THE UNIVERSITY OF CHICAGO

COUPLING TRANSCRIPTIONAL DYNAMICS AND DNA STRUCTURE WITH SINGLE MOLECULE LOCALIZATION MICROSCOPY

A THESIS SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHYSICS

BY
CLAYTON W. SEITZ

CHICAGO, ILLINOIS
SPRING 20XX

TABLE OF CONTENTS

ABSTRACT

Eukaryotic transcription is episodic, consisting of a series of transcriptional bursts, Bursty transcriptional dynamics are well-exemplified by the transient expression of pro-inflammatory guanylate binding proteins (GBPs) - a group interferon-inducible GTPases that restrict the replication of intracellular pathogens [XXX]. Classical models of gene regulation explain transcriptional bursts by invoking stochastic binding and unbinding of transcription factors, RNA polymerase and mediator proteins at enhancer or promoter sequences. However, more recent studies have pointed towards a more cooperative picture of transcriptional control where phase-separated aggregates of DNA, RNA, and proteins form higher-order structures to control gene expression. For example, both chromatin immunoprecipitation and super resolution imaging have captured the phase separation of super-enhancer-binding proteins MED1 and BRD4 in transcriptional condensates at the *Essrb* genomic locus [XXX]. Furthermore, fluorescence microscopy techniques have colocalized MED1 and BRD4 with the GBP gene cluster alongside a reduction in the degree of disorder of 3D chromatin structure in murine macrophages after infection with *Mycobacterium tuberculosis*. Taken together, these results suggest that phase separation may play a role in the reorganization of chromatin structure during trancriptional control of innate immune response genes [XXX]. Here, we hypothesize that phase separation reduces the entropy of chromatin structure in order to induce bursty gene expression. Using single molecule localization microscopy (SMLM) to obtain super-resolution images of the H2B protein, we intend to demonstrate simultaneous (i) loss of disorder in chromatin structure (ii) formation of transcriptional condensates containing MED1 and BRD4 and (iii) non-Poissonian gene expression. The following sections discuss recent the biological evidence in more detail and summarize the single molecule microscopy techniques and biophysical models we employ to study the interactions between transcriptional condensates and the chromatin scaffold.

CHAPTER 1

SINGLE MOLECULE LOCALIZATION MICROSCOPY

1.1 Point spread functions in single molecule localization microscopy

Gaussian approximation. Lateral and axial point spread functions, Potentially double helix point spread functions for 3D storm

Most detectors used for imaging have many elements (pixels) so that we can record an image projected onto the detector by a system of lenses. In fluorescence imaging, this is usually a relay consisting of an objective lens and a tube lens to focus the image onto the camera. Due to diffraction, any point emitter, such as a single fluorescent molecule, will be registered as a diffraction limited spot. The profile of that spot is often described as a Gaussian point spread function (Richardson and Wolf)

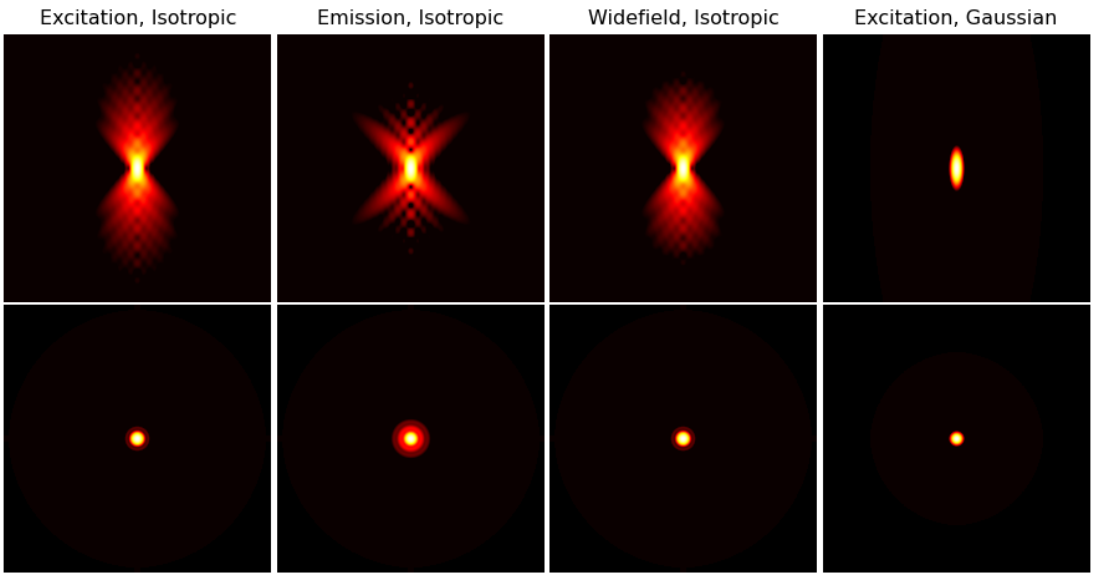
$$\mathrm{G}(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma^2}} + B_0 \tag{1.1}$$

