Influence of Gene Expression Gradients on Positional Information Content in Fly Embryos

by

Alasdair Hastewell

Submitted to the Department of Physics
in partial fulfillment of the requirements for the degree of

Bachelor of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Signature redacted

Author .....
Department of Physics
May 11, 2018

Signature redacted

Certified by .....
Professor Jörn Dunkel
Department of Mathematics
Thesis Supervisor

Signature redacted

Accepted by .....
Professor Scott Hughes
Interim Physics Associate Head, Department of Physics

Influence of Gene Expression Gradients on Positional Information Content in Fly Embryos

by

Alasdair Hastewell

Submitted to the Department of Physics
on May 11, 2018, in partial fulfillment of the
requirements for the degree of
Bachelor of Science

Abstract

The concept of positional information was introduced to qualitatively explain how individual cells are involved in forming patterns. Recent experimental and theoretical developments have made studying specific biological systems in a quantitative manner possible using the framework of positional information. Much previous work has focused on using the full gene expression profiles when calculating the available positional information. In an attempt to simplify the model a discretized version, where the gene expression profiles are simplified to a binary system, was proposed. Binarizing, however, results in a significant loss of information over using the full profiles. The question remains how coarsely can we discretize the full model without losing essential positional information.

Recent work has shown the importance of concentration gradients in impacting the folding of proteins during embryonic development. Based on this work we posit that the gradients of gene profiles might be an important addition to the discretized model. Using data provided by the Gregor lab at Princeton University we test this hypothesis on the gap gene network of *Drosophilia* embryos. In order to implement the addition of gradients to the positional information requires producing an algorithm that can efficiently take meaningful derivatives of noisy data, which is done using Chebyshev interpolation. An adaptation of Monte Carlo methods to find maxima of multidimensional functions is also implemented.

We find that the derivatives can account for over one bit of the information lost by the discretization process. Allowing the cells to locate themselves with an average precision close to one internuclear spacing. This suggests that a binary model using gradients may be almost as efficient as the model that uses the full gene profiles. We propose that a discrete model of positional information that includes gradients does not lose significant information over a model that uses full profiles.

Thesis Supervisor: Professor Jörn Dunkel
Title: Assistant Professor, Department of Mathematics

Acknowledgments

I would like thank Prof. Jörn Dunkel for the many hours of invaluable conversation and advice that went into this thesis. I would also like to thank Prof. Thomas Gregor and his lab at Princeton University for providing the data that were used in this thesis. Finally I would like to thank my parents for all their support over the last four years.

Contents

1 Introduction **13**

2 Mathematical preliminaries **17**

2.1 Positional information: from a qualitative to a quantitative description 17

2.1.1 Why mutual information? 19

2.2 Smoothing noisy data using polynomial interpolation 21

2.2.1 Polynomial interpolation 21

2.2.2 Chebyshev polynomials 22

2.2.3 Chebyshev interpolation 24

2.2.4 Using Chebyshev polynomials to smooth noisy data 27

2.2.5 Finding derivatives using Chebyshev polynomials 28

2.3 Markov chain Monte Carlo maximization 29

3 Application to fruit fly embryos **33**

3.1 The model organism and experimental results 33

3.2 Initial data processing 34

3.3 Smoothing the data 35

3.4 Calculating the thresholded mutual information 37

3.4.1 Including derivatives in the calculation of positional information 39

3.5 Interpreting the amount of information available 40

3.6 Time series and mutant data 41

4 Conclusions **47**

List of Figures

2-1 Graphical representation of how we can use an asymmetrical profile to divide a line into three sections using the ideas of positional information. The asymmetrical profile allows for the definition of thresholds that can be used to determine position. The power of the interpretation step is that based on how we interpret the profile we can get multiple "flags". 18

2-2 MCMC results using the absolute value of the derivative profiles. Left: zoomed in plot of the initial 5000 runs. Center: Full MCMC run. Right: zoomed in plot around the maximum value obtained. The maximum point is indicated with a red triangle. It can be seen that the algorithm frequently decreases before increasing to a larger maximum. This illustrates the importance of allowing the algorithm to occasionally accept a smaller value to escape a local maximum. Eventually it can be seen that the algorithm converges, oscillating between two maxima a local one and the global one. 31

3-1 Scalar product realignment on the kr gap gene profiles for 25 samples all taken in the same time window, 38 to 48 minutes into development. Comparing the figure on the left before optimization with the figure on the right after optimization shows the power of the scalar product to align the respective peaks and produce sharper edges in the data. . 35

3-2 Plot of the maximum value of the information found using MCMC techniques as a function of the threshold fraction that determines the cut-off N_c for the Chebyshev expansion. We can see that the curve begins to flatten off after a threshold fraction of 55. This suggests that at this threshold fraction all the features that are important in the calculation of positional information are retained. We, therefore, use a threshold fraction of 55 to determine the cut-off values for the Chebyshev interpolations when smoothing the data. 36

3-3 Demonstration of the smoothing process using truncated Chebyshev approximations. It can be seen that the final smoothed profile in the middle maintains most of the key features of the original profile on the left but that the noise has been significantly reduced. The figure on the right shows how the cut-off for the Chebyshev interpolation was calculated from the threshold fraction. The blue arrow points to the coefficient with the greatest magnitude and the horizontal red dashed lines shows one 55th of this value. All coefficients not in the upper left corner are ignored in the Chebyshev expansion in all further calculations. 37

3-4 Comparing the numerical derivative of the raw data, calculated using a second order finite difference scheme, on the left with the derivative calculated directly from the truncated Chebyshev approximation on the right shows the importance of removing noise from the data before taking derivatives. The noise adds jerky spikes to the derivative that result in a derivative profile that is not smooth, which means it is not useful in calculating the positional information. 38

3-5 The four Gap gene concentration profiles for WT *Drosophila* embryos. The top row shows the smoothed profiles using Chebyshev interpolation, the middle row shows the derivative profiles calculated directly from the Chebyshev expansion and the bottom row shows the absolute value of the derivative profiles. The dark curve shows the mean value and the lighter curves show the profiles from individual embryos. The lighter curves highlight the embryo to embryo variability that exists between embryos and that can be used to define the probability distribution for the profiles. The bars at the bottom of the figures show the final binary profiles after maximization using the thresholds shown with the dashed lines. It can be seen that the threshold acts to pick out the key peaks in each of the profiles effectively splitting the profiles into high and low regions. 42

3-6 Schematic of how the information in the gene profiles is calculated from the binarized gene profiles. 43

3-7 Resulting binary profiles after maximization using the four gene profiles without the gradients. The upper figure is taken from: Julien O Dubuis, Gačper Tkačik, Eric F Wieschaus, Thomas Gregor, and William Bialek. Positional information, in bits. *Proceedings of the National Academy of Sciences USA*, 110(41) : 1630116308, 2013. The information contained in each individual gene's bit string is shown on the left. The total information found using the the new method presented here is 2.90 bits which is consistent with the 2.92 bits found in the previous work using a different technique. 44

3-8 Resulting binary profiles after maximization using the four gene profiles and the absolute value of the gradients. The bars represent the region where the binary string equals 1. The information contained in the individual strings is shown on the left. Profiles are shown in the background for reference. The total information found is 4.02 bits which is enough to specify the precision along the length of the embryo to around one internuclear spacing. The result is close to the 4.27 bits that there is evidence for experimentally. This provides a significant improvement over the amount contained in the gene profiles alone seen in Figure 3-7 45

3-9 The figure shows how the binary strings evolve in time for each gene, colored coded as in Figure 3-6. The top row is concentrations and the second row the absolute value of the derivatives. The far right shows the number of distinct regions specified by the profiles at each time, the upper figure only using the 4 concentrations and bottom figure including derivatives. We can see new behavior emerges around 38 minutes into embryonic development, especially in kni and hb, which could be the cause of the information stabilizing and often peaking during the 14th nuclear cycle. 46

3-10 Histograms show the maximum information contained in the profiles between 38 and 48 minutes for each of the mutations considered. The far left bar is the WT data. Bcd dosage mutants are represented by 0.5 and 2 for half and double concentration respectively. Maternal mutants are represented by three numbers, 1 indicates presence and 0 represents removal. The corresponding time series information is next to the histogram. The red dashed horizontal lines indicate the 38 to 48 minute time window used to determine the thresholds. We can see that the information mostly peaks in the 38 to 48 minute period and generally decreases as more mutations are implemented. We also see that including the derivative adds information to all of the mutations. 46

Chapter 1

Introduction

The study of how embryos gain their patterns and shapes has been a topic of interest in developmental biology for many years. D'Arcy Thomson, reacting to the increasing rise in the popularity of Darwin's theory of natural selection, in 1917, was one of the first people to use fundamental principles to prove physical limits on possible biological shapes [2]. Thomson showed how simple physical and mathematical techniques could answer questions that would otherwise have required knowledge of complex biological systems [24]. Since then many physicists and mathematicians have tried to use physical reasoning to explain complex biological systems. Erwin Schrödinger [21] famously lay out another framework for considering biological systems in his short book *What is Life?* published in 1944. Schrödinger used concepts from quantum physics and statistical physics to try and understand a broad range of biological systems from consciousness to the hereditary process.

Inspired by Thomson's work, Alan Turing set the theoretical foundation for the field in 1952 when he showed that reaction diffusion equations of the general form,

$$\frac{\partial \mathbf{c}}{\partial t} = \boldsymbol{D}\nabla^2\mathbf{c} + \mathbf{R}(\mathbf{c}), \tag{1.1}$$

where $\boldsymbol{D}$ is the diffusion matrix and $\mathbf{R}(\mathbf{c})$ is a vector function of the local concentrations, could form patterned solutions [27]. As discussed in Green and Sharpe, Turing's results are counter intuitive since diffusion is normally associated with the destruc-

tion of peaks, which should lead to spatially homogeneous profiles [8]. The power of Turing's work was that it showed, only assuming basic properties of chemical reactions, that a spatially symmetric profile could evolve into an asymmetric one . Turing introduced the word morphogen to describe a chemical substance that mediates the production of a pattern. Turing's paper had little impact, however, until advances in experimental and computational techniques made detailed studies of gene profiles and the chemical reactions that govern them during embryonic development possible [7, 15]

Over a decade later, with Turing's work on pattern formation still not largely utilized, a new way of approaching pattern formation was proposed: positional information. First introduced in 1969 by Lewis Wolpert the framework of positional information does not aim to produce patterns from spatially homogeneous profiles. Instead the framework asks how simpler patterns can be used to build more complex ones [28]. The concept of positional information is to take previously existing inhomogeneities, for example a gene concentration profile along the length of an embryo and use these to encode information for producing a new pattern. A major difference between Turing's reaction diffusion systems and Wolpert's positional information is the introduction of an interpretation step in positional information, which in Wolpert's own words leads to "positional information in general preced[ing] and [being] independent of molecular differentiation" [28]. This is not the case in Turing's work where the pattern produced closely follows the pattern of the morphogens [29]. The interpretation step allows for the same positional information profile to ultimately generate different patterns. It is important to note that the concept of positional information does not attempt to answer how many parts of the development process happen, for example how the gradient profiles that positional information relies on are produced and maintained or how the interpretation step is implemented, instead accepting them as given and leaving understanding them to different theories. [29, 30, 3].

