UNCOVERING TRANSCRIPTIONAL BURSTING DYNAMICS FROM SPATIALLY RESOLVED SINGLE-CELL MICROSCOPY DATA

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF BIOLOGY, MEDICINE AND HEALTH

2022

Jonathan R. Bowles

School of Health Sciences

Contents

Word Count: 21070

List of Figures

Abstract

UNCOVERING TRANSCRIPTIONAL BURSTING DYNAMICS FROM SPATIALLY RESOLVED SINGLE-CELL MICROSCOPY DATA

Jonathan R. Bowles

A thesis submitted to The University of Manchester
for the degree of Doctor of Philosophy, 2022

Recent advances in live imaging technology, such as the MS2-GFP system, have enabled the recording of transcriptional data at the single-cell level at ever-greater temporal and spatial resolution. Whereas previously researchers had to rely on static 'snapshots' of developing embryos, such as those provided by Single Molecule Fluorescence in situ Hybridisation (smFISH), it is now possible to record fluorescence microscopy movies of developing embryos in the laboratory.

An example of one such live imaging technique is the MS2-GFP system, where gene editing is used to insert a transgene into a gene of interest. When the gene is transcribed, a noisy fluorescent time series is generated which acts as a proxy for transcriptional activity. The dorsal-ventral patterning system in the early *Drosophila* embryo provides an ideal system for studying transcription using live imaging. In this system, a single input, a member of the Bone Morphogenetic Protein (BMP) family, controls multiple target genes, each of which exhibit transcriptional bursting, where transcripts are produced stochastically in discrete 'bursts' of activity, rather than as a constant, Poissonian process.

The aim of analysing these movies is to gain insight into transcriptional regulation in the early embryo, i.e. the relation between the dynamics of mRNA production and cell developmental fate. BMP signalling is of particular interest due to the known involvement of misregulated BMP signalling in developmental defects and cancer. A key problem is how to process and analyse MS2 datasets in order to answer this question.

The main output of the thesis is the development of a novel type of Hidden Markov Model (HMM) for extracting kinetic parameters from MS2 movies, with the aim of establishing the relationship between BMP signalling and transcriptional bursting. We first provide an overview of BMP Signalling in *Drosophila*, followed by a summary of

previous theoretical definitions of biological noise and transcriptional bursting in the literature.

We then outline the details of the implementation of our algorithm. The algorithm demonstrates a significant improvement in computational efficiency relative to the current state of the art model for MS2 analysis, the Compound State Hidden Markov Model (cpHMM), while allowing for the inference of single-cell transcriptional parameters. Results are shown comparing our algorithm to the original algorithm in terms of computational speed and accuracy, using synthetic and experimental *Drosophila* data.

Finally, we present the in-depth results from using our algorithm to investigate the bursting dynamics of the *Drosophila ush* and *hnt* genes. We have been able to establish that regulation of bursting dynamics in this system is achieved through frequency modulation, i.e. by regulating the frequency of bursts, rather than burst duration or amplitude; burst frequency decreases as a function of distance from the embryo midline.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

I would like to thank my supervisors, Magnus Rattray and Hilary Ashe, for their support and guidance throughout my time at Manchester. The appearance of Covid halfway through the PhD has made this an eventful time to be a researcher, but thanks to them I managed to complete the project.

I would like to thank everyone in the Ashe and Rattray labs, especially Carol, who carried out the wet lab experiments that made this thesis possible and was always willing to answer my questions about the biology. Special thanks also to Lauren and Cath.

Thankyou to my parents for supporting me throughout the PhD, especially during the worst of Covid.

Chapter 1

Introduction

1.1 Modelling Stochastic Gene Expression

Proper regulation of gene expression, the process of transmitting genetic information from DNA to RNA and from RNA to proteins, is of fundamental importance to all living things, as misregulation of this process can lead to far-ranging negative consequences for both developing and adult organisms. Tight regulation of gene expression is necessary for determining where and when genes are expressed during development, thereby playing a crucial role in determining cell fate (Lee and Young, 2013). Although gene expression is tightly regulated, phenotypical heterogeneity does still occur; this phenomenon is referred to as biological 'noise' and has been identified as having multiple sources (Raser and O'Shea, 2005).

Biological noise has been theorised to play a number of possible roles in both embryonic development and evolutionary biology. Biological noise has been found to confer a possible evolutionary advantage through small fluctuations in protein levels (Raser and O'Shea, 2005), to allow for stochastic expression in a monoallelic population (Raser and O'Shea, 2005), to confer a fitness advantage in yeast populations through the ability to switch stochastically between metabolic states (Raj and van Oudenaarden, 2008), and to play a role in *Drosophila* eye development (Balázsi et al., 2011).

Mathematical frameworks have been developed (Swain et al., 2002; Paulsson, 2004; Raser and O'Shea, 2004) to formalise the description of noise in biological systems. The simplest possible model of transcription would describe mRNA production as a Poissonian process, where transcripts are produced stochastically at a constant rate. However, for many genes, this model is inadequate - the distribution of transcripts is

'Super-Poissonian', with the measured variance greater than the mean (Nicolas et al., 2017; Lenstra et al., 2016). One possible explanation for Super-Poissonian mRNA production is transcriptional bursting, whereby genes produce transcripts in discrete, stochastically distributed bursts of activity, rather than as a continuous process. Transcriptional bursting has been the subject of intense research, but there are still unanswered questions about how bursting is regulated and the nature of the relationship between bursting and tissue patterning. The so-called 'random telegraph' model (Peccoud and Ycart, 1995; Friedman et al., 2006) offers a simple mathematical framework for modelling transcriptional bursting, but improved imaging techniques have raised the possibility of more sophisticated models, given the high temporal and spatial resolution of live imaging data.

New imaging technologies have made possible the analysis of transcriptional dynamics in the developing embryo at the single-cell level (Munsky et al., 2012; Gregor et al., 2014). Imaging techniques fall into two broad categories: static fluorescence microscopy images and live imaging, where movies are recorded of transcription *in vivo*. Single Molecule Fluorescence in situ Hybridization (smFISH) allows for visualisation of single transcripts within fixed tissue at single-cell resolution through the attachment of short, fluorescently labelled DNA oligonucleotide probes to a complementary mRNA sequence (Trcek et al., 2017; Lyubimova et al., 2013). smFISH has been used successfully to visualise transcription in a wide range of different systems and applications (Boettiger and Levine, 2013; So et al., 2011; Skinner et al., 2013; Corrigan et al., 2016; Bahar Halpern and Itzkovitz, 2016). Mapping from static mRNA distributions to kinetic parameters is computationally challenging; software packages such as BayFish (Gómez-Schiavon et al., 2017) have been developed for analysis of static smFISH 'snapshots'.

Live imaging techniques, such as the MS2-GFP system, offer an alternative to inferring kinetic parameters from static snapshots of mRNA in fixed tissue. The MS2-GFP system, now widely adopted by developmental biology researchers, involves the insertion, through genetic modification, of a reporter transgene into a gene of interest. This transgene codes for a sequence of stem loops which bind to a protein fused to Green Fluorescent Protein when the gene is transcribed, generating a fluorescent time series that acts as a proxy for transcriptional activity (Gregor et al., 2014). Originally developed in 1998 (Bertrand et al., 1998) and optimised in 2004 (Golding and Cox, 2004), the MS2 system has since been used in a wide range of biological systems, including *Drosophila* (Boettiger and Levine, 2013; Corrigan et al., 2016; Garcia et al.,

2013; Bothma et al., 2014; Desponds et al., 2016; Fukaya et al., 2016; Berrocal et al., 2018). While earlier attempts at modelling MS2 data involved fitting relatively simple mathematical models to the data (Fukaya et al., 2016; Bothma et al., 2014), machine learning has more recently been used to extract kinetic parameters from time series data (Corrigan et al., 2016; Berrocal et al., 2018; Lammers et al., 2020).

1.2 Motivation and Objectives

Understanding transcriptional regulation is of fundamental importance due to the central role of disordered transcriptional regulation in both birth defects and disease in the adult organism; misregulation of Bone Morphogenetic Protein signalling, the experimental system focussed on in this thesis, is known to play a central role in cancer biology (Blanco Calvo et al., 2009; Bach et al., 2018). Live imaging offers a chance to observe transcriptional dynamics *in vivo*, but modelling MS2 data presents a number of computational challenges (Gregor et al., 2014).

One key difficulty involved in analysing MS2 data is the presence of a kind of persistence, or lag, in the recorded time series (Berrocal et al., 2018; Lammers et al., 2020). When the gene becomes active and transcription initiates, polymerase begins to travel down the gene body. After a certain period of time, the gene becomes inactive and transcription ceases. At this point, however, the MS2 fluorescence does not immediately drop to zero, as the polymerase that already initiated transcription are still in transit down the gene body, causing the MS2 signal to slowly fall. This persistence in an already noisy signal complicates inference, as there is no longer a direct correspondence between promoter activity and the recorded signal at a given time point.

This particular kind of system, where an underlying stochastically switching state generates an observed signal, lends itself well to analysis with Hidden Markov Models (HMMs). Standard HMMs, however, are not able to deal with the fluorescence persistence problem. Lammers et al. (2020) developed an adapted form of HMM, the Compound State Hidden Markov Model (cpHMM), which introduced the concept of compound states in order to model this specific system. In the Lammers model, although the transitions between promoter states are still Markovian, i.e. probability of the promoter being active or inactive depends only on the previous time state, the observations are dependent upon each other. The cpHMM introduced the concept of a compound promoter state, where the current observation at time t depends upon the previous W promoter states and W, or window size, depends upon factors such as

gene length and elongation rate. The model was used successfully to analyse anterior-posterior patterning in *Drosophila*.

A key drawback with the Lammers model, however, is the exponential relationship between gene length and computational time, as an exponentially increasing number of compound states is required by the model as the gene length is increased - inference of the current most likely promoter state requires an ever-larger window size as gene length increases. This makes use of the algorithm infeasible for many systems, including the *Drosophila* dorsal-ventral patterning system that is the focus of this thesis. In order to circumvent these computational problems, we have developed a dynamic form of the Lammers model, which uses a truncated state space, removing the exponential relationship between gene length and computational time. The dynamic model is able to provide similar results to the full model while greatly reducing computational time for longer genes. After validating the model on synthetic data, we have applied it to the dorsal-ventral patterning system in *Drosophila* (Hoppe et al., 2020)) focussing on the relationship between a member of the Bone Morphogenetic Protein family, Decapentaplegic, and two of its target genes, *u-shaped (ush)* and *hindsight (hnt)*.

The main objective of the thesis is to develop a computationally efficient algorithm for inferring kinetic parameters from MS2 imaging data. The example biological system used in this thesis is the dorsal-ventral patterning system in *Drosophila*, but the the aim is for the algorithm we have developed to be applicable to other systems, as our truncated model can in theory be applied to much longer genes than the original model. The list of aims can be summarised as follows:

- Develop a scalable algorithm for inferring kinetic parameters from MS2 data. The running time of the algorithm should not depend upon gene length.
- Implement the algorithm in software and publish the code in an Open Source repository, with documentation, for other researchers to use.
- Apply the algorithm to the dorsal-ventral system in *Drosophila*. The early *Drosophila* embryo represents an ideal experimental system due to highly reproducible developmental boundaries, a shared signalling medium and the presence of a single layer of syncytial nuclei just beneath the egg cortex (Gregor et al., 2014).
- Extend the software package to include scripts for inferring single-cell kinetic parameters. To our knowledge, the is the first time that kinetic parameters have been calculated on a single-cell basis.

1.3 Thesis outline

In Chapter Two, an outline of BMP Signalling and dorsal-ventral Patterning in the early *Drosophila* embryo is given, followed by a review of previous research in the field of transcriptional dynamics. Details of research into both mathematical modelling and experimental techniques used for quantifying biological noise, such as smFISH and MS2-GFP, are given. Finally, a mathematical overview of Hidden Markov Models, with a particular focus on Lammers et al.'s cpHMM is included.

Chapter Three includes the mathematical details of the dynamic Hidden Markov Model, along with results from applying the model to both synthetic and experimental *Drosophila* data. This chapter is based upon the Bowles et al. (2022) methods paper. The final section includes details of the *burstInfer* Python software package implementing the model.

In Chapter Four, based upon Hoppe et al. (2020), details are provided of the application of the algorithm to modelling MS2 data from the *Drosophila ush* and *hnt* genes. *burstInfer* was used in this paper to establish the key regulated parameter in the *Drosophila* dorsal-ventral patterning system through determining which model parameters varied significantly as one moves away from the embryo dorsal midline. Examples are given of the algorithm's ability to infer single-cell transcriptional parameters. Further applications from a different signalling system, using data still being collected in the lab, are also shown.

Chapter Five contains the discussion along with suggestions for further work. Specifically, suggestions are made for extending the model to take into account temporal as well as spatial variations in transcription into account, for generalising the model and for developing a model based upon Mean Field Variational Bayes Methods.

Chapter 2

Background

In this chapter we first describe the details of signalling and developmental patterning in the early *Drosophila* embryo, followed by an overview of theoretical and experimental tools for quantifying biological noise. We then focus on techniques for analysing MS2-GFP data. The theory of Hidden Markov Models is then outlined, along with the mathematical details of Lammers et al.'s Compound State Hidden Markov Model. The algorithm described in the remainder of the thesis extends and adapts this model.

2.1 *Drosophila* Early Development

The following sections provide an outline of patterning of the *Drosophila* embryo during early development (up to nuclear cycle 14), including the signal transduction mechanisms responsible for coordination of development.

2.1.1 Dorsal-Ventral Patterning

Following fertilisation, spatially varying gradients of maternal transcription factors deposited within the egg during oogenesis establish dorsal-ventral (DV) and anterior-posterior polarity through activation of a number of signal transduction pathways within the developing embryo (Belvin and Anderson, 1996; Kanodia et al., 2009; Levine and Davidson, 2005). These signalling pathways then act to subdivide the embryo into several different tissue types independently along the dorsal-ventral and anterior-posterior axes. DV polarity in *Drosophila* is established by the NF-kB-like (Belvin and Anderson, 1996) maternal transcription factor Dorsal, which forms a nuclear concentration

gradient along the DV axis, with peak levels in the ventral-most region of the embryo (O'Connor, 2005; Umulis et al., 2010; Hill, 2009; Raser and O'Shea, 2005). The graded distribution of Dorsal is established through differential activation of the toll signalling pathway arising from events prior to egg-laying (Umulis et al., 2010; Hill, 2009). Intermediate levels of Dorsal enter nuclei in lateral regions of the embryo, whereas Dorsal is absent in the dorsal-most region of the embryo (O'Connor, 2005; Umulis et al., 2010). This graded distribution of Dorsal leads to the differential expression of nearly 50 target genes in the DV system (Hill, 2009).

The thresholded response of the Dorsal target genes *snail* (*sna*), *short gastrulation* (*sog*) and *decapentaplegic* (*dpp*) to this smooth concentration gradient partitions the embryo into three basic tissue types – the mesoderm, the neurogenic ectoderm and the dorsal ectoderm (Hill, 2009). The dorsal ectoderm is further subdivided into two tissues types, the dorsal epidermis and the amnioserosa (Kanodia et al., 2009; Hill, 2009). The mesoderm is specified in the ventral-most regions of the embryo by *sna*, where levels of Dorsal are highest (Kanodia et al., 2009; O'Connor, 2005). Activation of *sog* by low levels of Dorsal specifies the neuro-ectoderm (Belvin and Anderson, 1996; Levine and Davidson, 2005; Hill, 2009), whereas the absence of Dorsal (and subsequent absence of the Dpp inhibitor Sog) in the dorsal-most regions of the embryo specifies, through the activation of a graded distribution of the bone morphogenetic protein (BMP) *dpp*, the dorsal epidermis and the amnioserosa, a contractile extraembryonic membrane (Levine and Davidson, 2005) believed to play a role in germ band elongation and dorsal closure.

Each of these genes act within their respective regions of expression to further subdivide these areas into different tissue types. This process is aided by the absence of cellular membranes in the early *Drosophila* embryo; for the first two hours following egg-laying, the embryo does not possess cell membranes, forming a syncytial blastoderm (Chalancon et al., 2012; Raj and van Oudenaarden, 2008). In addition to their role during early development, many of these genes have an additional role in later stages of development and tissue maintenance in the adult organism (Elowitz et al., 2002).

2.1.2 BMP Signalling in *Drosophila*

BMPs are a member of the TGF-B family of growth factors – in addition to Dpp, two other members of the BMP family are present in *Drosophila* – Glass Bottom Boat (Gbb) and Screw (Scw) (Meyers and Kessler, 2017; Deignan et al., 2016; Sutherland,

2003). Dpp is the functional orthologue of BMPs 2 and 4 in vertebrates, whereas Gbb is a member of the BMP 5, 6, 7 subgroup (O'Connor, 2005). BMPs in the early *Drosophila* embryo are secreted from a broad region in the upper 40% of the embryo and are then dynamically concentrated into a narrow region at the dorsalmost region of the embryo (O'Connor, 2005; Umulis et al., 2010). Dpp and Scw exist in three different forms in the early embryo – as a Dpp homodimer, as a Scw homodimer, and as a Dpp/Scw heterodimer (O'Connor, 2005). A complex formed of laterally secreted Sog and dorsally secreted Twisted Gastrulation (Tsg) preferentially transports Dpp/Scw heterodimers (the form of Dpp in this system with greatest signalling potency) to the dorsalmost region of the embryo (Hill, 2009).

Cleavage of the Sog/Tsg complex by the protease Tolloid (Tld) releases Dpp for signalling, resulting in a smooth gradient of free Dpp-Scw peaking at the midline of embryo (Lacy and Hutson, 2016). Binding of Dpp/Scw heterodimers to heterotetrameric Punt / Saxophone (Sax) receptors produces an intracellular signalling cascade through the phosphorylation of the transcription factor Mad. Phosphorylated Mad (pMad) associates with the co-Smad Medea before translocating to the nucleus and activating epidermal target genes such as *U-shaped* (*ush*) and *hindsight* (*hnt*) (Umulis et al., 2010). Additionally, binding of the pMad / Medea complex to the zinc finger protein Schnurri represses expression of neuronal genes – this dual action of BMP signalling results in the requirement of a much higher signalling threshold for epidermal genes rather than neuronal genes (Levine and Davidson, 2005).

2.2 Transcriptional Dynamics

The following sections summarise research into the underlying biology behind sources of biological noise, control of transcriptional bursting, mathematical modelling of transcription bursting and laboratory techniques for data acquisition.

2.2.1 Heterogeneity in Biological Systems

The so-called central dogma of molecular biology may be summarised as stating that the process of gene transcription and translation is essentially a question of information flow – genetic information encoded in DNA base sequences is transferred to messenger RNA in a process known as transcription; this information is then transferred to proteins through the process of translation (Crick, 1970). This overall process is known

as gene expression. Gene expression is of fundamental importance to the organism, as both correct development and the proper functioning of the adult organism require the ability to control which genes are expressed, the location of the expressed gene and the amount of gene product ultimately created. Gene regulation is therefore a question of not only spatial, but also temporal, control of gene activity (Balázsi et al., 2011; Gregor et al., 2014). The process of gene expression must therefore be tightly controlled by the organism; misregulation of gene expression is commonly found both in developmental disorders and in diseases of the adult organism (Lee and Young, 2013).

However, despite this apparent requirement to strictly control gene expression, phenotypic variation, or biological heterogeneity, is ubiquitous in biology; remarkable phenotypic diversity may be found even within a population of genetically identical cells (Swain et al., 2002). Research over the last fifteen years has focussed on attempting to discover the source of and to model this phenotypic variation between isogenic cells within a shared environment, which is commonly described as biological 'noise' (Munsky et al., 2012). Raser and O'Shea (2005) identified four possible sources of variation in gene expression: i) the inherent stochasticity in biochemical systems involving a small number of individual molecules, and therefore infrequent reaction events; (ii) variation in gene expression within a population of cells due to predictable changes in global processes such as the cell cycle; (iii) environmental variation due to signalling cues such as a morphogen gradients and (iv) genetic mutation. Each of these possible sources of noise contributes to the recorded noise within a cell population.

Further research, as summarised by Chalancon et al. (2012), has established multiple biological sources for both extrinsic and intrinsic noise. Factors contributing to intrinsic noise include transcriptional bursting (including variation in promoter sequence, nucleosome occupancy and positioning, along with the degree of transcriptional pausing), nuclear architecture, chromatin epigenetics and rates of translation, mRNA degradation and protein degradation. Factors contributing to extrinsic noise include the availability of basic gene expression machinery within the cell, pathway-specific propagation of noise, microfluctuations in the cellular environment and asymmetries arising from cell division. Additionally, noise propagation within a population of cells may also be considered from the perspective of gene regulatory network architecture (Chalancon et al., 2012).

In addition to attempting to mathematically model and characterise the biological mechanisms underlying biological noise, research has focussed on possible functional

roles for noise in areas such as evolutionary theory and developmental biology. Small fluctuations in protein levels may confer either a biological advantage or disadvantage to the organism (Raser and O'Shea, 2005). Heterozygous individuals may express neither alleles, either allele or both alleles of a given gene due to fluctuations in protein count arising from intrinsic noise. A period of intrinsic expression noise followed by negative feedback has been theorised to allow for stochastic expression in a stable monoallelic population (Raser and O'Shea, 2005). The ability to utilise intrinsic noise to stochastically switch between states has also been theorised to lie behind switching between metabolic states in yeast, where a fitness advantage arises from part of a population being able to activate metabolic networks in anticipation of food (Raj and van Oudenaarden, 2008); sacrificing part of a given population to sub-optimal growth while retaining fast response times may be preferable to consuming energy through relying on activation of the sensing apparatus instead (Raj and van Oudenaarden, 2008). Intrinsic noise has also been associated with the stochastic switching of *B. subtilis* from a vegetative to a 'competent' state (Raj and van Oudenaarden, 2008), the onset of meiosis in yeast (Raj and van Oudenaarden, 2008) and frequency-modulated stochastic nuclear localisation of the transcription factor Crz1 in yeast (Eldar and Elowitz, 2010).

A notable example of a role for stochastic gene expression in development is the proposed model for odorant receptor choice in olfactory neurones. Murine olfactory neurones express a single allele of an odorant receptor gene out of a possible selection of around 1500 odorant receptor genes; expression of odorant receptors is mutually exclusive (Raj and van Oudenaarden, 2008). Developing a regulatory network capable of computing the optimal choice of odorant receptor to express has been theorised to be too complex, so a 'Monte Carlo'-type strategy is instead adopted where each neurone randomly expresses a given odorant receptor. Noise is also believed to play a role in photoreceptor development in the *Drosophila* eye, where the decision by optical units known as ommatidia to express one of two possible pairs of photoreceptors has been shown to be stochastic, and in haematopoiesis, where commitment of stem cells to either an erythroid or myeloid lineage has been demonstrated to be dependent upon stochastic fluctuations in levels of the stem cell marker Sca-1 (Balázsi et al., 2011; Raj and van Oudenaarden, 2008).

2.2.2 Quantifying Biological Noise

A classic paper by Elowitz et al. (2002) outlined both a formal system for classification of biological noise, which was separated into extrinsic and intrinsic noise, and an

experimental method for disentangling these two sources of noise. The authors defined the noise present in the distribution of the amount of protein produced by a given cell as η_{tot}, the standard deviation divided by the mean, and that the noise could be divided into two components – extrinsic noise, η_{ext}, fluctuations in the amount or activity of regulatory polymerases and proteins, and intrinsic noise, η_{int}, inherent stochasticity within the system arising from the 'discrete nature of the biochemical process of gene expression' (Elowitz et al., 2002). Intrinsic noise represents a fundamental limit on the precision of gene regulation – even in the presence of a carefully managed extracellular environment, fluctuations in the Brownian motion of individual molecules within a given cell place limits on the similarity of two genetically identical cells within the same population. Experiments designed to disentangle extrinsic and intrinsic noise outlined in the paper involved introducing two identical copies of a promoter into *Escherichia Coli*, each regulating a distinct fluorescent reporter gene producing either cyan fluorescent protein or yellow fluorescent protein. The use of fluorescent reporter genes allowed the authors to visualise cell-to-cell variability in fluorescence and therefore gene activity; extrinsic sources of noise, such as limited availability of Pol II, would be expected to affect both promoters within a given cell equally, generating correlated noise between the two promoters. Intrinsic sources of noise, however, would be expected to vary in extent between the two promoters and therefore generate uncorrelated sources of noise. The authors found that both sources of noise could be significant, depending upon the exact experimental conditions. This work built upon earlier work by the main author in the field of synthetic biology (Elowitz and Leibier, 2000), where biological noise was initially encountered as a potential confounding variable when designing the repressilator – a synthetic network of repressors designing to induce oscillations in gene expression of programmable duration. The periodic oscillations generated by the repressilator network were found to contain significant levels of biological noise, which was conjectured to emanate from biological noise within the individual components of the system. From this initial beginning as an unexplained source of noise in a pioneering synthetic biology experiment biological noise and stochasticity have become substantial fields of active research.

Swain et al. (2002) built upon Elowitz et al.'s landmark study by deriving analytical expressions for intrinsic and extrinsic noise, which were then validated through simulation of a repressor protein. The derived expressions allowed quantification of deviation from Poisson statistics. Transcription was found to dominate the intrinsic noise when the average number of proteins produced per mRNA transcript was greater

than 2; below this level, translational effects had to be taken into account. Experimental validation of the source of intrinsic noise was provided by Ozbudak et al. (2002), who varied the rate of transcription and translation of a single fluorescent reporter gene in *Bacillus Subtilus*; transcription rates were modified through the use of an inducible promoter, whereas translation rates were controlled through the introduction of a number of mutations into the ribosomal binding site. Changes in phenotype were recorded while the rates of transcription and translation were varied. The results obtained indicated that measured noise depended inversely upon the rate of transcription but was not dependent on the rate of translation; this provided early evidence for the production of proteins in stochastic bursts.

Theoretical work by Paulsson (2004) further expanded the mathematical description of noise in biological systems by providing a single unified equation for extrinsic and intrinsic noise. Raser and O'Shea (2004) introduced the concept of a two-state transcriptional model, allowing both permissive and non-permissive transcriptional states, into a flow cytometry study of noise in budding yeast. They hypothesised that the phenotypical heterogeneity permitted by the presence of biological noise in a given population of cells may provide an evolutionary advantage, promoting adaptation to a variable environment through allowing a heterozygous population to express a wider range of phenotypes than otherwise possible.

The concept of transcriptional bursting as an important source of biological noise has since become increasingly central to our understanding of the physical mechanisms underpinning intrinsic noise; more recent work (summarised by Munsky et al. (2012) and expanded upon below) has focussed on using laboratory imaging techniques (such as Single Molecule Fluorescence in situ Hybridisation and the MS2-GFP system) to generate in vivo datasets for computational analysis, with the aim of extracting the kinetic parameters of transcriptional bursting from the data. While transcriptional bursting is now accepted as a key source of intrinsic noise within the cell, the exact details of the spatio-temporal regulation of bursting remain unclear.

2.2.3 Experimental Methodologies

Localisation and quantification of individual mRNA transcripts on a single-cell basis within the *Drosophila* embryo requires accurate, high-resolution microscopy data derived from imaging of both fixed and living embryos. New techniques for computational image analysis, along with refined wet-lab data acquisition protocols, have allowed for the generation of static microscopy images and videos capable of resolving

individual transcripts and transcription foci within the developing embryo. The following sections outline these experimental and computational techniques. Live imaging, in particular, is capable of providing insights into bursting dynamics at high temporal and spatial resolution. The algorithm and modelling results presented in the following two chapters focus on the MS2 system, a particular type of real-time imaging system.