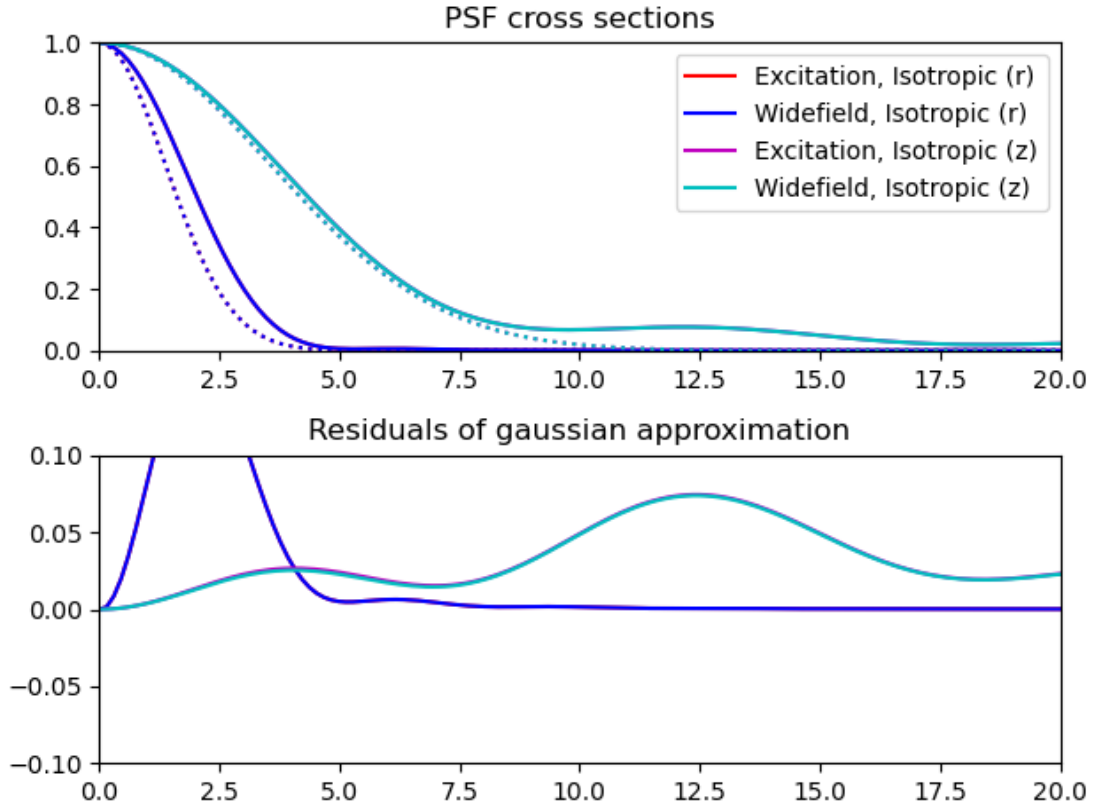
which has units of $[W/m^2]$.

1.1.1 The physics of semiconductor cameras

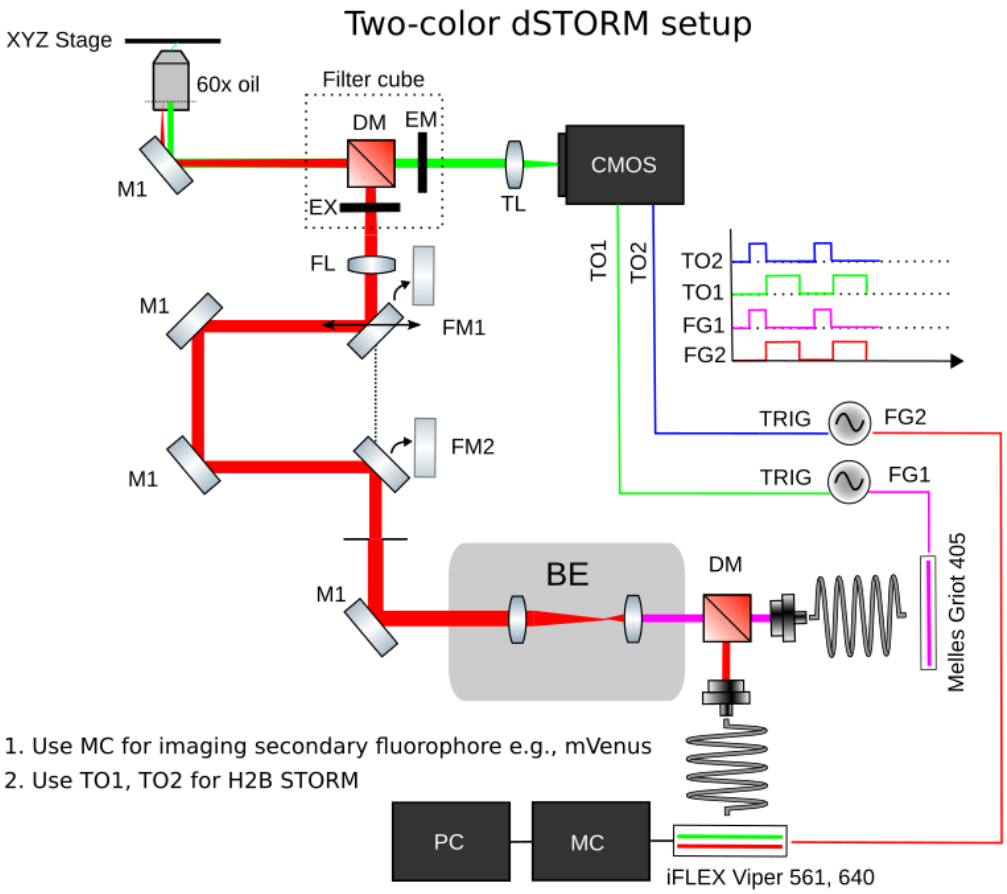
Modern cameras used in light microscopy, such as scientific complementary metal oxide semiconductor (sCMOS) cameras, are ultimately powered by the photoelectric effect. This occurs when electrons within the material absorb energy from the photons and have enough energy to escape their bonds and become ejected from the material. The materials used in camera sensors are typically semiconductors such as silicon or gallium

In essence, an image captured by a camera can be loosely thought of as histogram of photon arrivals and a discretized form of the density $\mathrm{PSF}(x,y)$ over an integration time τ.