While positional information provides a distinct new frame work to interpret pattern formation, it was only used qualitatively for many years and it was not until the last decade that a quantitative definition has been given [25]. With the advent

of more advanced experimental techniques that allow for individual gene expression levels to be measured over time, this new mathematical framework can be applied to specific model organisms to test further the idea of positional information [6, 9, 5, 18]. Indeed Dubuis *et al.* found that, using the full gene expression profiles, the amount of positional information available is consistent with cells being able to self locate with a high degree of precision along the anterior/posterior length of the embryo [6]. They have also extended this work by proposing models that describe how the cells can actually decode this positional information [18, 25].

Another major consideration when considering patterns in biological systems is the reproducibility of the pattern from embryo to embryo and also fluctuations within the embryos themselves. There has been significant theoretical and experimental work on understanding the role of noise and variability in biological systems in recent years [26]. Lestas *et al.* [13], using an information theoretic framework, proved fundamental bounds on the ability of a biological system to suppress variation that changes as the quartic root of the number of signaling events in the control chain. Experimentally, Gregor *et al.* have shown that there is reproducibility in the profiles across *Drosophila* embryos [9]. It has also been shown, however, that around 20% of variability between gap gene expression levels in *Drosophila* embryos cannot be accounted for by experimental error [5]. This means that biological systems while reproducible also contain inherent noise and any mathematical model must be robust to noise or even exploit it. Indeed many systems have been shown to use noise to their advantage, for example in stochastic resonance [10, 17]. The new mathematical model for positional information explicitly take into account the role of variability in determining positional information.

As mentioned, previous work on positional information has utilized the full gene expression levels. There has been some suggestion, however, that cells might not use the full profiles but instead use a thresholding mechanism [6]. With this mechanism a region is considered only to be on or off based on the expression level. This is supported by experimental work in *C. elegans* [23] and by the large role of allosteric induction in biological signaling networks [19]. While potentially simpler to model, as

expected binary models have decreased information available and it has been shown that using a binary threshold on the gap gene expression levels of *Drosophila* embryos cannot account for the required amount of positional information for cells to locate themselves [6]. The question remains what the most discretized model that still maintains a significant amount of the positional information is.

Today the way that people see the relationship between Turing's reaction diffusion systems and Wolpert's positional information is shifting. While once thought of as opposing ideas that could not work in harmony, new work has proposed that the two mechanisms may be able to work in conjunction [8]. The idea has many appealing attributes: reaction diffusion can describe the generation of an asymmetrical system, which has always been a challenge for positional information and positional information brings flexibility and transferability, which reaction diffusion systems lack. Turing's reaction diffusion model, described by equation (1.1), suggests that both local interactions described through the **R** term and non-local interactions characterized by the derivative term are both important for the production of patterns. This indicates that perhaps gradients should also be taken into account when considering positional information.

In fact recent joint experimental and theoretical work between the groups of Adam Martin and Jörn Dunkel has shown the importance of gradients for the creation of ventral furrows in *Drosophila* embryos [11]. Based on this work the importance of gradients has also been noted in recent continuum model simulations of embryogenisis [12]. Inspired by these findings and equation (1.1) we posit that gradients might be able to account for some of the information that is lost by binarizing the gene profiles. Using data provided by Thomas Gregor's lab at Princeton university we can test this hypothesis on the gap genes of *Drosophila* embryos using the recently developed mathematical framework of positional information.

Chapter 2

Mathematical preliminaries

2.1 Positional information: from a qualitative to a quantitative description

A simple example of how positional information works is the French flag problem first introduced by Wolpert [28]. The problem can be formulated as follows: consider a one dimensional system which we want to divide into three equal sections each having a different fate. Here the fates are the colors of the French flag (red, white and blue). We assume that we have a monotonically decreasing concentration gradient caused by some unspecified mechanism in the problem. We can define a co-ordinate system based on this gradient using the left edge as the reference point and using the slope of the line to tell us the polarity (direction from the reference point). A negative slope means that we are moving left to right and a positive slope means that we are moving right to left. In the French flag problem we need the coordinate system to specify three distinct regions. Therefore we choose two thresholds say T_1 and T_2 such that we can define the coordinate system as,

$$x_{\text{pos}} = \begin{cases} b & \text{if } c > T_1 \\ w & \text{if } T_1 < c < T_2 \\ r & \text{if } c < T_2 \end{cases} .$$

Clearly this will reproduce the French flag. The role of interpretation is clear to see here as well. By using the redefinition $b \to g$ and $r \to o$ we can get the Irish flag or with $b \to g$ the Italian flag. We demonstrate this whole process graphically in Figure 2-1. Clearly the same profile can be used to generate different patterns that require the same amount of information as each other.

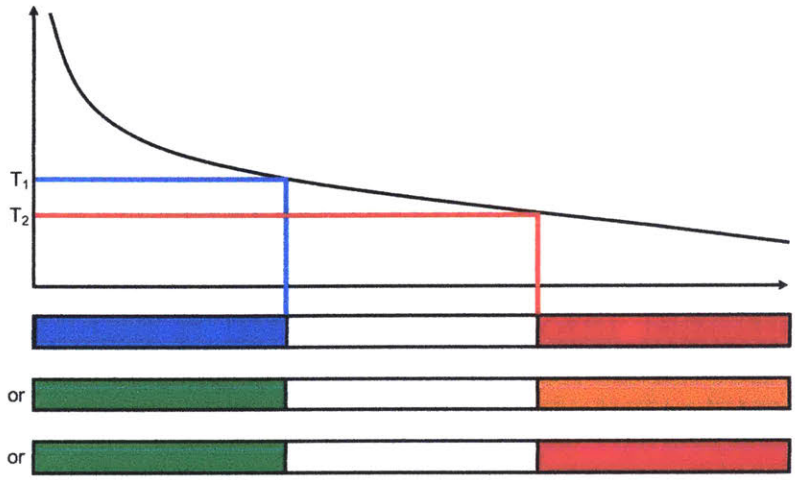

Figure 2-1: Graphical representation of how we can use an asymmetrical profile to divide a line into three sections using the ideas of positional information. The asymmetrical profile allows for the definition of thresholds that can be used to determine position. The power of the interpretation step is that based on how we interpret the profile we can get multiple "flags".

In computer science and statistical physics information is normally measured in terms of entropy introduced by Shannon in 1948 [22]. For every binary decision ('on' or 'off') we are able to make we associate one bit of information. If we had only chosen one threshold then we would have split the domain in half which corresponds to one bit of information; if, however, we were able to distinguish four total regions then we would have a total of two binary decisions or 2 bits of information – we can think of each of the four regions as being represented by a two bit string 00, 01, 10 or 11. In general if we have I bits of information then the are able to make out 2^I distinct regions. As will be helpful later we can also reformulate this in the following way, given that we want to make N decisions then we need $I = \log_2(N)$ bits of information.

In reality, however, as discussed in the introduction, there is variation in the distribution that can limit the total amount of information available. For example two states that differ by less than the variation would be indistinguishable. There may also be more than one concentration profile that contributes to the total amount of information. In this case we also have to consider the correlations between the profiles. Perfectly correlated profiles would provide no new information making the second profile redundant. As is discussed by Tkačik *et al.* combining all of these factors into a single number that can be defined as the positional information leads to a formulation using mutual information [25].

2.1.1 Why mutual information?

The aim is to transform the qualitative description above into a mathematical one. We can do this by defining the problem probabilistically. We consider a system where N genes are of interest with gene expressions $g_i(x)$ that are functions of position. Then due to the variability in expression we can think of the profiles as being drawn from a probability distribution for $\mathbf{g}$ at a given value of x, $P(\mathbf{g}|x)$, where $P(A|B)$ represents the conditional probability of A given B defined as, $P(A|B) = P(A,B)/P(B)$. Experimentally we can produce from simultaneous measurements an N dimensional histogram that gives us the probability of the total gene expression $\mathbf{g}$ at a given x. From this probability distribution we can associate with each gene a mean value $\bar{g}_i(x)$, a variance $\sigma_i^2(x)$ and between genes we can calculate a covariance matrix. These are defined mathematically as,

$$\bar{g}_i(x) = \int g_i P(\mathbf{g}|x)\, \mathrm{d}^N\mathbf{g} \tag{2.1}$$

$$\sigma_i^2(x) = \int (g_i - \bar{g}_i)^2 P(\mathbf{g}|x)\, \mathrm{d}^N\mathbf{g} \tag{2.2}$$

$$C_{ij}(x) = \int (g_i g_j - \bar{g}_i\bar{g}_j)^2 P(\mathbf{g}|x)\, \mathrm{d}^N\mathbf{g}\,. \tag{2.3}$$

It is worth noting that these are all functions of position. Thus $P(\mathbf{g}|x)$ contains all of the information that was discussed in the previous section. Since we are interested in

the information content of this probability distribution a natural number to calculate is the information entropy, the average amount of uncertainty we expect to have given a measurement from the distribution, measured in bits. It is well known that a uniform distribution, $P(x) = 1/L$ – the distribution that intuitively should carry the least amount of information – maximizes entropy to $\log_2(L)$. Mathematically entropy was defined in Shannon's paper as [22],

$$S[p(\mathbf{x})] = -\int \mathrm{d}\mathbf{x}\, p(\mathbf{x}) \log_2(p(\mathbf{x})). \tag{2.4}$$

Returning to our intuitive picture, positional information tells us how much our knowledge increases, or entropy decreases, when we take into account the gene profiles. When we know nothing about the genes the best guess we can make about the position comes from the original distribution of cells $P_x(x)$, which we will often take to be uniform. After taking into account the genes the probability becomes, $P(x|\mathbf{g})$. The gain in knowledge from the gene profiles at a given gene expression level is then,

$$S[P_x(x)] - S[P(x|\mathbf{g})] \tag{2.5}$$

which we can average over $\mathbf{g}$ to find the average total gain in information,

$$I(\mathbf{g} \to x) = \int \mathrm{d}^N\mathbf{g} P_g(\mathbf{g})(S[P_x(x)] - S[P(x|\mathbf{g})]). \tag{2.6}$$