Single Molecule Fluorescence in situ Hybridization

Visualisation of individual mRNA transcripts in fixed tissue can be achieved through Single Molecule Fluorescence in situ Hybridization (smFISH), allowing for localisation and quantification of transcripts at the single-cell level (Gregor et al., 2014; Trcek et al., 2017; Lyubimova et al., 2013). Conventional FISH allows visualisation of the spatial distribution of mRNA transcripts within fixed tissue; however, limited image contrast and dynamic range limits this approach to being non-quantitative. Development of the smFISH protocol by Femino et al. (1998) and subsequent refinement by Raj et al. (2008) allowed for visualisation at the single-cell level.

The technique involves hybridisation of 50-100 (Lyubimova et al., 2013) short, singly-labelled DNA oligonucleotide probes to a complimentary mRNA sequence of interest. Each of the probes is fluorescently labelled, allowing for detection of as little as a single mRNA transcript per cell using automated image analysis software (Trcek et al., 2017). The full image acquisition pipeline requires multiple stages of sample preparation and data analysis; an smFISH protocol for use with *Drosophila* published by Trcek et al. (2017) outlines four stages of preparation – embryo collection and fixation, embryo hybridization with commercial Stellaris probes, imaging and single-mRNA detection and counting. Crucially, the technique does not require genetic modification, greatly reducing the time required for data acquisition relative to live imaging techniques such as MS2-GFP (Lyubimova et al., 2013). Staining of the cell and nuclear membranes allows for binning of the detected fluorescent mRNA dots into individual cells, providing an estimate of transcriptional activity on a single-cell basis.

smFISH has been utilised across a wide range of published research, both in *Drosophila* and other organisms – investigation of the invagination of the mesoderm during *Drosophila* gastrulation (Boettiger and Levine, 2013), quantification of transcriptional bursting in *Escherichia Coli* (So et al., 2011; Skinner et al., 2013), quantification of mRNA degradation rates (Horvathova et al., 2017; Bahar Halpern and Itzkovitz, 2016), quantification of transcription bursting in *Dictyostelium* (Corrigan et al., 2016), measurement of mRNA nuclear retention in the mouse liver, gut and Beta cells (Bahar Halpern et al.,

2015a) and quantification of transcriptional bursting in *Drosophila* (Garcia et al., 2013; Zoller et al., 2018; Little et al., 2013). Automated processing of smFISH images poses numerous challenges, as outlined by Mueller et al. (2013), Trcek et al. (2017), Bahar Halpern and Itzkovitz (2016), Gregor et al. (2014) and others. Despite these computational challenges, smFISH images provide a rich source of data for researchers wishing to infer the kinetic parameters of transcription.

MS2-GFP

While it is possible to infer kinetic parameters relating to transcriptional bursting from static snapshot images, the application of live imaging techniques, such as the use of the MS2-GFP system, allows for the visualisation of transcription in living cells in real-time. The technique involves introduction of a reporter transgene into a gene of interest through genetic modification; the transgene codes for a sequence of repeated stem loops which bind to a protein fused to Green Fluorescent Protein (GFP) (Gregor et al., 2014). A widely-adopted specific implementation of this technique utilises the MS2 bacterial stem loop and corresponding GFP-tagged MS2 coat protein. Transcription of the gene containing the transgene construct generates nascent mRNA bound to GFP tags, allowing for automated computational detection of transcription foci through analysis of fluorescence microscopy images (Gregor et al., 2014). Example MS2 data is show in Figures 2.2 and 2.1.

The use of GFP for live imaging in *Saccharomyces cerevisiae* was originally implemented by Bertrand et al. (1998) and optimised by Golding and Cox (2004); Golding et al. (2005). The technique has been used to investigate Pol II dynamics (Darzacq et al., 2007; Fukaya et al., 2017), transvection (Lim et al., 2018), co-ordination of gastrulation (Boettiger and Levine, 2013; Lim et al., 2017) and control of transcriptional dynamics in *Drosophila* (Corrigan et al., 2016; Garcia et al., 2013; Bothma et al., 2014; Desponds et al., 2016; Fukaya et al., 2016; Berrocal et al., 2018), among other applications. Recent work by Lammers et al. (2020) and Berrocal et al. (2018) has highlighted the suitability of MS2 data for machine learning-based computational modelling of the regulation of transcription in *Drosophila*, allowing for a quantitative, rather than purely phenomenological, approach to understanding transcription in eukaryotes.

Analysis of MS2 images provides measurements in units of arbitrary fluorescence per transcription foci, rather than absolute measurements in terms of individual transcripts, as with smFISH. In addition to allowing for snapshot quantification of the distribution of mRNA transcripts at the single-cell scale, smFISH images may be used

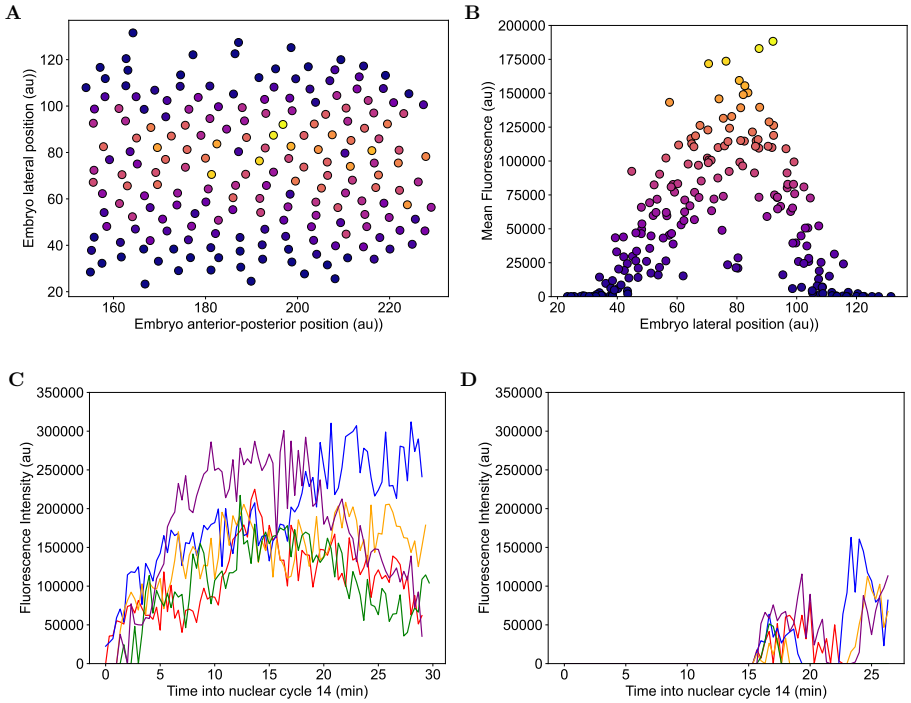

Figure 2.1: Experimental MS2 data from the *Drosophila ush* gene, which is discussed in details in chapters 3 and 4. **A**: Two-dimensional plot of the posterior portion of the embryo. Each point represents a transcription site. Microscope co-ordinates are in arbitrary units. The heatmap corresponds to the mean fluorescence for each trace. **B**: Plot of mean fluorescence for each trace as a function of lateral position. The peaked stripe of expression corresponds to peak levels of Decapentaplegic signalling at the embryo midline. **C**: Five example traces from the dataset, plotted as a function of time into nuclear cycle 14. These traces have been selected from around the embryo midline, where BMP signalling, and therefore transcriptional levels, are at their peak. The noisiness of the data is apparent. **D**: Example traces from the outermost region of the embryo, where signalling levels are at their weakest.

to calibrate MS2 imaging data (Gregor et al., 2014; Bothma et al., 2014; Berrocal et al., 2018; Lammers et al., 2020). Computational processing of MS2 videos generates time series data quantified in terms of arbitrary fluorescence as a function of time; while this allows for visualisation of transcriptional activity, inference of Pol II production requires conversion of the MS2 data into Pol II production as a function of time, rather than arbitrary fluorescence. Fluorescence data extracted from smFISH images may be used to calibrate MS2 videos to provide an estimate of Pol II production at the single-cell level; MS2 data can then simply be divided by a conversion coefficient to provide an estimate of Pol II production as a function of time for each cell. This technique, originally pioneered by Garcia et al. (2013), has since been used in a number of other studies (Berrocal et al., 2018; Lammers et al., 2020; Bothma et al., 2014).

Two variants have been described in the literature. In the original version described by Garcia et al. (2013), the obtained smFISH profile is overlaid with the total amount of mRNA produced inferred from the normalised MS2 profile. The integrated fluorescence intensity corresponding to the transit of one polymerase molecule along the gene is then inferred. Dividing the calculated integrated fluorescence intensity by the elongation time provides a value for the average fluorescence intensity per polymerase. The MS2 traces may then be calibrated in terms of the number of polymerase per transcription site. However, the systematic error associated with this approach may be as high as 29%.

Lammers et al. (2018) have provided an alternative expression for the MS2 calibration factor, given by:

$$F_{RNAP} = \frac{v_{elong} F_{MS2}}{N_{FISH}} \frac{1}{(L_I + L_{II})} \tag{2.1}$$

Where v_{elong} is the elongation time, F_{MS2} is the total fluorescence per nucleus, N_{FISH} is the number of mRNA per nucleus, L_I is the length of the MS2 loops and L_{II} is the distance between the end of the MS2 cassette and the 3' end of the gene. The authors provide an estimate of the calibration error as 13% and a calibration factor of 13 AU/RNAP $\pm$ 1.7 using their equipment. This technique was used to calibrate the MS2 data shown in Chapters 3 and 4, where AirLocalize (Trcek et al., 2017) analysis of smFISH data was used to determine the mRNA output for a single allele. Combining this with the measured mean integrated MS2 fluorescent signal from nuclei at the embryo midline allowed for calculation of F_{RNAP}. See Hoppe et al. (2020) for further details.

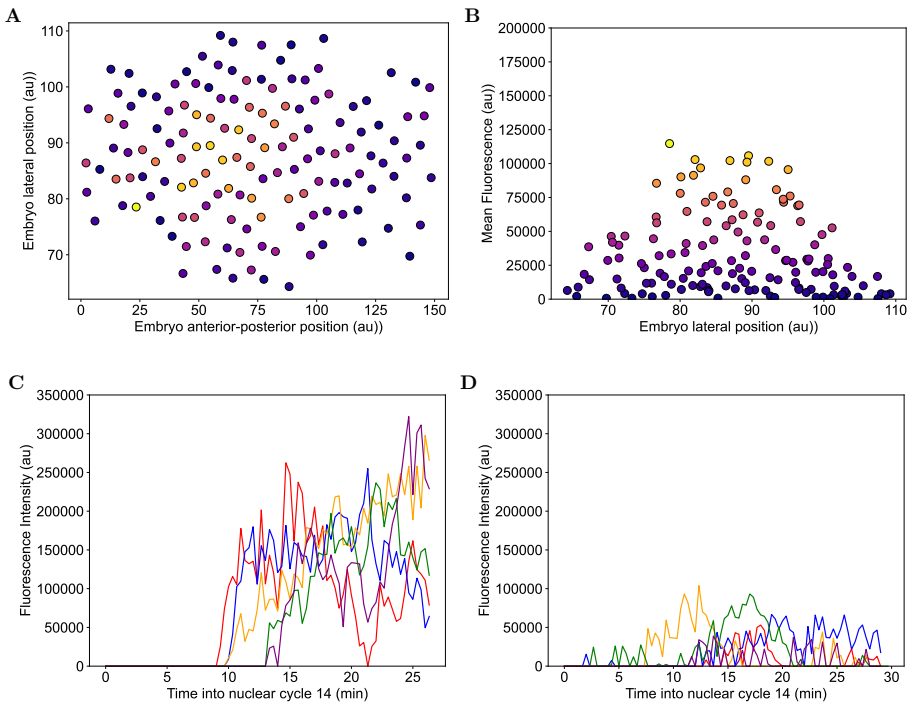

Figure 2.2: Experimental MS2 data from the *Drosophila hnt* gene. **A**: Two-dimensional plot of the posterior portion of the embryo. Each point represents a transcription site. Microscope co-ordinates are in arbitrary units. The heatmap corresponds to the mean fluorescence for each trace. **B**: Plot of mean fluorescence for each trace as a function of lateral position. The peaked stripe of expression corresponds to peak levels of Decapentaplegic signalling at the embryo midline. Note the reduced mean fluorescence levels relative to *ush*. **C**: Five example traces from the dataset, plotted as a function of time into nuclear cycle 14. **D**: Example traces from the outermost region of the embryo.

2.3 Mathematical Modelling of Transcriptional Bursting

In the simplest possible model of gene expression transcripts would be produced and degraded at a constant rate, described by a first-order reaction with production and degradation rate constants. In a system with many molecular components, averaging effects would be expected to smooth out the stochastic creation and degradation of transcripts and proteins. However, in many systems, the limited number of individual components of transcription within a given cell – potentially less than ten transcripts or proteins per cell for a given gene (Paulsson, 2004) – means that molecular reactions must instead be modelled as discrete events, where the averaging effect no longer holds (El Samad et al., 2005).

Mathematically, the continuous production and degradation of mRNA transcripts at random intervals would result in a Poisson distribution of mRNA transcripts; the mean of the distribution would be expected to be equal to the variance (Gregor et al., 2014). Waiting times between the production of individual transcripts would be exponentially distributed, with a most likely time interval of zero and the long tail of distribution representing less likely longer waiting times (Lenstra et al., 2016). While this model has been reported to successfully describe a number of systems such as several housekeeping genes in yeast (Chen and Larson, 2016), the temporal distribution of mRNA transcripts for many genes in both prokaryotes and eukaryotes is not well explained by this particular model. For many genes, snapshot measurements of transcript number have indicated that the distribution of transcripts is 'Super-Poissonian' – the measured variance is greater than the mean (Lenstra et al., 2016; Nicolas et al., 2017). The Super-Poissonian distribution of transcripts indicates that many genes do not produce transcripts as a continuous process – the molecular machinery of transcription must be behaving in such a way so as to explain the Super-Poissonian distribution of transcripts. Additionally, live imaging of transcription has provided direct evidence for discontinuous transcript production in many genes (Fukaya et al., 2016).

An alternative model to Poissonian transcript production is therefore required. A simple alternative model is the so-called random telegraph model (Peccoud and Ycart, 1995; Friedman et al., 2006). In this model, the promoter stochastically switches between active and inactive transcriptional states; a number of molecular mechanisms, such as changes in chromatin structure, have been proposed to biologically explain these changes in state (Nicolas et al., 2017). While in the active state, Pol II elongation commences and mRNA transcripts are produced at a constant rate. While in the inactive state no transcription takes place. This model, while simple, has been used to

describe the dynamics of transcription in a number of different systems; crucially, the distribution of transcripts produced by a system behaving according to the random telegraph model is non-Poissonian; the model is able to generate temporal transcriptional distributions following a wide range of different profiles.

Fitting a mathematical model of transcription, such as the random telegraph model, involves inference of the model parameters through training using biological data. Inference techniques typically focus either on inferring model parameters from mRNA distributions generated using techniques such as Single Molecule Fluorescence in situ Hybridisation or from live imaging approaches, such as the MS2-GFP system, as described in section 2.2.3. In the case of a model of transcriptional dynamics, such as the random telegraph model, inference of the model parameters directly corresponds to inference of the kinetic parameters of the molecular machinery of transcription itself, such as the rate of changes in promoter state and Pol II loading rate; calculation of these parameters may then give insight into both the regulation of transcription within the system and the response of the system to genetic and environmental perturbations.

Research in regulation of transcriptional dynamics has uncovered a range of possible physical parameters under regulatory control. Molecular processes occurring prior to recruitment of Pol II to the promoter, the formation of the transcriptional complex and elongation of Pol II may all be regulated steps (Lenstra et al., 2016; Coulon et al., 2013). Bentovim et al. (2017) identified three broad categories of transcriptional regulation, each consisting of separate sub-categories: the location of transcription (combinatorial control by cis-acting elements and regulation of boundary formation), the level of transcriptional activity (burst size and frequency modulation, physical interaction of regulatory elements and shadow enhancers) and the timing of transcription (promoter synchronisation via promoter-proximal pausing, temporal co-ordination of enhancers and alteration of chromatin structure via pioneer factors). Each of these particular factors may contribute to the transcriptional dynamics of a given gene, which in turn shapes the temporal and spatial characteristics of the gene's expression domain. Quantification of these parameters via statistical inference provides an insight into regulatory control of a given gene.

Three parameters are necessary to completely specify the characteristics of the random telegraph model: k_{on}, the rate of promoter activation, k_{off}, the rate at which the promoter enters the inactive state and r, the polymerase loading rate while in the active state. Transcriptional regulation may be achieved through modulation of any combination of these parameters; research in both prokaryotes and eukaryotes has indicated

regulation via burst frequency (regulation of k_{on}), burst size (regulation of k_{off}) and burst amplitude (regulation of r). A range of different mechanistic interpretations have been proposed for each of these regulatory strategies. Nicolas et al. (2017) have proposed seven physical parameters which may play a role in modulating burst size and frequency: the local chromatin environment, nucleosome occupancy, histone modifications, the number and binding affinity of cis-regulatory elements, DNA looping and transcription factor availability. The extent to which each of these parameters is responsible for regulation of transcriptional bursting for both a given organism and a given gene is unclear. Determination of the regulatory strategy employed for a given gene requires accurate quantification of transcriptional activity; quantitative analysis of individual mRNA transcripts and transcription foci extracted from imaging data offers a relatively new opportunity to infer transcriptional activity.

2.3.1 Inferring Kinetic Parameters from mRNA Distributions

Zenklusen et al. (2008) used FISH to provide exact mRNA counts in *Saccharomyces cerevisiae* for the first time, demonstrating transcriptional bursting for a key gene involved in pre-ribosomal processing. So et al. (2011) proposed a novel theoretical perspective based on Shannon's mutual information function to analyse transcriptional bursting in *S. cerevisiae*, concluding that transcriptional time series contained information transmitted from an outside stimulus, such as the extracellular concentration of inducer molecules. Additionally, they concluded that the mRNA expression level was modulated through varying the gene 'off' rate. Garcia et al. (2013) pioneered the use of smFISH in calibration of MS2 videos in terms of polymerase molecules (discussed at length in the MS2 and image analysis sections). Little et al. (2013) focussed on using FISH to measure intrinsic noise relating to transcription of the *hunchback* gene in the early *Drosophila* embryo, concluding that precise developmental boundaries were achieved through simple spatio-temporal averaging in the absence of feedback, despite the presence of intrinsic noise in the system.

Jones et al. (2014) constructed a set of synthetic promoters in *Escherichia coli* and used FISH to evaluate the effect of varying promoter strength, transcription factor binding strength and transcription factor copy number on variability in gene expression, concluding that the ability of their model to predict the observed variability in gene expression indicated that transcription noise is tuneable and therefore represents an evolutionarily accessible parameter. Senecal et al. (2014) investigated the control of transcriptional bursting of the proto-oncogene c-Fos in human U2OS cells, using

the FISH-quant protocol originally developed by Mueller et al. (2013). MAPK induction was found to control the frequency of transcriptional bursts through variation in transcription factor concentration levels. Synthetic transcription factors were used to tune the parameters of the transcriptional bursts, implying a role for the strength of the transactivation domain in regulation of polymerase initiation frequency and transcription factor lifetime in controlling burst duration.

In a key paper Xu et al. (2015) combined smFISH and immunofluorescence to analyse the relationship between the *hunchback* gene and the bicoid transcription factor in the *Drosophila* embryo. Analysis of transcription factor binding revealed a Hill function-type relationship between *hunchback* and bicoid, uncovering the gene regulation function linking the gene and its activator. Maximum likelihood estimation of the kinetic parameters of the random telegraph model indicated that regulation of the activation rate alone, i.e. burst frequency, was able to explain the observed distribution of mRNA. This particular form of analysis may be relevant in determining the relationship between Dpp and genes in the *ush* group involved in dorsal-ventral patterning, although the practical difficulties involved in estimating the distribution of Dpp (Umulis et al., 2010) may limit its application.

Bahar Halpern et al. (2015b) concluded from analysis of smFISH images of mammalian liver tissue that genes with short mRNA lifespans were associated with an increased burst fraction, allowing a rapid transcriptional response while reducing 'burst-associated noise'. They also noted that bursting may reduce the statistical likelihood of transcription factor misbinding events. The lab protocol associated with the same paper has since been published as a separate methodology (Bahar Halpern and Itzkovitz, 2016), allowing estimation of mRNA degradation (in addition to production) rates. Labelling of introns was used to identify transcription sites, as introns are generally spliced co-transcriptionally; labelling of exons was used to calculate average Pol II occupancies. Taken in conjunction, the average number of cytoplasmic mRNA molecules per cell and the average polymerase occupancy of a transcription site were used to calculate the mRNA degradation rate.

Bartman et al. (2016) revealed, through FISH experiments, that increasing the frequency of chromatin contact through forcing of enhancer contact using an LCR-β-globin promoter chromatin loop increased burst fraction, but not burst size in mammalian cells, providing further evidence for enhancer control of transcriptional burst frequency. A recent paper has integrated Pol II ChIP-seq with smFISH to argue for a 'multi-step' regulatory process, where transcriptional burst frequency is the main

parameter under control (Bartman et al., 2019). Zoller et al. (2018) advanced an elegant mathematical argument for transcriptional burst frequency regulation in anterior-posterior patterning of the *Drosophila* embryo, offering a re-parameterisation of the standard formulation of transcriptional kinetic parameters. The authors analysed the effect of solving the master equation for the anterior-posterior system while allowing each of k_{ini}, k_{off} and k_{on} to vary. In this manner, they were able to compare the measured noise to the predicted noise levels from varying each of the inputs.

2.3.2 Inferring Kinetic Parameters from Live Imaging

As an alternative to the inference of the kinetic parameters of transcription from static snapshots of the distribution of mRNA transcripts within fixed tissue, live imaging may be used to infer kinetic parameters *in vivo* through the use of the MS2-GFP system. As described in section 2.2.3, this technique involves the use of fluorescence microscopy to quantify fluorescent light emitted from Green Fluorescent Protein bound to hairpin loops contained within transcripts of a genetically edited gene of interest. Approaches to extracting kinetic parameters from MS2 data have typically focussed on either manual fitting of simple mathematical models to recorded data (Fukaya et al., 2016; Bothma et al., 2014) or the inference of the parameters of the random telegraph model through maximum likelihood methods (Suter et al., 2011). In recent years, techniques from machine learning have been applied to MS2 data with promising results (Corrigan et al., 2016; Berrocal et al., 2018; Lammers et al., 2020).

Early work by Golding et al. (2005) utilised MS2-GFP to provide evidence of transcriptional bursting in *Escherichia coli*. Following induction, the mRNA count within a population of cells was recorded at 30 second intervals. mRNA production within each cell was found to scale as a Poisson process, but with larger fluctuations than expected for a simple Poisson process (larger variance than the mean), consistent with transcriptional bursting. Kinetic parameters were compared to the random telegraph model, where superposition of exponential waiting times between periods of transcript production and Poissonian production of transcripts during a period of activity results in a geometrically distributed number of transcripts produced during each 'on period'. Observed data was found to be consistent with the random telegraph model, as measured through comparison of observed data with data simulated using a Gillespie Algorithm formulation of the random telegraph model.

Darzacq et al. (2007) used MS2-GFP to measure kinetic parameters of transcription in a mammalian cell line, allowing quantification of not only promoter initiation,

dissociation and escape constants but also the dynamics of transcriptional pausing. Muramoto et al. (2010) investigated the role of methylation of H3K4 in inheritance of activate transcriptional states between mother and daughter cells through imaging of transcription in *Dictyostelium*. Both transcriptional pulse length and firing rate were found to be inherited through epigenetic mechanisms, providing direct evidence of transcriptional burst and frequency modulation. Evidence of frequency modulation of transcriptional bursting was provided by Larson et al. (2013) through light-sensitive control of a single steroid-responsive gene in human U2-OS cells; the duration of on and off-times was inferred using a Hidden Markov Model originally developed by Lee (2009).

A key paper by Garcia et al. (2013) (discussed in more detail later in this subsection, see also Lucas et al. (2013)) introduced the use of MS2 to quantification of transcription in the *Drosophila* embryo, allowing for association of transcriptional activity with formation of the expression domain of the *hunchback* gene. smFISH was used to calibrate MS2 fluorescence traces, allowing estimation of the number of Pol II on a single-cell basis. The authors concluded that not only the mRNA content of a given cell, but also the time period of transcriptional activity were key in determining cellular developmental decisions, introducing an averaging effect over space and time into expression domain formation. Calibration using smFISH allowed for estimation of the rate of polymerase loading, in addition to estimation of the length of promoter on and off-periods. Bothma et al. (2014) also investigated transcriptional dynamics in the *Drosophila* embryo; specifically, the dynamics of the formation of *eve* stripe 2 during nuclear cycle 14. The spatial distribution of mRNA transcripts on a single-cell basis was calculated through integration of MS2 fluorescent traces. The authors observed a highly dynamic pattern of transcription; burst cycles were estimated to last 4-10 minutes, with 20-100 mRNA transcripts produced during a single burst. Estimation of Pol II loading rates through integration of the M2 fluorescent signal produced a surprising result – the rate of Pol II loading appeared to be temporally regulated, with rates of Pol II loading ranging from 4 to 14 Pol II complexes per minute. The authors argued that this provided evidence against a simple two-state model, with the promoter instead adopting a multi-state model through switching between multiple rates of Pol II loading.

smFISH was used in conjunction with MS2 by Ochiai et al. (2014) to investigate

regulation of Nanog, a transcription factor associated with pluripotency, in mouse embryonic stem cells. The authors wished to understand the association between cell-to-cell heterogeneity in Nanog expression and embryonic stem cell fate decisions. Nanog was found to be transcribed in a pulsatile and stochastic fashion. The authors found statistically significant variation in both transcriptional burst frequency and duration. Super-Poissonian variability in the distribution of Nanog mRNA was observed, consistent with transcriptional bursting. The transcriptional response of β-actin to perturbation of extracellular signalling factors in mouse embryonic fibroblasts was analysed by Kalo et al. (2015), with the aim of establishing the relationship between signaling factor levels and transcriptional pulse fidelity. Inference of promoter transition times through maximum likelihood estimation of the parameters of a binomial mixture model trained using MS2 data revealed that the transition rate from an inactive to active transcriptional state was modified following serum induction.

An insight into the molecular processes underlying transcriptional bursting was provided by Tantale et al. (2016), who revealed the formation of 'convoys' of Pol II during transcriptional bursts in both synthesis of HIV-1 and cellular transcription. Analysis of MS2 data revealed bursts of Pol II separated by several hundred nucleotides, with the promoter activity regulated by two separate processes – minute-scale fluctuations in transcription regulated by Mediator and TBP-TATA box regulated fluctuations on a timescale of hours. Additional evidence of the structural phenomena behind transcriptional bursting has been provided by research in *Drosophila* by Fukaya et al. (2016), who examined the role of enhancer-promoter communication in the regulation of transcription. Insertion of different enhancers downstream of synthetic reporter genes was used to investigate the effect of varying enhancer 'strength' on transcriptional activity. Strong enhancers were found to increase the frequency of transcriptional bursting, with the amplitude and duration of bursts remaining unchanged. Insertion of insulators into the genome resulted in a reduction in burst frequency, suggesting a role for enhancer modulation of transcriptional bursting via frequency control. However, further validation outside of this specific synthetic application may be necessary.

The application of machine learning to the inference of promoter activity from MS2 data was pioneered by Corrigan et al. (2016), building upon earlier work by Lee (2009). A Hidden Markov Model was used to characterise transcription of the actin housekeeping gene in *Dictyostelium*. Due to the persistence of the fluorescence of Pol II while still in transit during elongation, a two-layer model was constructed with

the hidden states of the Hidden Markov Model corresponding to the Pol II initiation rate and number of Pol II initiated at a given time point, uncoupling the modelling of the rate of Pol II initiation from inference of gene state. Following maximum likelihood training on synthetic data, application of the Akaike Information Criterion (AIC) model comparison technique indicated the presence of a 'continuum' of transcriptional states, rather than a small, discrete number of transcriptional states. A 'ladder' of possible transcriptional states was proposed by the authors to describe this continuum of initiation rates. Autocorrelation analysis of initiation rates revealed fluctuations in initiation rate on a timescale of 5-6 minutes. The introduction of point mutations in the TATA box of the act5 gene was used to investigation the effect of perturbations on transcriptional parameters inferred using the model; significant changes in Pol II initiation rate were found, whereas changes in burst frequency and duration were absent.

