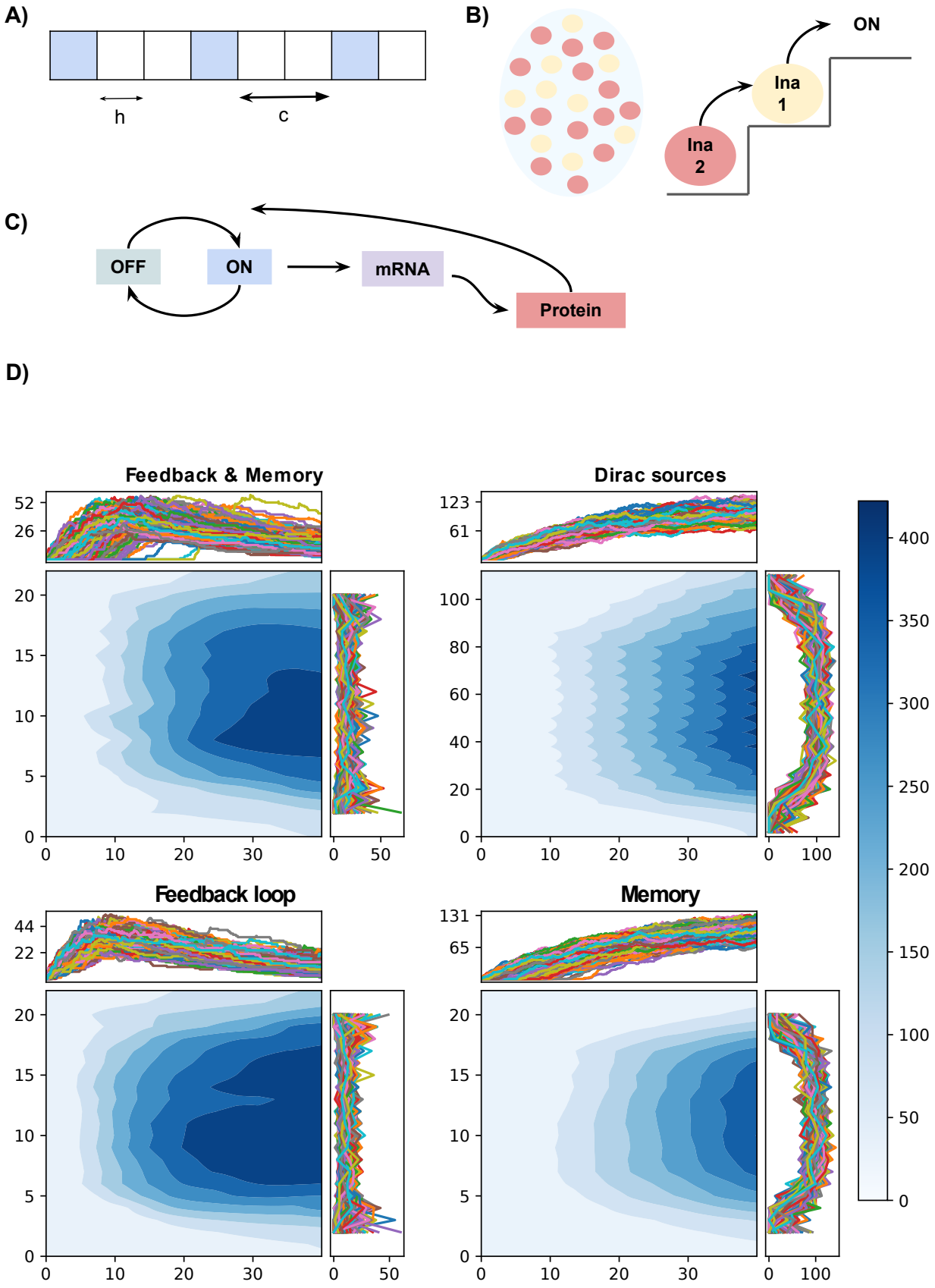


Figure 5.4: Additional biological layers.

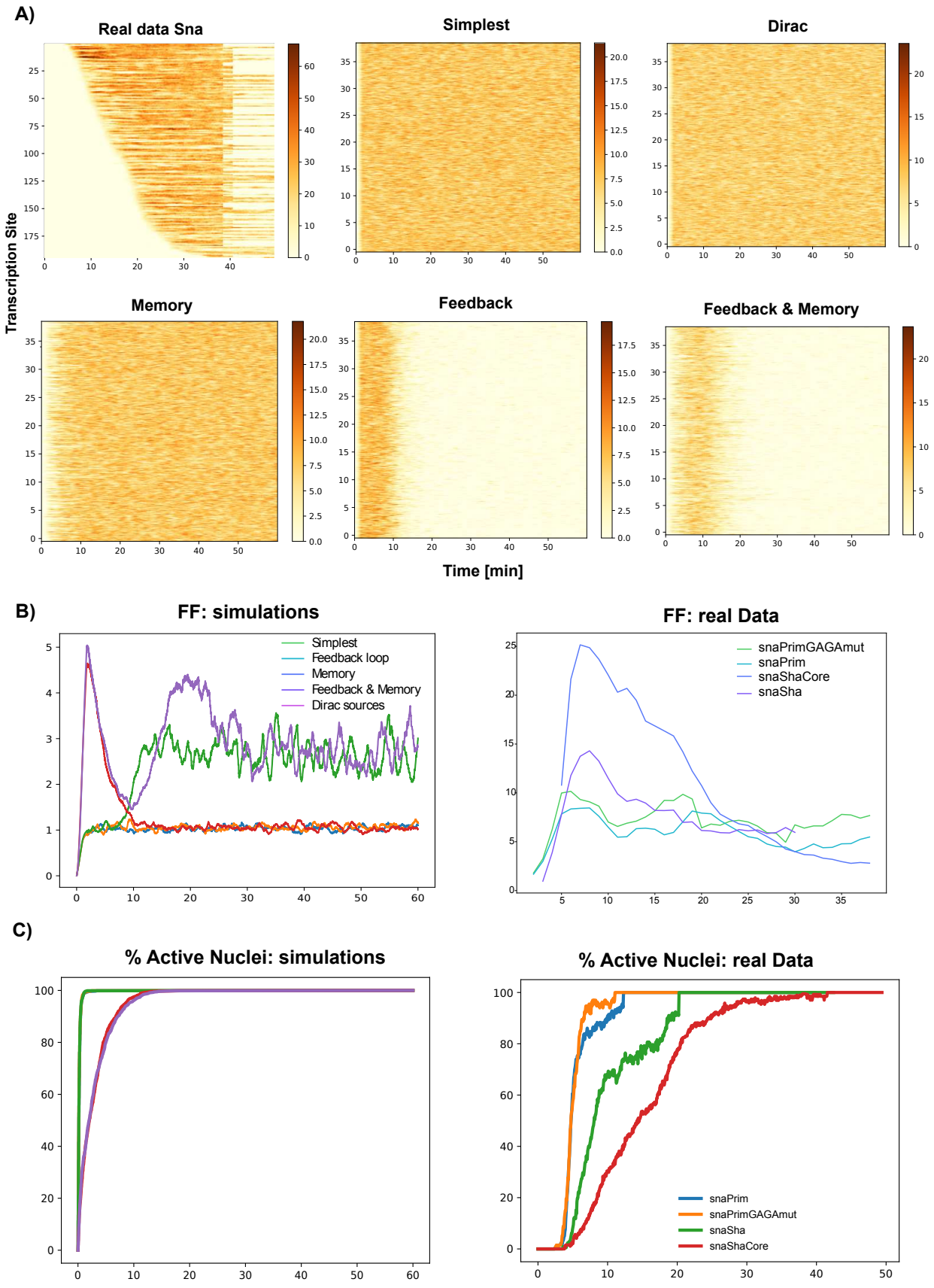


Figure 5.5: Visual comparison of the simulations output w.r.t. real data

APPENDIX A

List of simple Thomas decomposition systems

A.1.0 The Remaining Simple Systems for Model M1

$$\mathcal{S}_2 = \{ -L_2^2 S_1^2 - L_3 S_1^3 + 2 L_2 L_3 S_1 - L_3^2 + (L_2 S_1^3 - L_3 S_1^2) k_1 + (L_2 S_1^3 - L_3 S_1^2) L_1 = 0,$$
$$L_3 S_1 + (L_2 S_1 - L_3) k_3 = 0,\ k_2 S_1^2 + L_2 S_1 - L_3 = 0,\ k_4 - S_1 = 0,\ k_5 + S_1 = 0,$$
$$L_2 S_1 - L_3 \neq 0,\ S_1 \neq 0,\ S_2 = 0\},$$
$$\mathcal{S}_3 = \{k_1 k_2 S_1 S_2 + k_2^2 S_1 S_2 + L_3 S_2 + (L_2 S_2 - L_3 S_1) k_2 = 0,\ k_3 k_2 S_1 - L_3 = 0,\ k_2 \neq 0,$$
$$k_4 = 0,\ k_5 + S_1 = 0,\ L_1 S_1 S_2 - L_2 S_2 + L_3 S_1 - S_2^2 = 0,\ S_1^2 - S_2 = 0,\ S_2 \neq 0\},$$
$$\mathcal{S}_4 = \{k_1 S_2 + k_3 S_2 + L_2 S_1 = 0, k_2 = 0,\ k_4 S_1 - S_1^2 + S_2 = 0,\ k_5 + S_1 = 0,$$
$$L_1 S_1 S_2 - L_2 S_1^2 - S_2^2 = 0,\ L_3 = 0,\ S_1 \neq 0,\ S_2 \neq 0\},$$
$$\mathcal{S}_5 = \{k_1 + k_3 + L_1 = 0,\ k_2 = 0,\ k_4 - S_1 = 0,\ k_5 + S_1 = 0,\ L_2 = 0,\ L_3 = 0,$$
$$S_1 \neq 0,\ S_2 = 0\},$$