Using the definition of entropy and conditional probability, we can rewrite this as,

$$\begin{aligned} I(\{g_k\} \to x) &= \int \mathrm{d}^N\mathbf{g} P_g(\mathbf{g})(S[P_x(x)] - S[P(x|\mathbf{g})]) \\ &= \int \mathrm{d}x \left[-P_x(x) \log_2(P_x(x)) + \int \mathrm{d}^N\mathbf{g} P_g(\mathbf{g}) P(x|\mathbf{g}) \log_2(P(x|\mathbf{g})) \right] \\ &= \int \mathrm{d}^N\mathbf{g} \int \mathrm{d}x\, P(x,\mathbf{g}) \log_2 \left(\frac{P(x,\mathbf{g})}{P_g(\mathbf{g})} \right) - P(x,\mathbf{g}) \log_2(P_x(x)) \\ &= \int \mathrm{d}^N\mathbf{g} \int \mathrm{d}x\, P(x,\mathbf{g}) \log_2 \left(\frac{P(x,\mathbf{g})}{P_g(\mathbf{g}) P_x(x)} \right) \end{aligned} \tag{2.7}$$

which is symmetric in x and $\mathbf{g}$ and, therefore, we will label the positional information as $I(x, \mathbf{g})$ since the direction of the arrow does not matter. This is an excellent property since in a cell we want to go from gene profiles to position but in experiments we go from position to gene profiles. It does not matter with this definition, however, which way we go when we calculate the positional information. Shannon also proved that this quantity, commonly called mutual information, is the sole quantity that completely quantifies the dependency of probability distributions on each other and satisfies certain properties such as additivity [22, 25]. Mutual information, therefore, makes sense both on information theoretic grounds and intuitively as a mathematical description of positional information.

2.2 Smoothing noisy data using polynomial interpolation

In order to apply the concepts of positional information to experimental data it is important to ensure that as much noise as possible has been removed from the gene expression profiles. This will be especially important when we take derivatives where noise adds arbitrary spikes to the profile. We are not trying to remove the embryo to embryo variability here just identify a smooth underlying profile for each embryo. One way to do this, which then provides simple ways to manipulate the profiles, is with polynomial interpolation.

2.2.1 Polynomial interpolation

We assume that a gene profile can be described by some function $f(x)$, which we assume to be continuous on a region $[a, b]$. The goal is to find an analytic approximation to this function. Experimental data provide us with a list of function values $f(x_i)$ at a discrete number of points x_i for $i = 0, 1, \ldots, N$. Lagrange interpolation guarantees that there exists a polynomial of degree less than or equal to N that goes through all of the $f(x_i)$ and that this polynomial is unique. High order polynomial interpolation,

however, suffers from Runge's phenomenon, which results in large oscillations around the final few interpolation points [20]. Approximations to experimental data that provide on the order of 1000 interpolation points, like the data analyzed in this thesis, are particularly prone to these effects, which leads to highly inaccurate approximations.

There are several ways to avoid this phenomenon. For continuous, twice differentiable, 2π periodic functions, the Fourier series of the form,

$$f(x) \approx \sum_{n=-N}^{N} c_n e^{inx} = \frac{a_0}{2} + \sum_{n=1}^{N} [a_n \cos(nx) + b_n \sin(nx)] \tag{2.8}$$

can be proven to converge to $g(x)$ in the limit as $N \to \infty$. We can approximate the coefficients c_n from a discrete number of equally spaced function values using the discrete Fourier transform [1, 16]. We do not, however, expect that the gene profiles will be periodic and, therefore, cannot naïvely use Fourier transforms on the data.

2.2.2 Chebyshev polynomials

A simple approach can be used to make a function defined on the interval $[-1, 1]$ periodic; we consider instead the function $g(\theta) = f(\cos(\theta))$. If a function is defined on $[a, b]$ then we can use a transformation to convert its domain to $[-1.1]$. The function $g(\theta)$ returns $f(x)$ on the interval $\theta \in [\pi, 0]$ and is periodic on the interval $[0, 2\pi]$ due to the periodicity of cosine. We note that $g(\theta)$ is an even function and, therefore, the b_n coefficients vanish, so that

$$g(\theta) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\theta) \tag{2.9}$$

where we can write the formula for the coefficients a_n as,

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} g(\theta) \cos(n\theta)\, \mathrm{d}\theta = \frac{2}{\pi} \int_{0}^{\pi} g(\theta) \cos(n\theta)\, \mathrm{d}\theta\,. \tag{2.10}$$

The convergence of Fourier transforms then guarantees that the series will converge to $g(\theta)$. Making the transformation $x = \cos(\theta)$ in equations (2.9) and (2.10) yields,

$$g(\arccos(x)) = f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n \arccos(x)) \tag{2.11}$$

where, using $\mathrm{d}x = -\sin(\theta)\,\mathrm{d}\theta = -\sqrt{1-x^2}\,\mathrm{d}\theta$,

$$a_n = \frac{2}{\pi}\int_1^{-1} f(x)\cos(n\arccos(x))\frac{-1}{\sqrt{1-x^2}}\,\mathrm{d}x = \frac{2}{\pi}\int_{-1}^{1}\frac{f(x)\cos(n\arccos(x))}{\sqrt{1-x^2}}\,\mathrm{d}x \tag{2.12}$$

equations (2.11) and (2.12) define the Chebyshev expansion of a function $f(x)$. We define the nth Chebyshev polynomial as,

$$T_n(x) = \cos(n\arccos(x)). \tag{2.13}$$

These polynomials have several nice properties similar to sine and cosine that make them particularly useful to work with. Since cosine is bounded so are the Chebyshev polynomials, $|T_n(x)| < 1$. The Chebyshev polynomials satisfy an orthogonality condition,

$$\begin{aligned}\int_{-1}^{1}\frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\,\mathrm{d}x &= \int_{-1}^{1}\frac{\cos(n\arccos(x))\cos(m\arccos(x))}{\sqrt{1-x^2}}\,\mathrm{d}x \\ &= \frac{1}{2}\int_{-\pi}^{\pi}\cos(ny)\cos(my)\,\mathrm{d}y \\ &= \begin{cases} \pi & \text{for } m = n = 0 \\ \frac{\pi}{2} & \text{for } m = n \neq 0 \\ 0 & \text{for } m \neq n \end{cases}\end{aligned} \tag{2.14}$$

In fact this relationship can be written in a discrete form. Replacing the integral with a summation over the Chebyshev points, $x_l = \cos(l\pi/N)$ for $l = 0, \ldots, N$, we find the

discrete version of the Chebyshev orthogonality condition,

$$\begin{aligned}\sum_{l=0}^{N}{}''T_n(x_l)T_m(x_l) &= \sum_{l=0}^{N}{}'' \cos\left(\frac{nl\pi}{N}\right)\cos\left(\frac{ml\pi}{N}\right) \\ &= \frac{1+(-1)^{n+m}}{2} + \frac{1}{2}\sum_{l=1}^{N-1}\cos\left(\frac{(n-m)l\pi}{N}\right) + \cos\left(\frac{(n+m)l\pi}{N}\right) \\ &= \begin{cases} N & \text{for } n = m = 0, kN \\ \frac{N}{2} & \text{for } n = m + 2kN = 2kN - m \neq kN \\ 0 & \text{for } n \neq m \end{cases} \end{aligned} \tag{2.15}$$

where we use $\sum''$ to denote a summation where the first and last terms are halved and the summation result,

$$\begin{aligned}\sum_{l=1}^{N-1}\cos\left(\frac{nl\pi}{N}\right) &= \mathrm{Re}\sum_{l=0}^{N-1}\exp\left(\frac{inl\pi}{N}\right) - 1 \\ &= \mathrm{Re}\,\frac{1-\exp(in\pi)}{1-\exp\left(\frac{inl\pi}{N}\right)} - 1 \\ &= \begin{cases} -1 & \text{for } n \text{ even} \\ \mathrm{Re}\, i\cot\left(\frac{n\pi}{2N}\right) = 0 & \text{for } n \text{ odd} \\ N-1 & \text{for } n = 0, 2N, 4N, \ldots \end{cases}\end{aligned}$$

If our function is sampled at a finite number of points then we can use the discrete orthogonality relation, equation (2.15), to find an approximation to the Chebyshev coefficients, a_n.

2.2.3 Chebyshev interpolation

The Chebyshev interpolant of $f(x)$ is the Chebyshev polynomial $g(x)$ that equals $f(x_i)$ at every Chebyshev point x_i. We can therefore write the $N+1$ conditions on

the interpolating function as,

$$g(x_l) = f(x_l) = \sum_{n=0}^{\infty}{}' a_m T_m(x_l) \tag{2.16}$$

where a single prime represents a sum with the first term halved. Given a function evaluated at $N+1$ Chebyshev points we will construct the Nth degree Chebyshev interpolation function in the following way,

$$g^N(x) = \sum_{n=0}^{N}{}'' \alpha_n T_n(x) \tag{2.17}$$

where

$$\alpha_n = \frac{2}{N} \sum_{l=0}^{N}{}'' T_n(x_l) f(x_l). \tag{2.18}$$

We can relate α_n to the a_n's using equation (2.16). Writing $f(x_l)$ in terms of its Chebyshev expansion we find,

$$\begin{aligned}
\alpha_n &= \frac{2}{N} \sum_{l=0}^{N}{}'' T_n(x_l) f(x_l) \\
&= \frac{2}{N} \sum_{m=0}^{\infty}{}' a_n \sum_{l=0}^{N}{}'' T_n(x_l) T_m(x_l) \\
&= \bar{c}_n (a_n + a_{2N-n} + a_{2N+n} + a_{4N-n} + a_{4N+n} + \ldots)
\end{aligned} \tag{2.19}$$

where we have used the orthoganality condition to go from the second line to the third line and as a short hand introduced the notation $\bar{c}_n = 1$ when $n \neq 0, N$ and $\bar{c}_n = 2$ when $n = 0, N$. From these we can see that [1],

$$\frac{\alpha_0 - a_0}{2} = a_{2N} + a_{4N} + \ldots \tag{2.20a}$$

$$\alpha_n - a_n = a_{2N-n} + a_{4N-n} + \ldots \tag{2.20b}$$

$$\frac{\alpha_N}{2} - a_N = a_{3N} + a_{5N} + \ldots \tag{2.20c}$$

We can use these relationships to find an expression for the error from using interpolation through finitely many points to approximate a function . We define the error in approximating the true function $f(x)$ using interpolation as,

$$\begin{aligned} E_N(x) &= \left|f(x) - g^N(x)\right| \\ &= \left|{\sum_{n=0}^{\infty}}' a_n T_n - {\sum_{n=0}^{N}}'' \alpha_n T_n(x)\right| \\ &= \left|\frac{a_0 - \alpha_0}{2} + \sum_{n=1}^{N-1}(a_n - \alpha_n)T_n(x) + \left(a_N - \frac{\alpha_N}{2}\right) + \sum_{n=N+1}^{\infty} a_n T_n\right| \\ &\leq \left|\frac{a_0 - \alpha_0}{2}\right| + \sum_{n=1}^{N-1}|a_n - \alpha_n| + \left|a_N - \frac{\alpha_N}{2}\right| + \sum_{n=N+1}^{\infty}|a_n| \\ &\leq 2\sum_{n=N+1}^{\infty}|a_n|. \end{aligned} \tag{2.21}$$

where we have used the triangle inequality and the fact that $|T_n(x)| \leq 1$ to get from the third line to the fourth line and the triangle inequality applied to equations (2.20) to get the last line.