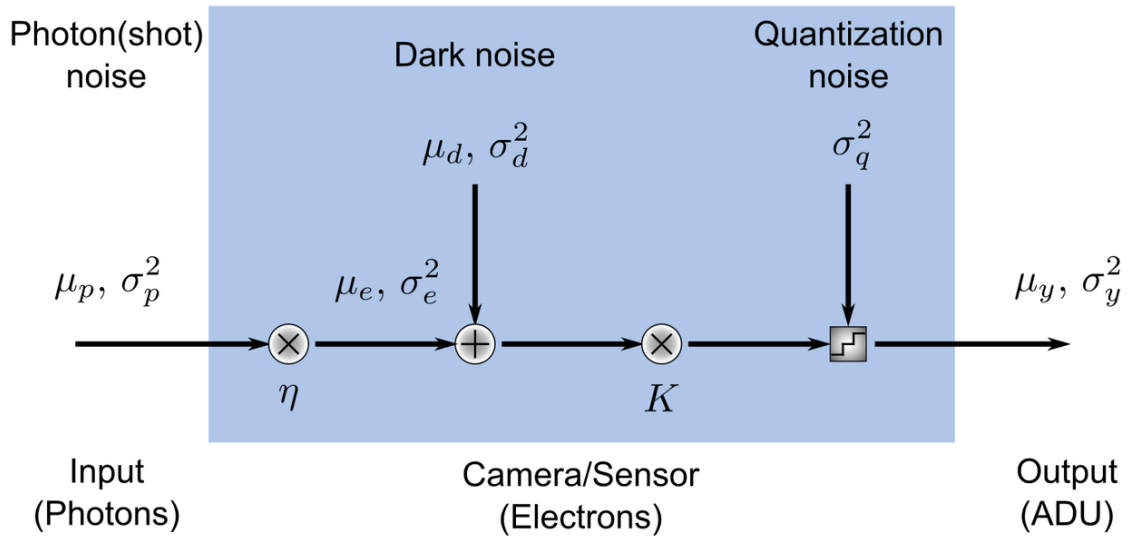


If PSF(x, y) can be approximated as constant in time, the value at a pixel approaches an integral of this density over the pixel:



$$\mu_k = \lambda_k \tau = \frac{\eta \tau}{\hbar \omega} \int G(x, y) dA \tag{1.2}$$

The parameter η is called the *quantum efficiency*, and without loss of generality, we will assume $\eta = 1$. The variable λ_k at a pixel k defines the probability of observing a photon per unit time and the *true signal* is Poisson: $S_k \sim P(\lambda_k)$. In this model, notice that S_k may represent multiple sources e.g., fluorescence from a single molecule plus a background source as shown in (1.1). Furthermore, we say that the detection process is *shot-noise limited* when Poisson noise dominates. It would be convenient to obtain an analytical expression for λ_k given the parameters of the point spread function $\theta = (x_0, y_0, \sigma)$. Since the 2D Gaussian is symmetric, it is separable, and we can write



$$\lambda_k = \frac{1}{2\pi\sigma^2}\left(\int_{x_k}^{x_{k+1}} e^{-\frac{(x-x_0)^2}{2\sigma^2}}\,dx\right)\left(\int_{y_k}^{y_{k+1}} e^{-\frac{(y-y_0)^2}{2\sigma^2}}\,dy\right)$$

We can then express the Gaussian integrals over a pixel by making use of the following property of the error function

$$\frac{1}{\sqrt{2\pi}\sigma}\int_a^b e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \frac{1}{2}\left(\mathrm{erf}\left(\frac{b-\mu}{\sqrt{2}\sigma}\right) - \mathrm{erf}\left(\frac{a-\mu}{\sqrt{2}\sigma}\right)\right)$$

Suppose the particle is known to be located at (x_0, y_0) and some square pixel k is centered on coordinates (x, y) and has a width a. Then μk at this pixel is

$$\mu(x_k, y_k) = \mu_x(x_k)\mu_y(y_k) \tag{1.3}$$

where

$$\mu_x(x_k) = \frac{\tau}{2}\left(\text{erf}\left(\frac{x + a/2 - x_0}{\sqrt{2}\sigma}\right) - \text{erf}\left(\frac{x - a/2 - x_0}{\sqrt{2}\sigma}\right)\right)$$
$$\mu_y(y_k) = \frac{\tau}{2}\left(\text{erf}\left(\frac{y + a/2 - y_0}{\sqrt{2}\sigma}\right) - \text{erf}\left(\frac{y - y/2 - y_0}{\sqrt{2}\sigma}\right)\right)$$

However this noise model is incomplete, because detectors often suffer from dark noise, which may refer to readout noise or dark current, and contributes to a nonzero signal even in the absence of incident light. Dark current is due to statistical fluctuations in the photoelectron count due to thermal fluctuations. Readout noise is introduced by the amplifier circuit during the coversion of photoelectron charge to a voltage. Here, we use the Hamamatsu ORCA v3 CMOS camera, which is air cooled to -10C and has very low dark current - around 0.06 electrons/pixel/second - and can be safely ignored for low exposure times. Readout noise has been often neglected in localization algorithms because its presence in EMCCD cameras is small enough that it can be ignored within the tolerances of the localization precision. In the case of sCMOS cameras, however, the readout noise of each pixel is significantly higher and, in addition, every pixel has its own noise and gain characteristic with dramatic pixel-to-pixel variations.