$$\mathcal{S}_6 = \{k_1\,k_2 + k_2^2 + k_4^2 + k_4\,L_1 + (2\,k_4 + L_1)\,k_2 + L_2 = 0,\ k_3\,k_2 - k_2\,k_4 - k_4^2 - k_4\,L_1 - L_2 = 0,$$
$$k_2 \neq 0,\ k_5 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$
$$\mathcal{S}_7 = \{k_1 + k_3 + k_4 + L_1 = 0,\ k_2 = 0,\ k_4^2 + k_4\,L_1 + L_2 = 0,\ k_5 = 0,\ L_1^2 - 4\,L_2 \neq 0,$$
$$L_2 \neq 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$
$$\mathcal{S}_8 = \{k_1 + k_3 = 0,\ k_2 = 0,\ k_4 + L_1 = 0,\ k_5 = 0,\ L_1 \neq 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$
$$\mathcal{S}_9 = \{k_1 + k_3 + L_1 = 0,\ k_2 = 0,\ k_4 = 0,\ k_5 = 0,\ L_1 \neq 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$
$$\mathcal{S}_{10} = \{2\,k_1 + 2\,k_3 + L_1 = 0,\ k_2 = 0,\ 2\,k_4 + L_1 = 0,\ k_5 = 0,\ L_1^2 - 4\,L_2 = 0,\ L_2 \neq 0,\ L_3 = 0,$$
$$S_1 = 0,\ S_2 = 0\},$$
$$\mathcal{S}_{11} = \{k_1 + k_3 = 0,\ k_2 = 0,\ k_4 = 0,\ k_5 = 0,\ L_1 = 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\}.$$

A.2.0 The Remaining Simple Systems for Model M6

$$\mathcal{S}_2 = \{k_1\,k_3\,S_1 + k_3^2\,S_1 + k_3\,L_1\,S_1 + L_3 = 0,\ k_2\,k_3\,S_1 - L_3 = 0,\ k_3 \neq 0,\ k_4 - S_1 = 0,$$
$$k_5 + S_1 = 0,\ L_2\,S_1 - L_3 = 0,\ S_1 \neq 0,\ S_2 = 0\},$$
$$\mathcal{S}_3 = \{k_1\,S_2 + k_2\,S_2 + L_2\,S_1 = 0,\ k_3 = 0,\ k_4\,S_1 - S_1^2 + S_2 = 0,\ k_5 + S_1 = 0,$$
$$L_1\,S_1\,S_2 - L_2\,S_1^2 - S_2^2 = 0,\ L_3 = 0,\ S_1 \neq 0,\ S_2 \neq 0\},$$
$$\mathcal{S}_4 = \{k_1 + k_2 + L_1 = 0,\ k_3 = 0,\ k_4 - S_1 = 0,\ k_5 + S_1 = 0,\ L_2 = 0,\ L_3 = 0,$$
$$S_1 \neq 0,\ S_2 = 0\},$$
$$\mathcal{S}_5 = \{k_1\,k_3 + k_3^2 + k_4^2 + k_4\,L_1 + (k_4 + L_1)\,k_3 + L_2 = 0,\ k_2\,k_3 - k_4^2 - k_4\,L_1 - L_2 = 0,$$
$$k_3 \neq 0,\ k_5 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$
$$\mathcal{S}_6 = \{k_1 + k_2 + k_4 + L_1 = 0,\ k_3 = 0,\ k_4^2 + k_4\,L_1 + L_2 = 0,\ k_5 = 0,\ L_1^2 - 4\,L_2 \neq 0,$$
$$L_2 \neq 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$
$$\mathcal{S}_7 = \{k_1 + k_2 = 0,\ k_3 = 0,\ k_4 + L_1 = 0,\ k_5 = 0,\ L_1 \neq 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,$$
$$S_2 = 0\},$$
$$\mathcal{S}_8 = \{k_1 + k_2 + L_1 = 0,\ k_3 = 0,\ k_4 = 0,\ k_5 = 0,\ L_1 \neq 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,$$

$S_2 = 0\big\},$

$\mathcal{S}_9 = \big\{2\,k_1 + 2\,k_2 + L_1 = 0,\ k_3 = 0,\ 2\,k_4 + L_1 = 0,\ k_5 = 0,\ L_1^2 - 4\,L_2 = 0,\ L_2 \neq 0,$

$L_3 = 0,\ S_1 = 0,\ S_2 = 0\big\},$

$\mathcal{S}_{10} = \big\{k_1 + k_2 = 0,\ k_3 = 0,\ k_4 = 0,\ k_5 = 0,\ L_1 = 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\big\}.$

A.3.0 The Remaining Simple Systems for Model M7

$\mathcal{S}_2 = \big\{L_1\,S_1\,S_2 - L_2\,S_1^2 - S_2^2 + (S_1^3 - S_1\,S_2)\,k_1 = 0,\ k_3\,S_1 + L_1\,S_1 - S_2 = 0,$

$-L_1\,S_1\,S_2 + L_2\,S_1^2 + S_2^2 + (S_1^3 - S_1\,S_2)\,k_2 = 0,\ k_4\,S_1 - S_1^2 + S_2 = 0,\ k_5 + S_1 = 0,$

$L_1\,S_1 - S_2 \neq 0,\ L_3 = 0,\ S_1^3 - S_1\,S_2 \neq 0,\ S_2 \neq 0\big\},$

$\mathcal{S}_3 = \big\{L_1\,S_1^2 - L_2\,S_1 - S_1\,S_2 + (S_1^2 - S_2)\,k_1 = 0,\ -L_1\,S_1\,S_2 + L_2\,S_1^2 + S_2^2 + (S_1^3 - S_1\,S_2)\,k_2 = 0,$

$k_3 = 0,\ k_4\,S_1 - S_1^2 + S_2 = 0,\ k_5 + S_1 = 0,\ L_1\,S_1 - S_2 \neq 0,\ L_3 = 0,\ S_1^3 - S_1\,S_2 \neq 0,\ S_2 \neq 0\big\},$

$\mathcal{S}_4 = \big\{-2\,L_2\,S_1^2 - S_1^2\,S_2 + 2\,L_3\,S_1 - S_2^2 + (2\,S_1^3 - 2\,S_1\,S_2)\,k_1 + (S_1^3 + S_1\,S_2)\,L_1 = 0,$

$-L_1\,S_1\,S_2 + L_2\,S_1^2 - L_3\,S_1 + S_2^2 + (S_1^3 - S_1\,S_2)\,k_2 = 0,\ 2\,k_3\,S_1 + L_1\,S_1 - S_2 = 0,$

$k_4\,S_1 - S_1^2 + S_2 = 0,\ k_5 + S_1 = 0,\ L_1^2\,S_1^2 - 2\,L_1\,S_1\,S_2 - 4\,L_3\,S_1 + S_2^2 = 0,$

$L_3 \neq 0,\ S_1^3 - S_1\,S_2 \neq 0,\ S_2 \neq 0\big\},$

$\mathcal{S}_5 = \big\{-L_2\,S_1 + (S_1^2 - S_2)\,k_1 = 0,\ L_2\,S_1 + (S_1^2 - S_2)\,k_2 = 0,\ k_3 = 0,\ k_4\,S_1 - S_1^2 + S_2 = 0,$

$k_5 + S_1 = 0,\ L_1\,S_1 - S_2 = 0,\ L_3 = 0,\ S_1^3 - S_1\,S_2 \neq 0,\ S_2 \neq 0\big\},$

$\mathcal{S}_6 = \big\{k_1\,S_1^2 + k_3\,S_1^2 + L_1\,S_1^2 - L_2\,S_1 + L_3 = 0,\ k_2\,S_1^2 + L_2\,S_1 - L_3 = 0,\ k_4 - S_1 = 0,$

$k_3^2\,S_1 + k_3\,L_1\,S_1 + L_3 = 0,\ k_5 + S_1 = 0,\ L_1^2\,S_1 - 4\,L_3 \neq 0,\ L_3 \neq 0,\ S_1 \neq 0,\ S_2 = 0\big\},$

$\mathcal{S}_7 = \big\{k_1\,S_1 - L_2 = 0,\ k_2\,S_1 + L_2 = 0,\ k_3 + L_1 = 0,\ k_4 - S_1 = 0,\ k_5 + S_1 = 0,\ L_1 \neq 0,\ L_3 = 0,$

$S_1 \neq 0,\ S_2 = 0\big\},$

$\mathcal{S}_8 = \big\{k_1\,S_1 + L_1\,S_1 - L_2 = 0,\ k_2\,S_1 + L_2 = 0,\ k_3 = 0,\ k_4 - S_1 = 0,\ k_5 + S_1 = 0,\ L_1 \neq 0,$

