Quantifying positional information during early embryonic development

Julien Dubuis

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Physics
Advisers: William Bialek and Thomas Gregor

November 2012

Abstract

During the development of multicellular organisms, cells acquire information about their position in the embryo in response to morphogens whose concentrations vary along the anteroposterior axis. In this thesis, we provide an information-theoretic definition of positional information and demonstrate how it can be quantified from experimental data. We start by setting up the mathematical framework and qualitatively discuss which features of expression patterns can contribute to positional information. Then, using the four major gap genes (Hunchback, Krüppel, Giant, and Knirps) of *Drosophila* as an example, we focus on the experimental standards that need to be met to accurately compute positional information from imunofluorescence stainings. We show that imunofluorescence makes it possible to extract not only very accurate mean profiles but also statistical noise and noise correlations from gene expression profile distributions and we use this analysis to extract gap gene profile dynamics with 1-2 min precision and to quantify their profile reproducibility. Finally, we describe how to quantify positional information, in bits, from the experimental gap gene profiles previously generated. Our results show that any individual gene carries nearly two bits of information and that, taken together, these four gap genes carry enough information to define a cell's location along the anteroposterior axis of the embryo with an error bar of half the intercellular distance. This precision is nearly constant along the length of the embryo and nearly enough for each cell to have a unique identity. We argue that this constancy is a signature of optimality in the transmission of information from primary morphogen inputs to the output of the gap gene network.

Acknowledgments

First and foremost, I first would like to thank my advisors: William Bialek, for giving me the opportunity to come to Princeton and offering me unconditional freedom during my Ph.D, and Thomas Gregor, for his unrelenting energy and daily support with my research. I am grateful to Eric Wieschaus for his insatiable enthusiasm and for teaching me the fundamentals of development with inspiring humility and passion. I would also like to thank Vincent Hakim who looked over my shoulder over the past five years and shared insightful comments on my work.

My sincere gratitude goes to Reba Samanta and Joseph Goodhouse for initiating me into their art of immunofluorescence staining and confocal microscopy respectively. I am most thankful to all the postdoctoral fellows who shared their accumulated wisdom and answered my myriad of questions, in particular Gordon Berman, Stefano Di Talia, Shawn Little, Attilio Pane, Kanaka Rajan, and Gasper Tkačik. I want to acknowledge and thank all the members of the Gregor Lab for the many fun and enjoyable experiences both in and out of the lab : Bryan Chun, Hernan Garcia, Sri Iyer Biswas, Feng Liu, Alex Morrison, Troy Mestler, Martin Scheeler, Allyson Sgro, Eric Smith, and Mikhail Tikhonov. I am also thankful to all the members the Biophysics Theory Group with whom I have had the fortune to collaborate over the past five years: Michele Castellana, Michael Desai, Matthias Kashube, Justin Kinney, Dima Krotov, Pankaj Metha, Anand Murugan, Thierry Mora, Stephanie Palmer, Eva-Maria Schötz, David Schwab, Audrey Sederberg, Greg Stephens, Thibault Taillefumier, Sam Taylor, and Aleksandra Walczak.

Last but not least, a special thanks to Rebecca Khalandovsky and Mariela Petkova for reading and correcting this manuscript.

Publications and preprints associated with this thesis

1. **Julien O. Dubuis**, Gasper Tkačik, Eric F. Wieschaus, Thomas Gregor, and William Bialek (2012). Positional Information, in bits.
   arXiv:1201.0198v1 [q-bio.MN]

2. **Julien O. Dubuis**, Reba Samanta, and Thomas Gregor (2012). Accurate measurements of dynamics and reproducibility in small genetic networks.

3. Timon Idema, **Julien O. Dubuis**, M. Lisa Manning, Philip Nelson, and Andrea J. Liu (2011). Wavefront propagation and mechanical signaling in early Drosophila embryos.

4. **Julien O. Dubuis**, Alexander H Morrison, Martin Scheeler, and Thomas Gregor (2010). Quantifying the Bicoid morphogen gradient in living fly embryos.
   arXiv:1003.5572v2 [q-bio.MN]

   (This article appeared as a book chapter in *Imaging in Developmental Biology: A Laboratory Manual, CSH press, 2010*)

I would like to separately thank my collaborators Gasper Tkačik, Reba Samanta, Timon Idema, Lisa Manning, Philip Nelson, Andrea Liu, Alex Morrison, and Martin Scheeler for these projects.