Readout noise is not negligible, and is represented as a random variable W_k which is *a statistical fluctuation of the number of ADU during the amplification process.* It is important to note that we cannot measure the contribution by readout noise W_k before amplification and therefore it must be expressed in units of ADU. This is in contrast to the Poisson variable S_k, which *can* be expressed in units of photoelectrons, because we are able to estimate the emission rate of the fluorophore analytically and we know the quantum efficiency η. The number of photoelectrons S_k is multiplied by a gain factor g_k which has units of [ADU/e^-], which generally must be measured for each pixel. For W_k, the gain factor is implied. Here, we will always assume that readout noise is Gaussian with some pixel-specific offset o_k and

variance σ_k^2 i.e. $W_k \sim \mathcal{N}(o_k, \sigma_k^2)$. Ultimately we write out distributions of the variable we observe H_k which has units of ADU and is a sum of shot noise and readout noise. A fundamental result in probability theory is that the distribution of H_k is the convolution of the distributions of S_k and W_k,

$$\begin{aligned} P(H_k|\theta) &= P(S_k) \circledast P(W_k) \\ &= A \sum_{q=0}^{\infty} \frac{1}{q!} e^{-\mu_k} \mu_k^q \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(H_k - g_k q - o_k)}{2\sigma_k^2}} \end{aligned}$$

In practice, this expression is difficult to work with, so we look for an analytical approximation. For sufficiently large μ_k, the Poisson distribution is well-approximated by a normal distribution $P(\mu_k) \approx \mathcal{N}(\mu_k, \mu_k)$. Under the conditions that this approximation is valid, $P(H_k)$ is easily calculated, because the sum of two Gaussian variables is another Gaussian variable:

$$P(H_k|\theta) \approx \frac{1}{\sqrt{2\pi(\lambda_k \tau + \sigma_k^2)}} e^{-\frac{(H_k - \lambda_k \tau - o_k)}{2(\lambda_k \tau + \sigma_k^2)}} \tag{1.4}$$

Assuming each pixel is an independent random variable, then the likelihood of the dataset (or equivalently the joint distribution on pixels) is simply a product $\mathcal{L}(H|\theta) = \prod_k P(H_k|\theta)$. Now suppose we are given a sample from $\mathcal{L}(H|\theta)$ with unknown θ. A general task in Bayesian inference is to determine θ from the data under the model $\mathcal{M}$. We may then ask - does the likelihood $\mathcal{L}$ vary as we vary the parameters? If the likelihood is flat, all parameter sets are equally likely and the data does not appear to carry much information about the parameters. Moreover, if $\mathcal{L}$ has a number of bumps or inflection points, then we expect that maybe some parameter sets are more likely that others. The "bumpiness" of the likelihood surface is called the Fisher information - a fundmantal concept in information geometry. The Fisher

information matrix $I(\theta)$ can be directly related to the curvature of the KL-Divergence over the parameter space

$$\begin{aligned}
\nabla^2_{\theta'} D_{KL}[\mathcal{L}(H|\theta) \parallel \mathcal{L}(H|\theta')] &= -\nabla_{\theta'} \int \mathcal{L}(H|\theta) \nabla_{\theta'} \log \mathcal{L}(H|\theta')\, dH \\
&= -\int \mathcal{L}(H|\theta) \nabla^2_{\theta'} \log \mathcal{L}(H|\theta')\, dH \\
&= -\mathbb{E}_\theta[\nabla^2_{\theta'} \log \mathcal{L}(H|\theta')] \\
&= \mathbf{I}(\theta)
\end{aligned}$$

We often call the Hessian matrix the *score*. The Fisher information is the result of averaging the score over the parameter space. To be clear, the score is a function of the *data*, not the parameters. It is a measure of sensitivity of the likelihood to changes in the parameters. Of course, the Fisher information matrix also depends on the parameterization chosen. Looking at (1.4), the likelihood is a hierarchical function that maps a vector space Θ to a vector space Λ to a scalar value. Formally, we define $T : \Theta \to \Lambda$ and $W : \Lambda \to \mathbb{R}$. The parameter vector $(x_0, y_0, \sigma) \in \Theta$, the Poisson rate vector $\boldsymbol{\lambda} \in \Lambda$ and $\mathcal{L} \in \mathbb{R}$. To get the Hessian, we need the chain-rule for Hessian matrices.

$$\mathbf{H}_{(L,\theta)} = \mathbf{J}^T_{(\lambda,\theta)} \mathbf{H}_{(L,\lambda)} \mathbf{J}_{(\lambda,\theta)} + (J_{(L,\lambda)} \otimes I_n) \mathbf{H}_{(\lambda,\theta)}$$

In the second term of the equation, we have a Kronecker product between the Jacobian matrix of the likelihood with respect to the parameters of the hierarchical model $(J_{(L,\lambda)})$ and the $n \times n$ identity matrix (I_n), denoted as $J_{(L,\lambda)} \otimes I_n$. This Kronecker product gives a diagonal matrix where the elements along the diagonal are the elements of the vector $J_{(L,\lambda)}$. This provides a concise way of summarizing the operation:

$$((J_{(L,\lambda)} \otimes I_n)\mathbf{H}_{(\lambda,\theta)})_{ij} = \sum_k \mathbf{J}^{ij}_{(L,\lambda)}\mathbf{H}^{ijk}_{(\lambda,\theta)}$$

Note that we sum over the last index to get a 3 x 3 matrix, which is not directly obvious in the compact notation.