$L_3 = 0,\ S_1 \neq 0,\ S_2 = 0\big\},$

$\mathcal{S}_9 = \big\{2\,k_1\,S_1^2 + L_1\,S_1^2 - 2\,L_2\,S_1 + 2\,L_3 = 0,\ k_2\,S_1^2 + L_2\,S_1 - L_3 = 0,\ 2\,k_3 + L_1 = 0,\ k_4 - S_1 = 0,$

$k_5 + S_1 = 0,\ L_1^2\, S_1 - 4\, L_3 = 0,\ L_3 \neq 0,\ S_1 \neq 0,\ S_2 = 0\big\},$

$\mathcal{S}_{10} = \big\{k_1\, S_1 - L_2 = 0,\ k_2\, S_1 + L_2 = 0,\ k_3 = 0,\ k_4 - S_1 = 0,\ k_5 + S_1 = 0,\ L_1 = 0,\ L_3 = 0,$

$S_1 \neq 0,\ S_2 = 0\big\},$

$\mathcal{S}_{11} = \big\{k_1\, S_1\, S_2 + k_2\, S_1\, S_2 + k_3\, S_1\, S_2 + L_2\, S_2 - L_3\, S_1 = 0,\ k_4 = 0,\ k_5 + S_1 = 0,$

$k_3^2\, S_1\, S_2 + L_3\, S_2 + (L_2\, S_2 - L_3\, S_1)\, k_3 = 0,\ L_1\, S_1\, S_2 - L_2\, S_2 + L_3\, S_1 - S_2^2 = 0,$

$L_2^2\, S_2 - 2\, L_2\, L_3\, S_1 - 4\, L_3\, S_1\, S_2 + L_3^2 \neq 0,\ L_3 \neq 0,\ S_1^2 - S_2 = 0,\ S_2 \neq 0\big\},$

$\mathcal{S}_{12} = \big\{k_1 + k_2 = 0,\ k_3\, S_1 + L_2 = 0,\ k_4 = 0,\ k_5 + S_1 = 0,\ L_1\, S_1 - L_2 - S_2 = 0,\ L_2 \neq 0,$

$L_3 = 0,\ S_1^2 - S_2 = 0,\ S_2 \neq 0\big\},$

$\mathcal{S}_{13} = \big\{k_1\, S_1 + k_2\, S_1 + L_2 = 0,\ k_3 = 0,\ k_4 = 0,\ k_5 + S_1 = 0,\ L_1\, S_1 - L_2 - S_2 = 0,\ L_2 \neq 0,$

$L_3 = 0,\ S_1^2 - S_2 = 0,\ S_2 \neq 0\big\},$

$\mathcal{S}_{14} = \big\{2\, k_1\, S_1\, S_2 + 2\, k_2\, S_1\, S_2 + L_2\, S_2 - L_3\, S_1 = 0,\ 2\, k_3\, S_1\, S_2 + L_2\, S_2 - L_3\, S_1 = 0,\ k_4 = 0,$

$k_5 + S_1 = 0,\ L_1\, S_1\, S_2 - L_2\, S_2 + L_3\, S_1 - S_2^2 = 0,\ L_2^2\, S_2 - 2\, L_2\, L_3\, S_1 - 4\, L_3\, S_1\, S_2 + L_3^2 = 0,$

$L_3 \neq 0,\ S_1^2 - S_2 = 0,\ S_2 \neq 0\big\},$

$\mathcal{S}_{15} = \big\{k_1 + k_2 = 0,\ k_3 = 0,\ k_4 = 0,\ k_5 + S_1 = 0,\ L_1\, S_1 - S_2 = 0,\ L_2 = 0,\ L_3 = 0,$

$S_1^2 - S_2 = 0,\ S_2 \neq 0\big\},$

$\mathcal{S}_{16} = \big\{k_1\, k_4 - k_3^2 - k_3\, L_1 - L_2 = 0,\ k_2\, k_4 + k_3^2 + k_4^2 + k_4\, L_1 + (k_4 + L_1)\, k_3 + L_2 = 0,\ k_4 \neq 0,$

$k_5 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\big\},$

$\mathcal{S}_{17} = \big\{k_1 + k_2 + k_3 + L_1 = 0,\ k_3^2 + k_3\, L_1 + L_2 = 0,\ k_4 = 0,\ k_5 = 0,\ L_1^2 - 4\, L_2 \neq 0,\ L_2 \neq 0,$

$L_3 = 0,\ S_1 = 0,\ S_2 = 0\big\},$

$\mathcal{S}_{18} = \big\{k_1 + k_2 = 0,\ k_3 + L_1 = 0,\ k_4 = 0,\ k_5 = 0,\ L_1 \neq 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\big\},$

$\mathcal{S}_{19} = \big\{k_1 + k_2 + L_1 = 0,\ k_3 = 0,\ k_4 = 0,\ k_5 = 0,\ L_1 \neq 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\big\},$

$\mathcal{S}_{20} = \big\{2\, k_1 + 2\, k_2 + L_1 = 0,\ 2\, k_3 + L_1 = 0,\ k_4 = 0,\ k_5 = 0,\ L_1^2 - 4\, L_2 = 0,\ L_2 \neq 0,$

$L_3 = 0,\ S_1 = 0,\ S_2 = 0\big\},$

$\mathcal{S}_{21} = \big\{k_1 + k_2 = 0,\ k_3 = 0,\ k_4 = 0,\ k_5 = 0,\ L_1 = 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\big\}.$

A.4.0 The Remaining Simple Systems for Model M8

$$\mathcal{S}_2 = \{k_1 = 0,\ -L_1 S_1^2 + L_2 S_1 + S_1 S_2 + (L_1 S_1 - S_2) k_3 = 0,$$
$$L_1 S_1 S_2 - L_2 S_1^2 - S_2^2 + (L_1 S_1^2 - S_1 S_2) k_4 = 0,\ L_1 S_1 + S_1 k_2 - S_2 = 0,\ k_5 + S_1 = 0,$$
$$L_1 S_1 - S_2 \neq 0,\ L_3 = 0,\ S_1 \neq 0\},$$
$$\mathcal{S}_3 = \{(L_1 S_1 - S_2) k_1 + 2 L_3 = 0,$$
$$2 L_2 S_1^2 + S_1^2 S_2 - 2 L_3 S_1 + S_2^2 + (L_1 S_1^2 - S_1 S_2) k_3 + (-S_1^3 - S_1 S_2) L_1 = 0,$$
$$2 L_1 S_1 S_2 - 2 L_2 S_1^2 + 2 L_3 S_1 - 2 S_2^2 + (L_1 S_1^2 - S_1 S_2) k_4 = 0,\ L_1 S_1 + 2 S_1 k_2 - S_2 = 0,$$
$$k_5 + S_1 = 0,\ L_1^2 S_1^2 - 2 L_1 S_1 S_2 - 4 L_3 S_1 + S_2^2 = 0,\ L_3 \neq 0,\ S_1 \neq 0\},$$
$$\mathcal{S}_4 = \{L_2 S_1 + S_2 k_1 = 0,\ -S_1^2 + S_1 k_3 + S_1 k_4 + S_2 = 0,\ k_2 = 0,\ k_5 + S_1 = 0,$$
$$L_1 S_1 S_2 - L_2 S_1^2 - S_2^2 = 0,\ L_3 = 0,\ S_1 \neq 0,\ S_2 \neq 0\},$$
$$\mathcal{S}_5 = \{k_1 + L_1 = 0,\ k_3 + k_4 - S_1 = 0,\ k_2 = 0,\ k_5 + S_1 = 0,\ L_2 = 0,\ L_3 = 0,\ S_1 \neq 0,\ S_2 = 0\},$$
$$\mathcal{S}_6 = \{k_1 + k_3 + k_4 + k_2 + L_1 = 0,$$
$$k_3^2 + k_4^2 + k_2^2 + k_2 L_1 + (2 k_4 + k_2 + L_1) k_3 + (2 k_2 + L_1) k_4 + L_2 = 0,$$
$$-L_1^2 + 2 L_1 k_2 + 3 k_2^2 + 4 k_2 k_4 + 4 L_2 \neq 0,\ k_2 \neq 0,\ k_5 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$
$$\mathcal{S}_7 = \{k_1 + k_3 + k_4 + L_1 = 0,\ k_3^2 + k_4^2 + k_4 L_1 + (2 k_4 + L_1) k_3 + L_2 = 0,\ k_2 = 0,\ k_5 = 0,$$
$$L_1^2 - 4 L_2 \neq 0,\ L_2 \neq 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$
$$\mathcal{S}_8 = \{k_1 = 0,\ k_3 + k_4 + L_1 = 0,\ k_2 = 0,\ k_5 = 0,\ L_1 \neq 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$
$$\mathcal{S}_9 = \{k_1 + L_1 = 0,\ k_3 + k_4 = 0,\ k_2 = 0,\ k_5 = 0,\ L_1 \neq 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$
$$\mathcal{S}_{10} = \{2 k_1 + k_2 + L_1 = 0,\ L_1^2 - k_2^2 + 4 k_2 k_3 - 4 L_2 = 0,$$
$$-L_1^2 + 2 L_1 k_2 + 3 k_2^2 + 4 k_2 k_4 + 4 L_2 = 0,\ k_2 \neq 0,\ k_5 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$
$$\mathcal{S}_{11} = \{2 k_1 + L_1 = 0,\ 2 k_3 + 2 k_4 + L_1 = 0,\ k_2 = 0,\ k_5 = 0,\ L_1^2 - 4 L_2 = 0,\ L_2 \neq 0,$$
$$L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$
$$\mathcal{S}_{12} = \{k_1 = 0,\ k_3 + k_4 = 0,\ k_2 = 0,\ k_5 = 0,\ L_1 = 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\}.$$