The important consequence of equation (2.21) is that for very simple assumptions on $f(x)$ we can prove that $g^N(x)$ converges uniformly to $f(x)$ so we do not have to worry about Runge phenomena at the end points. Returning to equation (2.10) and making the assumption that we are dealing with a function $f(x)$ that is k times differentiable we can integrate by parts to show that,

$$a_n = \frac{2}{\pi}\int_0^{\pi} g(\theta)\cos(n\theta)\,\mathrm{d}\theta \sim \mathcal{O}\left(\frac{1}{n^k}\right)$$

where we have used the fact that the nth derivative of $g(\theta)$ is periodic in 2π and assumed that the final integral over the nth derivative is finite. This equation justifies why α_n is a good approximation to a_n for suitably smooth functions. Substituting this into equation (2.21) we get that,

$$E_N \sim 2\sum_{n=N+1}^{\infty}|a_n| \sim \mathcal{O}\left(\sum_{n=N+1}^{\infty}\frac{1}{n^k}\right) \sim \mathcal{O}\left(\frac{1}{N^k}\right) \tag{2.22}$$

for large N. The importance of this equation is that by assuming the underlying gene expression profile is suitably smooth we can approximate the function with a high degree of accuracy using Chebyshev interpolation since the coefficients will decay suitably fast. These interpolation functions can be created quickly using the Chebfun package in MATLAB [4].

2.2.4 Using Chebyshev polynomials to smooth noisy data

Experimental data inherently have noise associated with them. The data points that we use will not be for the profile we are interested in, $f(x)$, but the profile and the noise together, $F(x)$,

$$F(x) = f(x) + \eta(x) \tag{2.23}$$

where $\eta(x)$ is the noise added to the signal. For the following analysis to make practical sense we must assume $|\eta(x)| \ll |f(x)|$, otherwise determining the signal would be untenable. When we work with normalized signals, as we will here, this condition can be expressed as $|f(x)| \sim \mathcal{O}(1)$ and $|\eta(x)| \ll 1$. Considering the Chebyshev expansion of equation (2.23) we can write,

$$\sum_{n=0}^{\infty} b_n T_n(x) = \sum_{n=0}^{\infty} (a_n + \xi_n) T_n(x) \tag{2.24}$$

where the b_n, a_n and ξ_n are defined through equation (2.12) with the suitable function $F(x)$, $f(x)$ or $\eta(x)$ used respectively. We can use equation (2.10) along with the triangle inequality to find a bound on the size of the Chebyshev coefficients ξ_n,

$$\xi_n = \frac{2}{\pi} \int_0^{\pi} \eta(\cos(\theta)) \cos(n\theta)\, \mathrm{d}\theta \leq \frac{2}{\pi} \int_0^{\pi} |\eta(\cos(\theta))|\, \mathrm{d}\theta \leq 2M \ll 1$$

where $M = \max_{-1 \leq x \leq 1} \eta(x) \ll 1$. We do not expect that the noise term will satisfy the same smoothness conditions as the main signal and will not, therefore, have high order continuous derivatives. As a result the coefficients will not decay as n is increased but just be uniformly small. In contrast we expect the first few coefficients a_n to be large and to decay as $\mathcal{O}\left(1/n^k\right)$ based on the smoothness of $f(x)$.

This splits the relation $b_n = a_n + \xi_n$ into three distinct regions. For small n we expect $a_n \gg \xi_n$ and therefore that $b_n \approx a_n$. For large n we expect that the a_n will have decayed suitably so that $\xi_n \gg a_n$ meaning $b_n \approx \xi_n$. There is also the intermediate region where $|a_n| \sim |\xi_n|$ where we have to use $b_n = a_n + \xi_n$. From these relationships we can see that the small n coefficients are dominated by the behavior of the underlying signal and that the large n coefficients are dominated by the noise.

In order to recreate the original base signal then we want to keep enough terms in the expansion that the shape and key features of the data are maintained but few enough that the noise dominated terms are ignored, which means the noise signal from the data is suppressed, similar to using a bandpass filter. This gives an expression for the final Chebyshev approximation of the smoothed signal,

$$f(x) \approx f_S(x) = {\sum_{n=0}^{N_c}}' \alpha_n T_n(x) \tag{2.25}$$

where the α_n's are calculated as in equation (2.18) and N_c is the chosen cut-off above which the coefficients in the expansion are ignored by setting them to 0. A suitable cut-off for a given Chebyshev interpolation can be determined by studying the magnitudes of the coefficients and how they decay .

2.2.5 Finding derivatives using Chebyshev polynomials

The power of approximating data using Chebyshev approximations is that given the approximation of a function as in equation (2.25) the derivative can be calculated from the coefficients analytically without the need for further numerical techniques and expressed as another Chebyshev series. This series is given by [16],

$$\frac{\mathrm{d}f_S}{\mathrm{d}x} = {\sum_{n=0}^{N_c-1}}' \beta_n T_n(x),$$

where the β_n are calculated from the a_n through the relationship,

$$\beta_{r-1} - \beta_{r+1} = 2r\alpha_r$$

which can be written explicitly as,

$$\beta_r = \sum_{\substack{k=r+1 \\ k-r \text{ odd}}}^{n+1} 2k\alpha_k$$

2.3 Markov chain Monte Carlo maximization

As part of the analysis we need to maximize the positional information over all possible thresholding values. This is a high dimensional function calculated numerically. In order to find an approximate maximum to such a function Monte Carlo methods are used. In this section we briefly review the theory of Markov Chain Monte Carlo (MCMC) and the implementation of the maximization algorithm.

Given a sequence of random variables, $X_1, X_2, \ldots, X_n$, a Markov chain is defined by,

$$P(X_{n+1} = i | X_1, \ldots, X_n) = P(X_{n+1} = i | X_n = j) \tag{2.26}$$

which means that the probability of the next random variable depends only on the previous random variable and not the entire history.

The main idea behind Monte Carlo optimization methods is to randomly sample the function and retain the maximal value that is found. If the function is naïvely sampled randomly from a uniform distribution over the domain then the algorithm will not converge towards a maximum. The solution is to use a different probability distribution that will converge to the maximum as the algorithm iterates. We choose the probability distribution,

$$P(I_{n+1} = i | I_n = j) = \min\left(1, \frac{j}{i}\right) \tag{2.27}$$

so that if $i > j$ we always accept the change and if $j < i$ we accept the change sometimes with a probability that depends on how much smaller the value is. This clearly satisfies the condition to be a Markov chain. The reason that we sometimes decrease the value is so that the algorithm does not get stuck in a local maximum since it can escape by coming back down. The new trial threshold values are determined

from the old ones using a Gaussian distribution centered on the old threshold value with variance σ^2. The process is outlined below:

1. Set maximal mutual information, $I_{\max} = 0$
2. Randomly choose a vector of threshold values for the profiles t_0
3. Threshold the profiles using t_n and calculate the mutual information, I_n
4. If $I_n > I_{\max}$ then set $I_{\max} = I_n$, store $t_{\max} = t_n$ and accept the new threshold vector t_n as the starting point for the next step
5. If $I_n < I_{\max}$ then calculate $\alpha = \frac{I_n}{I_{\max}} < 1$ and accept the new threshold vector t_n as the starting point for the next step with probability α
6. Create Gaussian profiles with mean at each of the values in t_n and variance σ^2 and calculate t_{n+1} by drawing from these distributions
7. Repeat steps 3 to 6
8. Output $I_{\max}$ and $t_{\max}$

Figure 2-2 shows how the algorithm jumps around randomly with an overall increase to find a global maximum using the positional information of the four gene concentrations and their derivatives.

The convergence of the algorithm is dependent on the value of σ. If the value is too high the algorithm jumps around too erratically and never climbs to a maximum. If, however, σ is too small then the algorithm converges too slowly. The maximization algorithm can be run multiple times with different values of σ for optimal efficiency. Initially a large value of σ is chosen so that the entire parameter space is sampled. Using the maxima found in the previous round as a new starting point σ can be decreased progressively so that the algorithm converges to the maximum value.

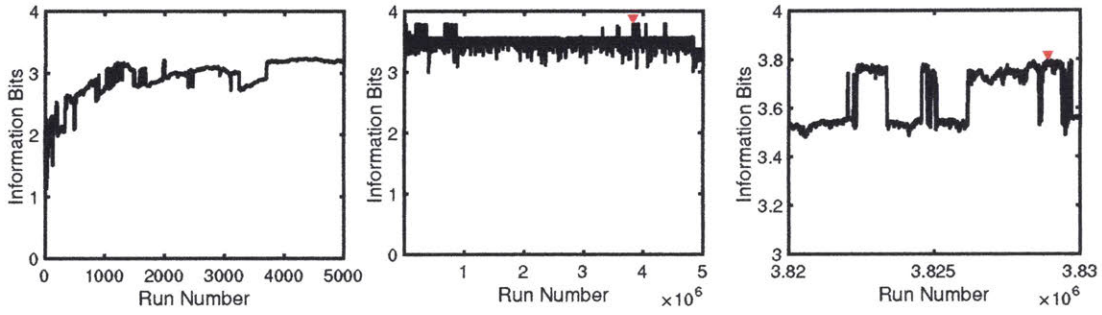

Figure 2-2: MCMC results using the absolute value of the derivative profiles. Left: zoomed in plot of the initial 5000 runs. Center: Full MCMC run. Right: zoomed in plot around the maximum value obtained. The maximum point is indicated with a red triangle. It can be seen that the algorithm frequently decreases before increasing to a larger maximum. This illustrates the importance of allowing the algorithm to occasionally accept a smaller value to escape a local maximum. Eventually it can be seen that the algorithm converges, oscillating between two maxima a local one and the global one.