Two companion papers (Berrocal et al., 2018; Lammers et al., 2020) adapting and extending the Hidden Markov Model approach of Corrigan et al. for application to modelling MS2 data collected from the early *Drosophila* embryo have provided a theoretical basis for computational modeling of DV patterning in *Drosophila*. Lammers et al. outlined a computational pipeline for analysis of MS2 movies of the transcriptional activity of stripe 2 of the *even-skipped* (*eve*) gene in the *Drosophila* embryo, with the aim of establishing if regulation of transcriptional burst frequency alone is sufficient to explain formation of the eve expression domain, i.e. if spatial modulation of burst frequency alone is able to re-create the formation of *eve* stripe 2. The authors used their Compound State Hidden Markov Model (cpHMM) to infer promoter state, along with Pol II loading rate, burst frequency and burst duration for different spatial regions of the *eve* expression domain and concluded that while burst frequency is the main regulated parameter, variation of burst frequency alone is insufficient to re-create the distribution of mRNA across the expression domain – instead, the duration of transcriptional activity, i.e. when which cells are active and for how long, is varied across the expression domain, resulting in a kind of 'digital' on-off control in conjunction with the analogue control provided by mRNA count – the so-called 'binary control' in the title of the paper.

The basic model of mRNA dynamics proposed in the paper is relatively simple. The mean rate of transcription is given by the product of the fraction of time spent in a transcriptionally active state and the rate of Pol II loading while in the active state:

$$\langle transcription\,rate \rangle (x,t) = r(x,t) \frac{k_{\text{on}}(x,t)}{k_{\text{on}}(x,t) + k_{\text{off}}(x,t)} \tag{2.2}$$

In order to fully describe the accumulation of mRNA at the single-cell scale, the model must also incorporate mRNA degradation. Incorporating the rate of mRNA degradation leads to the following equation:

$$\frac{dmRNA}{dt}(x,t) = r(x,t)\frac{k_{\text{on}}(x,t)}{k_{\text{on}}(x,t)+k_{\text{off}}(x,t)} - \gamma mRNA(x,t) \tag{2.3}$$

The authors then follow convention in assuming that the rate of bursting does not vary during the nuclear cycle, equivalent to assuming steady-state conditions (setting the derivative to zero), resulting in:

$$mRNA(x) = \frac{1}{\gamma}r(x)\frac{k_{\text{on}}(x)}{k_{\text{on}}(x)+k_{\text{off}}(x)} \tag{2.4}$$

This equation can then be used to compare model predictions to the actual amount of recorded mRNA. Note the presence of the three kinetic parameters: $r(x)$, k_{on} and k_{off}, corresponding to Pol II loading rate, burst frequency and burst duration. The cpHMM described in the paper performs inference over these three parameters.

The authors used the MS2 system to measure the rate of transcription of an *eve* reporter construct in multiple *Drosophila* embryos. smFISH was used to calibrate the MS2 video in terms of Pol II, rather than arbitrary fluorescence. The cpHMM was designed with a specific architecture – the ability to take persistence, or memory, in the data into account – due to the lack of one-to-one correspondence between the MS2 signal and the hidden variable of interest (promoter state); the recorded signal is the aggregate of all of fluorescent Pol II currently in transit down the gene. Validation using synthetic data indicated that the cpHMM was capable of carrying out inference on MS2 data. The authors specified a three-state model, where either allele, both alleles or neither alleles of a given pair could be active; the extra state was included so as to model the presence of sister chromatids in the imaging data. Parameter inference using the cpHMM indicated that while Pol II loading rate and the rate of transitions to an inactive state (burst duration) were not modulated across the width of the expression domain, the rate of transition from an inactive to active state (burst frequency) was significantly up-regulated in the centre of the expression domain, resulting in an increase of time spent in the active state in the centre of the expression domain; the authors conclude that this indicates that transcription factors regulate burst frequency in the developing embryo, consistent with Xu et al. (2015) and Fukaya et al. (2016).

The model in the paper assumes that the promoter may be in one of K possible states. Transitions between states at a given time point are assumed to be Markovian,

i.e. the current promoter state at a given time point is dependent only upon the promoter state at the previous time point. The probability of transitioning between the K possible states is encoded by a $K \times K$ transition probability matrix A. Each effective promoter state, z_K, is associated with a polymerase initiation rate, r. The persistence of fluorescence generated by Pol II in transit down the length of the gene is modelled through the inclusion of a window variable, W. The observed fluorescence y at a given time point is the combination of fluorescence from polymerase initiated in the previous W time steps, resulting in a dependence not only on the current hidden promoter state z_t, but also the hidden states in the previous W time steps (z_{ss}). The authors introduce the concept of a compound state, s_t, to model this dependency on previous time steps without violating the Markov condition, in conjunction with the set of model parameters, θ. We describe the cpHMM in greater detail in section 2.3.5 after first introducing Hidden Markov Models in the following section.

2.3.3 Hidden Markov Modelling of MS2 Data

The following sections give a general overview of the theory of Hidden Markov Models, followed by a more detailed analysis of the Compound State Hidden Markov Model proposed by Lammers et al. (2020). In the next chapter we introduce a new approximate inference scheme that allows the cpHMM to be applied to longer genes than the original formulation.

2.3.4 Markov Models

In many cases, data can be effectively described as independent and identically distributed (i.i.d). For many situations this assumption is sufficient, but it may not be appropriate for dealing with sequential data, where the time-dependence of the data should be taken into account. Many different types of model have been described in the scientific literature for modelling time series data. These models generally fall into into two main classes: deterministic and stochastic models (Murphy, 2012; Bishop, 2006; Durbin, 2006; Rabiner, 1989). For deterministic models, it can be assumed that the properties of the underlying signal can be captured by calculating the parameters of the signal, such as the amplitude, wavelength and phase of a sinusoidal signal. For stochastic models, on the other hand, a statistical approach is taken instead, with the aim of inferring the statistical parameters of the random process assumed to have generated the recorded signal.

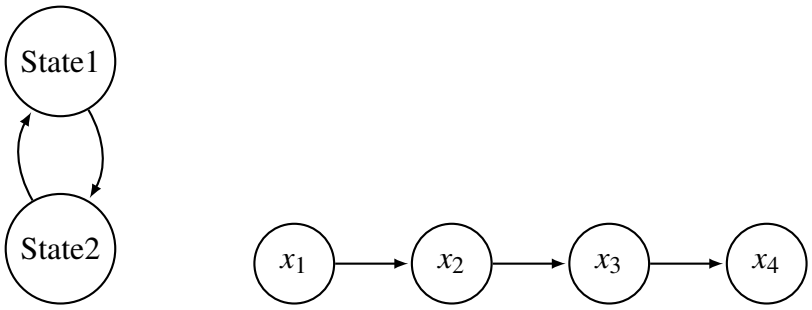

Figure 2.3: A simple first-order Markov Chain of observations x_t.

One relatively simple way of capturing time dependence in a series of observations is to consider a *Markov Model*. For a series of observations $x_1, ..., x_T$, the product rule can be used to express the joint distribution of the observations as

$$p(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1}) \tag{2.5}$$

If the conditional distributions on the right-hand side of equation 2.5 are assumed to be independent of all previous distributions apart except from x_{t-1} then a *First-Order Markov Chain* is obtained. A simple example of a first-order Markov Chain is shown in Figure 2.3. The system alternates stochastically between State 1 and State 2 at each time point t. An observation, x_t, is associated with each time point t.

The joint distribution of T observations under this model is then given by

$$p(x_1, \ldots, x_T) = p(x_1) \prod_{t=2}^{T} p(x_t | x_{t-1}) \tag{2.6}$$

Due to the product rule of probability, the conditional distribution for observation x_t given all observations up to time t is given by

$$p(x_t | x_1, \ldots, x_{t-1}) = p(x_t | x_{t-1}) \tag{2.7}$$

This equation implies that when predicting the next observation in a sequence, only the immediately preceding observation is taken into account; all previous observations are discounted. While this involves making strong assumptions about the time dependency characteristics of the data, Markov Models have been used successfully in practice to model time series data with simple short-range time dependencies. Longer-range time dependencies can be incorporated into the model by allowing x_t to be conditionally dependent upon observations further back in time. For example, allowing x_t to depend

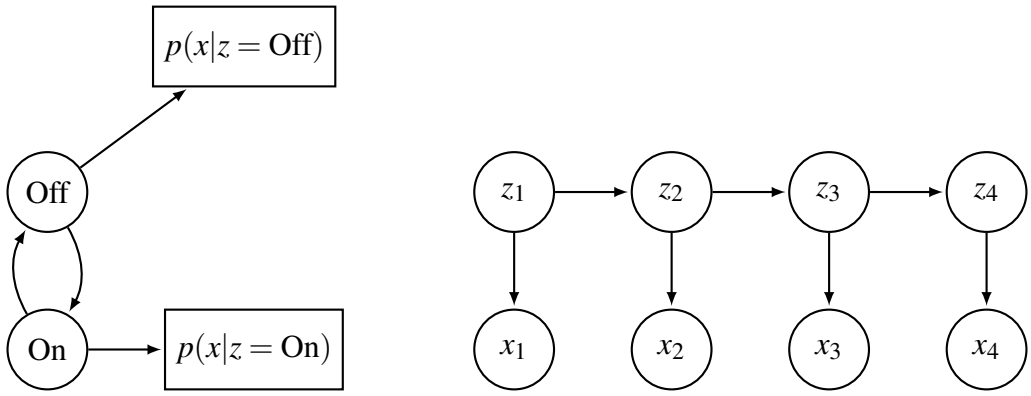

Figure 2.4: Ergodic and Trellis diagrams for a simple 2-state Hidden Markov Model, which cycles between 'Off' and 'On' latent states. The emission generated at each time point x_t is conditioned upon the latent state.

upon x_{t-1} and x_{t-2} results in a *Second-Order Markov Chain*. In practice, this can be be extended up to an M^{th} order Markov Chain, where the conditional distribution for a given variable depends upon the previous M variables. The penalty paid for introducing longer-range time dependencies is the increased complexity of the model - the exponential growth in the number of parameters as M increases results in this approach becoming computationally infeasible for large values of M.

Another approach to attempting to capture complexity in the data is to introduce the concept of latent variables, or *hidden states*. For each observation x_t there is a corresponding latent variable z_t. While the latent variables do satisfy the Markov assumption, the observations do not, as the observations are not conditionally independent. This general class of model is known as a *State Space Model*. When the latent variables are discrete, the term *Hidden Markov Model*, or HMM, is used. A simple example HMM is shown in Figure 2.4. The following sections describe the structure and algorithmic detail of HMM's.

The Structure of Hidden Markov Models

A standard HMM can be described by a number of essential elements. Firstly, the number of states of the model, or K. K generally has some kind of physical significance in relation to the data being modelled; for example, a HMM describing the sequence of DNA base pairs may have four states, corresponding to the four nucleotides. In the model described in following sections, K represents the state of the promoter at a given time point, which may be in either an 'active' or 'inactive', or 'on' and 'off' configuration, resulting in a 2-state model. The latent variables of the model at step t

take the form of a $1 \times K$ sequence $z_t = [z_{t1}, z_{t2} \ldots z_{tK}]$ with $K-1$ zeros and one element $z_{tk} = 1$ indicating which state is occupied at time t. The probability of switching between the hidden states of the model is described by a *transition matrix*, or A. A is a $K \times K$ matrix of transition probabilities. A corresponds to the conditional distribution $p(z_t|z_{t-1})$, which describes the dependence of the probability of the current latent state z_t on the previous latent state z_{t-1}. The transition probabilities may be written as $A_{jk} \equiv p(z_{tk} = 1|z_{t-1,j} = 1)$, allowing the conditional distribution of latent states to be written as

$$p(z_t|z_{t-1,A}) = \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{z_{t-1,j} z_{tk}} \tag{2.8}$$

The transition matrix has a physical significance in that the transition probabilities can be converted to rates. For the 2-state promoter model, the transition probabilities describe the probability of the promoter switching between active and inactive states, as shown in equation 2.9. For example, $k_{off \to on}$ describes the rate, or frequency, at which the promoter switches from an inactive or active state, and can be expressed in terms of transitions min^{-1}.

$$A = \begin{bmatrix} a_{00} & a_{10} \\ a_{01} & a_{11} \end{bmatrix} = \begin{bmatrix} k_{off \to off} & k_{on \to off} \\ k_{off \to on} & k_{on \to on} \end{bmatrix} \tag{2.9}$$

The vector of initial states, π_k, describes the probability of the hidden states starting in one of K configurations. π_k takes the form of a vector of probabilities representing the marginal distribution $p(z_1)$, such that

$$p(z_1|\pi) = \prod_{k=1}^{K} \pi_k^{z_{1k}} \tag{2.10}$$

with $\sum_k \pi_k = 1$.

An *observation model*, $p(x_t|z_t, \phi)$, is required to describe the probability of the observed data, where ϕ is a set of parameters belonging to the conditional distribution of observed variables. Many different probability distributions may be used as observation models, with both discrete and continuous observations being a possibility. These probabilities are known as *emission probabilities*, which may be represented as

$$p(x_t|z_t, \phi) = \prod_{k=1}^{K} p(x_t|\phi_k)^{z_{tk}} \tag{2.11}$$

The complete set of parameters describing the model are given by $X = \{x_1, \ldots, x_T\}$, $Z = \{z_1, \ldots, z_T\}$, $\theta = \{\pi, A, \phi\}$. Combining each of these components, the joint probability function over latent and observed variables for a standard HMM may therefore be given by

$$p(X, Z|\theta) = p(z_1|\pi) \left[\prod_{t=2}^{T} p(z_t|z_{t-1}, A) \right] \prod_{t=1}^{T} p(x_t|z_t, \phi) \tag{2.12}$$

For the Compound State Hidden Markov model described in following sections, a Gaussian observation model was used, parameterised by a mean (μ) and noise (σ) parameter. While many different variations upon the basic HMM template are possible, such as left-to-right and non-ergodic models (Rabiner, 1989), a HMM must include these basic components.

Determining the Likelihood of an Observation Sequence

Applying a HMM to real-world data involves solving three basic problems: how to quantify the likelihood of a given observation sequence, how to train the HMM so as to infer the most likely model parameters, and how to infer the most likely sequence of hidden states for a particular observation sequence, given the model parameters (Rabiner, 1989). The simplest way to determine the likelihood of a particular observation sequence would be to sum over all possible state sequences that could produce a particular observation sequence. However, such an approach would be computationally infeasible, as it would require K^T calculations for a HMM with K hidden states and observation sequences of length T.

A much more computationally efficient approach is therefore required. This is the *forward algorithm*, a form of *dynamic programming*. Rather than summing over all possible state sequences, an intermediate variable known as the *forward variable*, $\alpha(z_t)$, is used instead. $\alpha(z_t)$, the joint probability of observing $x_1, \ldots, x_T$ and being in state z_t, may be defined as

$$\alpha(z_t) \equiv p(x_1, \ldots, x_t, z_t) \tag{2.13}$$

Starting from an initial condition given by

$$\alpha(z_1) = p(x_1, z_1) = p(z_1)p(x_1|z_1) \tag{2.14}$$

A recursive process, based on conditional independence properties, and dependent

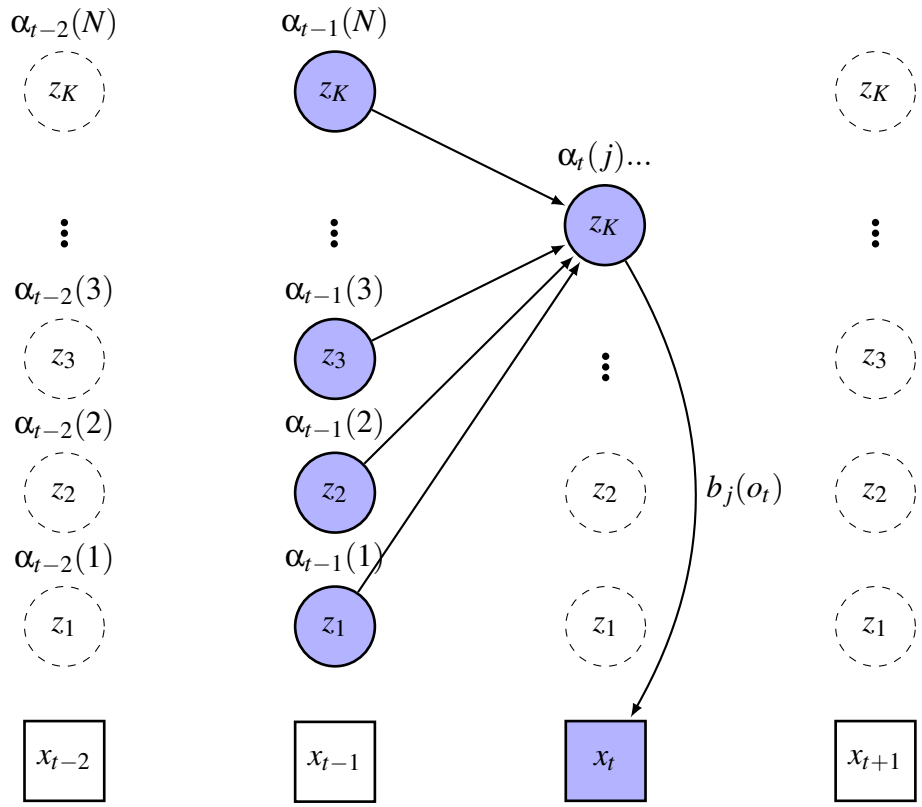

Figure 2.5: Calculation of the forward variable, $\alpha(z_t)$. At each time point t, $\alpha(z_t)$ is calculated by summing up the previous $\alpha(z_t-1)$ values, weighted by their associated observation likelihoods and transition probabilities.

upon the sum and product rules, can be used to express $\alpha(z_t)$ in terms of $\alpha(z_{t-1})$:

$$\alpha(z_t) = p(x_t|z_t) \sum_{z_{t-1}} \alpha(z_{t-1}) p(z_t|z_{t-1}) \tag{2.15}$$

Starting from the initial state, forward variables are calculated as part of an iterative process (Figure 2.5) which can be expressed as a form of message passing (Bishop, 2006). Computing the forward variables in this manner takes $O(K^2T)$ time, rather than K^T time, making calculation of the observation sequence likelihood much more computationally efficient.

Training Hidden Markov Models

In a similar manner to the forward algorithm, it is possible to define the *backward algorithm*, the conditional probability of a future observation $x_{t+1},\ldots,x_t$, given that the hidden state is currently z_t (Figure 2.6). Through a similar recursive process to the

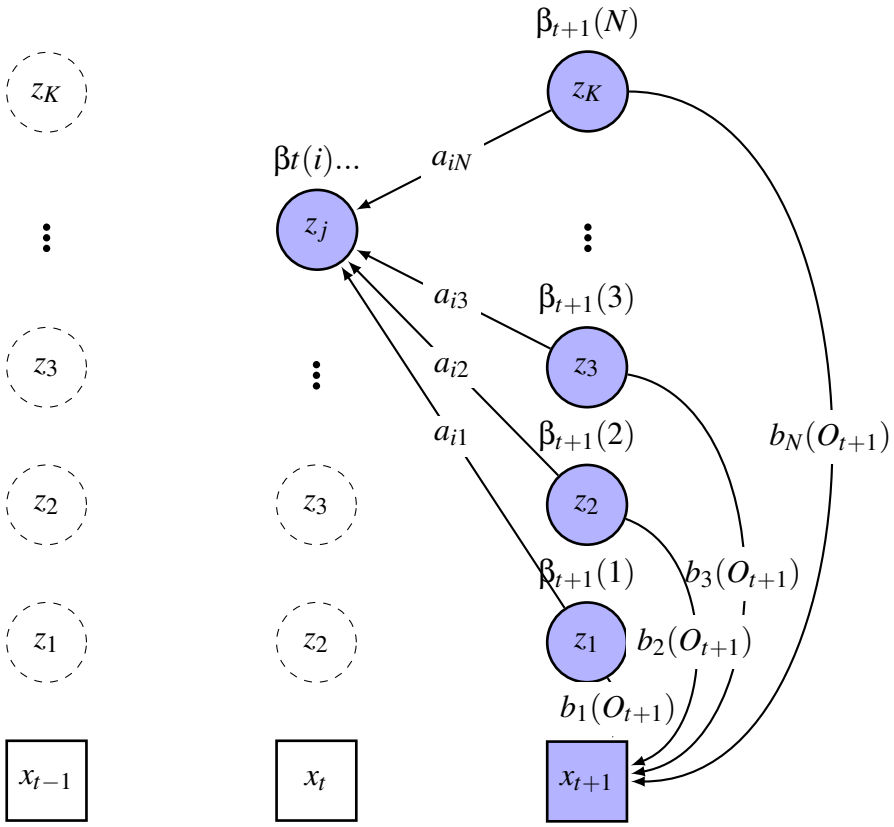

Figure 2.6: Calculation of the backward variable, $\beta(z_t)$. In a similar manner to the forward algorithm, $\beta(z_t)$ is calculated by summing over the contribution of all input $\beta(z_{t+1})$ terms, weighted by the observation and transition probabilities.

forward algorithm, we can define a backward variable $\beta(z_t)$

$$\beta(z_t) \equiv p(x_{t+1},\ldots,x_T|z_t) \tag{2.16}$$

Starting this time from the initial condition

$$\beta(z_T) = 1 \tag{2.17}$$

$\beta(z_t)$ is expressed in terms of $\beta(z_t+1)$ as

$$\beta(z_t) = \sum_{z_{t+1}} p(x_{t+1}|z_{t+1})p(z_{t+1}|z_t) \tag{2.18}$$

Taken in conjunction, the forward and backward algorithms can be used to train a HMM through the *forward-backward algorithm*, or *Baum-Welch algorithm*, a form of Expectation Maximisation (Bishop, 2006). The likelihood for the HMM can be

formulated as

$$p(X|\phi) = \sum_{Z} p(X,Z|\phi) \tag{2.19}$$

Following marginalisation over the latent variables Z. This expression does not factorise over t, preventing summations over z_t being carried out independently. In addition, performing all of the summations involved would involve a total of K^T terms. This exponential growth in computational time, similar to the problem faced when calculating the likelihood of a particular state sequence explicitly, requires that an alternative approach to be taken.

Expectation Maximisation (EM) is an iterative process, beginning with an initial estimate for the model parameters, designated Q^{old}. In the E-step, the posterior distribution over the latent variables, $p(Z|X,\theta^{old})$ is found. In the M-step, the parameters are updated using the values inferred during the E-step. This process is repeated until the change in likelihood calculated as part of the E-step falls below a certain threshold. Formally, the E-step involves evaluation of the Q-function, given by

$$Q(\theta,\theta^{old}) = \sum_{Z} p(Z|X,\theta^{old}) \ln p(X,Z|\theta) \tag{2.20}$$

Two intermediate terms are commonly defined, $\gamma(z_t)$ and $\zeta(z_{t-1},z_t)$, such that

$$\gamma(z_n) = p(z_t|X_\theta^{old}) \tag{2.21}$$

$$\zeta(z_{t-1},z_t) = p(z_{t-1},z_t|X,\theta^{old}) \tag{2.22}$$

Substituting these expressions (represented graphically in Figures 2.7 and 2.8) into the definition for the joint distribution, the Q-function can be written as

$$Q(\theta,\theta^{old}) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{t=2}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} \zeta(z_{t-1,j},j,z_{tk}) \ln A_{jk} + \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma(z_{tk}) \ln p(x_t|\phi_k) \tag{2.23}$$

For a HMM, this expression has a closed form and can be evaluated explicitly using Expectation Maximisation. During the E-step, $\gamma(z_t)$ and $\zeta(z_{t-1},z_t)$ are evaluated, followed by the M-step, where $Q(\theta,\theta^{old})$ is maximised with respect to the model parameters. In terms of $\gamma(z_t)$ and $\zeta(z_{t-1},z_t)$, the model parameters π and A can be expressed as

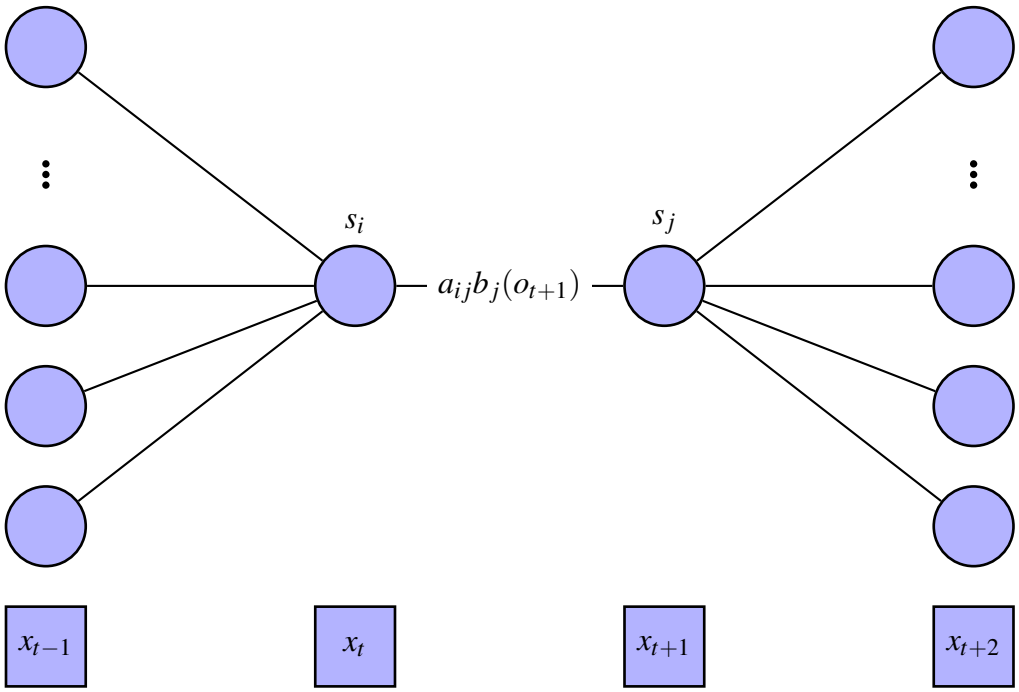

Figure 2.7: Calculation of the joint probability, $\zeta(z_{t-1}, z_t)$, the probability of being in state i at time t and state j at time $t+1$, given the observation sequence and the model.