A.5.0 The Remaining Simple Systems for Model M3

$$\mathcal{S}_2 = \{ -L_1 S_1^2 + L_2 S_1 + S_1 S_2 + (L_1 S_1 - S_2) k_3 = 0,\ k_1 = 0,$$
$$L_1 S_1 S_2 - L_2 S_1^2 - S_2^2 + (L_1 S_1^2 - S_1 S_2) k_4 = 0,\ k_2 S_1 + L_1 S_1 - S_2 = 0,\ k_5 + S_1 = 0,$$
$$L_1 S_1 - S_2 \neq 0,\ L_3 = 0,\ S_1 \neq 0\},$$

$$\mathcal{S}_3 = \{L_1 S_1 S_2 - L_2 S_1^2 - S_2^2 + (L_1 S_1^2 - S_1 S_2) k_3 = 0,\ k_1 S_1 + L_1 S_1 - S_2 = 0,$$
$$-L_1 S_1^2 + L_2 S_1 + S_1 S_2 + (L_1 S_1 - S_2) k_4 = 0,\ k_2 = 0,\ k_5 + S_1 = 0,\ L_1 S_1 - S_2 \neq 0,$$
$$L_3 = 0,\ S_1 \neq 0\},$$

$$\mathcal{S}_4 = \{k_3 S_1 + k_4 S_1 - S_1^2 + S_2 = 0,\ L_2 S_1 + (S_1^2 + S_2) k_1 - L_3 = 0,\ k_5 + S_1 = 0,$$
$$L_2 S_1 + (S_1^2 + S_2) k_2 - L_3 = 0,\ -2 L_2 S_1^2 - S_1^2 S_2 + 2 L_3 S_1 - S_2^2 + (S_1^3 + S_1 S_2) L_1 = 0,$$
$$L_2^2 S_1^3 - 2 L_2 L_3 S_1^2 + L_3^2 S_1 + (-S_1^4 - 2 S_1^2 S_2 - S_2^2) L_3 = 0,\ L_3 \neq 0,\ S_1^3 + S_1 S_2 \neq 0,\ S_2 \neq 0\},$$

$$\mathcal{S}_5 = \{k_3 S_1 + k_4 S_1 - S_1^2 + S_2 = 0,\ k_1 = 0,\ k_2 = 0,\ k_5 + S_1 = 0,\ L_1 S_1 - S_2 = 0,\ L_2 = 0,$$
$$L_3 = 0,\ S_1^3 + S_1 S_2 \neq 0,\ S_2 \neq 0\},$$

$$\mathcal{S}_6 = \{k_3 + k_4 - S_1 = 0,\ k_1 S_1^2 + L_2 S_1 - L_3 = 0,\ k_2 S_1^2 + L_2 S_1 - L_3 = 0,\ k_5 + S_1 = 0,$$
$$L_1 S_1^2 - 2 L_2 S_1 + 2 L_3 = 0,\ L_2^2 S_1^2 - L_3 S_1^3 - 2 L_2 L_3 S_1 + L_3^2 = 0,\ L_3 \neq 0,\ S_1 \neq 0,\ S_2 = 0\},$$

$$\mathcal{S}_7 = \{k_3 + k_4 - S_1 = 0,\ k_1 = 0,\ k_2 = 0,\ k_5 + S_1 = 0,\ L_1 = 0,\ L_2 = 0,\ L_3 = 0,\ S_1 \neq 0,$$
$$S_2 = 0\},$$

$$\mathcal{S}_8 = \{k_3 S_1 + k_4 S_1 + 2 S_2 = 0,\ 2 k_1 S_1 + L_1 S_1 - S_2 = 0,\ 2 k_2 S_1 + L_1 S_1 - S_2 = 0,\ k_5 + S_1 = 0,$$
$$L_1^2 S_2 + 2 L_1 S_1 S_2 + 4 L_3 S_1 - S_2^2 = 0,\ L_2 S_2 + L_3 S_1 = 0,\ L_3 \neq 0,\ S_1^2 + S_2 = 0,\ S_2 \neq 0\},$$

$$\mathcal{S}_9 = \{k_3 S_1 + k_4 S_1 + 2 S_2 = 0,\ k_1 = 0,\ k_2 = 0,\ k_5 + S_1 = 0,\ L_1 + S_1 = 0,\ L_2 = 0,\ L_3 = 0,$$
$$S_1^2 + S_2 = 0,\ S_2 \neq 0\},$$

$$\mathcal{S}_{10} = \{k_3 k_4 + k_4^2 + k_2^2 + k_2 L_1 + (2 k_2 + L_1) k_4 + L_2 = 0,\ k_1 k_4 - k_4 k_2 - k_2^2 - k_2 L_1 - L_2 = 0,$$
$$k_4 \neq 0,\ k_5 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$

$$\mathcal{S}_{11} = \{k_3 + k_1 + k_2 + L_1 = 0,\ k_4 = 0,\ k_2^2 + k_2 L_1 + L_2 = 0,\ k_5 = 0,\ L_1^2 - 4 L_2 \neq 0,\ L_2 \neq 0,$$
$$L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$

$$\mathcal{S}_{12} = \{k_3 + k_1 = 0,\ k_4 = 0,\ k_2 + L_1 = 0,\ k_5 = 0,\ L_1 \neq 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$

$$\mathcal{S}_{13} = \{k_3 + k_1 + L_1 = 0,\ k_4 = 0,\ k_2 = 0,\ k_5 = 0,\ L_1 \neq 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$

$$\mathcal{S}_{14} = \{2\,k_3 + 2\,k_1 + L_1 = 0,\ k_4 = 0,\ 2\,k_2 + L_1 = 0,\ k_5 = 0,\ L_1^2 - 4\,L_2 = 0,\ L_2 \neq 0,$$
$$L_3 = 0,\ S_1 = 0,\ S_2 = 0\},$$

$$\mathcal{S}_{15} = \{k_3 + k_1 = 0,\ k_4 = 0,\ k_2 = 0,\ k_5 = 0,\ L_1 = 0,\ L_2 = 0,\ L_3 = 0,\ S_1 = 0,\ S_2 = 0\}.$$

Bibliography

Odd O Aalen. Phase type distributions in survival analysis. *Scandinavian journal of statistics*, pages 447–463, 1995.

Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Samaneh Aminikhanghahi and Diane J. Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 2017. ISSN 0219-3116. doi: 10.1007/s10115-016-0987-z. URL https://doi.org/10.1007/s10115-016-0987-z.

Steven S Andrews and Dennis Bray. Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. *Physical Biology*, 1(3):137, aug 2004. doi: 10.1088/1478-3967/1/3/001. URL https://dx.doi.org/10.1088/1478-3967/1/3/001.

Steven S Andrews, Nathan J Addy, Roger Brent, and Adam P Arkin. Detailed simulations of cell biology with smoldyn 2.1. *PLoS Comput Biol*, page e1000705, March 2010.

L Arnold and M Theodosopulu. Deterministic limit of the stochastic model of chemical reactions with diffusion. *Advances in Applied Probability*, 12(2):367–379, 1980.

Ludwig Arnold. Mathematical models of chemical reactions. In Michiel Hazewinkel and Jan C. Willems, editors, *Stochastic Systems: The Mathematics of Filtering and Identification*

and Applications, pages 111–134, Dordrecht, 1981. Springer Netherlands. ISBN 978-94-009-8546-9.

Azam Asanjarani, Benoit Liquet, and Yoni Nazarathy. Estimation of semi-markov multi-state models: a comparison of the sojourn times and transition intensities approaches. *The International Journal of Biostatistics*, 18(1):243–262, 2021.

Malbor Asllani, Tommaso Biancalani, Duccio Fanelli, and Alan J McKane. The linear noise approximation for reaction-diffusion systems on networks. *The European Physical Journal B*, 86:1–10, 2013.

Søren Asmussen and Mogens Bladt. Renewal theory and queueing algorithms for matrix-exponential distributions. *Lecture notes in pure and applied mathematics*, pages 313–342, 1996.

T. Bächler and M. Lange-Hegermann. *AlgebraicThomas and DifferentialThomas: Thomas Decomposition for algebraic and differential systems*, 2008-2012. (`https://www.art.rwth-aachen.de/cms/MATHB/Forschung/Mathematische-Software/~lqnwi/`).

T. Bächler, V.P. Gerdt, M. Lange-Hegermann, and D. Robertz. Algorithmic Thomas decomposition of algebraic and differential systems. *J. Symbolic Comput.*, 47(10):1233–1266, 2012.

Keren Bahar Halpern, Inbal Caspi, Doron Lemze, Maayan Levy, Shanie Landen, Eran Elinav, Igor Ulitsky, and Shalev Itzkovitz. Nuclear retention of mRNA in mammalian tissues. *Cell Rep*, December 2015.