Chapter 3

Application to fruit fly embryos

3.1 The model organism and experimental results

It was believed early on that the Bicoid (bcd) gene in *Drosophila* embryos is one of the best candidates for a morphogen that encodes positional information [29]. We, therefore, use the *Drosophila* embryo as a model system for understanding the role of positional information. In *Drosophila* embryos the body plan is determined during the first 3 hours of development using a sequence of interacting genes [25]. Experimental work by Gregor *et al.* showed the reproducibility of expression profiles between embryos [9]. We are concerned with the expression profiles of four gap genes hunchback (hb), krüppel (kr), giant (gt) and knirps (kni) which can be used to define a chemical co-ordinate system containing positional information. We use time series data from Thomas Gregor's lab at Princeton University that simultaneously measured the expression levels of these four gap genes. The system, full experimental set up and initial data interpretation are described in detail in Dubuis *et al.* [5]. They showed that during the 14th nuclear cycle the patterns generated were reproducible to within one internuclear distance, which is why we predominantly study the system during this time frame. The images recorded experimentally are clearest in the middle 80% of the length of the embryo so we will often consider the data in this region and then scale back to the full length.

3.2 Initial data processing

The profiles for the four genes, kni, kr, hb and gt, are taken between 38 and 48 minutes into the embryonic development of wild type embryos so that we are in the 14th nuclear cycle. In total 25 different embryos are used. Each gene is represented by a vector of values consisting of 1000 evenly spaced spatial points along the normalized length from 0 to 1. As part of the imaging process there are small misalignments in the different embryos. In order to account for alignment errors between different embryos when they are imaged the profiles are shifted using a scalar product optimization. We can interpret the scalar product between two vectors as a projection of one of the vectors onto the other. From this we can see that the closer two vectors are to each other the larger the scalar product between them. A buffer of zeros is added to the start and end of each vector to allow for shifting the vector to the right and the left. Starting with the first sample the following method is used,

1. Shift c_{i+1} to the left by 50 values to give c'_{i+1}
2. Take the scalar product with c_i and c'_{i+1}
3. If the scalar product is greater than previously keep the new vector
4. Decrease the shift by one place and repeat the process until you have shifted 50 to the right
5. Choose the shift value that maximizes the scalar product
6. Repeat the process with the next sample

The shift values are determined from one gene and then used for all genes since the misalignment should be the same across all of the profiles because of the simultaneous imaging. The results of this process are shown in Figure 3-1. After the realignment the concentration profiles are normalized so that the mean profile for each gene ranges from 0 to 1.

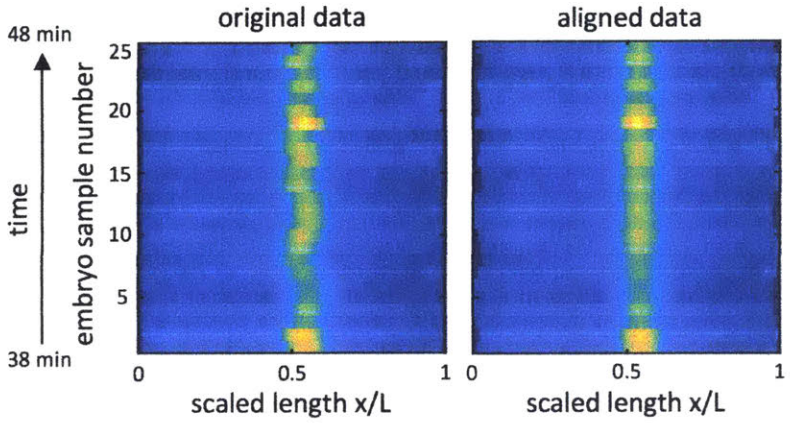

Figure 3-1: Scalar product realignment on the kr gap gene profiles for 25 samples all taken in the same time window, 38 to 48 minutes into development. Comparing the figure on the left before optimization with the figure on the right after optimization shows the power of the scalar product to align the respective peaks and produce sharper edges in the data.

3.3 Smoothing the data

The data at this point contain noise, see the first image in Figure 3-3. Performing numerical differentiation on the profile at this stage results in very jerky profiles that cannot reliably be used in calculating positional information. The data need to be smoothed before derivatives are taken in order for meaningful results to be obtained. We, therefore, smooth the data using the Chebyshev technique outlined earlier. First the edge values were discarded, since they contained values that were too low to be detected reliably and, therefore, were recorded as Not A Number (NaN) in the vector. A Chebyshev interpolant through the data for each sample and gene was then found using the chebfun package in MATLAB [4].

In order to determine the exact value of the cut-off N_c to use in the smoothed Chebyshev approximation, as in equation (2.25), the mutual information between the profiles was calculated for different values of the cut-off and the lowest value for which the curve started to flatten off was used. The cut-offs were determined using a threshold fraction. The cut-off is the smallest coefficient whose magnitude is greater than the threshold fraction of the maximum coefficient magnitude. This curve can be seen in Figure 3-2. How the mutual information was calculated will be discussed

in the next section. From the results in Figure 3-2 we choose the threshold at a 55th

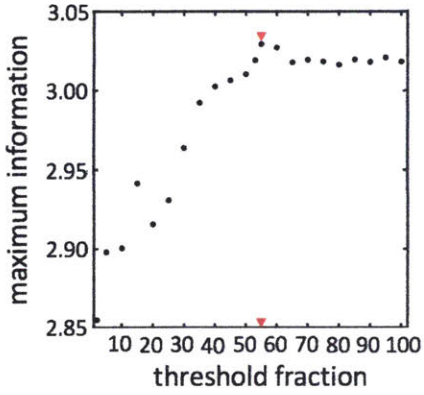

Figure 3-2: Plot of the maximum value of the information found using MCMC techniques as a function of the threshold fraction that determines the cut-off N_c for the Chebyshev expansion. We can see that the curve begins to flatten off after a threshold fraction of 55. This suggests that at this threshold fraction all the features that are important in the calculation of positional information are retained. We, therefore, use a threshold fraction of 55 to determine the cut-off values for the Chebyshev interpolations when smoothing the data.

of the maximum value of the coefficients. The smoothing process is demonstrated in Figure 3-3. The key features are maintained in the resulting profiles used for further analysis but the noise is reduced.

After the data have been smoothed using the Chebyshev polynomials the derivatives of the data are then calculated directly. To illustrate the importance of smoothing the data we also show the derivative of the unsmoothed profile and smoothed profile in Figure 3-4. It can be seen that using the smoothed data for the derivatives helps to remove unnecessary fluctuations due to noise.

The final aligned, smoothed and normalized profiles used for calculating the positional information are shown in Figure 3-5 along with the thresholds that are chosen later to binarize the profiles with maximal information and the resulting binary string.

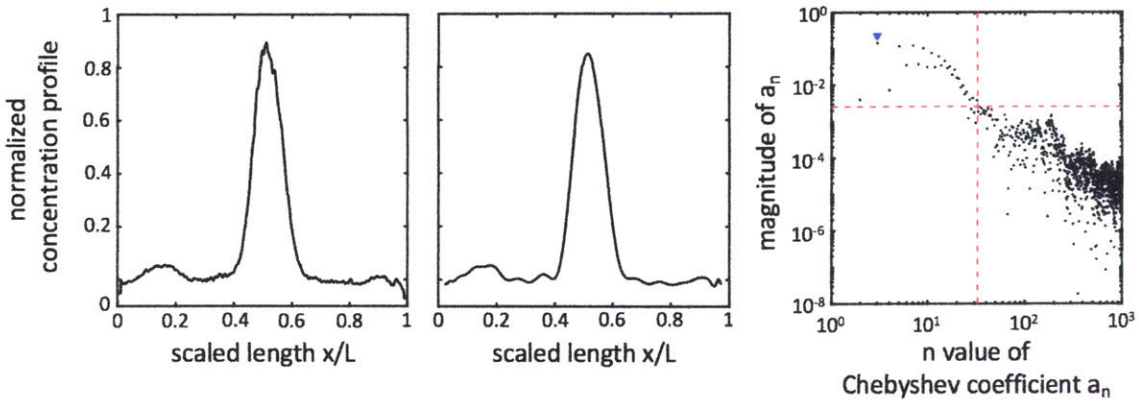

Figure 3-3: Demonstration of the smoothing process using truncated Chebyshev approximations. It can be seen that the final smoothed profile in the middle maintains most of the key features of the original profile on the left but that the noise has been significantly reduced. The figure on the right shows how the cut-off for the Chebyshev interpolation was calculated from the threshold fraction. The blue arrow points to the coefficient with the greatest magnitude and the horizontal red dashed lines shows one 55th of this value. All coefficients not in the upper left corner are ignored in the Chebyshev expansion in all further calculations.

3.4 Calculating the thresholded mutual information

The threshold for each concentration is chosen in a way that maximizes the overall information stored in the gene profiles. This technique was applied in Dubuis *et al.* [6]. We recreate the result of 2.9 bits being available here and then repeat the calculation including gradients as well.