1.1.2 Fisher information for Poisson random variables

Suppose we think of a set of image pixels as a family of independent Poisson random variables each with rate μ_k. The joint distribution is simply

$$\mathcal{L}(n_1, n_2, ..., n_N) = \prod_k \frac{e^{-\mu_k}\mu_k^n}{n!}$$

Therefore, the log-likelihood is

$$\ell(n_1, n_2, ..., n_N) = \sum_k n_k \log \mu_k - \log n_k! - \mu_k$$

The Jacobian in the first sum consists of two elements:

$$\frac{\partial \mu_k(x_i, y_i)}{\partial x_0} = \mu_y(x_i, y_i)\sqrt{\frac{2}{\pi\sigma^2}}\left(e^{-\frac{\left(-\frac{a}{2}+x_k-\mathrm{x0}\right)^2}{2\sigma^2}} - e^{-\frac{\left(\frac{a}{2}+x_k-\mathrm{x0}\right)^2}{2\sigma^2}}\right)$$
$$\frac{\partial \mu_k(x_i, y_i)}{\partial y_0} = \mu_x(x_i, y_i)\sqrt{\frac{2}{\pi\sigma^2}}\left(e^{-\frac{\left(-\frac{a}{2}+y_k-\mathrm{y0}\right)^2}{2\sigma^2}} - e^{-\frac{\left(\frac{a}{2}+y_k-\mathrm{y0}\right)^2}{2\sigma^2}}\right)$$

The four elements of the Hessian matrix in the first sum is

$$\frac{\partial \mu_k^2}{\partial x_0^2} = \mu_y(x_i, y_i) \left(\frac{\sqrt{\frac{2}{\pi}} \left(-\frac{a}{2} + x - \mathrm{x0}\right) e^{-\frac{\left(-\frac{a}{2}+x-\mathrm{x0}\right)^2}{2\sigma^2}}}{\sigma^3} - \frac{\sqrt{\frac{2}{\pi}} \left(\frac{a}{2} + x - \mathrm{x0}\right) e^{-\frac{\left(\frac{a}{2}+x-\mathrm{x0}\right)^2}{2\sigma^2}}}{\sigma^3} \right)$$

$$\frac{\partial \mu_k^2}{\partial y_0^2} = \mu_x(x_i, y_i) \left(\frac{\sqrt{\frac{2}{\pi}} \left(-\frac{a}{2} + y - \mathrm{y0}\right) e^{-\frac{\left(-\frac{a}{2}+y-\mathrm{y0}\right)^2}{2\sigma^2}}}{\sigma^3} - \frac{\sqrt{\frac{2}{\pi}} \left(\frac{a}{2} + y - \mathrm{y0}\right) e^{-\frac{\left(\frac{a}{2}+y-\mathrm{y0}\right)^2}{2\sigma^2}}}{\sigma^3} \right)$$

$$\begin{aligned} \frac{\partial \mu_k^2}{\partial x_0 \partial y_0} &= \frac{\partial \mu_k^2}{\partial x_0 \partial y_0} \\ &= \frac{\tau}{4} \left(\frac{\sqrt{\frac{2}{\pi}} e^{-\frac{\left(-\frac{a}{2}+x-\mathrm{x0}\right)^2}{2\sigma^2}}}{\sigma} - \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{\left(\frac{a}{2}+x-\mathrm{x0}\right)^2}{2\sigma^2}}}{\sigma} \right) \left(\frac{\sqrt{\frac{2}{\pi}} e^{-\frac{\left(-\frac{a}{2}+y-\mathrm{y0}\right)^2}{2\sigma^2}}}{\sigma} - \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{\left(\frac{a}{2}+y-\mathrm{y0}\right)^2}{2\sigma^2}}}{\sigma} \right) \end{aligned}$$

The element of the Jacobian in the first sum is

$$\frac{\partial \ell}{\partial \mu_k} = n_k / \mu_k(x_i, y_i) - 1$$

The Hessian matrix element in the second sum is

$$\frac{\partial^2 \ell}{\partial \mu_k^2} = -n_k / \mu_k(x_i, y_i)^2$$

1.1.3 Gaussian case

If the pixels are just Gaussian variables, the joint distribution $\mathcal{L}$ is really just a multivariate normal distribution:

$$\mathcal{L}(H_1, H_2, ..., H_n) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp -\frac{1}{2}(H-\mu)^T \Sigma^{-1}(H-\mu)$$

The Fisher information matrix when we are dealing with the multivariate normal distribution with parameters (μ, Σ) which are themselves functions of another vector of parameters $\boldsymbol{\theta}$ has been well-studied (Malago and Pestone; 2015)

CHAPTER 2

INFERRING DNA STRUCTURE FROM SUPER-RESOLUTION MICROSCOPY IMAGES

CHAPTER 3

DISCRETE MODELING OF TRANSCRIPTIONAL KINETICS

3.1 The ergodic theorem

3.2 The chemical master equation

The central assumption underlying a Markov process, is the memoryless property

$$P(X_t|X_{t-1}, X_{t-2}, ..., X_{t-N}) = P(X_t|X_{t-1})$$

A single Markov chain is the set of states $\boldsymbol{X} = \{X_1, X_2, ..., X_N\}$. Such a set can be generated provided that $P(X_t|X_{t-1})$ is known. To capture $P(X_t|X_{t-1})$ for all possible pairs X_t and X_{t-1}, we define a square transition matrix $T \in \mathcal{R}^{N\times N}$ where $N = |\Omega|$. As such, the elements of T represent the probability of a transition from a state ω_j to ω_i in a unit time

$$T_{ij} = \Pr\left(X_t = \omega_i, \mid X_{t-1} = \omega_j\right)$$

Under these definitions, the row T_i represents the present time, and is a conditional probability distribution $P(\omega|X_{t-1} = \omega_j)$ which requires that