Daniel Barry and John A Hartigan. Product partition models for change point problems. *The Annals of Statistics*, pages 260–279, 1992.

Anthony F Bartholomay. *A stochastic approach to chemical reaction kinetics*. PhD thesis, 1957.

Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Control of transcript variability in single mammalian cells. *Cell*, 163:1596–1610, 12 2015. doi: 10.1016/j.cell.2015.11.018.

Martin Baurmann, Thilo Gross, and Ulrike Feudel. Instabilities in spatially extended predator–prey systems: Spatio-temporal patterns in the neighborhood of Turing–Hopf bifurcations. *Journal of Theoretical Biology*, pages 220–229, 2007. doi: https://doi.org/10.1016/j.jtbi.2006.09.036. URL `https://www.sciencedirect.com/science/article/pii/`

S0022519306004280.

Maëlle Bellec, Ovidiu Radulescu, and Mounia Lagha. Remembering the past: Mitotic bookmarking in a developing embryo. *Current Opinion in Systems Biology*, 11:41–49, 2018.

Roman Belousov, Adrian Jacobo, and A. J. Hudspeth. Fluctuation theory in space and time: White noise in reaction-diffusion models of morphogenesis. *Phys. Rev. E*, Nov 2018. doi: 10.1103/PhysRevE.98.052125. URL https://link.aps.org/doi/10.1103/PhysRevE.98.052125.

M P Belvin and K V Anderson. A conserved signaling pathway: the drosophila toll-dorsal pathway. *Annu Rev Cell Dev Biol*, 12:393–416, 1996.

E Bertrand, P Chartrand, M Schaefer, S M Shenoy, R H Singer, and R M Long. Localization of ASH1 mRNA particles in living yeast. *Mol Cell*, 1998.

Albert T Bharucha-Reid. *Elements of the Theory of Markov Processes and their Applications*. Courier Corporation, 1997.

Jonathan Bieler, Christian Pozzorini, and Felix Naef. Whole-embryo modeling of early segmentation in drosophila identifies robust and fragile expression domains. *Biophysical journal*, 101(2):287–296, 2011.

Mogens Bladt. A review on phase-type distributions and their use in risk theory. *ASTIN Bulletin: The Journal of the IAA*, 35(1):145–161, 2005.

Moisés Blanco Calvo, Victoria Bolós Fernández, Vanessa Medina Villaamil, Guadalupe Aparicio Gallego, Silvia Díaz Prado, and Enrique Grande Pulido. Biology of bmp signalling and cancer. *Clinical and Translational Oncology*, 2009. ISSN 1699-3055. doi: 10.1007/S12094-009-0328-8. URL https://doi.org/10.1007/S12094-009-0328-8.

Douglas Blount. Law of large numbers in the supremum norm for a chemical reaction with diffusion. *The annals of applied probability*, pages 131–141, 1992.

Douglas Blount. Limit theorems for a sequence of nonlinear reaction-diffusion systems. *Stochastic Processes and their Applications*, 45(2):193–207, 1993.

Alistair Nicol Boettiger and Michael Levine. Rapid transcription fosters coordinate snail expression in the drosophila embryo. *Cell Rep*, 2013.

Jacques P Bothma, Hernan G Garcia, Emilia Esposito, Gavin Schlissel, Thomas Gregor, and Michael Levine. Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living drosophila embryos. *Proc Natl Acad Sci U S A*, 111(29):10598–10603, July 2014.

Ashe Bowles, Hoppe and Rattray. Scalable inference of transcriptional kinetic parameters from ms2 time series data. *Bioinformatics (Oxford, England)*, 38(4):1030–1036, 2021.

Susanne C. Brenner and Carsten Carstensen. *Finite Element Methods*, chapter 4. John Wiley & Sons, Ltd, 2004. ISBN 9780470091357. doi: https://doi.org/10.1002/0470091355.ecm003. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/0470091355.ecm003`.

Peter Buchholz, Jan Kriege, and Iryna Felko. *Input modeling with phase-type distributions and Markov models: theory and applications*. Springer, 2014.

Michael Bulger and Mark Groudine. Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 2011.

Yang Cao and Linda Petzold. Accuracy limitations and the measurement of errors in the stochastic simulation of chemically reacting systems. *Journal of Computational Physics*, 2006. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2005.06.012. URL `https://www.sciencedirect.com/science/article/pii/S0021999105003074`.

Matthew Cobb. 60 years ago, francis crick changed the logic of biology. *PLoS Biol*, 2017.

Christian Commault and Stéphane Mocanu. Phase-type distributions and representations: some results and open problems for system theory. *International Journal of Control*, 76(6): 566–580, 2003.

Adam M Corrigan, Edward Tunnacliffe, Danielle Cannon, and Jonathan R Chubb. A continuum model of transcriptional bursting. *Elife*, 5:e13051, 2016.

David Cottrell, Peter S. Swain, and Paul F. Tupper. Stochastic branching-diffusion models for gene expression. *Proceedings of the National Academy of Sciences*, 2012. doi: 10.1073/pnas.1201103109. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1201103109`.

Adrien Coulier, Stefan Hellander, and Andreas Hellander. A multiscale compartment-based model of stochastic gene regulatory networks using hitting-time analysis. *J Chem Phys*, 2021.

Antoine Coulon and David R Larson. Fluctuation analysis: dissecting transcriptional kinetics with signal theory. In *Methods in enzymology*, volume 572, pages 159–191. Elsevier, 2016.

Patrick Cramer. Organization and regulation of gene transcription. *Nature*, 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1517-4. URL https://doi.org/10.1038/s41586-019-1517-4.

A. Crudu, A. Debussche, A. Muller, and O. Radulescu. Convergence of stochastic gene networks to hybrid piecewise deterministic processes. *The Annals of Applied Probability*, 2012. ISSN 10505164. URL http://www.jstor.org/stable/41713399.

Alina Crudu, Arnaud Debussche, and Ovidiu Radulescu. Hybrid stochastic simplifications for multiscale gene networks. *BMC systems biology*, 3:1–25, 2009.

Joseph O. Dada and Pedro Mendes. Multi-scale modelling and simulation in systems biology. *Integr. Biol.*, 2011. doi: 10.1039/C0IB00075B. URL http://dx.doi.org/10.1039/C0IB00075B.

M.H.A. Davis. *Markov Models and Optimization*. 1993. ISBN 9780203748039. doi: 10.1201/9780203748039.

Stefano De Renzis, Olivier Elemento, Saeed Tavazoie, and Eric F Wieschaus. Unmasking activation of the zygotic genome using chromosomal deletions in the drosophila embryo. *PLOS Biology*, 2007. doi: 10.1371/journal.pbio.0050117. URL https://doi.org/10.1371/journal.pbio.0050117.

Arnaud Debussche and Mac Jugal Nguepedja Nankep. A law of large numbers in the supremum norm for a multiscale stochastic spatial gene network. *The International Journal of Biostatistics*, 15(2):20170091, 2019.

Max Delbrück. Statistical fluctuations in autocatalytic reactions. *Journal of Chemical Physics*, 8:120–124, 1940. URL https://api.semanticscholar.org/CorpusID:97531241.

Jonathan Desponds, Huy Tran, Teresa Ferraro, Tanguy Lucas, Carmina Perez Romero, Aurelien Guillou, Cecile Fradin, Mathieu Coppey, Nathalie Dostatni, and Aleksandra M Walczak. Precision of readout at the hunchback gene: analyzing short transcription time traces in living fly embryos. *PLoS computational biology*, 12(12), 2016.

Renaud Dessalles. *Stochastic models for protein production: the impact of autoregulation, cell cycle and protein production interactions on gene expression*. Theses, École Polytechnique, January 2017. URL `https://hal.inria.fr/tel-01482087`.

Maria Douaihy, Rachel Topno, Mounia Lagha, Edouard Bertrand, and Ovidiu Radulescu. BurstDECONV: a signal deconvolution method to uncover mechanisms of transcriptional bursting in live cells. *Nucleic Acids Research*, page gkad629, 07 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad629. URL `https://doi.org/10.1093/nar/gkad629`.

Julien O. Dubuis, Gašper Tkačik, Eric F. Wieschaus, Thomas Gregor, and William Bialek. Positional information, in bits. *Proceedings of the National Academy of Sciences*, 2013. doi: 10.1073/pnas.1315642110. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1315642110`.

Jeremy Dufourt, Antonio Trullo, Jennifer Hunter, Carola Fernandez, Jorge Lazaro, Matthieu Dejean, Lucas Morales, Saida Nait-Amer, Katharine N Schulz, Melissa M Harrison, et al. Temporal control of gene expression by the pioneer factor zelda through transient interactions in hubs. *Nature communications*, 10(1):1–1, 2018.

Daniel Dufresne. Fitting combinations of exponentials to probability distributions. *Applied Stochastic Models in Business and Industry*, 23(1):23–48, 2007.

Bruce A Edgar and Gerold Schubiger. Parameters controlling transcriptional activation during early drosophila development. *Cell*, 44(6):871–877, 1986.

Stefan Engblom, Lars Ferm, Andreas Hellander, and Per Lötstedt. Simulation of stochastic reaction-diffusion processes on unstructured meshes. *SIAM Journal on Scientific Computing*, 2009. doi: 10.1137/080721388. URL `https://doi.org/10.1137/080721388`.