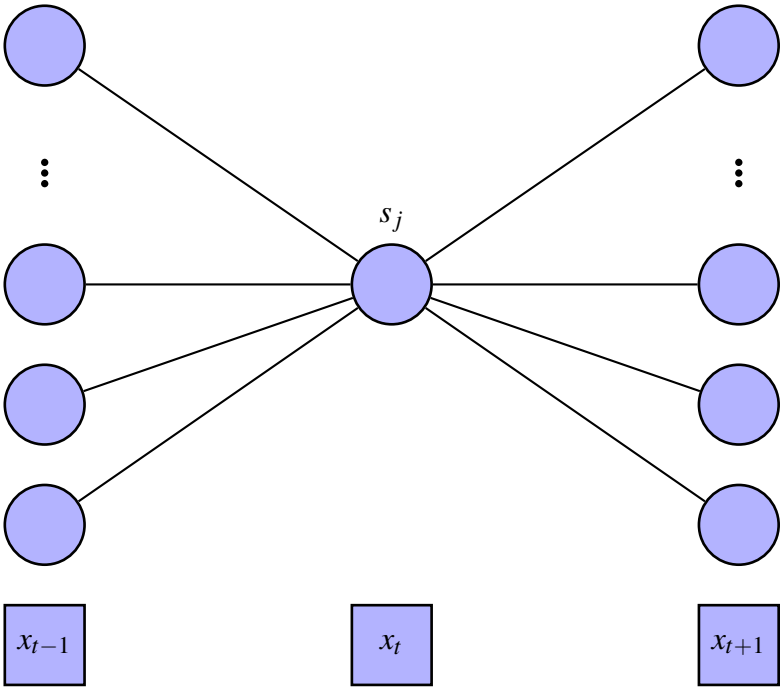

Figure 2.8: Calculation of $\gamma(z_t)$. $\gamma(z_t)$ represents the probability of being in state j at time t.

$$\pi_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} \tag{2.24}$$

$$A_{jk} = \frac{\sum_{t=2}^{T} \zeta(z_{t-1,j,z_{tk}})}{\sum_{l=1}^{K} \sum_{t=2}^{T} \zeta(z_{n-1,j,z_{tl}})} \tag{2.25}$$

Assuming a Gaussian observation likelihood, the mean and covariance of the emission density are given by

$$\mu_k = \frac{\sum_{t=1}^{T} \gamma(z_{tk}) x_t}{\sum_{t=1}^{T} \gamma(z_{tk})} \tag{2.26}$$

$$\sum_k = \frac{\sum_{t=1}^{T} \gamma(z_{tk})(x_t - \mu_k)(x_t - \mu_k)^T}{\sum_{t=1}^{T} \gamma(z_{tk})} \tag{2.27}$$

It is now possible to write $\gamma(z_t)$ and $\zeta(z_{t-1}, z_t)$ in terms of $\alpha(z_t)$ and $\beta(z_t)$

$$\gamma(z_t) = \frac{\alpha(z_t)\beta(z_t)}{p(X)} \tag{2.28}$$

$$\zeta(z_{t-1}, z_t) = \frac{\alpha(z_{t-1})\beta(z_t)p(z_t|z_{t-1})p(x_t|z_t)}{p(X)} \tag{2.29}$$

Along with the model parameters, θ

$$\pi_k = \frac{\alpha(z_{1k})\beta(z_{1k})}{\sum_{j=1}^{K} \alpha(z_{1j})\beta(z_{1j})} \tag{2.30}$$

$$A_{jk} = \frac{\sum_{t=2}^{T} \alpha(z_{t-1,j})\beta(z_{tk})p(x_t|\phi_k)A_{jk}}{\sum_{l=1}^{K} \sum_{t=2}^{T} \alpha(z_{t-1,j})\beta(z_{tl})p(x_t|\phi_l)A_{jl}} \tag{2.31}$$

$$\phi_{ik} = \frac{\sum_{t=1}^{T} \alpha(z_{tk})\beta(z_{tk})x_{ti}}{\sum_{t=1}^{T} \alpha(z_{tk})\beta(z_{tk})} \tag{2.32}$$

$\alpha(z_t)$ and $\beta(z_t)$ are inferred using the forward and backward algorithms. The whole process is then as follows: following an initial estimate of the model parameters, carry out the forward and backward algorithm during the E-step to estimate $\alpha(z_t)$ and $\beta(z_t)$. These estimates are then used to update the model parameters during the M-step until the change in likelihood with each successive pass of the algorithm falls below a certain threshold.

State Sequence Decoding

The final problem associated with HMM's is the decoding problem: finding the optimal sequence of observations associated with a particular sequence of states. This may be expressed as two similar problems: finding the most probable sequence of latent states and finding the set of states that are individually most probable. The latter problem is known as *posterior decoding*, and is solved by maximising the latent variable marginals $\gamma(z_t)$ inferred using the forward-backward algorithm. The former problem can be solved efficiently through the use of the *Viterbi algorithm*, a form of dynamic programming algorithm (Figure 2.9). As with attempting to determine the likelihood of a given state sequence, attempting to quantify the probability of all possible paths of latent states would be computationally too expensive. A backtracking approach is taken instead, using the emission and transition probabilities calculated during the forward-backward algorithm. At each time step t, only a record of the state associated with highest Viterbi probability needs to be stored. At the end of each pass of the algorithm these records, or backpointers, can be used to find the most likely sequence of hidden states. Formally, the problem is that given X, we wish to find Z^*, the overall most likely explanation of X, such that $Z_* = \arg\max_Z p(X,Z|\theta)$. $p(X,Z^*)$ can be expressed as

$$\begin{aligned} p(X,Z^*) &= \max_Z p(X,Z) = \max_{z_1,\ldots,z_T} p(x_1,\ldots,x_T,z_1,\ldots,z_T) \\ &= \max_{Z_T} \max_{z_1,\ldots z_{T-1}} p(x_1,\ldots,x_T,z_1,\ldots,z_T) \\ &= \max_{Z_T} \omega(z_T) \\ z_N^* &= \arg\max_{z_N} \omega(z_N) \end{aligned} \tag{2.33}$$

Where $\omega(z_N = \max_{z_1,\ldots,z_T} p(x_1,\ldots,x_T,z_1,\ldots,z_T)$ is the probability of the most likely sequence of states $z_1,\ldots,z_t$ ending in z_t generating the observations $x_1,\ldots,x_t$.

2.3.5 The Compound State Hidden Markov Model

In the Lammers model, the promoter at time point t may be in one of K effective states. Transitions between effective promoter states z_t at time point t are assumed to be Markovian and are modelled by the $K \times K$ transition matrix $A = p(z_t|z_{t-1})$. Due to the persistence in the MS2 signal, the observation y_t at time t depends upon not

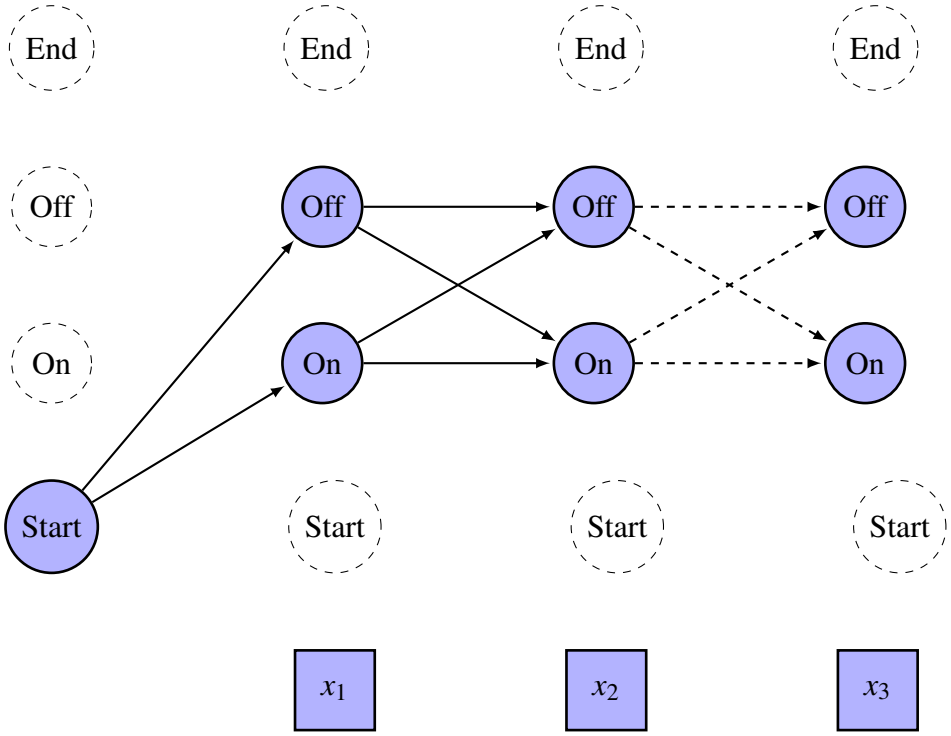

Figure 2.9: Outline of the Viterbi algorithm. At each time point, a variable known as a backpointer is stored, allowing for computation of the most likely path through the state trellis.

only upon the current hidden promoter state z_t, but also on the previous w hidden promoter states $\{z_t, z_{t-1}, \ldots, z_{t-w+1}\}$, where $w = \frac{\tau_{\text{elong}}}{\Delta\tau}$ is the window size of the model. τ_{elong} is the polymerase elongation time and $\Delta\tau$ is the time resolution of the system. τ_{elong} depends upon the Pol II elongation rate and the gene length. $\Delta\tau$ depends upon the experimental setup, but was 20s for both the results presented in Lammers et al. (2020) and in Chapters 3 and 4. Estimates of the elongation rate vary broadly between different organisms (Zenklusen et al., 2008; Gómez-Schiavon et al., 2017). Garcia et al. (2013) estimated the elongation rate in *Drosophila* to be 1.5 kb min^{-1}. More recent work has estimated an elongation rate of 2.4 - 3 kb min^{-1} (Fukaya et al., 2017). Lammers et al. used autocorrelation analysis to calculate an elongation rate of 2.8 kb min^{-1}, which is consistent with this estimate. In the results presented in Chapters 3 and 4 we have used 2.8 kb min^{-1} to calculate the window size.

To describe the system Lammers et al. introduced the concept of a compound state $s_t = \{z_t, z_{t-1}, \ldots, z_{t-w+1}\}$, where each compound state can take on one of K^w different values. The set of all possible compound states $s_t \in \{1, \ldots, K^w\}$ is K^w in size. At each time point the most recent $w-1$ promoter states are passed from one compound state to the next, resulting in the last $w-1$ promoter states in $s_{t+1} = \{z_{t+1}, z_t, \ldots, z_{t-w+2}\}$ being included in s_t. This deterministic passing of the previous promoter states results in only K different transitions being allowed between each compound state at time t.

The emission probabilities associated with each hidden compound state are modelled using Gaussian distributions with standard deviation σ. ν represents a $K \times 1$ vector, with each row of the vector corresponding to the emission associated with a particular effective state. The model noise parameter, σ, is a scalar (Lammers et al., 2020). The joint probability of hidden compound states and fluorescence values is given by:

$$p(\mathbf{y}, \mathbf{s}|\theta) = p(s_1)|\pi) \prod_{t=1}^{T} p(y_t|s_t, \nu, \sigma) \prod_{t=2}^{T} p(s_t|s_{t-1}, \mathbf{A}) \tag{2.34}$$

Training the model requires finding an estimate of the model parameters, $\hat{\theta}$, which maximise the likelihood of observing the MS2 fluorescence data:

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{y}|\theta) \tag{2.35}$$

The likelihood may be calculated through marginalisation of the joint probability distribution, i.e. summing over compound states:

$$p(\mathbf{y}|\theta) = \sum_{s=\{s_1,s_2,.....s_T\}} p(\mathbf{y},\mathbf{s}|\theta) \tag{2.36}$$

Exact inference in this situation is intractable, due to the very large number of summations required to manually sum-out the hidden state variable. Instead, as is common with Hidden Markov Model inference, the authors instead use the Expectation-Maximisation (EM) algorithm (or Baum-Welch algorithm in the context of Hidden Markov Models) to train the model, based on the following expression for the logarithm of the joint probability distribution:

$$\log p(\mathbf{y},\mathbf{s}|\theta) = \log p(s_1|\pi) + \sum_{t=1}^{T} \log p(y_t|s_t,\nu,\sigma) + \sum_{t=2}^{T} \log p(s_t|s_{t-1},\mathbf{A}) \tag{2.37}$$

Inferred model parameters may then be used to calculate transcriptional kinetic parameters within a given spatial region. While there is a clear conceptual link between the Hidden Markov Model framework and the underlying physical process of an unobserved binary promoter state generating continuous fluorescence time series data, a key limitation of the cpHMM as implemented in this paper is the scaling of computational cost with the W window size variable. W increases with gene length, as the variable is responsible for capturing the length of time needed for Pol II to travel down the length of the gene. The *eve* reporter construct used in the paper is of a relatively short length, resulting in a window size of 7. Many other *Drosophila* genes, however, such as *ush*, would require a much larger window size – 19 in the case of *ush*. This is computationally intractable using the model as published in the paper, due to the exponential scaling of the algorithm with window size. While the underlying mathematics is sound, re-formulating the algorithm to scale less severely (ideally linearly) with gene length would allow for inference of promoter state in any gene of interest in *Drosophila*, and may also form the basis of a general computational tool for inference in other organisms. Initial modelling results using the Matlab code published with the paper have indicated that while the algorithm is effective at modelling dorsal-ventral patterning, the scaling of computational cost with window size is a factor prohibiting use with other genes.

Calculating the likelihood involves solving equation 2.37. In order to break the problem down into its constituent parts, Lammers et al. introduced several notations:

$$\langle s_t^i \rangle = \sum_{\{s=s1,s2,\ldots,s_T\}} s_t^i p(s|y,\hat{\theta}_k) \tag{2.38}$$

$$\langle s_t^i, s_{t-1}^j \rangle = s_{t-1}^j p(s|y,\hat{\theta}_k) \tag{2.39}$$

With these additional terms, equation 2.37 can be rewritten as:

$$\log p(s_1|\pi) = \sum_{i=1}^{K_w} \sum_{k=1}^{K} s_1^i C_{ki} \log \pi_k \tag{2.40}$$

$$\log p(y_t|S_t,\mu,\sigma) = \frac{1}{2} \sum_{i=1}^{K^w} s_t^i (\log \lambda - \log(2\pi) - \lambda(y_t - V_i(\mu))^2) \tag{2.41}$$

$$\log p(s_t|s_{t-1},A) = \sum_{i,j=1}^{K^w} \sum_{k,l=1}^{K} B_{ij} s_t^i s_{t-1}^j C_{ki} C_{lj} \log A_{kl} \tag{2.42}$$

Where $V_i(v)$ represents the aggregate fluorescence and $\lambda = \frac{1}{\sigma^2}$ represents the precision of the gaussian observation distribution. In these equations $B_{s',s} = 1$ if and only if the transition $s \to s'$ is allowed and $C_{zs} = 1$ if and only if $\Delta(s,1) = z$. A is the transition matrix introduced in equation 2.9. The authors introduced the concept of aggregate fluorescence in order to describe the dependence of the fluorescence contribution of each polymerase on its position on the gene body. Following initiation of transcription, a finite amount of time is taken for the polymerase to transit along the MS2 probe. The fluorescent contribution of each polymerase is therefore dependent upon its position within the time window, W. The aggregate fluorescence is calculated as

$$V_i(v) = F_{i,:} \tag{2.43}$$

F is a $K^W \times K$ matrix, where the i^{th} row of F represents the number of times each promoter state is present in the i^{th} compound state, weighted by the position of the polymerase within the time window W.

The terms defined in equations 2.40, 2.41 and 2.42, along with equations 2.38 and 2.39, allow the log likelihood to be written as:

$$\begin{aligned}\mathcal{L}(\theta|y,\hat{\theta}_k) &= \sum_{i=1}^{K_w}\sum_{k=1}^{K} s_1^i C_{ki}\log\pi_k \\ &+\log p(y_t|S_t,\mu,\sigma) = \frac{1}{2}\sum_{i=1}^{K^w} s_t^i(\log\lambda - \log(2\pi) - \lambda(y_t - V_i(\mu))^2) \\ &+\log p(s_t|s_{t-1},A) = \sum_{i,j=1}^{K^w}\sum_{k,l=1}^{K} B_{ij}s_t^i s_{t-1}^j C_{ki}C_{lj}\log A_{kl}\end{aligned} \tag{2.44}$$

The terms $p(s_t|y,\hat{\theta}_k)$ and $p(s_{t-1},s_t|y,\hat{\theta}_k)$ may be expressed in terms of the α and β HMM forward-backward algorithm parameters mentioned in the earlier section as:

$$p(s_t|y,\hat{\theta}_k) = \frac{\alpha_t(s_t)\beta_t(s_t)}{p(y|\hat{\theta})_k} \tag{2.45}$$

$$p(s_{t-1},s_t|y,\hat{\theta}_k) = \frac{\alpha_{t=1}(s_{t-1})p(y_t|s_t,\hat{\theta}_k)p(s_t|s_{t-1},\hat{\theta}_k)\beta_t(s_t)}{p(y|\hat{\theta}_k)} \tag{2.46}$$

With $\alpha_t(i)$ and $\beta_t(i)$ representing the joint probability of being in the i^{th} compound state at time step t while observing the emission values in the first t time steps and the conditional probability of observing emission values from time point $(t+1)$ up to the end of the time series, given that the compound state at time t is i, respectively:

$$\alpha_t(i) = p(y_l,\ldots,y_t,s_t = i|\hat{\theta}_k) \tag{2.47}$$

$$\beta_t(i) = p(y_{t+1},\ldots,y_T|s_t = i,\hat{\theta}_k) \tag{2.48}$$

Collecting these terms together, the parameters of the cpHMM may finally be expressed as:

$$\hat{\pi}_m = \frac{\sum_{i=1}^{K_w}\langle s_1^i\rangle C_{mi}}{\sum_{k=1}^{K}\sum_{i=1}^{K^w}\langle s_1^i\rangle C_{ki}} \tag{2.49}$$

$$\hat{\mu} = M^{-1}b \tag{2.50}$$

$$M_{mn} = \sum_{t=1}^{T}\sum_{i=1}^{K^w}\langle s_t^i\rangle F_{in}F_{im} \tag{2.51}$$

$$b_m = \sum_{t=1}^{T} \sum_{i=1}^{K^w} \langle s_t^i \rangle y_t F_{im} \tag{2.52}$$

$$\frac{1}{\hat{\lambda}} = \hat{\sigma}^2 = \frac{\sum_{t=1}^{T} \sum_{i=1}^{K^w} \langle s_t^i \rangle (y_t - F_{i,:}\hat{\mu})^2}{\sum_{t=1}^{T} \sum_{i=1}^{K^w} \langle s_t^i \rangle} \tag{2.53}$$

$$\hat{A}_{mn} = \frac{\sum_{t=1}^{T} \sum_{i,j=1}^{K^w} B_{ij} \langle s_t^i s_{t-1}^j \rangle C_{mi} C_{nj}}{\sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{i,j=1}^{K^w} \langle B_{ij} s_t^i s_{t-1}^j \rangle C_{ki} C_{nj}} \tag{2.54}$$

These parameters are inferred and updated at each step of the EM algorithm until convergence. In the original cpHMM formulation of the model, the forward-backward algorithm scales as TK^{2w}, in contrast to the standard HMM forward-backward scaling of TK^2. This represents a significant increase in computational time, particularly for longer genes, which require longer window sizes.

2.4 Conclusion

Mathematical and computational modelling of gene expression has received extensive attention in the scientific literature. While several approaches have been proposed for extracting kinetic parameters from gene expression datasets, the most recent proposed model for modelling MS2 data would become computationally infeasible for analysing many *Drosophila* genes involved in dorsal-ventral patterning. The following chapter outlines an approach allowing for inference of kinetic parameters using genes of arbitrary length.

Chapter 3

Scalable Inference of Transcriptional Dynamics

In this chapter we present an algorithm for efficient inference of transcriptional kinetic parameters from MS2 data. The algorithm has been published in the Oxford University Press *Bioinformatics* journal (Bowles et al., 2022) and is available online as part of the *burstInfer* Python software package. The algorithm represents an improvement over the original cpHMM implementation in that it is possible to infer kinetic parameters for a gene of arbitrary length, as the problem with exponential scaling of computational time with gene length has been resolved. Examples are given of parameter inference using both synthetic and experimental *Drosophila* MS2 data.

3.1 Introduction

Recent advances in *in vivo* live imaging technologies (Pichon et al., 2018) have created a pressing need for algorithms capable of analysing large, complex biological datasets. Live imaging techniques, such as the MS2-MCP system, have been of particular interest to the developmental biology community due to the ability to visualise transcription at single-cell resolution *in vivo*. As correct spatial and temporal control of gene expression is of fundamental importance during both normal development and disease, the ability to analyse the rich datasets generated by live imaging approaches is vital.

The MS2-MCP system allows for the quantification of transcription in real-time through the introduction of hairpin structures into a gene of interest (Pichon et al., 2018). Following the entry of the promoter into an active state, elongation of RNA Polymerase II (Pol II) along the gene body results in the production of nascent mRNA

transcripts containing hairpin stem-loops. Binding of the MCP fluorescent protein to this hairpin structure allows for detection of the resulting fluorescent signal by fluorescence microscopy (Bertrand et al., 1998; Qureshi et al., 1978; Lucas et al., 2013). Quantification of this fluorescent signal results in a fluorescent time series, which acts as a proxy for transcriptional output at each transcription site (Lucas et al., 2013; Qureshi et al., 1978; Bertrand et al., 1998). The ability to track the fluorescence of accumulated nascent mRNA at transcription foci (and therefore levels of transcriptional activity) over time and at single-cell resolution opens up the possibility of investigating spatial and temporal transcriptional dynamics in model organisms, in addition to the response of tissue culture cells to external stimuli (Pichon et al., 2018). The use of the MS2-MCP system allows for the collection of temporal transcriptional data, an advantage over the static 'snapshots' of transcription generated using techniques such as single molecule fluorescent *in situ* hybridisation (smFISH) (Pichon et al., 2018).

Transcription is now understood to be a highly dynamic process, with many genes producing transcripts in discrete pulses, or 'bursts', of transcriptional activity (Coulon et al., 2013; Raj and van Oudenaarden, 2008; Chubb et al., 2006; Golding et al., 2005). Transcriptional bursting has been observed in organisms ranging from *Drosophila* to vertebrates and is implicated in both normal development and disease (Raj and van Oudenaarden, 2008; Eldar and Elowitz, 2010); bursting is of particular interest to the gene regulation community, as many key developmental genes appear to exhibit bursting-like behaviour (Lenstra et al., 2016). Mathematical modelling of transcriptional bursting may be described by a set of kinetic parameters which report the frequency, amplitude and duration of transcriptional bursts (Zoller et al., 2018; Li et al., 2018; Dar et al., 2012; Raj et al., 2006; Fukaya et al., 2016; Corrigan et al., 2016). Previous work on mathematical modelling of transcriptional bursting has focused on inference of these transcriptional parameters through analysis of either static smFISH snapshots (Mueller et al., 2013; Bahar Halpern et al., 2015a; So et al., 2011; Gómez-Schiavon et al., 2017) or MS2-MCP time series data (Corrigan et al., 2016; Qureshi et al., 1978; Fukaya et al., 2016; Berrocal et al., 2018; Lammers et al., 2020; Tantale et al., 2016; Bothma et al., 2014). The ability to infer these kinetic parameters opens up the possibility of providing a deeper insight into the spatio-temporal regulation of bursting at single-cell resolution.

While MS2-MCP time series data allows for visualisation of nascent transcription at single-cell resolution in real-time, inference of kinetic parameters from MS2-MCP data presents a number of unique challenges (Gregor et al., 2014). Crucially,

the presence of persistent fluorescence within the signal complicates inference of transcriptional kinetic parameters (Corrigan et al., 2016; Lammers et al., 2020). Upon the promoter entering an active state, RNA Polymerase (Pol II) commences elongation along the gene body, leading to a fluorescent signal through MCP-fluorescent protein binding. When the promoter becomes inactive, the fluorescent signal does not immediately cease. Pol II molecules are still in transit along the gene body and the incomplete mRNA transcripts are bound by MCP-fluorescent proteins. Inference of kinetic parameters therefore requires an algorithm capable of taking this persistence into account.

Lammers et al. (2020) incorporated the persistence of the MS2 signal through implementing a compound state hidden Markov model (cpHMM), building on an earlier hidden Markov model for MS2-GCP parameter inference (Corrigan et al., 2016). The transition probabilities and emission values of the model correspond to the promoter switching frequencies and Pol II loading rate, respectively, which together are sufficient to describe the bursting dynamics of the system. The promoter switches between active and inactive states according to the transition matrix, loading polymerase onto the gene while in the active state at a rate determined by the model emission parameter (Figure 3.1 A). Persistence in the signal is dealt with through the inclusion of a window parameter, W, that models the dependence of the recorded fluorescence on the previous W promoter states, each of which may take one of K (here 2) values. The inclusion of the window parameter results in K^W compound states to fully describe the system. This exponential scaling becomes problematic when dealing with long genes, as the dependence of the window parameter on elongation time (and therefore gene length) may lead to infeasible computational times.

In this chapter we present a modified form of the cpHMM referred to as *burstInfer*, for fast inference of kinetic parameters from MS2-MCP data. This new method can model genes of arbitrary length through the use of a time-adaptive truncated compound state space. The truncated state space provides a good approximation to the full state space by retaining the most likely set of states at each time during the forward pass of the algorithm. The algorithm represents a significant speed boost over the original cpHMM technique when applied to long genes, removing the exponential time-scaling of the technique with gene length. Results indicate that the use of a reduced compound state space is sufficient to accurately infer kinetic parameters relative to the original model, while significantly reducing computational time for longer genes, making inference of kinetic parameters for genes of all sizes feasible.

3.2 Implementation of Algorithm

3.2.1 Model Formulation

Following insertion of the MS2 stem-loop sequences into the gene of interest, elongation of Pol II along the length of the gene body results in the generation of a fluorescent time series signal. We intend to model the dynamics of these recorded fluorescent signals, with the aim of extracting the kinetic parameters driving expression of the target gene. Following the cpHMM formulation derived by Lammers et al. (2020), whose method this chapter extends, we denote an individual fluorescent signal (corresponding to one transcription foci) as $y = \{y_1, y_2..., y_T\}$, with T denoting the number of time points within the individual trace (Figure 2). We assume that the promoter may be in one of K = 2 effective states, i.e. active or inactive. The promoter switches between hidden states z at time step t according to the $K \times K$ transition matrix, $A = p(z_t|z_{t-1})$. A_{kl} represents the probability of making the transition from hidden promoter state k to hidden promoter state l during time step t. Transitions between hidden promoter states z_t are assumed to satisfy the Markov property, i.e. the hidden promoter state at a given time point depends only upon the hidden promoter state at the previous time point (Lammers et al., 2020).

Each effective state z_t is associated with a polymerase initiation rate, $r(k)$, representing the number of Pol II molecules loaded onto the gene in a given minute. The fluorescence data presented here are shown in terms of arbitrary units of fluorescence. Quantification of the transcriptional output of cells using smFISH may be used to calibrate the signal in terms of Pol II number instead (Qureshi et al., 1978; Lammers et al., 2020; Hoppe et al., 2020). The fluorescence emission per time step t for each effective state is defined as $v(k) = Fr(k)$, where F is a calibration factor used to convert the units of arbitrary fluorescence to units of Pol II (Lammers et al., 2020).