$$\sum_j T_{ij} = \sum_j P(X_t = \omega_j|X_{t-1} = \omega_i) = 1$$

The matrix T is not necessarily symmetric $T_{ij} \neq T_{ji}$. One should note that the columns T_j *do not* define a probability distribution $P(X_t = \omega_i|X_{t-1} = \omega_j)$ and therefore do not necessarily sum to unity. The probability $P(X_t = \omega_i|X_{t-1} = \omega_j)$ has no meaning in this context, since we have defined the rows to represent a probability of the future given the present. We simply sample $X_t \sim P(X_t = \omega_j|X_{t-1} = \omega_i)$, assign $i = j$, and repeat. It directly follows from the fundamental rules of probability, the first order dynamics for a

particular state ω_i: $P(\omega_i, t)$ is given by

$$P(\omega_i, t + dt) = P(\omega_i, t) + \mathcal{J}_i dt \tag{3.1}$$

The net probability current $\mathcal{J}_i$ must be

$$\mathcal{J}_i = \sum_i T_{ij} P(\omega_j, t) - \sum_j T_{ij} P(\omega_i, t)$$

The first is a sum on a column and the second a sum on a row. This can be simplified further by noticing that the normalization condition implies

$$\begin{aligned} T_{ij} &= 1 - \sum_j T_{ij}(1 - \delta_{ij}) \\ &= 1 - \sum_j T_{ij} + \sum_j T_{ij}\delta_{ij} \end{aligned}$$

$$\begin{aligned} \mathcal{J}_i &= \sum_i T_{ij} P(\omega_j, t) - \sum_j T_{ij} P(\omega_i, t) \\ &= \sum_i \left(1 - \sum_j T_{ij} + \sum_j T_{ij}\delta_{ij} \right) P(\omega_j, t) - \sum_j T_{ij} P(\omega_i, t) \\ &= |\Omega| - |\Omega| + \sum_i \sum_j T_{ij} P(\omega_j, t)\delta_{ij} - \sum_j T_{ij} P(\omega_i, t) \\ &= \sum_j T_{ji} P(\omega_j, t) - T_{ij} P(\omega_i, t) \end{aligned}$$

Notice that the Kronecker delta effectively just swaps the index. Taking the limit of (1.1), we arrive at the **master equation**

$$\frac{\partial P(\omega_i)}{\partial t} = \sum_j T_{ji} P(\omega_j, t) - T_{ij} P(\omega_i, t)$$

It is common to then define an operator $\boldsymbol{W}$ s.t. $W_{ij} = T_{ij}$ and $W_{ii} = -\sum_j T_{ij}$

$$\frac{dP(\omega_i)}{dt} = \sum_j W_{ij} P(\omega_j) \rightarrow \frac{dP(\boldsymbol{\omega})}{dt} = \mathcal{J}(\boldsymbol{\omega}) = \mathbf{W} P(\boldsymbol{\omega})$$

This operator form has a solution in terms of a matrix exponential

$$P(\boldsymbol{\omega}, t) = \exp(\mathcal{J}(\boldsymbol{\omega}))$$

This matrix exponential is intractable for large $|\Omega|$. However, in the Finite State Projection algorithm, it is possible to truncate the state space $\Omega \rightarrow \tilde{\Omega}$ and obtain good estimates $\tilde{P}(\boldsymbol{\omega}, t)$ with some certificate of accuracy.

3.3 ODE model for the ensemble average

We can define the following system of ODEs for a autorepressive gene circuit

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \frac{\beta_m}{1 + (p/k)^n} - \gamma_m m, \tag{3.2}$$

$$\frac{\mathrm{d}r}{\mathrm{d}t} = \beta_r m - \gamma_r r, \tag{3.3}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \beta_p r - \gamma_p p, \tag{3.4}$$

We can greatly reduce the number of parameters by nondimensionalizing the system. To that end, we define a characteristic time scale t_0 and characteristic "length" scales for

each variable: m_0, r_0, p_0. The choice of these characteristic scale will become apparent after writing the nondimensionalized ODEs out for a general case. The derivatives transform as

$$\frac{dm}{dt} \to \frac{m_0}{t_0}\frac{dm'}{d\tau}, \quad \frac{dr}{dt} \to \frac{r_0}{t_0}\frac{dr'}{d\tau}, \quad \frac{dp}{dt} \to \frac{p_0}{t_0}\frac{dp'}{d\tau}$$

Our general system of nondimensionalized ODEs reads:

$$\begin{aligned}
\frac{\mathrm{d}m'}{\mathrm{d}\tau} &= \frac{t_0\beta_m}{m_0(1+(p_0p'/k)^n)} - \gamma_m t_0 m' \\
\frac{\mathrm{d}r'}{\mathrm{d}\tau} &= \frac{\beta_r t_0 m_0 m'}{r_0} - \gamma_r t_0 r' \\
\frac{\mathrm{d}p'}{\mathrm{d}\tau} &= \frac{\beta_p t_0 r_0 r'}{p_0} - \gamma_p t_0 p'
\end{aligned}$$

Define the characteristic scales as: $t_0 = 1/\gamma_m$, $m_0 = \gamma_m^2 k/\beta_r\beta_p$, $r_0 = \gamma_m k/\beta_p$, and $p_0 = k$. Making these substitutions, we have

$$\begin{aligned}
\frac{\mathrm{d}m'}{\mathrm{d}\tau} &= \frac{\beta}{(1+(p')^n)} - m' \\
\frac{\mathrm{d}r'}{\mathrm{d}\tau} &= m' - \frac{\gamma_r}{\gamma_m} r' \\
\frac{\mathrm{d}p'}{\mathrm{d}\tau} &= r' - \gamma p'
\end{aligned}$$

3.3.1 Telegraph model of gene expression

We will begin by writing the the system of ODEs describing the dynamics of the first moment

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \beta_m(1-m) - \gamma_m m \tag{3.5}$$
$$\frac{\mathrm{d}r}{\mathrm{d}t} = \beta_r m - \gamma_r r \tag{3.6}$$
$$\tag{3.7}$$