The thresholding process changes the calculations of the entropies to a discrete process. The profile for a single gene can now only be 0 or 1. As a result the entropy of the gene profile described by equation (2.4) becomes,

$$S[P_g(g)] = -[P(g=0)\log_2(P(g=0)) + P(g=1)\log_2(P(g=1))] \tag{3.1a}$$

$$S[P(g|x)] = -[P(g=0|x)\log_2(P(g=0|x)) + P(g=1|x)\log_2(P(g=1|x))]. \tag{3.1b}$$

When we consider the combined gene profiles there are now 16 possible values for the g_i's, of the form '0010' etc. In order to calculate the information stored in the thresholded values we perform a transformation that maps the four bit string to a

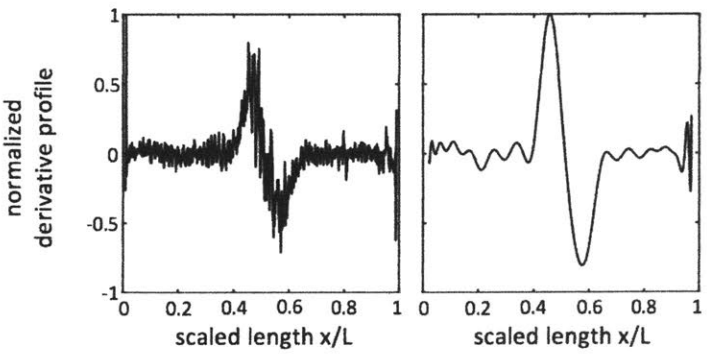

Figure 3-4: Comparing the numerical derivative of the raw data, calculated using a second order finite difference scheme, on the left with the derivative calculated directly from the truncated Chebyshev approximation on the right shows the importance of removing noise from the data before taking derivatives. The noise adds jerky spikes to the derivative that result in a derivative profile that is not smooth, which means it is not useful in calculating the positional information.

number between 0 to 15 as follows,

$$\text{String Number} = 2^0 * (\text{Kr Bit}) + 2^1 * (\text{Kni Bit}) + 2^2 * (\text{Kni Bit}) + 2^3 * (\text{Kni Bit})$$

this turns the four binary thresholded matrices for each gene into a single matrix with values from 0 to 15. From this matrix we can produce a histogram and find the probability of each string occurring. This allows us to calculate the new joint mutual information contained in all 4 gene profiles, with the new entropies,

$$S[P_g(\mathbf{g})] = -\sum_{n=0}^{15} P(g=n)\log_2(P(g=n)) \tag{3.2a}$$

$$S[P(\mathbf{g}|x)] = -\sum_{n=0}^{15} P(g=n|x)\log_2(P(g=n|x)). \tag{3.2b}$$

From experimental data the symmetry of mutual information allows us to calculate the positional information using,

$$I(x,\mathbf{g}) = \int \mathrm{d}x\, P_x(x)(S[P_g(\mathbf{g})] - S[P(\mathbf{g}|x)]) \tag{3.3}$$

where the integral over x is approximated using the trapezoidal rule. We assume that

the cells are evenly distributed along the length of the embryo so that $P_x(x) = 1/L$.

Applying this technique on the smoothed expression profiles we find that there are 2.90 bits of information available to the cells. This is consistent with the value found, using other techniques, in Dubuis *et al.* of 2.92 bits of information [6]. This corresponds to being able to differentiate $2^{2.9} = 7.5$ unique states in the embryo. The resulting binary profiles that were found are shown in Figure 3-7.

3.4.1 Including derivatives in the calculation of positional information

As mentioned previously, the derivatives are calculated directly from the Chebyshev expansion and are then normalized to lie between $[-1, 1]$. Because the derivatives can go in either direction (positive or negative) it is suggested that if the cells only care about whether they are in a region of high and low change then the absolute value of the derivatives should be used to calculate the information content. It was found that squaring the derivatives instead of taking the absolute value overly suppresses key features in the derivative profiles leading to lower information content. These profiles are the bottom row of Figure 3-5. Including gradients in the calculation of the information means that we now have 8 threshold values used to binarize the data and we have to modify the string number and equations 3.2a and 3.2b to,

$$\begin{aligned}\text{String Number} = {} & 2^0 * (\text{Kr Bit}) + 2^1 * (\text{Kni Bit}) + 2^2 * (\text{Kni Bit}) + 2^3 * (\text{Kni Bit}) \\ & + 2^4 * (\text{Kr Div Bit}) + 2^5 * (\text{Kni Div Bit}) + 2^6 * (\text{Kni Div Bit}) \\ & + 2^7 * (\text{Kni Div Bit})\end{aligned}$$

$$S[P_g(\mathbf{g})] = -\sum_{n=0}^{255} P(g = n) \log_2(P(g = n)) \tag{3.4a}$$

$$S[P(\mathbf{g}|x)] = -\sum_{n=0}^{255} P(g = n|x) \log_2(P(g = n|x)). \tag{3.4b}$$

The same MCMC algorithm is run to find the maximum possible value for the information and the results are shown in Figure 3-8. We find a maximum of 4.02 bits of

information significantly higher than the value we find using only the gene profiles.

3.5 Interpreting the amount of information available

Throughout the analysis so far we have modeled the embryo as a continuum system. To interpret the available information however it is important to remember that cells are actually discrete. There are on average 59 cells in the middle 80% of the embryo during the 14th nuclear cycle, which are approximately evenly distributed [25, 6]. To be able to make out each of these individually would require $\log_2(59) = 5.9$ bits of information. This is not, however, the level of detail that cells appear to have. Experimental evidence instead suggests that the precision with which cells can determine their position is $\sigma_x/L \sim 1\%$ [14, 6]. Using the assumption that $P_x(x) = 1/L$ is uniform and that the uncertainty in the position is constant along the length of the embryo we can write [6, 25],

$$I(\mathbf{g}, x) \approx \log_2\left(\frac{L}{\sigma_x\sqrt{2\pi e}}\right) \tag{3.5}$$

which we can rearrange to find,

$$\sigma_x \approx \frac{L}{\sqrt{2\pi e}} 2^{-I}. \tag{3.6}$$

From this we can see that a precision of 1.0% corresponds to the cells having access to 4.27 bits of information available and corresponds to a probability of an error greater that the internuclear distance of $P = 0.08$ assuming that the distribution is Gaussian [6]. The 4.02 bits of information calculated from the binarized data with gradients corresponds to a precision of 1.7% and the 2.90 bits from the binarized data without gradients corresponds to a precision of 3.6%. The internuclear distance can be calculated from $0.8/N = 0.014$, which means that one internuclear spacing corresponds to a precision of 1.75%. We, therefore, see that using a binarized model with derivatives provides enough information to specify precision to around one internuclear distance, with the probability of an error greater than one internuclear distance $P = 0.30$ down

considerably from the probability without gradients of $P = 0.62$. This shows the importance of gradients in retaining information when binary profiles are used.

3.6 Time series and mutant data

In this section we look at some early work on possible extensions of the binary analysis done here. Up until this point we have only considered wild type gene concentration profiles between 38 and 48 minutes. One question is what happens to the positional information when an embryo has a mutation? Another is how do the gene profiles and the information they contain change with time? Figure 3-9 shows how the binary strings for each gene evolve in time and how the distinct regions they specify change. We notice new activity around 38 minutes into the development which also corresponds to when the gene profiles are most uniform and the available positional information appears to peak. As can be seen in the figures using binary profiles provides a simplified view of the time evolution of gap genes during embryonic development, which could potentially lead to a model for the time evolution.

The same thresholding and maximization process is also run on 10 different mutants. The wild type, 2 bcd dosage mutants and 7 maternal mutants. Again considering gene concentration profiles in the 14th nuclear cycle we can run the same algorithms as above to find maximal thresholds for each mutation. We can then use this threshold to determine the information content between the profiles as a function of time. The results for all mutants and time series data are shown in Figure 3-10. We can see that as we might expect the more mutations that an embryo undergoes the fewer bits of information they have access to. We can also see that the information contained in the profiles generally increases as the embryo ages. The inclusion of gradients in the calculation of the positional information changes which of the mutations have the most significant impact on the positional information. This suggests that a binary model of positional information could be useful in understanding the consequences of mutations.

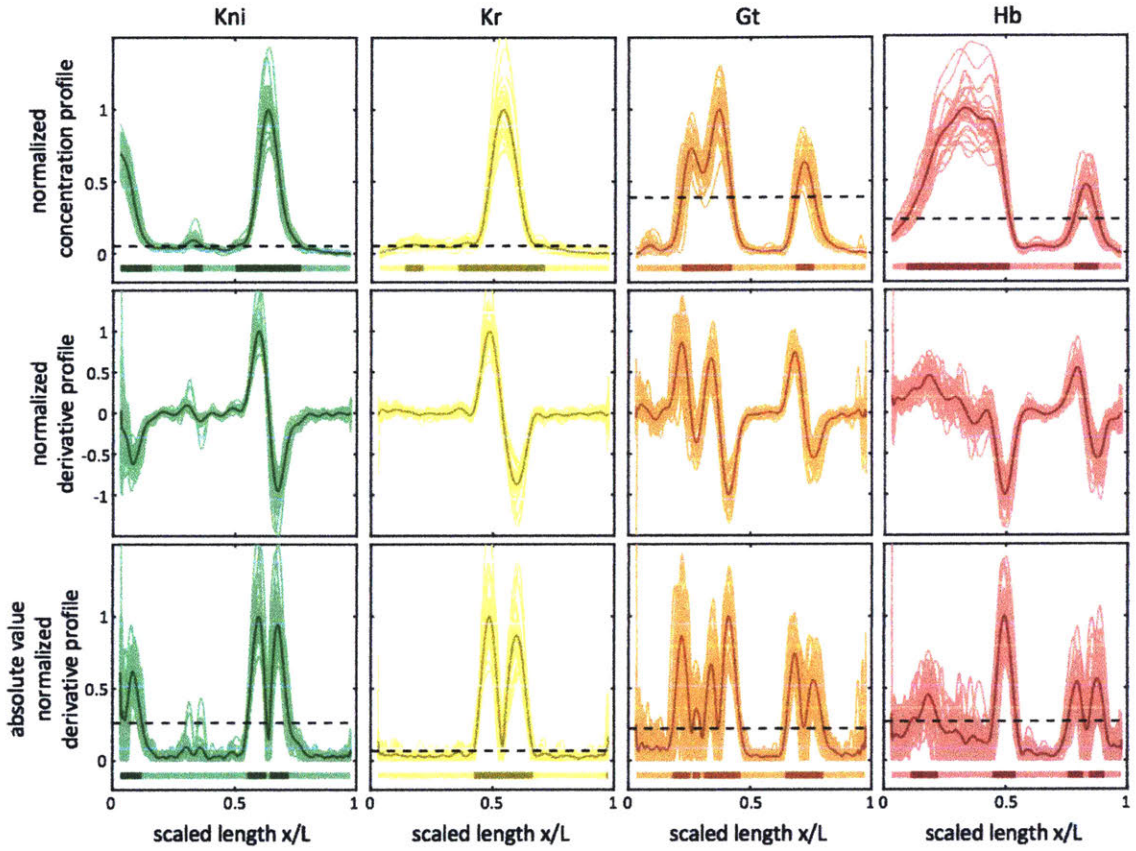

Figure 3-5: The four Gap gene concentration profiles for WT *Drosophila* embryos. The top row shows the smoothed profiles using Chebyshev interpolation, the middle row shows the derivative profiles calculated directly from the Chebyshev expansion and the bottom row shows the absolute value of the derivative profiles. The dark curve shows the mean value and the lighter curves show the profiles from individual embryos. The lighter curves highlight the embryo to embryo variability that exists between embryos and that can be used to define the probability distribution for the profiles. The bars at the bottom of the figures show the final binary profiles after maximization using the thresholds shown with the dashed lines. It can be seen that the threshold acts to pick out the key peaks in each of the profiles effectively splitting the profiles into high and low regions.