The recorded fluorescence intensity at a given time point (Figure 3.1 B) depends upon not only the fluorescence generated during the previous time step, but also the cumulative fluorescence generated by Pol II in transit along the length of the gene during previous time steps. To model this dependence upon previous time steps the concept of a sliding window, W, is introduced into the model. This window, or memory, represents the dependence of the observation y_t at time point t on not only the hidden promoter state z_t at the current time point but also the previous W hidden promoter states (depicted in Figure 3.2). The value of W is gene-dependent and is calculated as $W = \frac{\tau_{elong}}{\Delta\tau}$, where τ_{elong} is the elongation time and $\Delta\tau$ is the size of an individual time

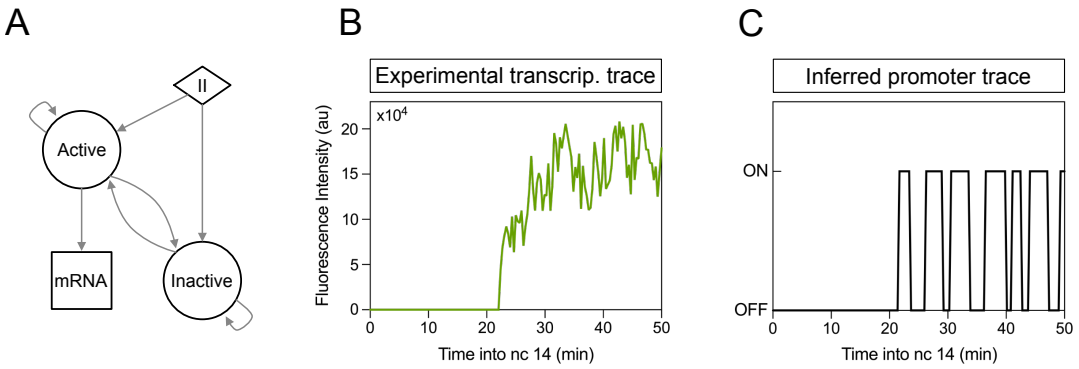

Figure 3.1: The model structure and basic principle behind *burstInfer*. **A**: Dynamic compound state hidden Markov model state diagram. At the beginning of the time sequence the promoter is in either the active or inactive state (π). Over the course of the time series the promoter switches stochastically between the active and inactive states according to the k_{on} and k_{off} burst parameters. While in the active state Pol II molecules are loaded onto the gene and mRNA transcripts are produced at a rate determined by the model emission parameter. **B**: Example MS2 fluorescence time series trace for a single nucleus in a *Drosophila* embryo showing nascent *ush* transcription. **C**: The promoter sequence inferred by the model corresponding to the fluorescent trace in B. These promoter traces can be used to generate single-cell parameters.

step, i.e. the time resolution of the data. Hidden promoter states falling outside the previous W time points can be assumed not to contribute to the recorded fluorescence at time point t, as Pol II initiated at that particular time point is no longer in transit along the gene.

To model this dependency of the observed fluorescence at time point t on the previous W hidden promoter states z_t, the concept of a compound state $s_t = \{z_t, z_{t-1}, \ldots, z_{t-W+1}\}$ is introduced. s_t, a $1 \times W$ vector, encodes the sequence of W hidden promoter states up to and including the current hidden promoter state at time point t. At each given time point the previous $W-1$ promoter states are deterministically passed to the new compound state, becoming the $1 \ldots W-1$ elements of the new compound state vector, with the W^{th} compound state at time point t being determined stochastically by the state transition matrix A. In the original cpHMM model, each compound state takes one of K^W different values, as each of W hidden promoter states may take one of K values (Lammers et al., 2020). This exponential scaling with window size W imposes a significant computational burden. How our model addresses this is detailed in the following section. As in the original cpHMM model, the emissions of the Hidden Markov Model are described by a Gaussian distribution with mean μ and standard deviation σ. The initial hidden promoter states at time $t = 0$ are given by a $1 \times K$ vector π. The joint

distribution of compound states and observed fluorescence values is given by:

$$p(y,s|\theta) = p(s_1|\pi)\prod_{t=1}^{T} p(y_t|s_t,\mu,\sigma)\prod_{t=2}^{T} p(s_t|s_{t-1},A) \tag{3.1}$$

Expectation Maximisation is used to infer the Hidden Markov Model parameters, $\hat{\theta} = \{\hat{\pi},\hat{\mu},\hat{A},\hat{\sigma}\}$. The use of an approximate inference technique renders inference of the model parameters computationally tractable. However, the exponential scaling of computation time with window size represents a significant problem for longer genes.

3.2.2 Dynamic State Space Truncation

In order to circumvent the exponential scaling of the algorithm with window size we propose a dynamic reduced state space variant of the cpHMM, which uses a truncated state space to avoid exponential scaling in computational time. We illustrate the advantages of this approach using a specific example implementation of the cpHMM model with $K = 2$ promoter states and a window size of 19, as would be required to model the *Drosophila melanogaster* gene *u-shaped (ush)*, which is 16825 base pairs in length (isoform C). Nascent transcription was captured at 20s time resolution. This results in a compound state which may take on $K^W = 2^{19} = 524288$ values. Repeated manipulation of the resulting $K^W \times t$ state matrix while performing expectation maximization requires a significant amount of computational time, which cannot be improved significantly by increasing available computational power.

The required computational time may be reduced by observing that although 524288 possible compound state values are required to fully specify the model, the majority of these compound states will have very low (often negligible) associated probability values, and can therefore be excluded from the model without impacting predictive performance. For example, during portions of the fluorescence signal recorded during the initiation of a transcriptional burst, compound states associated with inactive promoter states during the initial part of the compound state and active promoter states during the latter part of the compound state would be much more likely than compound states with sequences of promoter states associated with a very different observed fluorescence pattern, e.g. falling fluorescence levels or sustained inactivity.

Truncation in the model is enforced through the use of an allowed memory, M, with $M < K^w$. M is selected so as to reduce computational time without significantly impacting the performance of the algorithm. The use of M results in a reduced promoter state space, Φ_t, replacing s and reducing the scaling of the forward algorithm

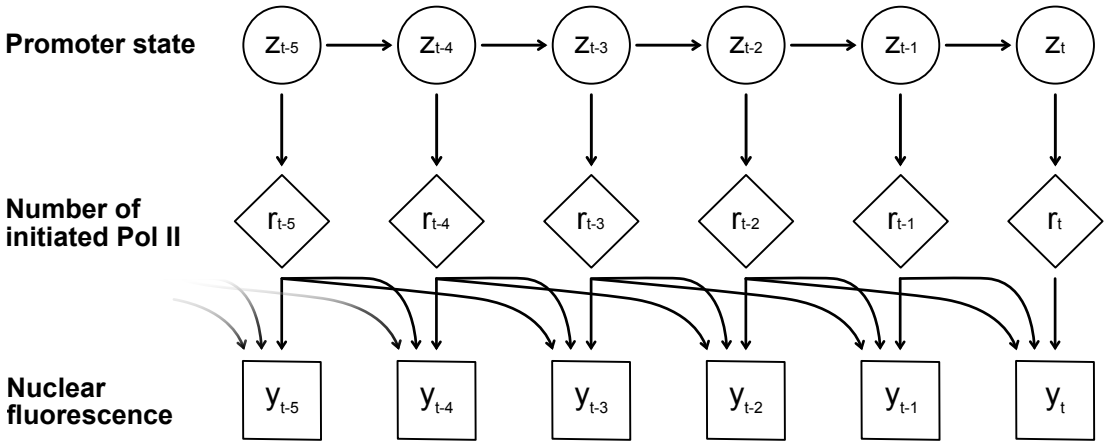

Figure 3.2: Diagram illustrating the dependence of the measured fluorescent signal at the present time, t, on both the present promoter state and previous promoter states falling within the observation time window, W. This time-dependence arises due to the persistence in the MS2 signal caused by Pol II still being in transit down the gene body following the promoter becoming inactive. The example shown here is for window size W=3.

with window size from exponential to linear scaling. To select a set of M likely compound states at time $t+1$ the forward algorithm is used to rank the $2M$ next possible states starting from M at time t. The forward algorithm computes the probability of the data up to the current time and being in each state, therefore the most likely states can be prioritized and the least likely are removed from the model until M distinct compound states remain. In practice, it is best to choose the maximum value of M that is computationally feasible, given the size of the dataset and the resources available. The state space will expand with $\Phi_t = 2^t$ until $\Phi_t = M$. Testing the model on synthetic data provides an indication of parameter estimation accuracy for given M and gene size.

An example of model truncation using a single trace of *ush* MS2 data is shown in Figure 3.3, with an allowed memory of 4 states specified for illustration purposes. Each box represents an individual state, with the leftmost number giving the binary representation of the promoter state (1 for on and 0 for off) and the rightmost number giving the log forward variable associated with each state. The state space expands during the forward algorithm until the allowed value of M is reached at $t = 1$ (for this particular example with a very small value of M). Forward variables are calculated for each allowed transition (the previous promoter state with either a 0 or 1 inserted at the rightmost bit) and are ranked. The least likely forward variables are eliminated (red outline), with the most likely states becoming the new reduced state space (blue outline). The process is repeated until the end of the trace (here $t = 3$).

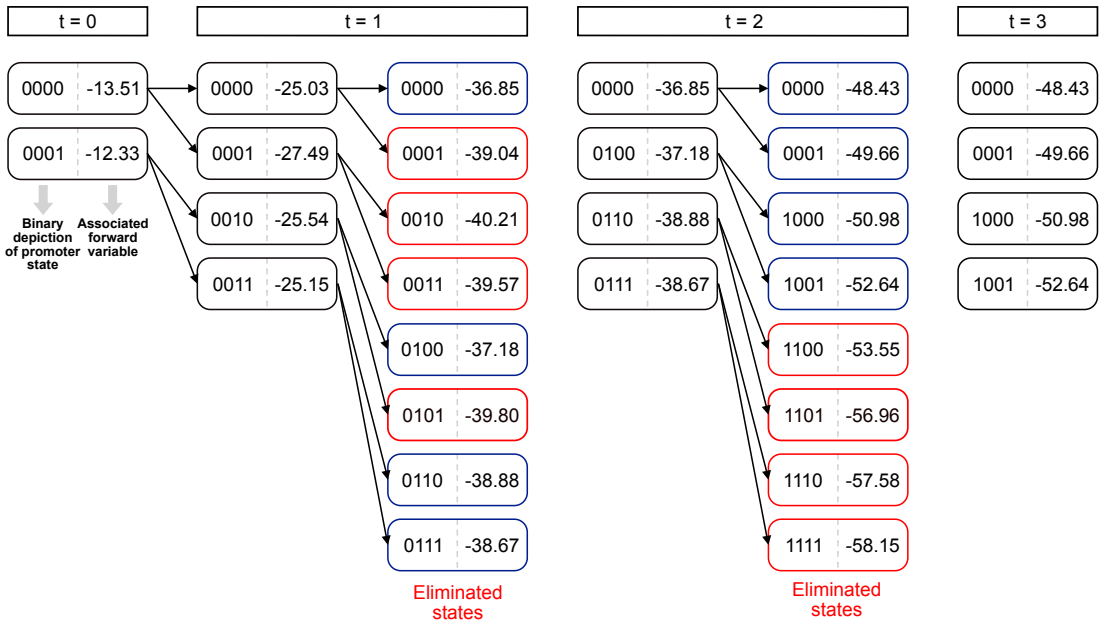

Figure 3.3: Example illustrating state-space truncation carried out as part of the HMM forward algorithm, using example data derived from the *Drosophila ush* gene. Each oblong bubble represents a compound promoter state at a particular time point with the number on the left representing the binary representation of the promoter state and the number on the right showing the log probability associated with each forward variable. The promoter starts at time $t = 0$ in either the inactive (0000) or active (0001) state (the rightmost bit indicates the current state). At time $t = 1$, the promoter can switch to either of two states from each of these two states, causing the state space to expand from 2 to 4 possible compound states (i.e. inactive to inactive, inactive to active, active to inactive, active to active). At time $t = 2$, the possible state space doubles again to 8 compound states. At this point, truncation is carried out — the compound states are ranked according to probability and the least likely states are eliminated. The number of eliminated/retained states is set to $M = 4$ here so that elimination can be visualised. In practice, the highest number of allowed states that is computationally feasible is used instead. This process of truncation and elimination is carried out until the end of each trace contained in the entire dataset. This truncated graph then becomes the state space for the entire model.

3.2.3 Inferring Single-cell Transcriptional Parameters

In addition to inferring 'global' model estimates for burst amplitude, frequency and duration for a given dataset, our model can be used to infer single-cell transcriptional parameters, i.e. burst parameters for each individual cell within the expression domain, rather than a global estimate for the entire expression domain or region of interest. Although single-cell parameter estimates are associated with high levels of uncertainty, they can provide a useful view of how bursting parameters vary across the spatial domain.

Training the model using the forward-backward algorithm yields estimates of $\alpha_t(i) = p(y_1,...,y_t,s_t = i|\hat{\theta}_k)$, the joint estimate of the observed fluorescence up to time t and the compound hidden promoter state at time t and $\beta_t(i) = p(y_{t+1},...,y_T|s_t = i,\hat{\theta}_k)$, the conditional probability of the observations from $(t+1)$ to the end of each trace, given the current hidden promoter state. Combining these variables with the expression for the likelihood of the observed fluorescence values given the model parameters, $p(y|\hat{\theta}_k)$, gives the following:

$$p(s_t|y,\hat{\theta}_k) = \frac{\alpha_t(s_t)\beta_t(s_t)}{p(y|\hat{\theta}_k)} \tag{3.2}$$

where $p(s_t|y,\hat{\theta}_k)$ denotes the probability of the promoter being in an active or inactive state at a given time point t, given the observed fluorescence and inferred model parameters. Taking the argmax of Equation (3.2) at each time point gives a sequence of the most likely promoter states at each observed time step. As previously mentioned, the *Drosophila* gene *ush* is used here as an example. MS2 stem-loops were inserted into the endogenous *ush* gene 5'UTR region, allowing us to visualise transcription in the form of nascent MCP-GFP fluorescence (Figure 3.1 B). The inferred promoter trace calculated using Equation (3.2) corresponding to this time series is shown in Figure 3.1 C.

In addition to providing a way of visualising the model fit, these inferred promoter traces may be used to calculate single-cell transcriptional parameters, so that in addition to giving single maximum likelihood parameters estimates for a given dataset, i.e. a k_{on}, k_{off} and emission term for the set of traces used to train the model, each cell in the expression domain is assigned each of these parameters.

The calculation of the transition parameters is achieved through a simple counting-based technique, where the number of normalised on-to-off and off-to-on transitions is counted from the inferred promoter traces. These counts are used to create transition

matrices for each trace, which are then converted to transition rates (in a similar way to the calculation of the global parameters). The single-cell emission term is a reduced form of the emission term from the global model (see Lammers et al., 2020):

$$\hat{\mathbf{v}} = \mathbf{M}^{-\mathbf{1}}\mathbf{b} \tag{3.3}$$

$$\mathbf{M}_{mn} = \sum_{h=1}^{N}\sum_{t=1}^{T_h}\sum_{i=1}^{K^w}\langle s_t^i(h)\rangle \mathbf{F}_{in}\mathbf{F}_{im} \tag{3.4}$$

$$\mathbf{b}_m = \sum_{h=1}^{N}\sum_{t=1}^{T_h}\sum_{i=1}^{K^w}\langle s_t^i(h)\rangle y_t(h)\mathbf{F}_{im} \tag{3.5}$$

where the $\langle s_t^i(h)\rangle$ term becomes a delta function due to the state probabilities already being known.

Parameter estimates for single cells are much more uncertain than global estimates for an entire dataset. The aim, however, is to be able to visualise broad spatial trends across the expression domain. The supplementary material gives an example of inferred single-cell parameters for the *ush* gene. LOESS smoothing was used to smooth the data, allowing general spatial changes in expression level to be shown - in this case, a more peaked distribution in the probability of the promoter becoming active than inactive. Calculating confidence intervals for a binomial proportion revealed that while there was high uncertainty associated with the parameter estimates, particularly towards the edges of the expression domain, the general spatial trend for the parameters could still be detected.

3.3 Results

3.3.1 Visualising Inferred Promoter Traces

An example of the model output is show in Figure 3.4. A Markov Chain was used to generate two synthetic datasets of promoter sequences. Each dataset consisted of 100 traces of 100 time points each. A Python script was then used to convert each promoter sequence into a corresponding fluorescent signal, by specifying the emission and noise parameters. Two different window sizes of 5 and 13 were selected, resulting in a ‘short’ gene dataset and a ‘long’ gene dataset. The model was then trained on these datasets. Subfigures A and B show example fluorescent traces (black) and inferred model fit

(red) for the short and long genes, respectively, along with the 95% confidence intervals (shaded red). Subfigures C and D show the 'true' promoter sequence (black) and inferred promoter sequence (red) corresponding to the signals above. Overall, the inferred signal corresponds well to the original signal, with some small errors in the inferred promoter sequence.

Further comparisons are shown in Figures 3.4 and 3.5. In Figure 3.4, two synthetic datasets were generated, using the same transcriptional parameters (emission, noise, transition parameters) but different window sizes ($W = 5$ and $W = 13$). After training the model on these datasets, the inferred most likely promoter sequences were compared to the synthetic promoter sequences used to generate the MS2 training dataset. The short and long gene are plotted on the left and right sides of the figure, respectively. In Figure 3.5, a similar comparison has been carried out between low and high noise conditions for the long gene. Although the performance of the algorithm is reduced for the very noisy dataset, inference of the promoter sequence is still possible.

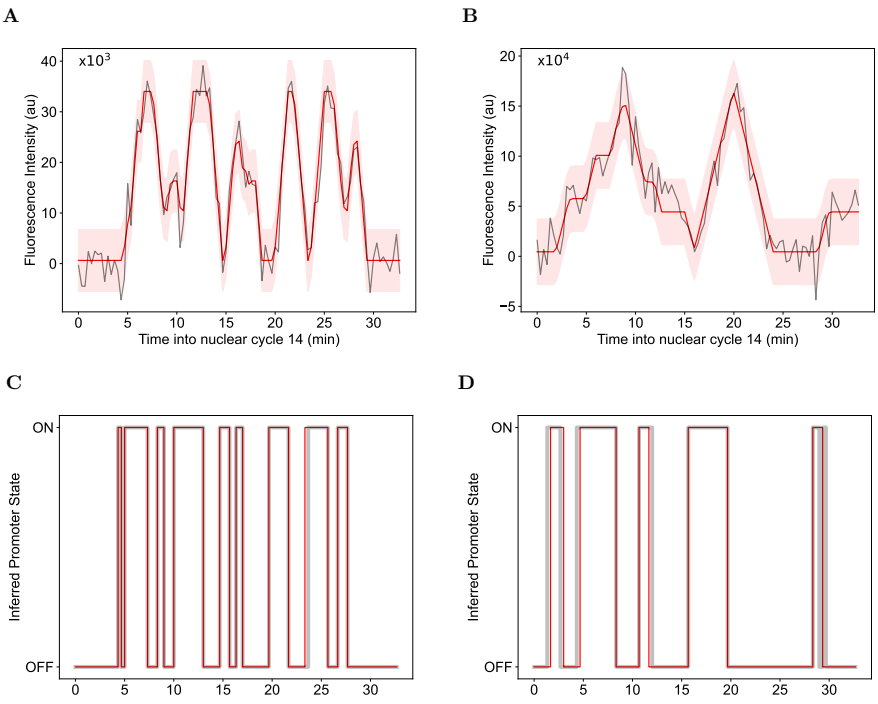

Figure 3.4: Visualising the model fit using synthetic MS2 data. **A**: Synthetic 'short' gene (Window Size 5) MS2 data generated using a Markov Chain (black) with the model fit overlaid in red. **B** Synthetic data and model fit for a 'long' gene (Window Size 13). **C**: Synthetic promoter sequence used to generate the 'long' gene data corresponding to the signal above. **D**: Synthetic promoter sequence for the 'short gene'. There is a small mismatch in the final inferred burst.

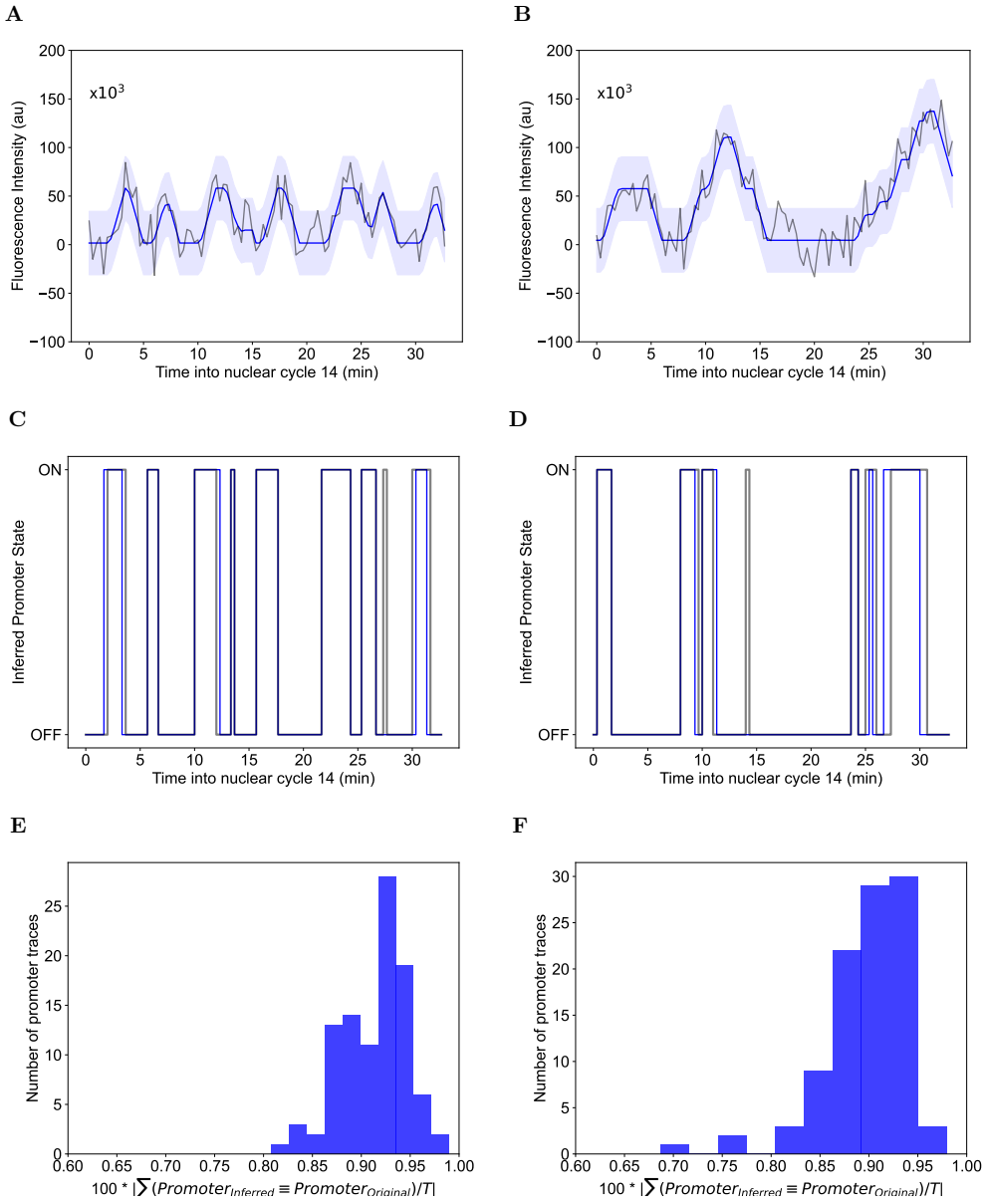

Figure 3.5: Visualising the model fit for synthetic genes with the same bursting parameters, but different window sizes. **A**: Plot of model fit (blue) and original synthetic 'low noise' MS2 data (black) for a synthetic gene with window size $W = 5$. **B** Fitted model for a dataset with the same bursting parameters as **A**, but with window size $W = 13$. **C**: Synthetic promoter trace used to generate fluorescence trace in **A** (black) and promoter sequence fitted by model (blue). **D**: Same as **C**, but corresponding to **B**. **E**: Agreement between promoter traces in training set and promoter traces inferred by model (Sum of number of times $Promoter_{Inferred} \equiv Promoter_{Original}/T$) for the short gene. **F**: Same as **E**, for the long gene.

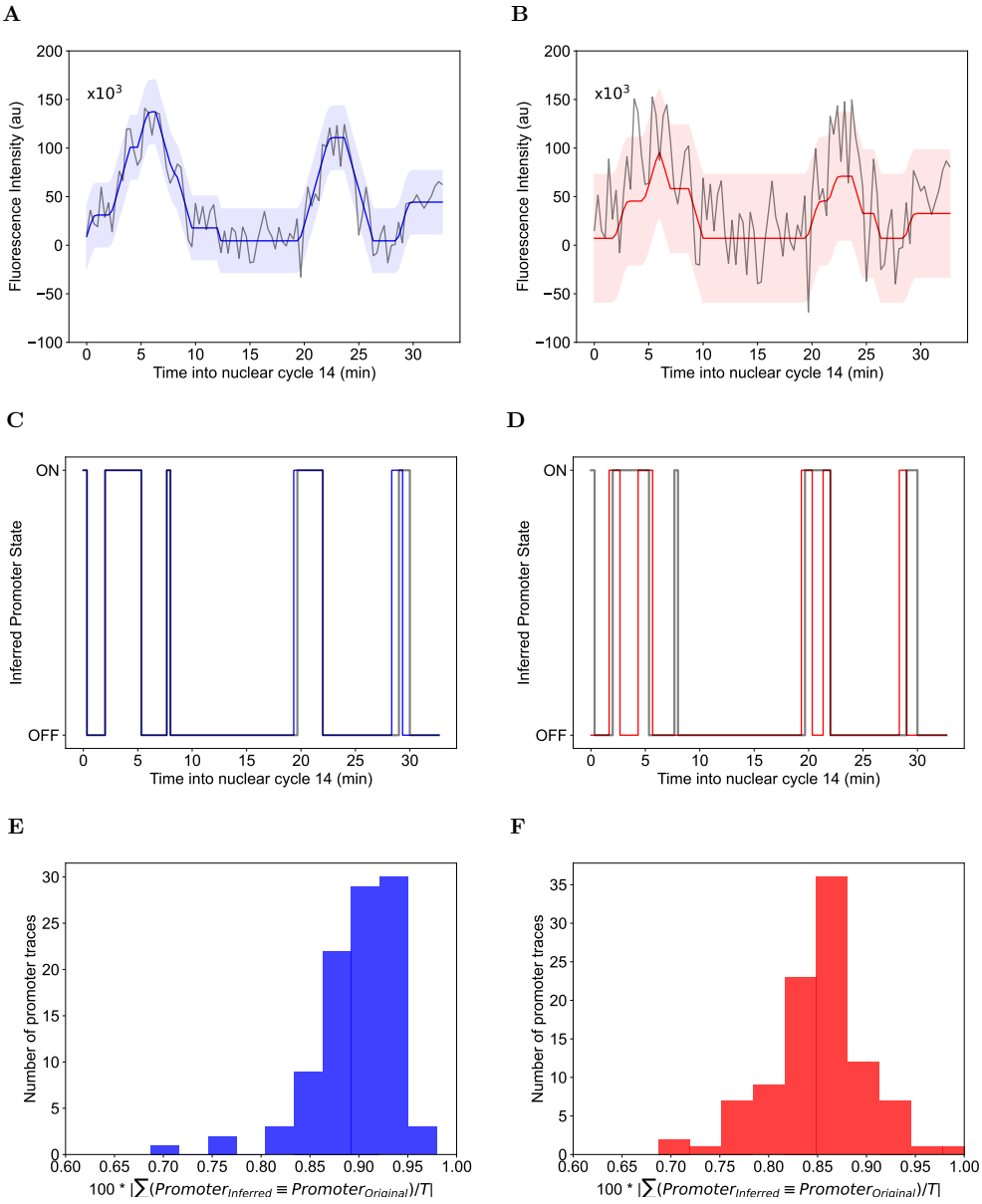

Figure 3.6: Visualising the model fit in low and high noise conditions for the long gene. **A**: Plot of model fit (blue) and original synthetic 'low noise' MS2 data (black). **B** Fitted model for a synthetic dataset identical to that in **A**, but with the noise increased. **C**: Synthetic promoter trace used to generate fluorescence trace in **A** (black) and promoter sequence fitted by model (blue). **D**: Same as **C**, but high noise condition. **E**: Agreement between promoter traces in training set and promoter traces inferred by model (Sum of number of times $Promoter_{Inferred} \equiv Promoter_{Original}/T$) for low noise condition. **F**: Same as **E**, for high noise condition.

3.3.2 Assessing the model fit

To demonstrate the ability of the truncated model to approximate the results obtained using the full hidden Markov model, we created synthetic fluorescent traces for a gene of window size 11 and tested convergence between the truncated and full models for this dataset (Figure 3.7 A). Fifty different initialisations of the model were created using random HMM parameters, selecting the EM run with the highest likelihood as the most likely model. The relative error between the full and truncated models falls smoothly as the state space of allowed states is increased, with the relative error falling to less than 1% at $M = 128$ where the size of the full model here would be $2^{11} = 2048$ compound states.