Materials from this dissertation have been previously presented publicly as contributed talks at the following scholarly conferences:

1. International Conference on Biological Physics 2011, San Diego, CA

2. APS March Meeting 2010, Portland, OR

3. Q-Bio Conference 2009, Santa Fe, NM

Contents

List of Figures

Introduction

Multicellular organisms develop from a single fertilized egg that repeatedly divides, leading to a complex system of cells with distinct shapes and functions. This process, through which cells and tissues are structured to give the organism its characteristic shape and function, is called ‘morphogenesis’. During morphogenesis cells choose from a variety of fates. That is, they undergo ‘differentiation’. While all cells contain identical genetic information stored in their DNA, in order to differentiate, they express only a subset of genes [Alberts *et al.*, 2002, Huang *et al.*, 2005]. As an example in humans, the difference between a liver cell and a skin cell comes from the set of genes they are expressing.

Differentiation occurs in spatial patterns and that have different levels of reproducibility. For some tissues, cell differentiation does not have to be spatially reproducible from one individual to the other, for instance, in the case of the skin cells that form fingerprint patterns. In some other cases though, having similar fates at similar locations across individuals is crucial, like the position of the head, arms and legs with respect to the body plan. In the latter case, which is what we are interested in here, cells have to adopt fates that are correlated with their location in the embryo or, in other words, they need to acquire some information about their position before making any decision.

In order to understand embryonic development, it is thus crucial to question how the observed patterns of cell differentiation are established and to quantify the reproducibility up to which this differentiation happens. In particular, if cells acquire information about their position, we want to first identify the means by which this information is conveyed. In this introduction, we first explain how positional information is set up by signaling molecules called ‘morphogens’ that act in a concentration-dependent manner. Then, using early *Drosophila* development as an example, we describe how this positional information is transmitted in the course of the first three hours of embryogenesis, and eventually leads into different cell fates. Namely, we describe the successive types of genes that are involved in that transmission and their interactions. Finally, we list a set of quantitative questions that arise when looking at the system from a physicist's point of view.

An introduction to 'morphogens'

The discovery of the mechanisms that determine cell fate is in itself a thrilling scientific saga (see [Ephrussi & St Johnston, 2004] for a complete review). It was first suggested at the beginning of the 20th century by Wilson and Conklin that cells adopt different fates by 'sensing' the presence or absence of chemicals located outside of them in the cytoplasm. Conklin observed in sea squirt eggs that cells segregated in different lineages (muscle or notochord) depending on which region they where developing in, thus suggesting that these regions were containing fate-determining factors [Conklin, 1905]. Wilson observed in snail eggs that the presence of a protrusion called the polar lobe determined which cells formed the muscles, shell gland, mouth and foot, and concluded that the polar lobe contained some factors necessary for cell differentiation [Wilson, 1904]. Yet one had to wait until 1974 before Illmensee and Mahowald undoubtedly proved the existence of these cytoplasmic determinants by transplanting the pole plasm of *Drosophila* from the anterior to the posterior of the egg and observing the formation of pole cells [Illmensee & Mahowald, 1974]. At the same time, some experiments by Boveri and Morgan showed that specific regions of the embryo can have long-range effects on the development of the rest of the organism. Experiments from Boveri on sea urchin in 1901 suggested that its developmental pattern could be determined by opposite animal and vegetal 'gradients' [Boveri, 1901], while Morgan's experiments on regeneration of flat worms lead him to postulate the existence of 'formative substances' that influenced the developmental plan of the embryo [Morgan, 1904, Morgan, 1905].

In the middle of the last century, Turing coined the term 'morphogens' to describe these long-range effect cytoplasmic determinants [Turing, 1952]. He used mathematical models to show that a simple reaction-diffusion process could lead cell differentiation through a process of self-organization. In this class of models, which were further developed by Gierer and

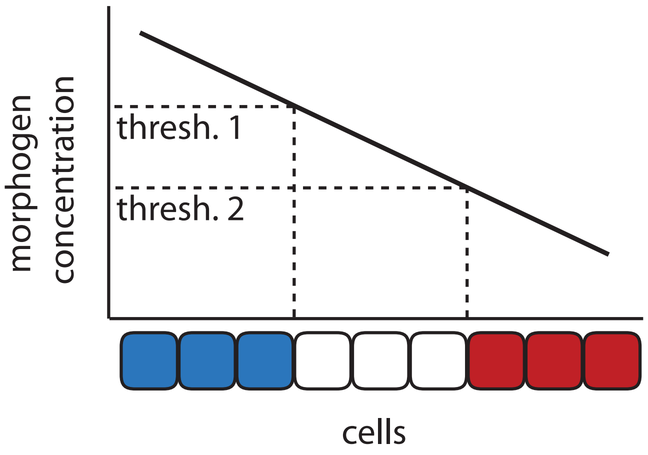

Figure 1: Illustration of Wolpert's French flag model (1969). Cells read out the morphogen concentration in a threshold-like manner. Originally, all cells are equivalent and could potentially adopt any of the three fates (blue, white, or red). After the gradient is established, they read and compare their concentrations with a set of thresholds. The fate they adopt then correlates with their position or, in other words, they acquire positional information from the morphogen gradient.