This system has 4 parameters, making it difficult to directly visualize interesting relationships between parameterization and dynamics. Therefore, we will nondimensionalize this system

$$\frac{\mathrm{d}m'}{\mathrm{d}t} = \frac{t_0}{m_0}\beta_m - \frac{t_0}{m_0}\beta_m m_0 m' - \frac{t_0}{m_0}\gamma_m m_0 m' \tag{3.8}$$
$$\frac{\mathrm{d}r'}{\mathrm{d}t} = \frac{t_0}{m_0}\beta_r m_0 m' - \frac{t_0}{m_0}\gamma_r r_0 r' \tag{3.9}$$
$$\tag{3.10}$$

Let $t_0 = 1/\gamma_m$, $m_0 = \beta_m/\gamma_m$, $r_0 = \beta_m\beta_r/\gamma_m^2$, $\gamma = \gamma_r/\gamma_m$, $\beta = \beta_m/\gamma_m$, and we have

$$\frac{\mathrm{d}m'}{\mathrm{d}t} = 1 - \beta m' - m' \tag{3.11}$$
$$\frac{\mathrm{d}r'}{\mathrm{d}t} = m' - \gamma r' \tag{3.12}$$
$$\tag{3.13}$$

Appendices

APPENDIX A

DERIVATION OF THE FOKKER PLANCK EQUATION

A.0.1 Variational Bayes

Variational inference attempts to approximate the true posterior distribution $p(\theta|x)$ with a variational distribution $q(\theta)$, which we assume to be Gaussian. We try to minimize the KL-divergence between the variational distribution and the true posterior:

$$\begin{aligned}
D_{\mathrm{KL}}(q(\theta)||p(\theta|x)) &= \mathop{\mathbb{E}}_{\theta\sim q(\theta)}\left(\log\frac{q(\theta)}{p(\theta|x)}\right) \\
&= \mathop{\mathbb{E}}_{\theta\sim q(\theta)}\left(\log\frac{q(\theta)p(x)}{p(\theta,x)}\right) \\
&= \log p(x) + \mathop{\mathbb{E}}_{\theta\sim q(\theta)}\left(\log\frac{q(\theta)}{p(x|\theta)p(\theta)}\right) \\
&= \log p(x) + \mathop{\mathbb{E}}_{\theta\sim q(\theta)}\left(\log q(\theta) - \log p(x|\theta) - \log p(\theta)\right) \\
&= \log p(x) + H(q) - \mathop{\mathbb{E}}_{\theta\sim q(\theta)}\left(\log p(x|\theta) + \log p(\theta)\right)
\end{aligned}$$

Clearly the KL-divergence is minimized by minimizing this expectation. It is common to define the evdience lower bound (ELBO).

$$\mathcal{L}(x,\theta) = -\mathop{\mathbb{E}}_{\theta\sim q(\theta)}\left(\log q(\theta) - \log p(x|\theta) - \log p(\theta)\right)$$

It is name such because

$$\log p(x) = D_{\mathrm{KL}}(q(\theta)||p(\theta|x)) + \mathcal{L}(x,\theta) \geq \mathcal{L}(x,\theta)$$

The ELBO can be minimized when the variational distribution is easy to sample from,

for example a multivariate normal distribution and when the likelihood is tractable. The gradients of the ELBO

$$\nabla_\Phi \mathcal{L}(x,\theta) = -\mathop{\mathbb{E}}_{\theta\sim q(\theta)}(\log q(\theta) - \log p(x|\theta) - \log p(\theta))$$

multivariate Gaussian distribution:

$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$

Where:

x is a k-dimensional random variable μ is the mean vector Σ is the covariance matrix k is the dimension of x And the gradient with respect to mean:

$\frac{\partial}{\partial\mu}\log p(x|\mu,\Sigma) = -\Sigma^{-1}(x-\mu)$

And the gradient with respect to covariance matrix:

$\frac{\partial}{\partial\Sigma}\log p(x|\mu,\Sigma) = -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}(x-\mu)(x-\mu)^T\Sigma^{-1}$

You can use these equations to calculate the gradient of the log-multivariate Gaussian distribution with respect to the mean and covariance matrix.

A.0.2 Kramers-Moyal Expansion

Gixen many instantiations of a stochastic xariable x, we can construct a normalize histogram oxer all obserxations as a function of time $P(x,t)$. Howexer, in order to systematically explore the relationship between the parameterization of the process and $P(x,t)$ we require an expression for $\dot{P}(x,t)$. If we make a fundamental assumption that the exolution of $P(x,t)$ follows a Markox process i.e. its exolution has the memoryless property, then we can write

$$P(x',t) = \int T(x',t|x,t-\tau)P(x,t-\tau)dx \tag{A.1}$$

which is known at the Chapman-Kolmogorox equation. The factor $T(x',t|x,t-\tau)$ is

known as the *transition operator* in a Markox process and determines the exolution of $P(x,t)$ in time. We proceed by writing $T(x',t|x,t-\tau)$ in a form referred to as the Kramers-Moyal expansion

$$\begin{aligned} T(x',t|x,t-\tau) &= \int \delta(u-x')T(u,t|x,t-\tau)du \\ &= \int \delta(x+u-x'-x)T(u,t|x,t-\tau)du \end{aligned}$$

If we use the Taylor expansion of the δ-function

$$\delta(x+u-x'-x) = \sum_{n=0}^{\infty} \frac{(u-x)^n}{n!} \left(-\frac{\partial}{\partial x}\right)^n \delta(x-x')$$