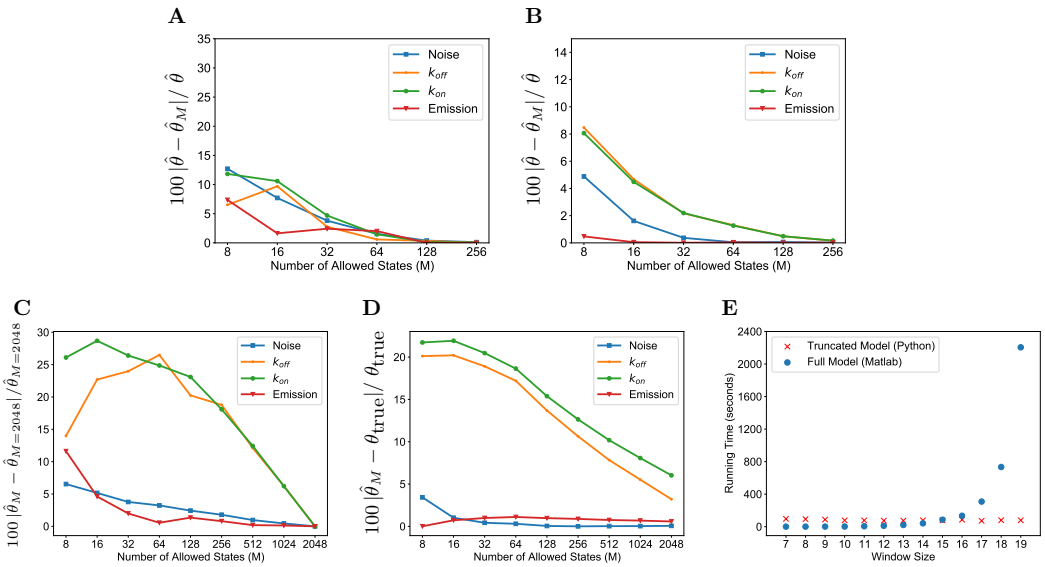

Figure 3.7: Assessing the model fit and running time on real and synthetic datasets. **A**: Relative difference between the maximum likelihood parameter estimates for the truncated $\hat{\theta}_M$ and full model $\hat{\theta}$ as a function of increasing M for data from the *Drosophila* gene *hnt*. This gene is short and only requires 512 compound states in the full model. **B**: Relative difference between the model parameters for the truncated and full model as a function of increasing M for synthetic data with window size 11. In this case $2^{11} = 2048$ compound states are necessary for the full model. **C**: Relative change in model parameters for the *Drosophila ush* gene as M is increased, compared to the value for $M = 2048$. For this gene $2^{19} = 524288$ compound states are required to specify the full model. **D**: Relative error in the model parameters between the truncated model and true model as a function of increasing M for synthetic data with window size 20. In this case $2^{20} = 1048576$ compound states would be necessary for the full model. **E**: Running time for a single EM step for both models.

To test the model on experimental MS2-MCP data where is it possible to fit the full model, both the full and truncated models were trained on a dataset of MS2 fluorescent traces for the *Drosophila melanogaster* gene *hindsight (hnt)*. The *hnt* gene has length of 7441 base pairs between the MS2 probe and the end of the gene body, in conjunction with an MS2 cassette length of 1290 base pairs, a window size of 9 was specified. The results of training the model using both the full and truncated models can be seen in Figure 3.7 B, a plot of relative error between the truncated and full ('true') model parameters as a function of increasing number of allowed states. Each curve represents a separate parameter of the model. The model was trained specifying 50 separate runs of expectation-maximisation for each value of M. The convergence of the truncated model parameters to the full model parameters is apparent from the diagram.

We then applied the model to datasets of MS2 traces recorded from longer genes, using both synthetic and real data. The change in model parameters as M is increased, compared to the parameter values at $M = 2048$, can be seen in 3.7 C for the *Drosophila* gene *ush*. Although 524288 compound states would be required for the full model, the parameters are converging with a much smaller subset of allowed states. The relative error between the inferred and true parameters as M is increased for a synthetic gene with window size 20 are shown in 3.7 D. Although 1048576 compound states would be used for the full model, with a subset of 2048 states the relative errors for the noise, k_{off}, k_{on} and emission parameters are 0.064%, 3.208%, 6.029% and 0.579%, respectively, showing that accurate parameter inference is still possible using the reduced model.

3.3.3 Computation Time

Next, we compared the scaling of computational time for a single step of the expectation maximisation algorithm for the truncated model and the full, original model (Matlab implementation). The dataset used in the comparison is a set of 50 MS2 fluorescence traces of the *ush* gene in a *Drosophila* embryo, where active transcription occurs during a 30 minute time window. A window size of 19 is required to model the fluorescence traces. The curve plotted in blue shows the result of increasing the window size upon the computational time required for a single expectation-maximisation step for the full model; the exponential scaling of the algorithm with window size is apparent. The computational time for the truncated model (red, M = 128 compound states, 90s per step) is essentially de-coupled from window size / gene length, allowing for application of the truncated model to a much wider set of window sizes (Figure 3.7 E). For short genes, the original version model is faster due to less computational

overhead associated with truncation, e.g. calculating and eliminating least likely states etc. The benefits of the truncated version of the model become apparent at longer gene lengths, where exponentially increasing computation time makes inference impractical. A window size of 30+ may be needed for both much longer *Drosophila* genes and vertebrate genes, making use of the full model infeasible.

3.3.4 Analysing EM Parameter Convergence

In order to validate the use of EM to train the algorithm, we have plotted the results of training the truncated model on synthetic data using 50 random restarts in Figure 3.8. Each panel shows a plot of the log likelihood against the relative error between the 'true' and inferred values. The same 'long gene' (W = 13) dataset from Figure 3.6 was used to train the model. The maximum likelihood parameter is highlighted in red. For $p_{off} \rightarrow p_{on}$ and the noise parameter, the maximum likelihood solution is clearly separated from low likelihood, high relative error results in the bottom right of the subfigures. For the emission parameter and $p_{on} \rightarrow p_{off}$, however, some results had relatively high likelihood despite also having high relative error. These results, however, were not chosen as the ML solution - for both parameters, there is a large cluster of overlapping datapoints in the top left that are near the ML solution, so the global optimum was still found. These results indicate the need to do a large number of random restarts in order to find the global optimum, as local optima appear to be a problem.

3.3.5 Estimating single-cell parameters

The *burstinfer* toolbox provides the ability to infer single-cell transcriptional parameters, as opposed to only a single set of parameters for an entire dataset or large subsets of cells. A model is initially trained using the whole dataset or subsets of the data (e.g. spatial domains). The learned parameters are then used to infer the most likely sequence of promoter states for each cell, i.e. the sequence of 1's and 0's that generated the observed data. Inferring single-cell parameters involves high levels of uncertainty, since very little data is available to estimate the transition frequencies for one cell. However, the aim is to help visualise the spatial trend at the single-cell level that may not be apparent from looking at a single spatial region or a small number of spatial regions side-by-side. This can be achieved by spatial smoothing of the single-cell estimates to infer the mean parameter change trend.

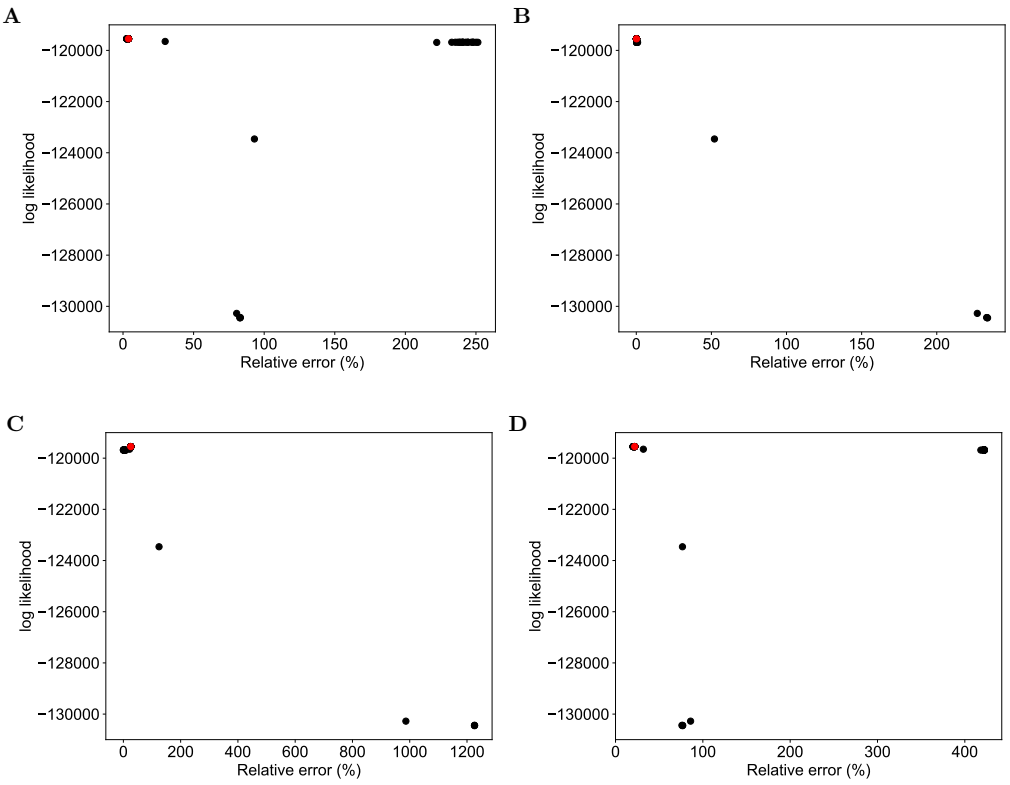

Figure 3.8: Plots of inferred EM parameters for the 'high noise' synthetic dataset from Figure 3.6. The log likelihood is plotted against relative error between the true and inferred parameters. Fifty random EM restarts were used. The Maximum likelihood parameter is highlighted in red. **A**: Relative error for the emission parameter. The top-left corner of the plot contains many overlapping points with similar log likelihood / relative error. **B**: Relative error for the noise parameter. **C**: Relative error for $p_{off} \rightarrow p_{on}$. **D**: Relative error for $p_{on} \rightarrow p_{off}$.

To give an example of how the single-cell parameters can be used to visualise spatial trends, Figure 3.9 shows the result of smoothing single-cell parameters for the *Drosophila* gene *ush*. Three separate models were trained for the outer, intermediate and central spatial regions of the embryo. These were then used to generate single-cell parameters. The left and right panels show $p_{\text{off}\to\text{on}}$ and $p_{\text{on}\to\text{off}}$, respectively: the probability of off-to-on and on-to-off transitions across the entire trace for each cell. The *loess* function from the Python *scikit-misc* library was used to smooth the data, as shown by the red curves. While there are large variations associated with the single-cell estimates, it is possible to visualise general spatial trends in the data, such as the decrease in the probability of promoter activation in cells further from the embryo midline as well as the flatter distribution of $p_{\text{on}\to\text{off}}$ in the central region.

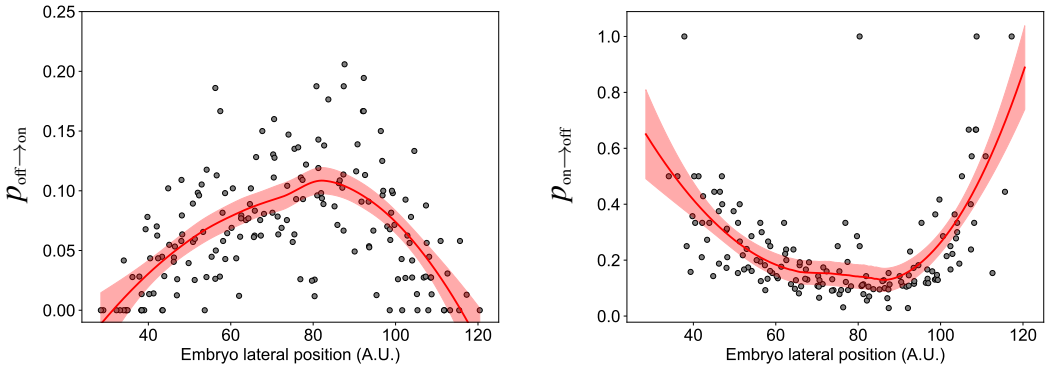

Figure 3.9: LOESS fit to inferred single-cell transition probability parameters for the *ush* gene against embryo lateral position.

In order to quantify the uncertainty associated with the parameter estimate for each cell, the confidence interval for a binomial proportion was calculated for each cell using the *proportion_confint* function from the Python *statsmodels* library. Figure 3.10 shows these error bars plotted on the same single cell data as the previous figure. While there is very large uncertainty associated with these estimates, the general spatial trend in the estimates can still be detected.

3.4 The *burstInfer* Software Package

burstInfer has been implemented in Python and is available on GitHub at `https://github.com/ManchesterBioinference/burstInfer`. While the original cpHMM model was written in Matlab, the decision was taken to use Python for the truncated

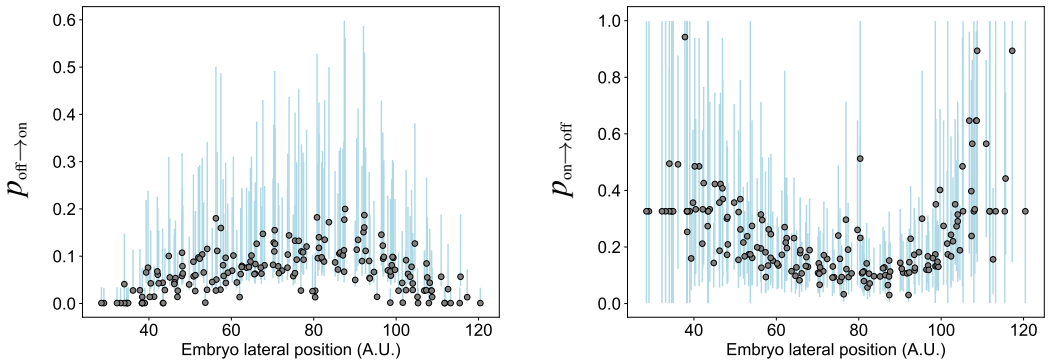

Figure 3.10: 95% confidence intervals for inferred single-cell transition probability parameters for the *ush* gene against embryo lateral position.

model due to the open source and non-proprietary philosophy of Python. The software package on GitHub includes examples using both synthetic MS2 data and data from Hoppe et al. (2020).

Figure 3.11 shows the basic structure of the software package. The main data folder (left) contains the MS2 data file, the main script and scripts containing auxiliary functions. Functions within the main.py and the auxiliary scripts call functions from the *burstInfer* library (right). main.py takes the csv file of raw MS2 data as input. process_raw_data.py is used to reshape the raw MS2 data into a usable format. Training is carried out via creation of a HMM object. The HMM class contains functions to randomly initialise parameters, run expectation maximisation and infer the single cell promoter traces. Several variants of expectation maximisation are included.

initialise_parameters provides initial estimates for the HMM parameters, based upon the raw MS2 data. The methods chosen are based upon those used by Lammers et al. for the cpHMM. The initial value for the emission parameters depends upon the maximum fluorescence value in the raw MS2 data, multiplied by a number drawn from a uniform distribution, whereas the noise parameter depends upon the mean fluorescence. The initial state estimate and transition parameters are drawn from a uniform distribution.

Following this initialisation, several options are available for training. Running *em_fixed* uses a variant of EM that fixes the transition parameters but allows the emission and noise parameters to vary, as is used in the original cpHMM implementation. This gives a more accurate estimate of the noise and emission parameters before running the full model (*em_with_priors*). This estimation step can be skipped by simply

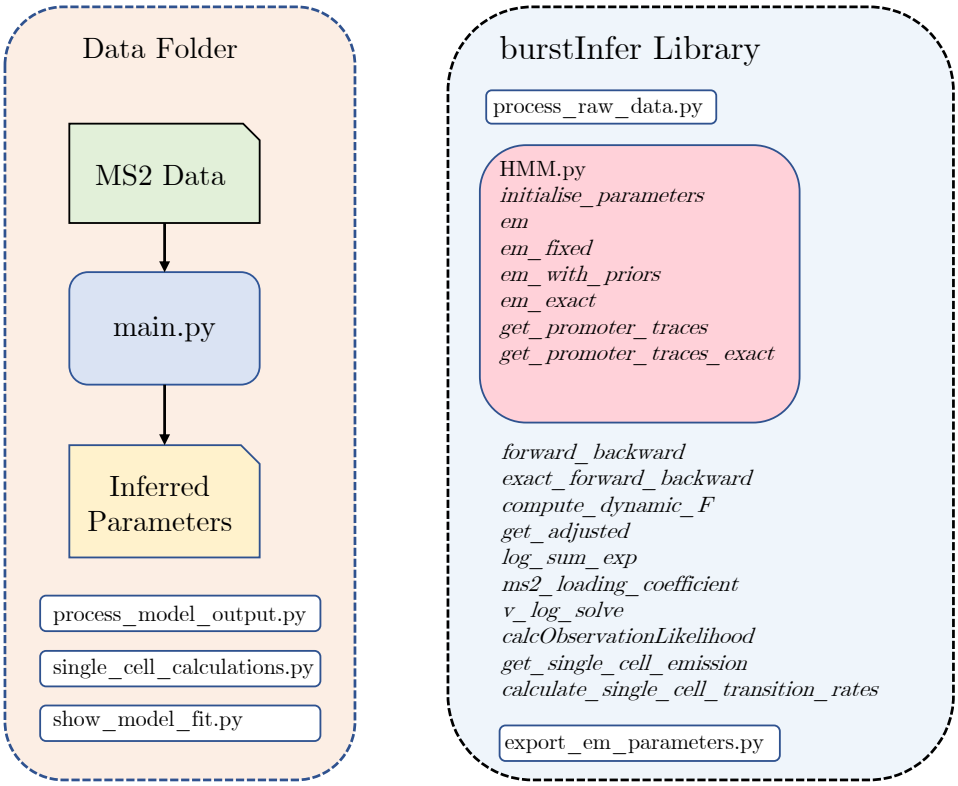

Figure 3.11: Structure of the *burstInfer* software package. Data files, the main script and data processing / visualisation scripts are kept together inside a data folder. Running main.py uses functions and class definitions contained inside the *burstInfer* library to create a HMM object, which infers and returns the kinetic parameters.

running *em* instead. *em_exact* runs a Python version of the non-truncated original model. Each of these functions uses external library functions containing the forward-backward algorithm (*forward_backward, exact_forward_backward*) along with a range of different helper functions designed to be computationally efficient (*log_sum_exp, v_log_solve*) or implement the observation model (*compute_dynamic_F, get_adjusted, ms2_loading_coefficient, calcObservationLikelihood)*. After the threshold for the minimum change in parameters with each EM step is met, the parameters are returned to main.py and exported as a csv by export_em_parameters.py.

Typically, the package is run on a computing cluster so as to allow for multiple instances of EM with random restarts to be run simultaneously. Once the results are ready for analysis, process_model_output.py is used to find the maximum likelihood parameters. Once these have been calculated, single_cell_calculations.py, which uses *get_single_cell_emission* and *calculate_single_cell_transition_rates*, is run to infer the single cell transcriptional parameters. The model fit can then be visualised using show_model_fit.py.

A key novel feature of our dynamic version of the algorithm is the use of binary encoding to store and calculate the promoter state. In the original model, the promoter state is represented as a Matlab array of '0's and '1's, indicating the position of promoter inactive and active states in the model sliding window. We chose to implement the promoter state as a binary number instead, allowing for the use of computationally efficient bitwise operations to calculate the next promoter state following a transition.

3.5 Conclusion

In this chapter we have presented the details of the theoretical background and software implementation of our scalable algorithm. The original cpHMM algorithm scales poorly due to the exponential relationship between gene length and the number of compound states necessary to model the data. Our implementation uses a form of dynamic state space truncation, whereby the model state space is allowed to expand until it reaches a pre-set allowed memory size, M. Only the most likely M compound states are retained at each time step. In this way the exponential scaling problem is avoided (Figure 3.3). We have shown, using both synthetic and experimental *Drosophila* data, that the truncated model provides a good approximation to the full model (Figure 3.7). In addition to inferring kinetic parameters for a given region of an embryo, the algorithm is able to infer single-cell transcriptional parameters, as shown in Figure 3.9.

The ability to visualise and statistically analyse these single-cell parameters can provide additional insight into the spatial regulation of bursting dynamics for a given gene.

Chapter 4

Inferring BMP Signalling Dynamics in *Drosophila*

The early *Drosophila* embryo represents an ideal system for gathering imaging data. In this chapter we outline the application of *burstInfer* to MS2 data derived from developing *Drosophila* embryos. The results from modelling dorsal-ventral patterning during Nuclear Cycle 14 have been published in Developmental Cell (Hoppe et al., 2020). The ability of the algorithm to infer single-cell transcriptional parameters was of key importance in this paper, allowing for inference of the key parameters regulating the Dorsal-Ventral system. The algorithm was also successfully used to model transcription in mutant *Drosophila* embryos.

4.1 Introduction

An example of using the model to infer single-cell parameters is shown in Figure 4.1, using example data from Hoppe et al. (2020) (different embryo to that highlighted in the original paper). The aim of the paper was to use the parameters inferred by *burstInfer* to investigate regulatory control of Bone Morphogenetic Protein (BMP) target genes in the early *Drosophila* embryo, focussing on dorsal-ventral patterning of the dorsal ectoderm and amnioserosa. MS2 imaging was used to generate movies of transcriptional activity of one of the BMP target genes studied in the paper, *ush*, during nuclear cycle 14. The expression domain of *ush* forms a broad stripe down the anterior-posterior axis on the dorsal side of the embryo (Ashe et al., 2000), which mirrors the expression levels of the BMP Decapentaplegic (Dpp) (Figure 4.1 A) (Bier and De Robertis, 2015; Deignan et al., 2016; Eldar et al., 2002; Umulis et al., 2010). Cells

falling within the Dpp gradient express Dpp target genes in a concentration-dependent manner - intermediate levels of signalling are sufficient to activate *ush*, for example.

To investigate spatial regulation of Dpp target genes, MS2 movies were recorded in the embryo during nuclear cycle 14. Each embryo was divided into three separate regions corresponding to different signalling levels, determined by either distance from the midline or through the use of a clustering-based approach. *burstInfer* was then trained on each of these three regions, giving estimates of k_{on}, k_{off} and Pol II loading rate (emission) for each section of the embryo. These regional parameters were then used to infer single-cell parameters (Figure 4.1 B) and promoter traces (Figure 4.1 C) for each cell within the expression domain (see Section 4.2.1 for further details). Figure 4.1 B shows heatmaps of mean expression and three example single-cell parameters for *ush* - the region shown here represents a subset of the expression domain shown in the cartoon in Figure 4.1 A. Mean expression corresponds to the mean recorded fluorescence for each cell, with the arbitrary fluorescence signals converted into number of Pol II. The single cell occupancy, k_{on} and k_{off} parameters were calculated using *burstInfer*. From these heatmaps the strong similarity between mean expression and occupancy is immediately apparent, along with the slightly weaker similarity between expression levels and k_{on} (Figure 4.1 B). In order to quantify the dependency of expression levels on each of these three 'kinetic' parameters (along with derived bursting parameters, such as burst duration and frequency), correlation analysis was carried out on the single-cell expression data and inferred parameters. The bursting parameters derived from the model kinetic parameters (k_{on}, k_{off} and the Pol II loading rate, k_{ini}) are shown below in Table 4.1.

Parameter Definitions	
Promoter switching on rate	k_{on}
Promoter switching off rate	k_{off}
Pol II loading rate	k_{ini}
Burst frequency	$\frac{k_{\text{on}}k_{\text{off}}}{k_{\text{on}}+k_{\text{off}}}$
Burst size	$k_{\text{ini}} \cdot \frac{1}{k_{\text{off}}}$
Burst duration	$\frac{1}{k_{\text{off}}}$
Promoter off period	$\frac{1}{k_{\text{on}}}$
Promoter occupancy	$\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}$

Table 4.1: Table of bursting parameter definitions. Adapted from (Zoller et al., 2018).

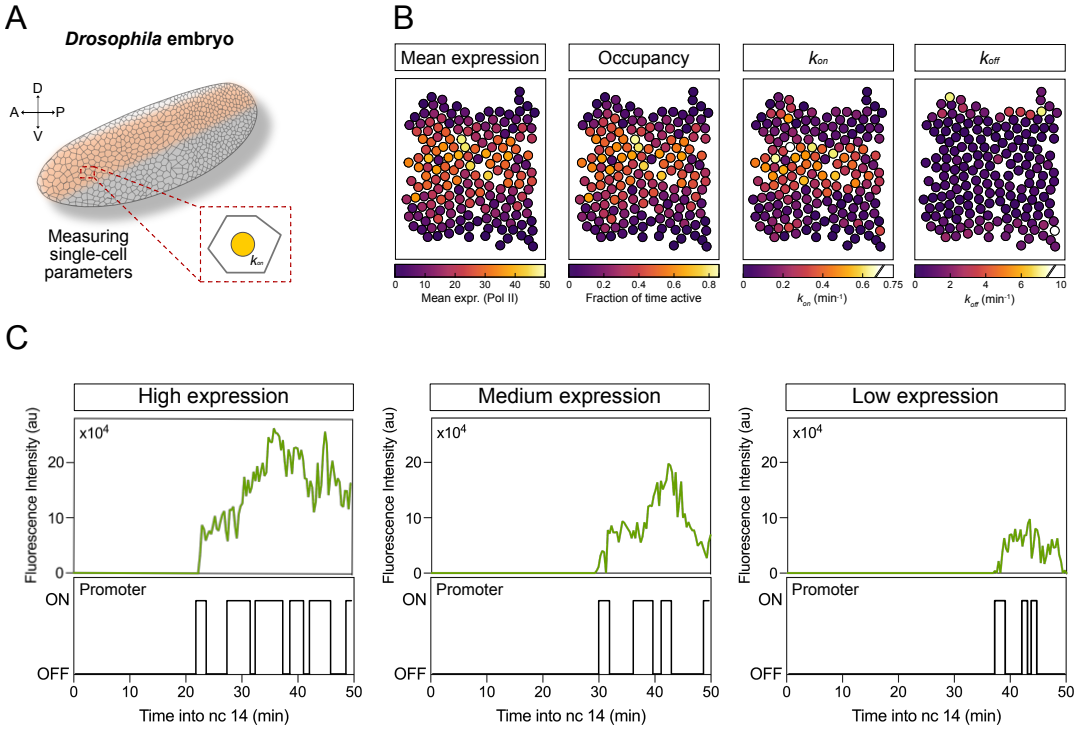

Figure 4.1: Example inferred single-cell parameter using *Drosophila ush* data from Hoppe et al. (2020). **A**: The expression domain of the *ush* gene shown in the cartoon was divided into three separate regions, corresponding to high, medium and low levels of expression, with the model trained separately on each of these three regions. The inferred global parameters for each region were used to infer the most likely promoter path corresponding to each fluorescent trace. **B**: Heatmaps of the measured mean expression level, along with the k_{on}, k_{off} and occupancy ($\frac{k_{\text{on}}}{k_{\text{on}}+k_{\text{off}}}$) parameters for each cell are shown. Analysis of single-cell parameters in this case revealed k_{on} as the main determinant of expression level. **C**: Example fluorescent traces and corresponding inferred promoter paths for each of the three regions. See Hoppe et al. (2020) and Bowles et al. (2022) for further details.