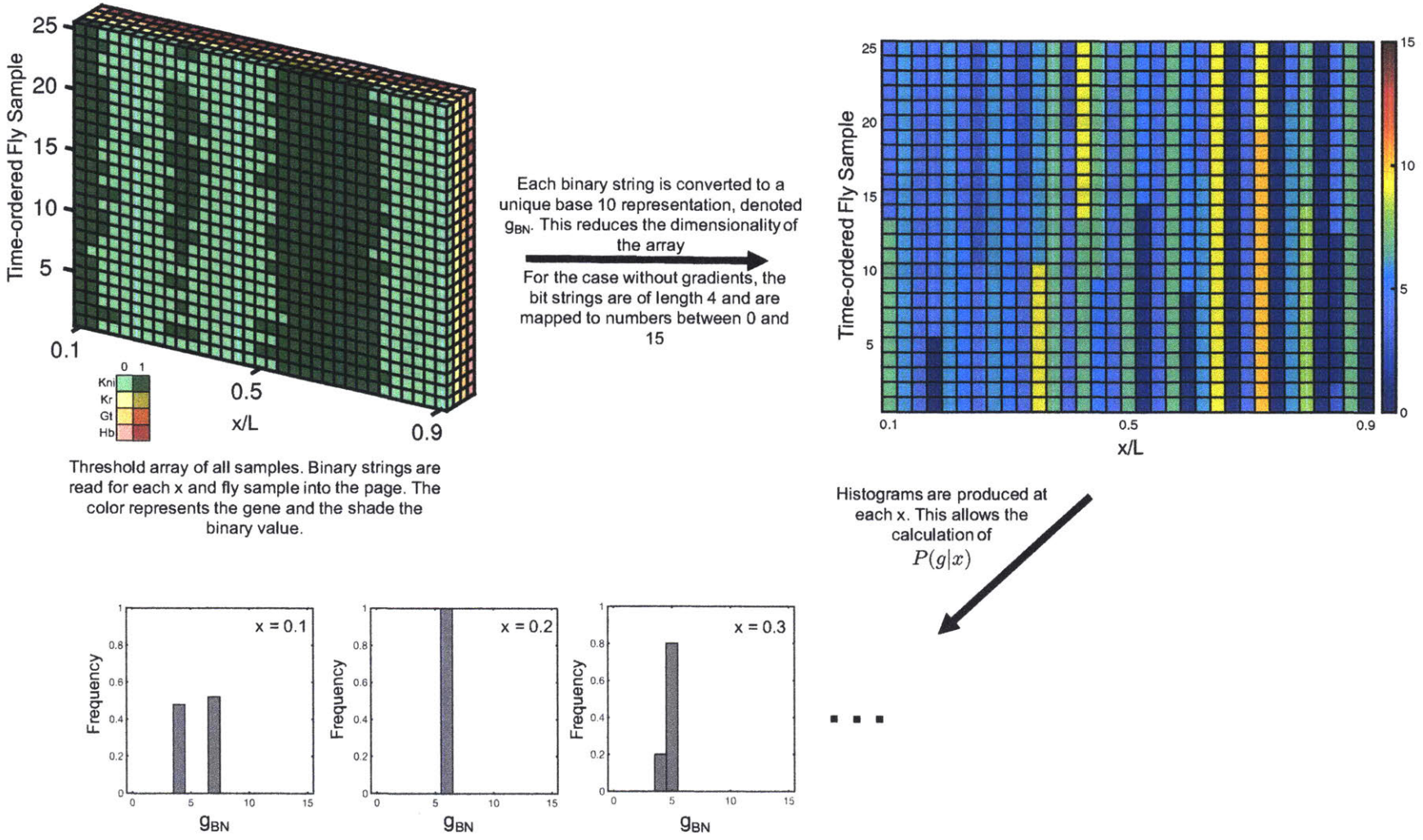

Figure 3-6: Schematic of how the information in the gene profiles is calculated from the binarized gene profiles.

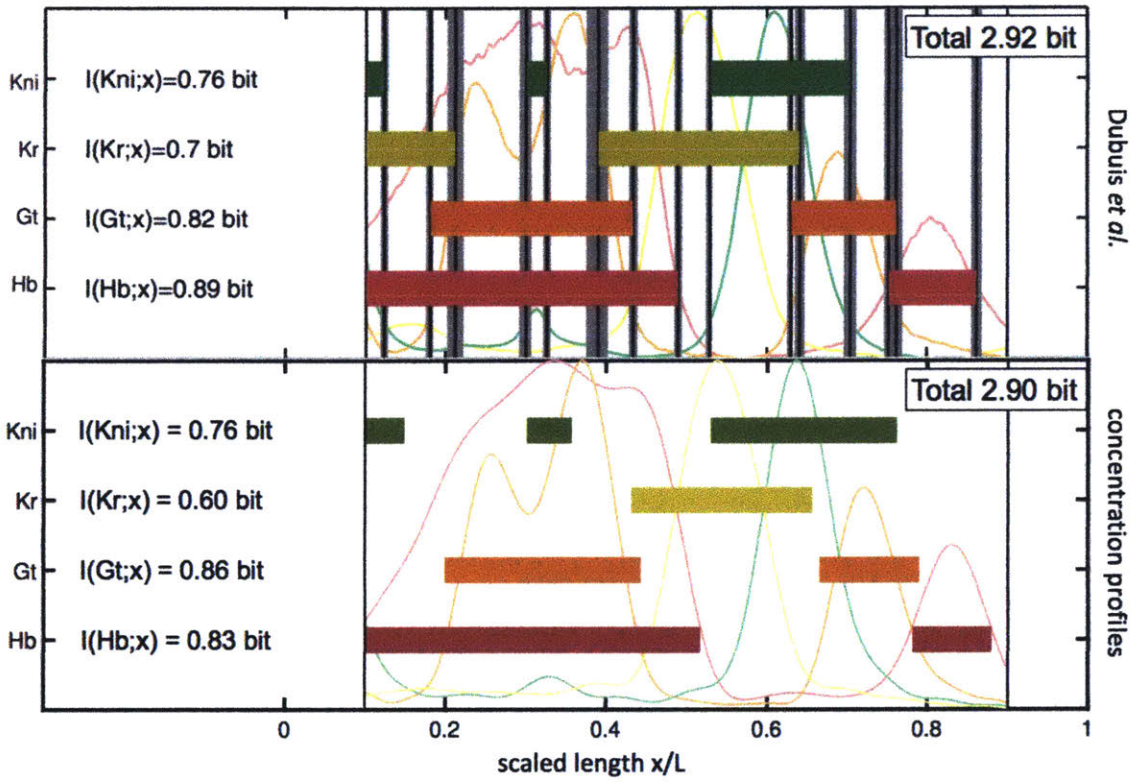

Figure 3-7: Resulting binary profiles after maximization using the four gene profiles without the gradients. The upper figure is taken from: Julien O Dubuis, Gačper Tkačik, Eric F Wieschaus, Thomas Gregor, and William Bialek. Positional information, in bits. *Proceedings of the National Academy of Sciences USA*, 110(41) : 16301-16308, 2013. The information contained in each individual gene's bit string is shown on the left. The total information found using the the new method presented here is 2.90 bits which is consistent with the 2.92 bits found in the previous work using a different technique.

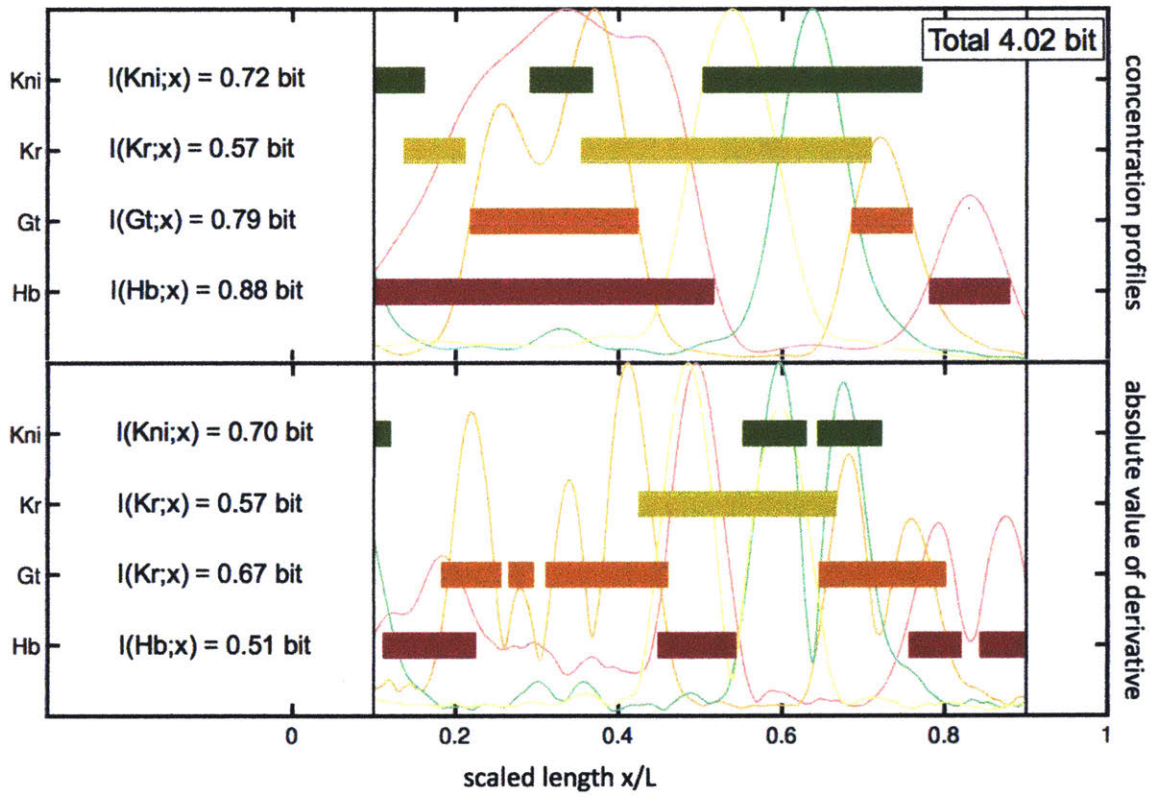

Figure 3-8: Resulting binary profiles after maximization using the four gene profiles and the absolute value of the gradients. The bars represent the region where the binary string equals 1. The information contained in the individual strings is shown on the left. Profiles are shown in the background for reference. The total information found is 4.02 bits which is enough to specify the precision along the length of the embryo to around one internuclear spacing. The result is close to the 4.27 bits that there is evidence for experimentally. This provides a significant improvement over the amount contained in the gene profiles alone seen in Figure 3-7