Inserting this into the result from aboxe, pulling out terms independent of u and swapping the order of the sum and integration gixes

$$T(x',t|x,t-\tau) = \sum_{n=0}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x}\right)^n \delta(x-x') \int (u-x)^n T(u,t|x,t-\tau)du \tag{A.2}$$

$$= \sum_{n=0}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x}\right)^n \delta(x-x') M_n(x,t) \tag{A.3}$$

noticing that $M_n(x,t) = \int (u-x)^n T(u,t|x,t-\tau)du$ is just the nth moment of the transition operator T. Plugging (2.6) back in to (2.4) gixes

$$P(x,t) = \int \left(1 + \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x}\right)^n M_n(x,t)\right) \delta(x - x') P(x, t-\tau) dx \tag{A.4}$$

$$= P(x', t-\tau) + \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x}\right)^n [M_n(x,t) P(x,t)] \tag{A.5}$$

Approximating the derixatixe as a finite difference and taking the limit $\tau \to 0$ gixes

$$\dot{P}(x,t) = \lim_{\tau \to 0} \left(\frac{P(x,t) - P(x, t-\tau)}{\tau}\right) \tag{A.6}$$

$$= \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x}\right)^n [M_n(x,t) P(x,t)] \tag{A.7}$$

which is formally known as the Kramers-Moyal (KM) expansion. The Fokker-Planck equation is a special case of (2.10) where we neglect terms $n > 2$ in the *diffusion approximation*.

Consider the following Ito stochastic differential equation

$$d\vec{x} = F(\vec{x}, t) + G(\vec{x}, t) dW$$

The SDE gixen aboxe corresponds to the Kramers-Moyal expansion (KME) of a transition density $T(x', t'|x, t)$ see (Risken 1989) for a full derixation.

$$\frac{\partial P(x,t)}{\partial t} = \sum_{n=1}^{\infty} \frac{1}{n!} \left(-\frac{\partial}{\partial x}\right)^n [M_n(x,t) P(x,t)] \tag{A.8}$$

where M_n is the nth moment of the transition density. In the diffusion approximation,

the KME becomes the Fokker-Planck equation (FPE) (Risken 1989). For the sake of demonstration, consider the univariate case with random variable x and the form of $T(x', t'|x, t)$ is a Gaussian with mean $\mu(t)$ and variance $\sigma^2(t)$. In this scenario, the FPE applies because $M_n = 0$ for all $n > 2$. Given that the drift $M_1(x, t) = \mu(t)$ and the diffusion $M_2(x, t) = \sigma^2(t)$, the FPE reads

$$\frac{\partial P(x,t)}{\partial t} = \left(-\frac{\partial}{\partial x} M^{(1)}(t) + \frac{1}{2} \frac{\partial^2}{\partial x^2} M^{(2)}(t) \right) P(x,t) \tag{A.9}$$

We can additionally define the term in parentheses as a differential operator acting on $P(x, t)$

$$\hat{\mathcal{L}}_{FP} = \left(-\frac{\partial}{\partial x} M^{(1)}(t) + \frac{1}{2} \frac{\partial^2}{\partial x^2} M^{(2)}(t) \right) \tag{A.10}$$

It is common to additionally define the probability current $J(x, t)$ as

$$J(x,t) = \left(M^{(1)}(t) - \frac{1}{2} \frac{\partial}{\partial x} M^{(2)}(t) \right) P(x,t) \tag{A.11}$$

This definition provides some useful intuition. The value of $J(x, t)$ is the net probability flux into the interval between x and $x + dx$ at at time t. This also allows us to write the FPE as a continuity equation

$$\frac{\partial P(x,t)}{\partial t} = -\frac{\partial J(x,t)}{\partial x} \tag{A.12}$$

3.1 Poisson processes

The Poisson process can be derived very quickly by noticing that it is simply the continuous-time limit of the Binomial distribution.

$$B(m;t) = \binom{n}{m} \lambda^m (1-\lambda)^{n-m}$$

Since λ is the fraction of successes, the expected number of successes is $\mu = n\lambda$

$$\begin{aligned} B(m;t) &= \binom{n}{m} \left(\frac{\mu}{n}\right)^m \left(1-\frac{\mu}{n}\right)^{n-m} \\ &= \binom{n}{m} \left(\frac{\mu}{n}\right)^m \left(1-\frac{\mu}{n}\right)^{n} \left(1-\frac{\mu}{n}\right)^{-m} \end{aligned}$$

$$\begin{aligned} B(m;t) &= \frac{n!}{m!(n-m)!} \left(\frac{\mu}{n}\right)^m \left(1-\frac{\mu}{n}\right)^{n} \left(1-\frac{\mu}{n}\right)^{-m} \\ &= \frac{n!}{(n-m)!} \left(\frac{1}{n}\right)^m \left(1-\frac{\mu}{n}\right)^{-m} \frac{\mu^m \left(1-\frac{\mu}{n}\right)^n}{m!} \end{aligned}$$

In the first term, we can take the first m subterms of the numerator $n! = n(n-1)...(n-m)$ and, since $n >> m$, each term will cancel with one factor of n from the term $1/n^m$. This leaves

$$B(m;t) = \frac{(n-m)!}{(n-m)!}\left(1-\frac{\mu}{n}\right)^{-m}\frac{\mu^m \exp(-\mu)}{m!}$$

We now take the continuous time limit i.e. $n \to \infty$ and, again, since $m << n$ we are left with

$$\lim_{n\to\infty} \; B(m;t) = \frac{\mu^m \exp(-\mu)}{m!}$$

If an event can be detected will probability γ, the rate of the Poisson process will be reduced by that factor i.e., $\lambda' = \gamma\lambda$. Therefore, the mean and variance of the process becomes $\mu = \gamma\lambda\Delta t$