This analysis revealed a very strong correlation between expression levels and occupancy, with effectively no correlation between expression and k_{off} (Hoppe et al., 2020). Pol II loading rate (the HMM emission parameter) was flat across the expression domain Figure (4.1 B). As occupancy depends upon both k_{on} (which did exhibit strong correlation) and k_{off}, the results indicated that expression levels were regulated through modulation of k_{on}, the promoter activation rate. Representative single cell fluorescence and promoter traces for each region show that nuclei experiencing high signalling produce more transcriptional bursts compared to other regions (Figure 4.1 C). The single-cell parameters extracted from quantification of traces like these were used to create the heatmaps shown in Figure 4.1 B. Code to re-create these figures is included in the *burstInfer* GitHub repository.

4.2 Modelling Results

4.2.1 Inferring Global Transcriptional Parameters

The following subsection gives further details of the application of the *burstInfer* algorithm to *Drosophila* embryonic development data, as outlined in Hoppe et al. (2020). Figures have been reproduced from Hoppe et al. (2020) where appropriate. *burstInfer* was used in this paper to investigate the relationship between BMP signalling and mRNA levels in the early *Drosophila* embryo. As stated in the introduction, Dorsal-Ventral patterning in the early *Drosophila* embryo is determined by a member of the BMP family known as decapentaplegic, or Dpp. Graded levels of BMP signalling act to partition the embryo into different tissue subtypes - cells respond to varying levels of BMP signalling by producing varying levels of mRNA.

In Hoppe et al. (2020), MS2 live imaging was used to visualise transcription in *Drosophila* embryos during Nuclear cycle 14. CRISPR gene editing allowed for visualisation of two Dpp target genes, *u-shaped* (*ush*) and *hindsight* (*hnt*). The datasets generated from visualisation of transcription of these genes was used to train the *burstInfer* algorithm, with the aim of understanding which kinetic parameters were responsible for regulation of bursting of BMP target genes, i.e. whether burst frequency, duration or amplitude were responsible for modulation of target gene bursting dynamics.

The results of training the model using *ush* and *hnt* data can be seen in Figure 4.2. Embryos were divided into three (*ush*) and two (*hnt*) regions corresponding to different

levels of BMP signalling. K-means clustering was used to divide the *ush* data, whereas the *hnt* embryos were divided based upon distance from the embryo midline. The model was then trained on each of these regions, giving 'global' parameter estimates for each region corresponding to k_{on}, k_{off} and the emission parameter. Figure 4.2 A and B show example traces from the central high signalling region for both *ush* and *hnt*. Inferred promoter traces (calculated by taking the argmax of $p(s_t|y,\hat{\theta}_k)$, the probability of the promoter being in an active or inactive state at time point t on a trace-by-trace basis) can be seen in the lower panel. The inferred global parameters for the two genotypes are shown in Figure 4.2 C and D (three replicates per genotype). Carrying out a one-way ANOVA revealed which parameters were varying significantly between high and low signalling regions.

For both genotypes, k_{on} was found to be the main regulated parameter, contributing to significant differences in the promoter occupancy ($\frac{k_{on}}{k_{on}+k_{off}}$) between high and low signalling regions. Differences in loading rate (emission) and k_{off} were not found to be statistically significant. These results indicated that burst frequency, not burst amplitude or duration, was the key regulated parameter. Although only three biological replicates were used, these results seem to indicate a strong decrease in k_{on} between the Medium and Low signalling regions. Figure 4.2 E gives a representation of simulated bursts using the inferred parameters. The two genes respond to high BMP signalling levels at the midline in different ways: *ush* is transcribed in less frequency, low amplitude, longer duration bursts, whereas *hnt* is transcribed in shorter, high amplitude, high frequency bursts.

4.2.2 Inferring Single Cell Transcriptional Parameters

The inferred global parameters were then used to produce single-cell transcriptional parameters. The single-cell promoter traces corresponding to $p(s_t|y,\hat{\theta}_k)$, as shown in Figure 4.2 A & B, were used to calculate single-cell values of k_{on}, k_{off} by counting the normalised number of $off \rightarrow on$ transitions in each inferred promoter sequence. These counts were then converted to a rate, giving values for k_{on} and k_{off} for each cell in the expression domain. Values for the single-cell occupancy and burst frequency were calculated from these parameters. The single-cell loading rate was calculated by applying the emission section of the main *burstInfer* algorithm on a per-cell basis.

Figure 4.3 details the results of the single cell analysis for *ush* and *hnt*. Each cell in the expression domain (Figure 4.3 A) is allocated an inferred single-cell parameters, as shown by the heatmaps in Figure 4.3 B. The correspondence between mean expression

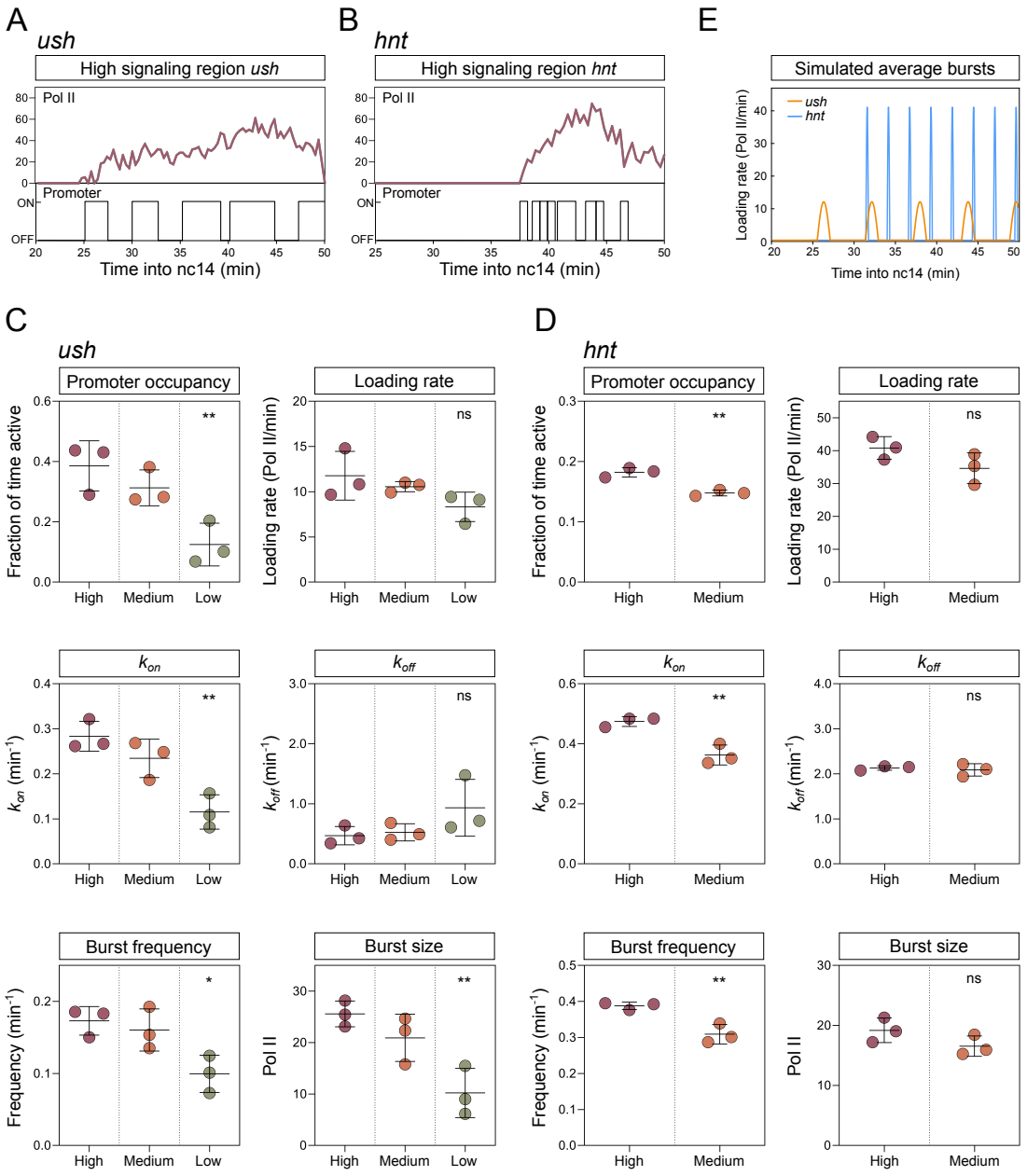

Figure 4.2: Inferred promoter traces and global bursting parameters using *burstInfer*. **A**: Example MS2 fluorescence trace from the high signalling region in an *ush* embryo at the embryo midline, along with inferred promoter sequence for the same cell. **A**: equivalent MS2 trace and inferred promoter sequence for *hnt*. **C**: Inferred global parameters for *ush*. Three biological replicates are shown. Whisker bars show mean $\pm$ sd. $* = p < 0.05$, $** = p < 0.01$, *ns* = not significant. A one-way ANOVA with a Dunnett's multiple comparisons test was carried out to test for significant differences relative to the high signalling region. **D**: Equivalent global parameters for *hnt*. Student's t test was used to test for significant changes in signalling relative to the midline. **E**: Simulated transcriptional bursts using inferred mean global parameters for *ush* and *hnt*. Adapted from Hoppe et al. (2020)

and occupancy can be seen in the heatmaps, as can the flatter distribution of the k_{off} parameters. In order to quantify the relationship between the single cell parameters the observed expression data, a correlation analysis was carried out between each of the parameters and the observed fluorescence, normalised to the mean number of Pol II engaged on the gene for *ush* (Figure 4.3 C).

This analysis revealed a strong correlation between k_{on} and mean number of Pol II engaged ($r = 0.84$) and a very strong correlation between the occupancy and number of Pol II ($r = 0.99$). The loading rate ($r = 0.07$) and k_{off} ($r = -0.62$) were found to be poorly correlated. This strong correlation between k_{on} and mean number of Pol II was found to also occur with *hnt* (Figure 4.3 D and E). These results further indicated that BMP signalling was responsible for regulation of target gene bursting through modulation of k_{on}.

4.2.3 Applying the Algorithm to Mutant Embryos

Experiments were then carried out to investigate the effect of additional BMP signalling on bursting (Figure 4.4). Ectopic signalling was introduced at the embryo midline through the insertion of a single copy of the st2-*dpp* transgene, which acts to misexpress *dpp* (Ashe et al., 2000). The inclusion of the transgene leads to a broader expression domain than *ush* wild type embryos (Figure 4.4 A). Compared to wild type embryos, st2-*dpp* embryos show an earlier onset time of transcription (Figure 4.4 B) and a higher number of Pol II on the gene body, although the time of peak transcription is similar to *ush* wild type embryos (Figure 4.4 C).

After dividing the embryos into four separate spatial regions based on distance from the midline (three regions equivalent to those used in wild type embryos, and an additional one due to the broader expression domain), *burstInfer* was used to infer global and single cell transcriptional parameters. Single cell parameters were separated into bins of one cell width, moving out from the dorsal midline (Figure 4.4 D). Plotting the data in single cell width bins indicated that k_{on} at the dorsal midline (cell widths 1 - 3) was not significantly altered by the addition of the ectopic Dpp. In rows 4 - 10, however, k_{on} was significantly increased relative to the wild type embryos, along with an accompanying increased in occupancy and Burst Frequency. k_{off} and Loading rate were not significantly altered, however.

These results indicate that k_{on} is close to saturation at the embryo midline, in agreement with previous research (Dorfman and Shilo, 2001; Mizutani et al., 2005), and that loading rate and k_{off} are not sensitive to changes in changes in Dpp levels, consistent

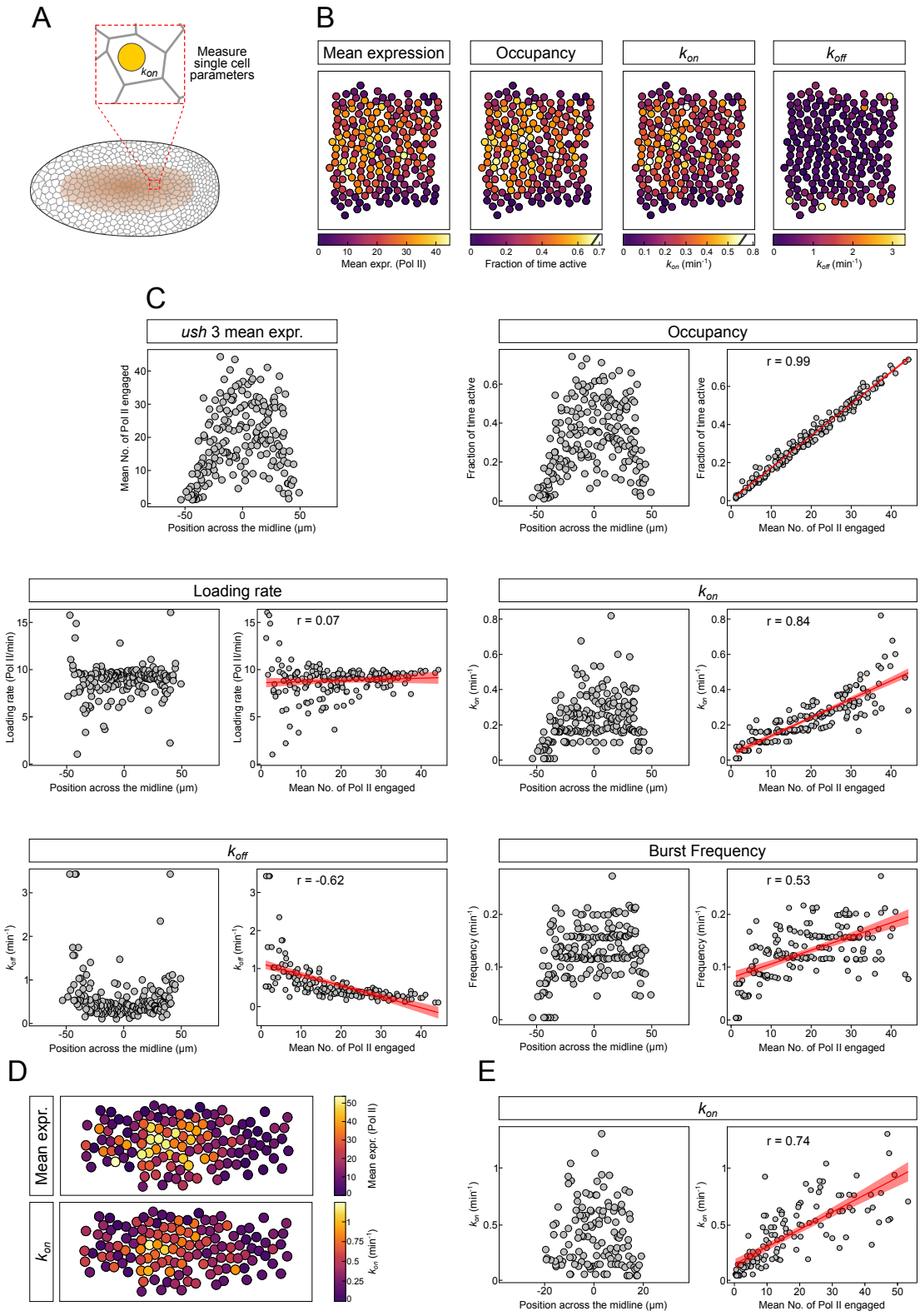

Figure 4.3: Visualisation and analysis of single-cell parameters inferred using *burstInfer*. **A**: Example cartoon of a *Drosophila* embryo with the *ush* expression domain overlaid in orange. Single-cell analysis allows for transcriptional parameters to be assigned to each cell in the expression domain, rather than on a 'global', or regional, basis. **B**: Heatmaps of the *ush* expression domain showing inferred single-cell parameters. Colour 'temperature' corresponds to single cell Mean expression, occupancy, k_{on} and k_{off}. **C**: Plots of single-cell parameters as a function of distance from the midline (left) and Mean number of Pol II engaged (right) for a representative *ush* embryo. The red line and shared region shows the results of a linear regression with $\pm$ 95% confidence intervals. Pearson correlation coefficient is also shown. **D**: Equivalent heatmaps from **B** for *hnt* Mean expression and k_{on}. **E**: Single-cell *hnt* k_{on} for a single representative embryo. Adapted from Hoppe et al. (2020).

with the findings shown above. smFISH was also used to measure mRNA levels in wild type and st2-*dpp* embryos (Figure 4.4 E), revealing that while peak numbers of mRNA at the dorsal midline were similar between st2-*dpp* and wild type embryos, a drop in mRNA levels in wild type embryos further away from the midline were observed. These results provide further evidence that target gene mRNA levels are determined through decoding of Dpp signalling levels via changes in k_{on}.

4.2.4 Model Verification

In order to verify the model, the distribution of 'on' and 'off' times extracted from inferred promoter traces (waiting times) was plotted for *ush* (Figure 4.5 D). As expected, the distribution of waiting times corresponds to a geometric distribution (fitted red line).

4.3 Discussion

We have presented an algorithm for efficient inference of transcriptional kinetic parameters, with the aim of improving upon an existing compound state Hidden Markov model (Lammers et al., 2020) by reducing the computational time required for inference. A method has also been provided for inferring single cell transcriptional parameters. The algorithm allows for the inference of burst amplitude, duration and frequency from MS2 data, which we expect to be of interest to researchers working on transcriptional regulation. The MS2-MCP system has provided researchers with high-quality data relating to transcriptional activity in individual cells, and has been used to provide insight into the dynamics of transcription. However, the persistence present within the MS2 signal presents a challenge when attempting to infer kinetic parameters using these particular datasets. Our algorithm allows efficient inference of kinetic parameters for longer genes than is currently possible.

A comparison of the running time for a single step of the expectation maximisation algorithm for both the full and truncated models demonstrated the reduction in computational time while using the truncated model on the *Drosophila* gene *ush*, which would require a window size of 19 for inference. The time taken for a single expectation-maximisation step at window size 19 (42 minutes) would render inference using the full model for this particular gene computationally infeasible, particularly if repeated likelihood computations, e.g. for statistical approaches such as bootstrapping

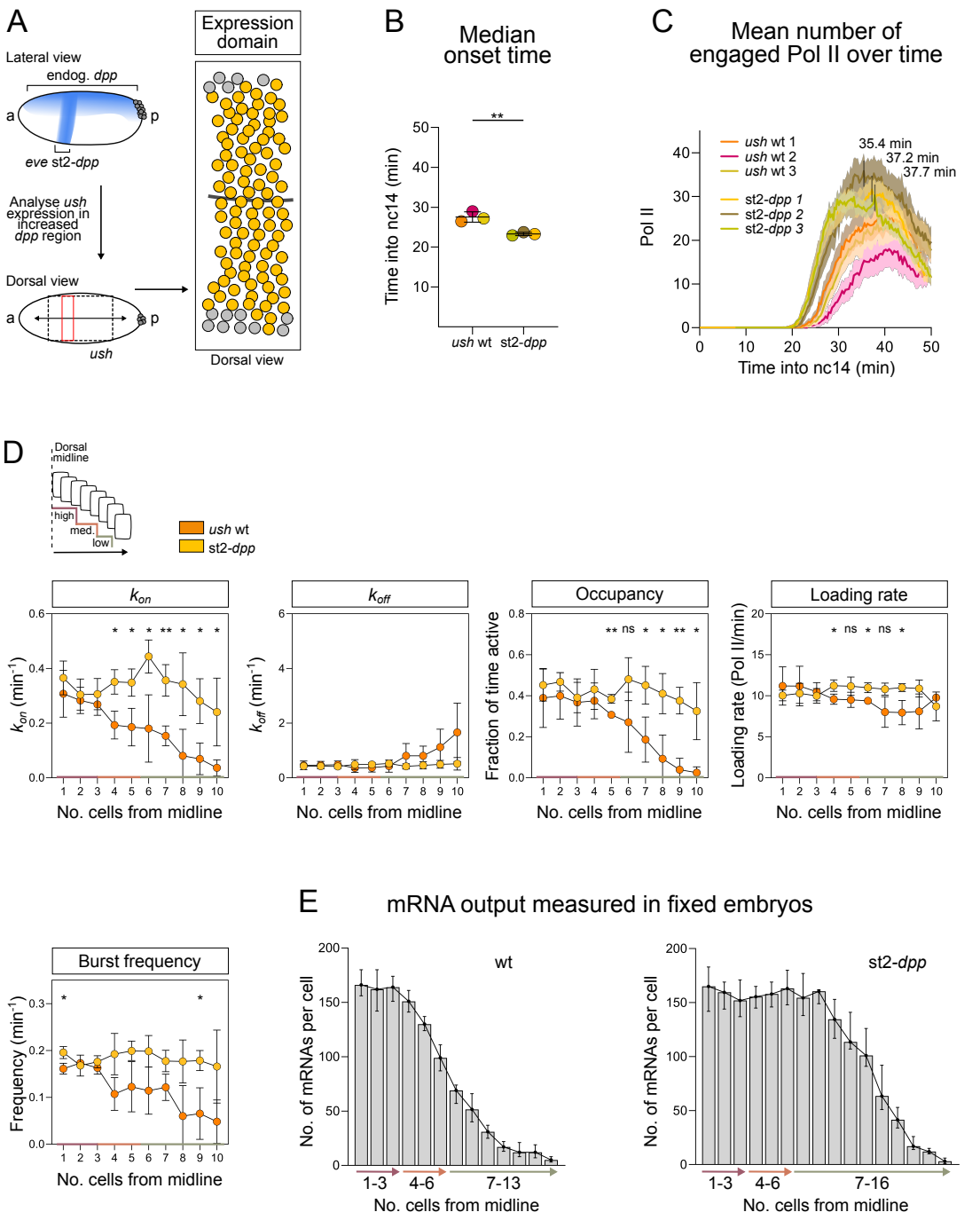

Figure 4.4: Analysis of single cell parameters from a mutant embryo. **A**: Diagram of ectopic dpp expression in dpp-*stp2* embryos, along with the broader expression domain in st2-*dpp* embryos relative to wild type. **B**: Median onset time occurs significantly earlier in dpp-*stp2* embryos. Results of Student's t test ($* = p < 0.05$, $** = p < 0.01$, *ns* = not significant) are shown. **C**: Comparison of mean transcription levels in WT and dpp-*stp2* embryos, along with maximum mean number of engaged Pol II for dpp-*stp2* embryos. **D**: Comparison of inferred single-cell parameters for WT and dpp-*stp2* embryos. Results of student's t test are shown, as in B. **E**: mRNA output for WT and dpp-*stp2* embryos as measured using smFISH. Median $\pm$ 95% confidence intervals are shown. Adapted from Hoppe et al. (2020).

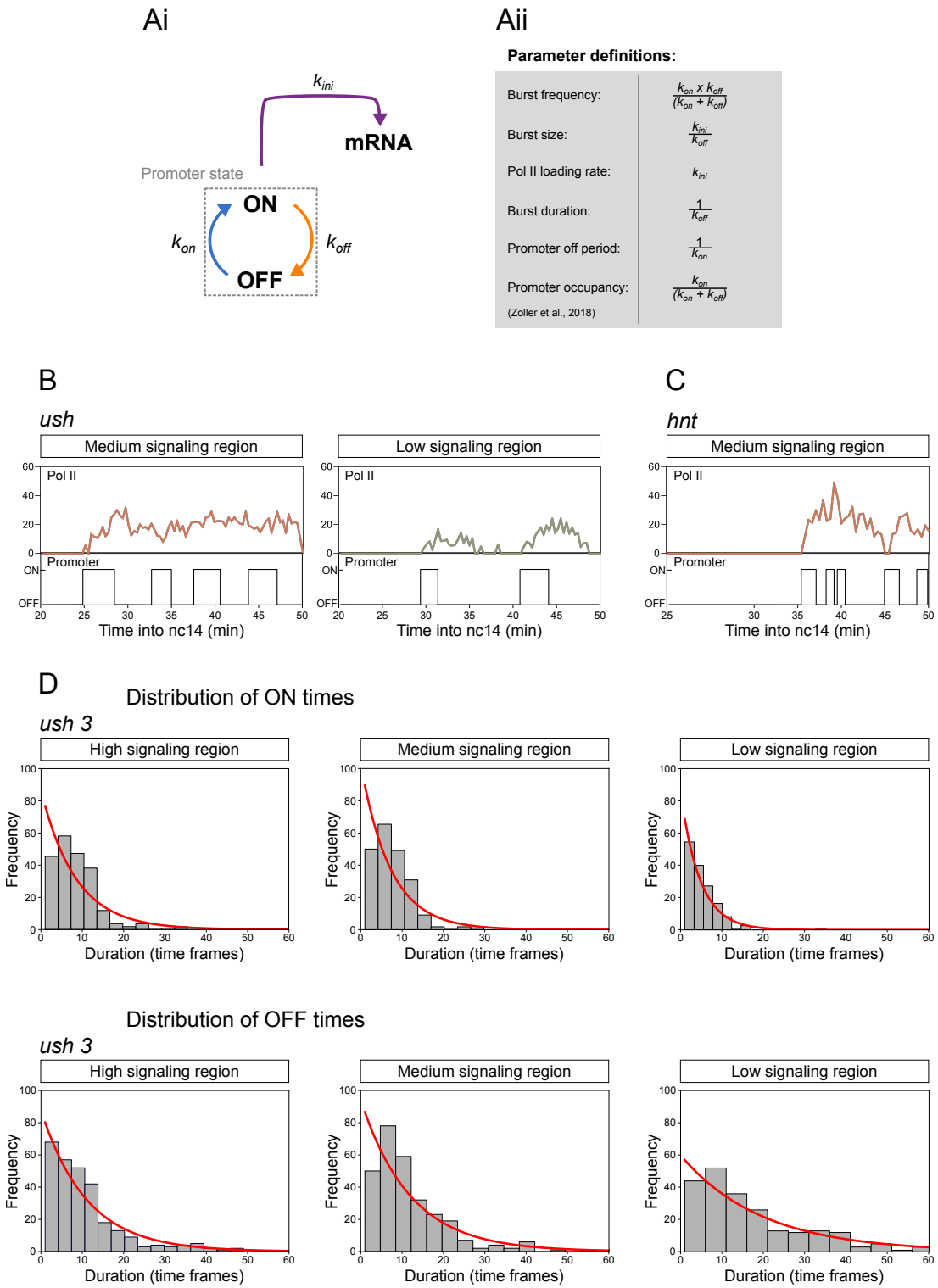

Figure 4.5: Distribution of model On and Off times. **A**: Diagram of the basic two state (random telegraph) transcriptional model, where the promoter alternates between an ON and OFF state according to k_{on} and k_{off}, producing mRNA transcripts at a rate k_{ini} while in the active state (**i**). The random telegraph model can be described using a set of six transcriptional parameters (**ii**). **B**: Representative *ush* MS2 fluorescence traces from the Medium and Low signalling regions, along with inferred promoter sequences. **C**: Equivalent representative trace and inferred promoter for *hnt*. **D**: Distribution of promoter On and Off times for *ush* embryo 3, along with fitted geometric distribution derived from the inferred transition matrix. Adapted from Hoppe et al. (2020).

or MCMC sampling, are required. The truncated model, in comparison, does not scale significantly with gene length and is instead primarily limited by a dependence on the size of the training dataset. This ability to model genes of arbitrary length should allow the model to be applied to more complex organisms, with longer genes, than *Drosophila*.

A limitation of the current implementation of the model is the restriction to $K = 2$ states. It has been observed in *Drosophila* that for some genes the transcriptional dynamics is better described by two rate-limiting steps (Pimmett et al., 2021) resulting in a model with three states. An additional limitation is that our model does not currently take into account the fact that the time series may be non-stationary. In our MS2 time series datasets typically there is an initial silent period (which may be truncated), followed by a rapid ramping up of fluorescence / transcriptional activity, followed by a period of sustained bursting. A non-stationary model would better capture these temporal dynamics. We are working on a non-stationary approach that takes this into account by fitting separate models to different sections of the time series. This requires sharing the emission parameter between different time sections but allowing the kinetic parameters to vary.