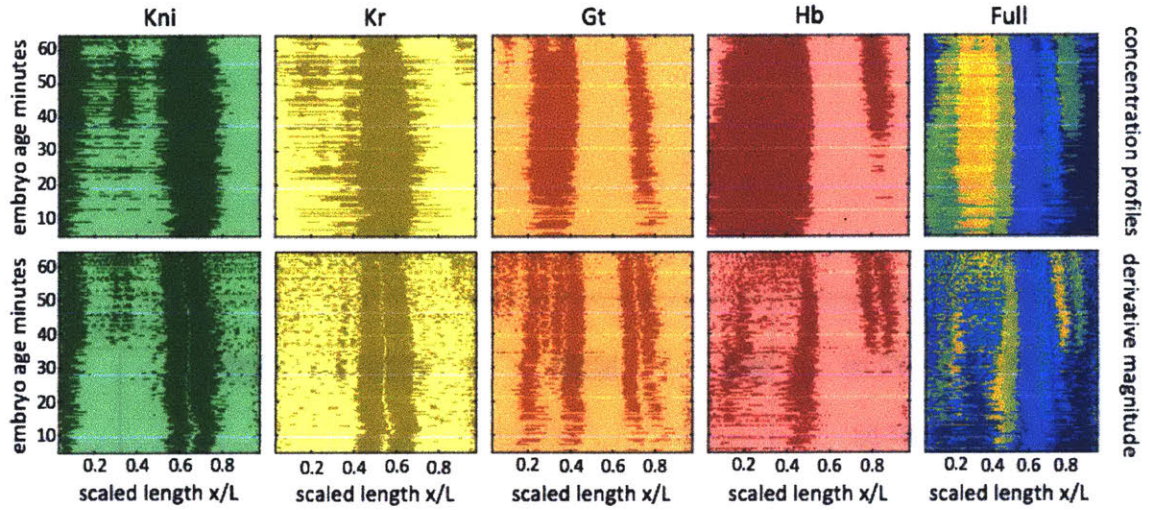

Figure 3-9: The figure shows how the binary strings evolve in time for each gene, colored coded as in Figure 3-6. The top row is concentrations and the second row the absolute value of the derivatives. The far right shows the number of distinct regions specified by the profiles at each time, the upper figure only using the 4 concentrations and bottom figure including derivatives. We can see new behavior emerges around 38 minutes into embryonic development, especially in kni and hb, which could be the cause of the information stabilizing and often peaking during the 14th nuclear cycle.

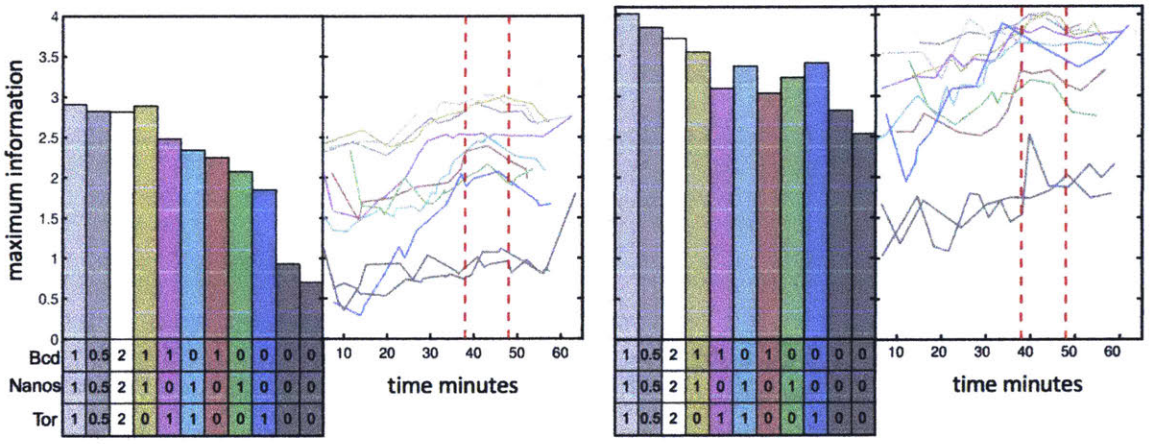

Figure 3-10: Histograms show the maximum information contained in the profiles between 38 and 48 minutes for each of the mutations considered. The far left bar is the WT data. Bcd dosage mutants are represented by 0.5 and 2 for half and double concentration respectively. Maternal mutants are represented by three numbers, 1 indicates presence and 0 represents removal. The corresponding time series information is next to the histogram. The red dashed horizontal lines indicate the 38 to 48 minute time window used to determine the thresholds. We can see that the information mostly peaks in the 38 to 48 minute period and generally decreases as more mutations are implemented. We also see that including the derivative adds information to all of the mutations.

Chapter 4

Conclusions

We have shown that by including gradients in the calculation of positional information we can account for 1.12 bits of the information that is lost by using binary profiles. This extra gain in information is enough for cells within the embryo to be able to localize along the length of the embryo with a precision that is close the internuclear distance. This suggests that a discrete model of positional information that includes gradients does not lose significant amounts of information over a model that uses the full profiles. The power of gradients to provide this missing information provides further evidence of the importance of gradients during pattern formation in embryonic development.

There are many possible future extensions of this work. A study on the effects of mutations on the binary positional information and the impact that gradients have on it was briefly touched on here but leaves open a more detailed analysis. We also touched on using binary profiles to simplify studying the time evolution of the gene profiles and positional information during embryonic development.

We posit as a result of this study that a discrete model of positional information that includes gradients does not lose significant information over a model that uses full profiles and provides a significant improvement over a discrete model that uses only concentrations.

Bibliography

[1] John P Boyd. *Chebyshev and Fourier spectral methods.* Courier Corporation, 2001.

[2] James Briscoe and Anna Kicheva. The physics of development 100 years after d'arcy thompson's "on growth and form". *Mechanisms of development*, 145:26–31, 2017.

[3] Francis Crick. Diffusion in embryogenesis. *Nature*, 225(5231):420, 1970.

[4] Tobin A Driscoll, Nicholas Hale, and Lloyd N Trefethen. Chebfun guide, 2014.

[5] Julien O Dubuis, Reba Samanta, and Thomas Gregor. Accurate measurements of dynamics and reproducibility in small genetic networks. *Molecular systems biology*, 9(1):639, 2013.

[6] Julien O Dubuis, Gašper Tkačik, Eric F Wieschaus, Thomas Gregor, and William Bialek. Positional information, in bits. *Proceedings of the National Academy of Sciences USA*, 110(41):16301–16308, 2013.

[7] Andrew D Economou and Jeremy BA Green. Modelling from the experimental developmental biologists viewpoint. In *Seminars in cell & developmental biology*, volume 35, pages 58–65. Elsevier, 2014.

[8] Jeremy BA Green and James Sharpe. Positional information and reaction-diffusion: two big ideas in developmental biology combine. *Development*, 142(7):1203–1211, 2015.

[9] Thomas Gregor, David W Tank, Eric F Wieschaus, and William Bialek. Probing the limits to positional information. *Cell*, 130(1):153–164, 2007.

[10] Peter Hänggi. Stochastic resonance in biology how noise can enhance detection of weak signals and help improve biological information processing. *ChemPhysChem*, 3(3):285–290, 2002.

[11] Natalie C Heer, Pearson W Miller, Soline Chanet, Norbert Stoop, Jörn Dunkel, and Adam C Martin. Actomyosin-based tissue folding requires a multicellular myosin gradient. *Development*, 144(10):1876–1886, 2017.

[12] Khaled Khairy, William Lemon, Fernando Amat, and Philipp J Keller. A preferred curvature-based continuum mechanics framework for modeling embryogenesis. *Biophysical journal*, 114(2):267–277, 2018.

[13] Ioannis Lestas, Glenn Vinnicombe, and Johan Paulsson. Fundamental limits on the suppression of molecular fluctuations. *Nature*, 467(7312):174, 2010.

[14] Feng Liu, Alexander H Morrison, and Thomas Gregor. Dynamic interpretation of maternal inputs by the drosophila segmentation gene network. *Proceedings of the National Academy of Sciences USA*, 110(17):6724–6729, 2013.

[15] Luciano Marcon and James Sharpe. Turing patterns in development: what about the horse part? *Current opinion in genetics & development*, 22(6):578–584, 2012.

[16] John C Mason and David C Handscomb. *Chebyshev polynomials*. CRC Press, 2002.

[17] Mark D McDonnell and Derek Abbott. What is stochastic resonance? definitions, misconceptions, debates, and its relevance to biology. *PLoS computational biology*, 5(5):e1000348, 2009.

[18] Mariela D Petkova, Gašper Tkačik, William Bialek, Eric F Wieschaus, and Thomas Gregor. Optimal decoding of information from a genetic network. *arXiv preprint arXiv:1612.08084*, 2016.

[19] Manuel Razo-Mejia, Stephanie L. Barnes, Nathan M. Belliveau, Griffin Chure, Tal Einav, Mitchell Lewis, and Rob Phillips. Tuning transcriptional regulation through signaling: A predictive theory of allosteric induction. *Cell Systems*, 6(4):456–469.e10, 2018/05/03 2018.

[20] Carl Runge. Über empirische funktionen und die interpolation zwischen äquidistanten ordinaten. *Zeitschrift für Mathematik und Physik*, 46(224-243):20, 1901.

[21] Erwin Schrödinger. *What is life?: With mind and matter and autobiographical sketches*. Cambridge University Press, 1992.

[22] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[23] Adrian Streit and Ralf J Sommer. Genetics: Random expression goes binary. *Nature*, 463(7283):891, 2010.

[24] Darcy Wentworth Thompson. *On growth and form*. Cambridge University Press, 1942.

[25] Gašper Tkačik, Julien O Dubuis, Mariela D Petkova, and Thomas Gregor. Positional information, positional error, and readout precision in morphogenesis: a mathematical framework. *Genetics*, 199(1):39–59, 2015.

[26] Lev S Tsimring. Noise in biology. *Reports on Progress in Physics*, 77(2):026601, 2014.

[27] Alan Mathison Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37–72, 1952.

[28] Lewis Wolpert. Positional information and the spatial pattern of cellular differentiation. *Journal of theoretical biology*, 25(1):1–47, 1969.

[29] Lewis Wolpert. Positional information revisited. *Development*, 107(Supplement):3–12, 1989.

[30] Lewis Wolpert. Positional information and patterning revisited. *Journal of Theoretical Biology*, 269(1):359–365, 2011.