Stewart N. Ethier and Thomas G. Kurtz. *Markov processes – characterization and convergence*. John Wiley & Sons Inc., New York, 1986. ISBN 0-471-08186-8.

Mark Fackrell. Modelling healthcare systems with phase-type distributions. *Health care*

management science, 12:11–26, 2009.

MJ Faddy. A structured compartmental model for drug kinetics. *Biometrics*, pages 243–248, 1993.

David Fange, Otto G. Berg, Paul Sjöberg, and Johan Elf. Stochastic reaction-diffusion kinetics in the microscopic limit. *Proceedings of the National Academy of Sciences*, 2010. doi: 10.1073/pnas.1006565107. URL https://www.pnas.org/doi/abs/10.1073/pnas.1006565107.

Jeffrey A Farrell and Patrick H O'Farrell. From egg to gastrula: how the cell cycle is remodeled during the drosophila mid-blastula transition. *Annu Rev Genet*, 2014.

William Feller. *An introduction to probability theory and its applications, Volume 2*. John Wiley & Sons, 1966.

Willy Feller. On the integro-differential equations of purely discontinuous markoff processes. *Transactions of the American Mathematical Society*, 48(3):488–515, 1940.

Matthew L Ferguson and Daniel R Larson. Measuring transcription dynamics in living cells using fluctuation analysis. In *Imaging gene expression*, pages 47–60. Springer, 2013.

Lars Ferm, Andreas Hellander, and Per Lötstedt. An adaptive algorithm for simulation of stochastic reaction-diffusion processes. *J. Comput. Phys.*, 2010. doi: 10.1016/j.jcp.2009.09.030. URL https://doi.org/10.1016/j.jcp.2009.09.030.

Teresa Ferraro, Emilia Esposito, Laure Mancini, Sam Ng, Tanguy Lucas, Mathieu Coppey, Nathalie Dostatni, Aleksandra M Walczak, Michael Levine, and Mounia Lagha. Transcriptional memory in the drosophila embryo. *Curr Biol*, 2015.

Hans Peter Fischer. Mathematical modeling of complex biological systems: from parts lists to understanding systems behavior. *Alcohol Res Health*, 31(1):49–59, 2008.

V E Foe and B M Alberts. Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in drosophila embryogenesis. *J Cell Sci*, 1983.

C. Forbes, M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. Wiley, 2011. ISBN 9781118097823. URL https://books.google.com.lb/books?id=YhF1osrQ4psC.

William Fulton and Joe Harris. *Representation theory. A first course*, volume 129. Springer, 1991.

Hernan G Garcia, Mikhail Tikhonov, Albert Lin, and Thomas Gregor. Quantitative imaging of transcription in living drosophila embryos links polymerase activity to patterning. *Curr Biol*, 2013.

C. W. Gardiner, K. J. McNeil, D. F. Walls, and I. S. Matheson. Correlations in stochastic theories of chemical reactions. *Journal of Statistical Physics*, pages 307–331, 1976. ISSN 1572-9613. doi: 10.1007/BF01030197. URL https://doi.org/10.1007/BF01030197.

Crispin W Gardiner et al. *Handbook of stochastic methods*, volume 3. springer Berlin, 1985.

Roman Garnett, Michael A Osborne, and Stephen J Roberts. Sequential bayesian prediction in the presence of changepoints. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 345–352, 2009.

V.P. Gerdt, M. Lange-Hegermann, and D. Robertz. The Maple package TDDS for computing Thomas decompositions of systems of nonlinear PDEs. *Comp. Phys. Comm.*, 234:202–215, 2019.

I. I. Gikhman and A. V. Skorokhod. *Introduction to the Theory of Random Processes*. Dover, 1969.

Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, Dec 1977.

Daniel T Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1-3):404–425, 1992.

Ido Golding and Edward C. Cox. Rna dynamics in live escherichia coli cells. *Proceedings of the National Academy of Sciences*, 101(31):11310–11315, 2004. doi: 10.1073/pnas.0404443101. URL https://www.pnas.org/doi/abs/10.1073/pnas.0404443101.

Natalie K. Gordon, Zhan Chen, Richard Gordon, and Yuting Zou. French flag gradients and turing reaction-diffusion versus differentiation waves as models of morphogenesis. *Biosystems*, 196:104169, 2020. doi: https://doi.org/10.1016/j.biosystems.2020.104169. URL https://www.sciencedirect.com/science/article/pii/S030326472030068X.

Jeremy B. A. Green and James Sharpe. Positional information and reaction-diffusion: two big ideas in developmental biology combine. *Development*, 2015. doi: 10.1242/dev.114991. URL https://doi.org/10.1242/dev.114991.

Thomas Gregor, Hernan G Garcia, and Shawn C Little. The embryo as a laboratory: quantifying transcription in drosophila. *Trends Genet*, 2014.

Ramon Grima and Santiago Schnell. Modelling reaction kinetics inside cells. *Essays Biochem*, 45:41–56, 2008.

Piyush B Gupta, Ievgenia Pastushenko, Adam Skibinski, Cedric Blanpain, and Charlotte Kuperwasser. Phenotypic plasticity: Driver of cancer initiation, progression, and therapy resistance. *Cell Stem Cell*, December 2018a.

Sanjana Gupta, Jacob Czech, Robert Kuczewski, Thomas M. Bartol, Terrence J. Sejnowski, Robin E. C. Lee, and James R. Faeder. Spatial stochastic modeling with mcell and cellblender. *arXiv: Quantitative Methods*, 2018b.

H. Haken. *Synergetics, 2nd edn.* Springer-Verlag, Berlin, 1978.

Lucy Ham, David Schnoerr, Rowan D. Brackston, and Michael P. H. Stumpf. Exactly solvable models of stochastic gene expression. *The Journal of Chemical Physics*, 2020. ISSN 0021-9606. doi: 10.1063/1.5143540. URL https://doi.org/10.1063/1.5143540.

Stefan Hellander, Andreas Hellander, and Linda Petzold. Reaction-diffusion master equation in the microscopic limit. *Phys. Rev. E*, Apr 2012. doi: 10.1103/PhysRevE.85.042901. URL https://link.aps.org/doi/10.1103/PhysRevE.85.042901.

Benjamin Hepp, Ankit Gupta, and Mustafa Khammash. Adaptive hybrid simulations for multiscale stochastic reaction networks. *The Journal of chemical physics*, 142(3), 2015.

Sjoerd Holwerda and Wouter De Laat. Chromatin loops, gene positioning, and gene expression. *Frontiers in Genetics*, 2012. ISSN 1664-8021. doi: 10.3389/fgene.2012.00217. URL https://www.frontiersin.org/articles/10.3389/fgene.2012.00217.

Joung-Woo Hong, David A Hendrix, and Michael S Levine. Shadow enhancers as a source of evolutionary novelty. *Science*, September 2008a.

Joung-Woo Hong, David A. Hendrix, Dmitri Papatsenko, and Michael S. Levine. How the dorsal gradient works: Insights from postgenome technologies. *Proceedings of the National Academy of Sciences*, 2008b. URL https://www.pnas.org/doi/abs/10.1073/pnas.0806476105.

Sara Hooshangi and Ron Weiss. The effect of negative feedback on noise propagation in transcriptional gene networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 16 (2):026108, 2006.

Somkid Intep, Desmond J. Higham, and Xuerong Mao. Switching and diffusion models for gene regulation networks. *Multiscale Modeling & Simulation*, 2009. doi: 10.1137/080735412. URL https://doi.org/10.1137/080735412.

Y T Ip, R E Park, D Kosman, K Yazdanbakhsh, and M Levine. dorsal-twist interactions establish snail expression in the presumptive mesoderm of the drosophila embryo. *Genes Dev*, 1992.

Samuel A. Isaacson. The reaction-diffusion master equation as an asymptotic approximation of diffusion to a small target. *SIAM Journal on Applied Mathematics*, 2009. doi: 10.1137/070705039. URL https://doi.org/10.1137/070705039.

Johannes Jaeger, John Reinitz, et al. Drosophila blastoderm patterning. *Current opinion in genetics & development*, 22(6):533–541, 2012.

Andrew R Joyce and Bernhard Ø Palsson. The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*, March 2006.

Philip K. Maini, Kevin J. Painter, and Helene Nguyen Phong Chau. Spatial pattern formation in chemical and biological systems. *J. Chem. Soc., Faraday Trans.*, 93, 1997. doi: 10.1039/A702602A. URL http://dx.doi.org/10.1039/A702602A. Publisher: The Royal Society of Chemistry.

Jitendra S. Kanodia, Richa Rikhy, Yoosik Kim, Viktor K. Lund, Robert DeLotto, Jennifer Lippincott-Schwartz, and Stanislav Y. Shvartsman. Dynamics of the dorsal morphogen gradient. *Proceedings of the National Academy of Sciences*, 106(51):21707–21712, 2009. doi: 10.1073/pnas.0912395106.

Jitendra S Kanodia, Hsiao-Lan Liang, Yoosik Kim, Bomyi Lim, Mei Zhan, Hang Lu, Chris-

tine A Rushlow, and Stanislav Y Shvartsman. Pattern formation by graded and uniform signals in the early drosophila embryo. *Biophysical Journal*, 102(3):427–433, 2012.