In order to further investigate the significance of the model parameters, *burstInfer* was used to infer single-cell values for k_{on}, k_{off}, the loading rate and promoter occupancy, along with the derived parameters burst frequency and burst size (Figure 4.3). Using the inferred global parameters to infer promoter traces on a cell-by-cell basis allowed for the degree of correlation between the inferred single cell parameters and the observed MS2 fluorescence to be calculated (Figure 4.3 C). The single-cell occupancy for *ush* was found to be almost perfectly correlated with the mean fluorescence (Pearson correlation coefficient = 0.99). While k_{on} was found to be strongly positively correlated (r = 0.84), burst frequency and k_{off} were less strongly correlated (r = 0.53 and -0.63, respectively), while the loading rate was weakly correlated (r = 0.07), essentially appearing flat when plotted as a function of position across the midline. A similar relationship was found for *hnt* (k_{on} r = 0.74). These results, taken as a whole, provide further evidence for the role of the promoter activation rate in regulating the DV system.

In order to gain further insight into regulation of the DV system, experiments were carried out to introduce ectopic signalling at the embryo midline through the insertion of the st2-*dpp* transgene. Introduction of the transgene results in a broader *ush* expression domain, relative to wild type, along with an earlier onset time of transcription but

similar time to peak transcription as a wild type. In a similar manner to wild type, *burstInfer* was used to infer global and single-cell transcriptional parameters (Figure 4.4). Single cell parameters were divided into bins of single cell width and significance testing was carried out between the wild type and mutant embryo parameters in each bin (4.4 D). k_{on} was not found to be significantly different from wild type near the embryo midline. However, in rows 4-10 k_{on} was found to be significantly increased in the mutant embryos relative to wild type, along with significantly increased occupancy and burst frequency, while k_{on} and loading rate were not found to be significantly altered. These results, taken in conjunction with previous research showing that BMP signalling is saturated at the dorsal midline, provided further evidence for the role of Dpp signalling in regulating transcriptional dynamics through the promoter activation rate. Experimental evidence for these findings was provided by smFISH experiments that showed increased mRNA output in mutant embryos in the regions corresponding to rows 4-10 relative to wild type.

Chapter 5

Conclusion

5.1 Discussion

In this thesis we have presented an algorithm for efficient inference of transcriptional parameters from MS2 imaging data, along with results showing the application of the algorithm to *Drosophila* embryonic development. The algorithm presented builds upon previous work by Lammers et al. (2020), who developed a computational model (the compound state hidden Markov model, or cpHMM) that, while a significant step forward in the field, scaled poorly with gene length due to the exponential relationship between gene length and the number of compound states required.

Transcription has been extensively described in the scientific literature as a random telegraph process, whereby the state of the promoter cycles stochastically between active and inactive states, with no long-range time dependency between states. Many biological systems, including BMP target genes in *Drosophila*, exhibit transcriptional bursting, whereby genes are transcribed in discrete bursts, rather than as a continuous process. This particular type of system, involving stochastic switching between states (promoter activity) that are not directly observed, each of which is associated with an observation (the recorded MS2 fluorescence), suggests the use of a hidden Markov model (HMM).

Lammers et al. developed a custom implementation of a HMM, the compound-state HMM, or cpHMM, which was designed specifically for modelling transcriptional bursting during early *Drosophila* development. A custom implementation was required due to the nature of the time series data generated by the MS2 system. As polymerase transits down the gene following promoter activation, a noisy fluorescent waveform is generated. When the promoter becomes inactive, the fluorescent signal does not

immediately disappear, but instead falls gradually, due to polymerase still in transit down the gene body following promoter deactivation. A conventional HMM is not able to capture this time dependency, as while the (in reality, continuous) transition between promoter states is Markovian, the fluorescence level, or observation at time point t, depends upon previous promoter states.

Lammers et al. introduced two innovations to deal with this time dependency: the window size, controlled by the parameter W, and the concept of compound states. W, determined by the gene length, time resolution of the system and assumed polymerase elongation rate, determines how many previous time points are taken into account at the current time point t. The concept of compound states was introduced in order to implement the window size as part of the model. Rather than simply 2 or 3 states, the cpHMM includes an expanded state space of 2^W compound states.

At the beginning of the time series, t_0, only two system configurations are possible: either the promoter is active or inactive. At the next time step, the state space doubles to 4 possible compound states, as the promoter may switch to either inactive or active from each initial state. At each subsequent time step the state space doubles in size until it reaches 2^W compound states in size. This process is repeated with each successive pass of the forward and backward algorithms.

It is straightforward to see that the number of compound states, and therefore computational time, increases exponentially with the number of compound states. For the specific set of relatively short genes that Lammers et al. studied, this did not represent a problem. However, the cpHMM approach quickly becomes intractable when dealing with longer genes.

The dynamic HMM (dHMM) algorithm presented in this thesis was developed as an attempt to build upon the Lammers model by allowing for much more efficient inference of transcriptional parameters from systems involving genes of arbitrary length. This was done by introducing a truncated state space. In the dynamic model, the state space expands at each time point until it reaches a pre-selected threshold size, which may be a fraction of the size of the original model. The state space stored at each subsequent time point does not exceed this size. This is possible due to the very small probability associated with a large number of transitions; discarding these states does not significantly alter the inferred parameters. Truncating the model in this way removes the exponential scaling, at the cost of increased computational time at smaller

window sizes relative to the original model due to increased overhead costs from sorting and removing compound states. We have demonstrated that, as the number of compound states increases, the truncated model converges towards to the original model. In fact, 128-256 compound states were found to be sufficient to approximate a model that requires $2^{19} = 524288$ compound states in the original model. Following validation on synthetic data, the algorithm has been published and released as an open source software package, *burstInfer*.

Another innovative feature of *burstInfer*, as outlined in Chapter 4, is the ability to infer single-cell transcriptional parameters. In addition to inferring kinetic parameters for the entire embryo, or for a particular region of the embryo, *burstInfer* allows for estimates of k_{off},k_{on}, loading rate and promoter occupancy for each cell in the expression domain. This is done through a simple counting-based approach, whereby the posterior state probability inferred as part of the forward-backward algorithm is used to generate a sequence of most likely promoter states for each cell. Counting the normalised number of off $\rightarrow$ on and on $\rightarrow$ off transitions allows for an estimate of k_{off} and k_{on}, which in turns allows for the single-cell occupancy to be estimated. The single-cell emission is calculated on a cell-by-cell basis, rather than for an entire dataset. The aim of including single-cell calculations is to reveal insights into bursting dynamics from the data that may not be obvious when looking at global parameters that are averaged across the entire expression domain or across the entire embryo.

As demonstrated in Chapter 4, *burstInfer* was used by Hoppe et al. (2020) to reveal bursting dynamics in BMP target genes in the early *Drosophila* embryo. Dorsal-ventral (DV) patterning in *Drosophila* is regulated by Decapentaplegic, or Dpp, a member of the Bone Morphogenetic Protein (BMP) family. Dpp has a number of target genes in the DV system, such as *u-shaped* (*ush*) and *hindsight* (*hnt*), that act to partition the embryo into different tissues types during nuclear cycle 14 . The exact regulatory mechanism underlying the relationship between Dpp, its target genes and cell fate has remained elusive for some time.

Following partitioning of the embryo into different regions based upon distance from the midline and clustering of the MS2 data, *burstInfer* was used to infer global transcriptional parameters for the genes *ush* and *hnt* (Figure 4.2). Carrying out significance testing on the inferred k_{off}, k_{on}, and emission (polymerase loading rate) parameters revealed that k_{on} was the key regulated parameter in the system. k_{off} and the emission term did not show significant changes between the intermediate and out regions. The promoter occupancy, given by $\frac{k_{on}}{k_{on}+k_{off}}$, also showed a significant change

between the intermediate and outer regions of the embryo. As k_{off} was not found to be a regulated parameter, this relationship must depend upon k_{on}. Additional evidence for this conclusion was provided by the single cell modelling results, where k_{on} was found to be strongly positively correlated with with expression (r=0.84).

5.2 Future Work

5.2.1 Non-Stationary Hidden Markov Models

A potential limitation of our dynamic model is the inability to capture temporal, aswell as spatial, changes in transcriptional dynamics. The model has allowed us to investigate spatial regulation of transcriptional dynamics across the expression domain, but there may also be temporal changes in transcription. In the time series data used in this thesis, transcription typically exhibits a 'ramp-up' period of around ten minutes, before settling into a more consistent bursting pattern. The parameters we have inferred do not reflect this, as a single parameter is inferred for the entire time series - the model is a form of homogeneous Hidden Markov Model.

The most simple extension to the model would be to use a sliding window to divide the time series into different sections, training the model on each section. A more sophisticated approach would be to assume a shared rate of polymerase loading rate (emission), but allow k_{off} and k_{on} to vary between different segments of the time series. Still more sophisticated approaches are possible, such as modelling the temporal parameter changes as an underlying stochastic process.

5.2.2 Multi-State Models

Our HMM implementation only considers two transcriptional states - active and inactive. Previous work in the literature has raised the possibility of needing multi-state models to properly model transcription. Corrigan et al. (2016) included multiple states with different initiation rates, as they concluded that a 2-state model was insufficient to explain their data. Lammers et al. included both 2-state and 3-state models in their paper (Lammers et al., 2018) due to each fluorescent MS2 spot in their data including two transcriptional loci. Extending the model to include 3 states could involve significant work, as our binary encoding system would no longer be appropriate. Computational time may also be significantly increased.

5.2.3 Hidden semi-Markov Models

Hidden semi-Markov Models are an extension to a standard HMM where the underlying hidden states are allowed to be of variable duration, with each hidden state outputting a variable number of observations. The chance of leaving a given hidden state depends upon the time spent in the state - the underlying stochastic process is assumed to be semi-Markovian. This relaxation of the Markov assumption leads to a more flexible model, which may be, in theory, more capable of capturing variable data. The biological process underlying our model, i.e. promoter state, is arguably represented in quite a rigid way in the existing model, as the promoter is not actually switching every 20 seconds or so between states. Relaxing the Markov assumption may lead to a model that better describes the data.

5.2.4 Mean Field Variational Bayes Methods

Other algorithms have also been developed for efficient inference in intractable probabilistic models. A popular class are variational Bayesian inference algorithms and these have previously been applied to generalised versions of hidden Markov models (Murphy, 2022).

A limitation of the approximation developed in this thesis is that while inference is drastically sped up for very long genes, relative the original approach, the computational time taken still scales linearly with the number of allowed states in the truncated model approximation. As the number of allowed states is increased, the truncated model converges to the true values, but the computational time also increases (Figure 3.7). Therefore it is possible that the inference results will be inaccurate for some very long genes.

Unpublished work in our group has investigated the use of Mean Field Variational Bayes Methods instead, where the model likelihood is approximated using the Evidence Lower Bound (ELBO). A mean-field assumption is used to approximate the posterior distribution of the latent state variables by a factorised distribution. This approach offers a more computationally efficient alternative to sampling-based techniques and does not suffer from the scaling issue we have when we increase our number of allowed states. Initial results indicate that the model is able to infer kinetic parameters reasonably well after being trained on MS2 data. However, further experiments are required to determine whether the factorisation of the variational approximation is able to provide a good approximation to the state variable posterior distribution.

Bibliography

Ashe, H. L., Mannervik, M., and Levine, M. (2000). Dpp signaling thresholds in the dorsal ectoderm of the Drosophila embryo. *Development*, 127(15):3305–3312.

Bach, D. H., Park, H. J., and Lee, S. K. (2018). The Dual Role of Bone Morphogenetic Proteins in Cancer. *Molecular Therapy - Oncolytics*, 8:1–13.

Bahar Halpern, K., Caspi, I., Lemze, D., Levy, M., Landen, S., Elinav, E., Ulitsky, I., and Itzkovitz, S. (2015a). Nuclear Retention of mRNA in Mammalian Tissues. *Cell Reports*, 13(12):2653–2662.

Bahar Halpern, K. and Itzkovitz, S. (2016). Single molecule approaches for quantifying transcription and degradation rates in intact mammalian tissues. *Methods*, 98:134–142.

Bahar Halpern, K., Tanami, S., Landen, S., Chapal, M., Szlak, L., Hutzler, A., Nizhberg, A., and Itzkovitz, S. (2015b). Bursty gene expression in the intact mammalian liver. *Molecular Cell*, 58(1):147–156.

Balázsi, G., Van Oudenaarden, A., and Collins, J. J. (2011). Cellular decision making and biological noise: From microbes to mammals. *Cell*, 144(6):910–925.

Bartman, C. R., Hamagami, N., Keller, C. A., Giardine, B., Hardison, R. C., Blobel, G. A., and Raj, A. (2019). Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation. *Molecular Cell*, 73(3):519–532.

Bartman, C. R., Hsu, S. C., Hsiung, C. C., Raj, A., and Blobel, G. A. (2016). Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. *Molecular Cell*, 62(2):237–247.

Belvin, M. P. and Anderson, K. V. (1996). A CONSERVED SIGNALING PATHWAY: The Drosophila Toll-Dorsal Pathway . *Annual Review of Cell and Developmental Biology*, 12(1):393–416.

Bentovim, L., Harden, T. T., and DePace, A. H. (2017). Transcriptional precision and accuracy in development: From measurements to models and mechanisms. *Development (Cambridge)*, 144(21):3855–3866.

Berrocal, A., Lammers, N., Garcia, H. G., and Eisen, M. B. (2018). Kinetic sculpting of the seven stripes of the Drosophila even-skipped gene. *bioRxiv*, 1:335901.

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. *Molecular Cell*, 2(4):437–445.

Bier, E. and De Robertis, E. M. (2015). BMP gradients: A paradigm for morphogen-mediated developmental patterning. *Science*, 348(6242).

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Blanco Calvo, M., Bolós Fernández, V., Medina Villaamil, V., Aparicio Gallego, G., Díaz Prado, S., and Grande Pulido, E. (2009). Biology of BMP signalling and cancer. *Clinical and Translational Oncology*, 11(3):126–137.

Boettiger, A. N. and Levine, M. (2013). Rapid Transcription Fosters Coordinate snail Expression in the Drosophila Embryo. *Cell Reports*, 3(1):8–15.

Bothma, J. P., Garcia, H. G., Esposito, E., Schlissel, G., Gregor, T., and Levine, M. (2014). Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living Drosophila embryos. *Proceedings of the National Academy of Sciences of the United States of America*, 111(29):10598–10603.

Bowles, J. R., Hoppe, C., Ashe, H. L., and Rattray, M. (2022). Scalable inference of transcriptional kinetic parameters from MS2 time series data. *Bioinformatics*, 38(4):1030–1036.

Chalancon, G., Ravarani, C. N., Balaji, S., Martinez-Arias, A., Aravind, L., Jothi, R., and Babu, M. M. (2012). Interplay between gene expression noise and regulatory network architecture. *Trends in Genetics*, 28(5):221–232.

Chen, H. and Larson, D. R. (2016). What have single-molecule studies taught us about gene expression? *Genes and Development*, 30(16):1796–1810.

Chubb, J. R., Trcek, T., Shenoy, S. M., and Singer, R. H. (2006). Transcriptional Pulsing of a Developmental Gene. *Current Biology*, 16(10):1018–1025.

Corrigan, A. M., Tunnacliffe, E., Cannon, D., and Chubb, J. R. (2016). A continuum model of transcriptional bursting. *eLife*, 5(February 2016):1–38.

Coulon, A., Chow, C. C., Singer, R. H., and Larson, D. R. (2013). Eukaryotic transcriptional dynamics: From single molecules to cell populations. *Nature Reviews Genetics*, 14(8):572–584.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Dar, R. D., Razooky, B. S., Singh, A., Trimeloni, T. V., McCollum, J. M., Cox, C. D., Simpson, M. L., and Weinberger, L. S. (2012). Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 109(43):17454–17459.

Darzacq, X., Shav-Tal, Y., De Turris, V., Brody, Y., Shenoy, S. M., Phair, R. D., and Singer, R. H. (2007). In vivo dynamics of RNA polymerase II transcription. *Nature Structural and Molecular Biology*, 14(9):796–806.

Deignan, L., Pinheiro, M. T., Sutcliffe, C., Saunders, A., Wilcockson, S. G., Zeef, L. A., Donaldson, I. J., and Ashe, H. L. (2016). Regulation of the BMP Signaling-Responsive Transcriptional Network in the Drosophila Embryo. *PLoS Genetics*, 12(7):e1006164.

Desponds, J., Tran, H., Ferraro, T., Lucas, T., Perez Romero, C., 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 Computational Biology*, 12(12).

Dorfman, R. and Shilo, B.-Z. (2001). Biphasic activation of the BMP pathway patterns the Drosophila embryonic dorsal region. *Development*, 128(6):965–972.

Durbin, R. (2006). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press.

El Samad, H., Khammash, M., Petzold, L., and Gillespie, D. (2005). Stochastic modelling of gene regulatory networks. *International Journal of Robust and Nonlinear Control*, 15(15):691–711.

Eldar, A., Dorfman, R., Weiss, D., Ashe, H., Shilo, D. Z., and Barkal, N. (2002). Robustness of the BMP morphogen gradient in Drosophila embryonic patterning. *Nature*, 419(6904):304–308.

Eldar, A. and Elowitz, M. B. (2010). Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173.

Elowitz, M. B. and Leibier, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186.

Femino, A. M., Fay, F. S., Fogarty, K., and Singer, R. H. (1998). Visualization of single RNA transcripts in situ. *Science*, 280(5363):585–590.

Friedman, N., Cai, L., and Xie, X. S. (2006). Linking stochastic dynamics to population distribution: An analytical framework of gene expression. *Physical Review Letters*, 97(16).

Fukaya, T., Lim, B., and Levine, M. (2016). Enhancer Control of Transcriptional Bursting. *Cell*, 166(2):358–368.

Fukaya, T., Lim, B., and Levine, M. (2017). Rapid Rates of Pol II Elongation in the Drosophila Embryo. *Current Biology*, 27(9):1387–1391.

Garcia, H. G., Tikhonov, M., Lin, A., and Gregor, T. (2013). Quantitative imaging of transcription in living Drosophila embryos links polymerase activity to patterning. *Current biology : CB*, 23(21):2140–2145.

Golding, I. and Cox, E. C. (2004). RNA dynamics in live Escherichia coli cells. *Proceedings of the National Academy of Sciences of the United States of America*, 101(31):11310–11315.

Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036.

Gómez-Schiavon, M., Chen, L. F., West, A. E., and Buchler, N. E. (2017). BayFish: Bayesian inference of transcription dynamics from population snapshots of single-molecule RNA FISH in single cells. *Genome Biology*, 18(1):164.

Gregor, T., Garcia, H. G., and Little, S. C. (2014). The embryo as a laboratory: Quantifying transcription in Drosophila. *Trends in Genetics*, 30(8):364–375.

Hill, C. S. (2009). Nucleocytoplasmic shuttling of Smad proteins. *Cell Research*, 19(1):36–46.

Hoppe, C., Bowles, J. R., Minchington, T. G., Sutcliffe, C., Upadhyai, P., Rattray, M., and Ashe, H. L. (2020). Modulation of the Promoter Activation Rate Dictates the Transcriptional Response to Graded BMP Signaling Levels in the Drosophila Embryo. *Developmental Cell*, pages 1–15.

Horvathova, I., Voigt, F., Kotrys, A. V., Zhan, Y., Artus-Revel, C. G., Eglinger, J., Stadler, M. B., Giorgetti, L., and Chao, J. A. (2017). The Dynamics of mRNA Turnover Revealed by Single-Molecule Imaging in Single Cells. *Molecular Cell*, 68(3):615–625.

Jones, D. L., Brewster, R. C., and Phillips, R. (2014). Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346(6216):1533–1536.

Kalo, A., Kanter, I., Shraga, A., Sheinberger, J., Tzemach, H., Kinor, N., Singer, R. H., Lionnet, T., and Shav-Tal, Y. (2015). Cellular levels of signaling factors are sensed by β-actin alleles to modulate transcriptional pulse intensity. *Cell Reports*, 11(3):419–432.

Kanodia, J. S., Rikhy, R., Kim, Y., Lund, V. K., DeLotto, R., Lippincott-Schwartz, J., and Shvartsman, S. Y. (2009). Dynamics of the Dorsal morphogen gradient. *Proceedings of the National Academy of Sciences*, 106(51):21707–21712.

Lacy, M. E. and Hutson, M. S. (2016). Amnioserosa development and function in Drosophila embryogenesis: Critical mechanical roles for an extraembryonic tissue. *Developmental Dynamics*, 245(5):558–568.

Lammers, N. C., Galstyan, V., Reimer, A., Medin, S. A., Wiggins, C. H., and Garcia, H. G. (2018). Binary transcriptional control of pattern formation in development. *bioRxiv*, page 335919.

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. *Proceedings of the National Academy of Sciences of the United States of America*, 117(2):836–847.

Larson, D. R., Fritzsch, C., Sun, L., Meng, X., Lawrence, D. S., and Singer, R. H. (2013). Direct observation of frequency modulated transcription in single cells using light activation. *eLife*, 2.

Lee, T. H. (2009). Extracting kinetics information from single-molecule fluorescence resonance energy transfer data using hidden markov models. *Journal of Physical Chemistry B*, 113(33):11535–11542.

Lee, T. I. and Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251.

Lenstra, T. L., Rodriguez, J., Chen, H., and Larson, D. R. (2016). Transcription Dynamics in Living Cells. *Annual Review of Biophysics*, 45(1):25–47.

Levine, M. and Davidson, E. H. (2005). Gene regulatory networks for development. *Proceedings of the National Academy of Sciences of the United States of America*, 4(4):177–179.

Li, C., Cesbron, F., Oehler, M., Brunner, M., and Höfer, T. (2018). Frequency Modulation of Transcriptional Bursting Enables Sensitive and Rapid Gene Regulation. *Cell Systems*, 6(4):409–423.

Lim, B., Heist, T., Levine, M., and Fukaya, T. (2018). Visualization of Transvection in Living Drosophila Embryos. *Molecular Cell*, 70(2):287–296.

Lim, B., Levine, M., and Yamakazi, Y. (2017). Transcriptional Pre-patterning of Drosophila Gastrulation. *Current Biology*, 27(2):286–290.

Little, S. C., Tikhonov, M., and Gregor, T. (2013). Precise developmental gene expression arises from globally stochastic transcriptional activity. *Cell*, 154(4):789–800.

Lucas, T., Ferraro, T., Roelens, B., De Las Heras Chanes, J., Walczak, A. M., Coppey, M., and Dostatni, N. (2013). Live Imaging of Bicoid-Dependent Transcription in Drosophila Embryos. *Current Biology*, 23(21):2135–2139.

Lyubimova, A., Itzkovitz, S., Junker, J. P., Fan, Z. P., Wu, X., and Van Oudenaarden, A. (2013). Single-molecule mRNA detection and counting in mammalian tissue. *Nature Protocols*, 8(9):1743–1758.

Meyers, E. A. and Kessler, J. A. (2017). TGF-β family signaling in neural and neuronal differentiation, development, and function. *Cold Spring Harbor Perspectives in Biology*, 9(8):1–26.

Mizutani, C. M., Nie, Q., Wan, F. Y., Zhang, Y. T., Vilmos, P., Sousa-Neves, R., Bier, E., Marsh, J. L., and Lander, A. D. (2005). Formation of the BMP activity gradient in the drosophila embryo. *Developmental Cell*, 8(6):915–924.

Mueller, F., Senecal, A., Tantale, K., Marie-Nelly, H., Ly, N., Collin, O., Basyuk, E., Bertrand, E., Darzacq, X., and Zimmer, C. (2013). FISH-quant: Automatic counting of transcripts in 3D FISH images. *Nature Methods*, 10(4):277–278.

Munsky, B., Neuert, G., and Van Oudenaarden, A. (2012). Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187.

Muramoto, T., Müller, I., Thomas, G., Melvin, A., and Chubb, J. R. (2010). Methylation of H3K4 Is Required for Inheritance of Active Transcriptional States. *Current Biology*, 20(5):397–406.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An introduction*. MIT Press.

Nicolas, D., Phillips, N. E., and Naef, F. (2017). What shapes eukaryotic transcriptional bursting? *Molecular BioSystems*, 13(7):1280–1290.

Ochiai, H., Sugawara, T., Sakuma, T., and Yamamoto, T. (2014). Stochastic promoter activation affects Nanog expression variability in mouse embryonic stem cells. *Scientific Reports*, 4(1):7125.

O'Connor, M. B. (2005). Shaping BMP morphogen gradients in the Drosophila embryo and pupal wing. *Development*, 133(2):183–193.

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. *Nature Genetics*, 31(1):69–73.

Paulsson, J. (2004). Summing up the noise in gene networks. *Nature*, 427(6973):415–418.

Peccoud, J. and Ycart, B. (1995). Markovian modeling of gene-product synthesis. *Theoretical Population Biology*, 48(2):222–234.

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. *Molecular Cell*, 71(3):468–480.

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. *Nature Communications*, 12(1).

Qureshi, N., Takayama, K., Jordi, H. C., and Schnoes, H. K. (1978). Characterization of the purified components of a new homologous series of ??-mycolic acids from Mycobacterium tuberculosis H37Ra. *Journal of Biological Chemistry*, 253(15):5411–5417.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286.

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):1707–1719.

Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods*, 5(10):877–879.

Raj, A. and van Oudenaarden, A. (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell*, 135(2):216–226.

Raser, J. M. and O'Shea, E. K. (2004). Control of stochasticity in eukaryotic gene expression. *Science (New York, N.Y.)*, 304(5678):1811–4.

Raser, J. M. and O'Shea, E. K. (2005). Molecular biology - Noise in gene expression: Origins, consequences, and control. *Science*, 309(5743):2010–2013.

Senecal, A., Munsky, B., Proux, F., Ly, N., Braye, F. E., Zimmer, C., Mueller, F., and Darzacq, X. (2014). Transcription factors modulate c-Fos transcriptional bursts. *Cell Reports*, 8(1):75–83.

Skinner, S. O., Sepúlveda, L. A., Xu, H., and Golding, I. (2013). Measuring mRNA copy number in individual Escherichia coli cells using single-molecule fluorescent in situ hybridization. *Nature Protocols*, 8(6):1100–1113.

So, L. H., Ghosh, A., Zong, C., Sepúlveda, L. A., Segev, R., and Golding, I. (2011). General properties of transcriptional time series in Escherichia coli. *Nature Genetics*, 43(6):554–560.

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(6028):472–474.

Sutherland, D. J. (2003). Stepwise formation of a SMAD activity gradient during dorsal-ventral patterning of the Drosophila embryo. *Development*, 130(23):5705–5716.

Swain, P. S., Elowitz, M. B., and Siggia, E. D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800.

Tantale, K., Mueller, F., Kozulic-Pirher, A., Lesne, A., Victor, J. M., Robert, M. C., Capozi, S., Chouaib, R., Bäcker, V., Mateos-Langerak, J., Darzacq, X., Zimmer, C., Basyuk, E., and Bertrand, E. (2016). A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting. *Nature Communications*, 7:12248.

Trcek, T., Lionnet, T., Shroff, H., and Lehmann, R. (2017). mRNA quantification using single-molecule FISH in Drosophila embryos. *Nature Protocols*, 12(7):1326–1347.

Umulis, D. M., Shimmi, O., O'Connor, M. B., and Othmer, H. G. (2010). Organism-Scale Modeling of Early Drosophila Patterning via Bone Morphogenetic Proteins. *Developmental Cell*, 18(2):260–274.

Xu, H., Sepúlveda, L. A., Figard, L., Sokac, A. M., and Golding, I. (2015). Combining protein and mRNA quantification to decipher transcriptional regulation. *Nature Methods*, 12(8):739–742.

Zenklusen, D., Larson, D. R., and Singer, R. H. (2008). Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature Structural & Molecular Biology*, 15(12):1263–1271.

Zoller, B., Little, S. C., and Gregor, T. (2018). Diverse Spatial Expression Patterns Emerge from Unified Kinetics of Transcriptional Bursting. *Cell*, 175(3):835–847.