Balakuntalam S Kasinath, Meenalakshmi M Mariappan, Kavithalakshmi Sataranatarajan, Myung Ja Lee, and Denis Feliers. mRNA translation: unexplored territory in renal science. *J Am Soc Nephrol*, pages 3281–3292, September 2006.

Dirk A Kleinjan and Veronica van Heyningen. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet*, 2004.

Jeremias Knoblauch and Theodoros Damoulas. Spatio-temporal bayesian on-line change-point detection with model selection, 2018.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Doubly robust bayesian inference for non-stationary streaming data with β-divergences, 2018.

Bernard Osgood Koopman. On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409, 1936.

Peter Kotelenez. Law of Large Numbers and Central Limit Theorem for Linear Chemical Reactions with Diffusion. *The Annals of Probability*, 14(1):173 – 193, 1986. doi: 10.1214/aop/1176992621. URL `https://doi.org/10.1214/aop/1176992621`.

Peter Kotelenez. Fluctuations near homogeneous states of chemical reactions with diffusion. *Advances in applied probability*, 19(2):352–370, 1987.

Peter Kotelenez. High density limit theorems for nonlinear chemical reactions with diffusion. *Probability theory and related fields*, 78(1):11–37, 1988.

ANATOL Kuczura. Piecewise markov processes. *SIAM Journal on Applied Mathematics*, 24 (2):169–181, 1973.

T. G. Kurtz. Limit theorems for sequences of jump markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 1971. ISSN 00219002. doi: 10.2307/3211904. URL `https://doi.org/10.2307/3211904`. Full publication date: Jun., 1971.

Thomas G. Kurtz. Solutions of ordinary differential equations as limits of pure jump markov processes. *Journal of Applied Probability*, 7(1):49–58, 1970. doi: 10.2307/3212147.

Jamie C Kwasnieski, Terry L Orr-Weaver, and David P Bartel. Early genome activation in drosophila is extensive with an initial tendency for aborted transcripts and retained introns. *Genome Res*, June 2019.

Mounia Lagha, Jacques P Bothma, Emilia Esposito, Samuel Ng, Laura Stefanik, Chiahao Tsui, Jeffrey Johnston, Kai Chen, David S Gilmour, Julia Zeitlinger, and Michael S Levine. Paused pol II coordinates tissue morphogenesis in the drosophila embryo. *Cell*, 2013.

Nicholas C Lammers, Vahe Galstyan, Armando Reimer, Sean A Medin, Chris H Wiggins, and Hernan G Garcia. Multimodal transcriptional control of pattern formation in embryonic development. *Proceedings of the National Academy of Sciences*, 117(2):836–847, 2020a.

Nicholas C Lammers, Yang Joon Kim, Jiaxi Zhao, and Hernan G Garcia. A matter of time: Using dynamics and theory to uncover mechanisms of transcriptional bursting. *Current opinion in cell biology*, 67:147–157, 2020b.

Markus Lange-Hegermann, Daniel Robertz, Werner M. Seiler, and Matthias Seiß. Singularities of algebraic differential equations. *Advances in Applied Mathematics*, 131:102266, 2021. ISSN 0196-8858. doi: https://doi.org/10.1016/j.aam.2021.102266. URL `https://www.sciencedirect.com/science/article/pii/S0196885821001044`.

Tong Ihn Lee and Richard A Young. Transcriptional regulation and its misregulation in disease. *Cell*, 2013.

Tineke L. Lenstra, Joseph Rodriguez, Huimin Chen, and Daniel R. Larson. Transcription dynamics in living cells. *Annual Review of Biophysics*, 45(1):25–47, 2016. doi: 10.1146/annurev-biophys-062215-010838. URL `https://doi.org/10.1146/annurev-biophys-062215-010838`.

Michael Levine and Eric H. Davidson. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences*, 102(14):4936–4942, 2005. doi: 10.1073/pnas.0408031102. URL `https://www.pnas.org/doi/abs/10.1073/pnas.0408031102`.

Mike Levine. Transcriptional enhancers in animal development and evolution. *Curr Biol*, 2010.

Benoit Liquet, Jean-François Timsit, and Virginie Rondeau. Investigating hospital heterogeneity with a multi-state frailty model: application to nosocomial pneumonia disease in intensive care units. *BMC medical research methodology*, 12(1):1–14, 2012.

Tanguy Lucas, Teresa Ferraro, Baptiste Roelens, Jose De Las Heras Chanes, Aleksandra M Walczak, Mathieu Coppey, and Nathalie Dostatni. Live imaging of bicoid-dependent transcription in drosophila embryos. *Current biology*, 23(21):2135–2139, 2013.

Anna Lyubimova, Shalev Itzkovitz, Jan Philipp Junker, Zi Peng Fan, Xuebing Wu, and Alexander van Oudenaarden. Single-molecule mRNA detection and counting in mammalian tissue. *Nature Protocols*, August 2013.

Philip K Maini, Ruth E Baker, and Cheng-Ming Chuong. Developmental biology. the turing model comes of molecular age. *Science*, December 2006.

Tatiana T. Marquez-Lago and Kevin Burrage. Binomial tau-leap spatial stochastic simulation algorithm for applications in chemical kinetics. *The Journal of Chemical Physics*, 2007. ISSN 0021-9606. doi: 10.1063/1.2771548. URL `https://doi.org/10.1063/1.2771548`. 104101.

Harley H. McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 1997. doi: 10.1073/pnas.94.3.814. URL `https://www.pnas.org/doi/abs/10.1073/pnas.94.3.814`.

Sarah McFann, Sayantan Dutta, Jared E Toettcher, and Stanislav Y Shvartsman. Temporal integration of inductive cues on the way to gastrulation. *Proceedings of the National Academy of Sciences*, 118(23):e2102691118, 2021.

Steven L McKnight and Oscar L Miller Jr. Post-replicative nonribosomal transcription units in d. melanogaster embryos. *Cell*, 17(3):551–563, 1979.

Donald A McQuarrie. Stochastic approach to chemical kinetics. *Journal of applied probability*, 4(3):413–478, 1967.

Heather Meyer and Adrienne Roeder. Stochasticity in plant cellular growth and patterning. *Frontiers in plant science*, 5, 2014. doi: 10.3389/fpls.2014.00420.

Takashi Miura and Kohei Shiota. Tgfb2 acts as an "activator" molecule in reaction-diffusion model and is involved in cell sorting phenomenon in mouse limb micromass culture.

Developmental Dynamics, 2000. doi: https://doi.org/10.1002/(SICI)1097-0177(200003)217:3<241::AID-DVDY2>3.0.CO;2-K.

Brian Munsky, Gregor Neuert, and Alexander van Oudenaarden. Using gene expression noise to understand gene regulation. *Science*, 2012.

Kevin P Murphy. Conjugate bayesian analysis of the gaussian distribution. *def*, 1(2σ2):16, 2007.

Marcel F Neuts. Probability distributions of phase type. *Liber Amicorum Prof. Emeritus H. Florin*, 1975.

Damien Nicolas, Nick Phillips, and Felix Naef. What shapes eukaryotic transcriptional bursting? *Mol. BioSyst.*, 13, 06 2017. doi: 10.1039/C7MB00154A.

G. Nicolis and I. Prigogine. *Self-organization in Nonequilibrium Systems*. Wiley, New York, 1977.

Colm Art O'Cinneide. Phase-type distributions: open problems and a few properties. *Stochastic Models*, 15(4):731–757, 1999.

Michael B O'Connor, David Umulis, Hans G Othmer, and Seth S Blair. Shaping BMP morphogen gradients in the drosophila embryo and pupal wing. *Development*, 2006.

Chin-Tong Ong and Victor G. Corces. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics*, 2011. ISSN 1471-0064. doi: 10.1038/nrg2957. URL `https://doi.org/10.1038/nrg2957`.

Johan Paulsson. Summing up the noise in gene networks. *Nature*, Jan 2004. ISSN 1476-4687. doi: 10.1038/nature02257. URL `https://doi.org/10.1038/nature02257`.

Michael W Perry, Alistair N Boettiger, Jacques P Bothma, and Michael Levine. Shadow enhancers foster robustness of drosophila gastrulation. *Curr Biol*, 2010.

Peter Pfaffelhuber and Lea Popovic. How spatial heterogeneity shapes multiscale biochemical reaction network dynamics. *Journal of the Royal Society Interface*, 12(104):20141106, 2015.

Virginia L Pimmett, Matthieu Dejean, Carola Fernandez, Antonio Trullo, Edouard Bertrand, Ovidiu Radulescu, and Mounia Lagha. Quantitative imaging of transcription in living drosophila embryos reveals the impact of core promoter motifs on promoter state dynamics. *Nature communications*, 12(1):1–16, 2021.

Edwin James George Pitman. Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the cambridge Philosophical society*, pages 567–579. Cambridge University Press, 1936.

L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986. doi: 10.1109/MASSP.1986.1165342.

Morris Rabinowitz. Studies on the cytology and early embryology of the egg of drosophila melanogaster. *Journal of Morphology*, 1941. doi: https://doi.org/10.1002/jmor.1050690102. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/jmor.1050690102`.

Ovidiu Radulescu, Aurélie Muller, and Alina Crudu. Théorèmes limites pour les processus de markov à sauts. *Tech. Sci. Informatiques*, 26(3-4):443–469, 2007.

V Ramaswami and David M Lucantoni. Algorithms for the multi-server queue with phase type service. *Stochastic Models*, 1(3):393–417, 1985.

V Ramaswami and Marcel F Neuts. Some explicit formulas and computational methods for infinite-server queues with phase-type arrivals. *Journal of Applied Probability*, 17(2): 498–514, 1980.

A. F. Ramos, G. C. P. Innocentini, and J. E. M. Hornos. Exact time-dependent solutions for a self-regulating gene. *Phys. Rev. E*, 83:062902, Jun 2011. doi: 10.1103/PhysRevE.83.062902. URL `https://link.aps.org/doi/10.1103/PhysRevE.83.062902`.

Jonathan M Raser and Erin K O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, May 2004.

Jonathan M Raser and Erin K O'Shea. Noise in gene expression: origins, consequences, and control. *Science*, pages 2010–2013, September 2005.

A Rényi. Betrachtung chemischer reaktionen mit hilfe der theorie der stochastichen prozesse. *Magyar TudAkadAlkalmMatIntKözl*, 2:93–101, 1954.

Joseph Rodriguez, Gang Ren, Christopher R Day, Keji Zhao, Carson C Chow, and Daniel R Larson. Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. *Cell*, 176(1-2):213–226, 2019.

Yunus Saatçi, Ryan D Turner, and Carl E Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 927–934, 2010.

Katharine N Schulz and Melissa M Harrison. Mechanisms regulating zygotic genome activation. *Nature Reviews Genetics*, 20(4):221–234, 2019.

Adrien Senecal, Brian Munsky, Florence Proux, Nathalie Ly, Floriane E. Braye, Christophe Zimmer, Florian Mueller, and Xavier Darzacq. Transcription factors modulate c-fos transcriptional bursts. *Cell Reports*, 2014. ISSN 2211-1247. doi: https://doi.org/10.1016/j.celrep.2014.05.053. URL `https://www.sciencedirect.com/science/article/pii/S2211124714004471`.

Igor Shafarevich. *Basic Algebraic Geometry 1*. Springer, 1972.

Cameron A. Smith and Christian A. Yates. Spatially extended hybrid methods: a review. *Journal of The Royal Society Interface*, 2018. doi: 10.1098/rsif.2017.0931. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2017.0931`.

Gordon D Smith, Gordon D Smith, and Gordon Dennis Smith Smith. *Numerical solution of partial differential equations: finite difference methods*. Oxford university press, 1985.

Stephen Smith and Ramon Grima. Single-cell variability in multicellular organisms. *Nature communications*, 9(1):345, 2018.

M von Smoluchowski. Study of a mathematical theory of the coagulation kinetics of colloidal solutions. *Z. Phys. Chem.*, 92:129, 1917.

Lok-Hang So, Anandamohan Ghosh, Chenghang Zong, Leonardo A Sepúlveda, Ronen Segev, and Ido Golding. General properties of transcriptional time series in escherichia coli. *Nat Genet*, 2011.

Thomas R. Sokolowski, Joris Paijmans, Laurens Bossen, Thomas Miedema, Martijn Wehrens, Nils B. Becker, Kazunari Kaizu, Koichi Takahashi, Marileen Dogterom, and Pieter Rein

ten Wolde. eGFRD in all dimensions. *The Journal of Chemical Physics*, 150(5), 2019. doi: 10.1063/1.5064867. URL https://doi.org/10.1063/1.5064867. 054108.

Kieran Stone, Reyer Zwiggelaar, Phil Jones, and Neil Mac Parthaláin. A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLOS Digital Health*, 1(4):e0000017, 2022.

Peter S. Swain, Michael B. Elowitz, and Eric D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 2002. doi: 10.1073/pnas.162041399. URL https://www.pnas.org/doi/abs/10.1073/pnas.162041399.

Wael Tadros and Howard D Lipshitz. The maternal-to-zygotic transition: a play in two acts. *Development*, September 2009.

Katjana Tantale, Florian Mueller, Alja Kozulic-Pirher, Annick Lesne, Jean-Marc Victor, Marie-Cécile Robert, Serena Capozi, Racha Chouaib, Volker Bäcker, Julio Mateos-Langerak, Xavier Darzacq, Christophe Zimmer, Eugenia Basyuk, and Edouard Bertrand. A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting. *Nat Commun*, 7:12248, July 2016.

Katjana Tantale, Encar Garcia-Oliver, Marie-Cécile Robert, Adèle L'Hostis, Yueyuxiao Yang, Nikolay Tsanov, Rachel Topno, Thierry Gostan, Alja Kozulic-Pirher, Meenakshi Basu-Shrivastava, Kamalika Mukherjee, Vera Slaninova, Jean-Christophe Andrau, Florian Mueller, Eugenia Basyuk, Ovidiu Radulescu, and Edouard Bertrand. Stochastic pausing at latent HIV-1 promoters generates transcriptional bursting. *Nature Communications*, 12, July 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-24462-5. URL https://www.nature.com/articles/s41467-021-24462-5.

Hamid Teimouri and Anatoly B. Kolomeisky. Power of stochastic kinetic models: From biological signaling and antibiotic activities to t cell activation and cancer initiation dynamics. *WIREs Computational Molecular Science*, 2022.

J.M. Thomas. *Differential Systems*. Colloquium Publications XXI. American Mathematical Society, New York, 1937.

J.M. Thomas. *Systems and Roots*. W. Byrd Press, 1962.

Gašper Tkačik and Thomas Gregor. The many bits of positional information. *Development*, 148(2), 02 2021. ISSN 0950-1991. doi: 10.1242/dev.176065. URL https://doi.org/10.1242/dev.176065. dev176065.

Tsz-Leung To and Narendra Maheshri. Noise can induce bimodality in positive transcriptional feedback loops without bistability. *Science*, 327(5969):1142–1145, 2010.

Tatjana Trcek, Timothée Lionnet, Hari Shroff, and Ruth Lehmann. mRNA quantification using single-molecule FISH in drosophila embryos. *Nat Protoc*, June 2016.

Tatjana Trcek, Samir Rahman, and Daniel Zenklusen. *Measuring mRNA Decay in Budding Yeast Using Single Molecule FISH*. United States, 2018.

A Trullo, J Dufourt, and M Lagha. MitoTrack, a user-friendly semi-automatic software for lineage tracking in living embryos. *Bioinformatics*, 36(4):1300–1302, 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz717. URL https://doi.org/10.1093/bioinformatics/btz717.

Edward Tunnacliffe and Jonathan R Chubb. What is a transcriptional burst? *Trends in Genetics*, 36(4):288–297, 2020.

Alan Mathison Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37–72, 1952. doi: 10.1098/rstb.1952.0012. URL https://royalsocietypublishing.org/doi/abs/10.1098/rstb.1952.0012.

Ryan Turner, Yunus Saatci, and Carl Edward Rasmussen. Adaptive sequential bayesian change point detection. In *Temporal Segmentation Workshop at NIPS*, pages 1–4. Citeseer, 2009.

David M Umulis, Osamu Shimmi, Michael B O'Connor, and Hans G Othmer. Organism-scale modeling of early drosophila patterning via bone morphogenetic proteins. *Dev Cell*, 2010.

Elizabeth A Urban and Robert J Johnston, Jr. Buffering and amplifying transcriptional noise during cell fate specification. *Front Genet*, 9:591, November 2018.

Gerrit JJ Van den Burg and Christopher KI Williams. An evaluation of change point detection

algorithms. *arXiv preprint arXiv:2003.06222*, 2020.

NG Van Kampen. Fluctuations in continuous systems. In *AIP Conference Proceedings*, volume 27, pages 153–186. American Institute of Physics, 1976.

Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.

Cornelius Joost Van Rijsbergen. Information retrieval. 2nd. newton, ma, 1979.

Jeroen S van Zon and Pieter Rein ten Wolde. Green's-function reaction dynamics: a particle-based approach for simulating biochemical networks in time and space. *J Chem Phys*, December 2005.

Jeroen S. van Zon, Marco J. Morelli, Sorin Tănase-Nicola, and Pieter Rein ten Wolde. Diffusion of transcription factors can drastically enhance the noise in gene expression. *Biophysical Journal*, 2006. ISSN 0006-3495. doi: https://doi.org/10.1529/biophysj.106.086157. URL https://www.sciencedirect.com/science/article/pii/S000634950672149X.

Leor S Weinberger, John C Burnett, Jared E Toettcher, Adam P Arkin, and David V Schaffer. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 tat fluctuations drive phenotypic diversity. *Cell*, July 2005.

L. Wolpert. Positional information and the spatial pattern of cellular differentiation. *Journal of Theoretical Biology*, 25(1):1–47, 1969. ISSN 0022-5193. doi: https://doi.org/10.1016/S0022-5193(69)80016-0. URL https://www.sciencedirect.com/science/article/pii/S0022519369800160.

L Wolpert. Positional information and pattern formation. *Curr Top Dev Biol*, 6(6):183–224, 1971.

Jianquan Xu and Yang Liu. Probing chromatin compaction and its epigenetic states in situ with single-molecule localization-based super-resolution microscopy. *Frontiers in Cell and Developmental Biology*, 9, 2021. ISSN 2296-634X. doi: 10.3389/fcell.2021.653077. URL https://www.frontiersin.org/articles/10.3389/fcell.2021.653077.