Meinhart, competition between local self-enhancement and long-range inhibition amplifies local inhomogeneities to generate a stable pattern [Gierer & Meinhardt, 1972, Meinhardt, 1989]. These mechanisms may indeed underlie the pigmentation patterns seen in certain animal coats like the stripe patterns in the zebra or the angelfish [Painter *et al.*, 1999, Wolpert *et al.*, 2002].

In parallel, an alternative line of thought suggested that the symmetry breaking, which appears during embryonic development, is predetermined by some prior, asymmetrically localized factor. In this view, morphogens are produced in cells that are located in a spatially restricted source region, diffuse along the main axis of the egg, and are degraded throughout the embryo or possibly only in a sink region located at some distance to the source [Lawrence, 1966, Stumpf, 1968]. This setup leads to the formation of a graded profile of morphogen concentration. In 1969, Wolpert postulated that the fates of cells are determined by the local concentrations of these graded profiles, which contain positional information about the distance to the morphogen source as shown on Figure 1 [Wolpert, 1969]. In other words, undifferentiated cells measure the concentration of a chemical substance that exists in a gradient along some embryonic axis. Different concentration thresholds induce different cell fates. This particular activity has come to define the concept of a morphogen in the modern sense of the term [Slack, 2001].

Up to this point, however, the concept of a morphogen remained a theoretical one, with no clear experimental evidence of its existence. One had to wait until 1988 and the groundbreaking work of Driever and Nüsslein-Volhard to identify the *Drosophila* maternal effect gene *bicoid* (bcd), the first gene displaying the characteristics of Wolpert's definition of a morphogen [Driever & Nüsslein-Volhard, 1988a, Driever & Nüsslein-Volhard, 1988b, Driever & Nüsslein-Volhard, 1989, Driever *et al.*, 1989]. Driever and Nüsslein-Volhard developed very pure antibodies against Bicoid that allowed the visualization of its protein distribution in the *Drosophila* embryo. They showed that it establishes a concentration gradient that extends from an anterior high point over more than 50% of the embryo. More importantly, they demonstrated that the dosage and shape of the gradient was correlated with the final cuticle pattern. Thus Bicoid acts as a 'transcription factor'. That is to say, it can regulate the activity of other genes at different concentration thresholds to pattern the anteroposterior axis of the embryo. This work ended over 80 years of speculation over whether morphogen gradients exist and the nature of the molecules involved. Subsequently, many more embryonic morphogens were discovered both in *Drosophila* [Neumann & Cohen, 1997] and in other organisms including vertebrates such as chick [McMahon *et al.*, 2003] and zebra fish [Chen & Schier, 2001].

Developmental plan of *Drosophila*

The concept of the morphogen is useful for visualizing cell differentiation in the embryo: a determinant spreads from a localized source, forms a concentration gradient across the embryo and cells adopt fates that are correlated with the concentration of this determinant. In other words, the morphogen provides an original chemical coordinate system that has to be 'read out' by the cells. Yet it is not clear at this stage by which means this reading process is happening.

Those means were actually discovered *before* the existence of morphogen gradients was proven. In the late '70s Nüslein-Volhard chose to identify the unknown molecules involved in the patterning of the embryo by a genetic approach rather than a biochemical one. Her successful analysis of maternal-effect mutations that affect the dorsoventral patterning of the embryo [Nüsslein-Volhard, 1977] lead her and Wieschaus to propose a systematic genetic approach to discover the genes responsible for *Drosophila* axis formation. They performed saturation mutageneses for zygotic mutants that affect the patterning of the embryonic cuticle [Nüsslein-Volhard & Wieschaus, 1980] and found that the embryo is patterned by a hierarchical network of about 40 interacting genes composed of three layers: the gap genes, the pair rule genes, and the segment polarity genes. Subsequent molecular analysis showed that these genes encode transcription factors, hence allowing the layers to interact to generate progressively finer patterns of gene expression [Lawrence, 1992].

In a broad overview, the developmental plan of *Drosophila* is as follows (summarized in Figure 2, see [Jaeger, 2011] for a complete review):

1. After fertilization, bcd and nos mRNAs, which are sequestered into the anterior and posterior poles of the egg, are translated and their proteins diffuse towards the posterior and the anterior regions, respectively. The Bcd protein prevents translation of maternal cad mRNAs, which are originally uniformly distributed throughout the whole embryo [Dubnau & Struhl, 1996, Rivera-Pomar *et al.*, 1996]. At the same time, Nos protein prevents translation of the maternal hb mRNAs, which are also originally uniformly distributed [Murata & Wharton, 1995]. Thus the embryo is roughly divided into two broad regions: the anterior half containing Bcd and maternal Hb proteins, and the posterior half containing Cad proteins (and Nos, which is believed to have no role in AP patterning other than being a translational repressor of maternal Hb [Irish & Lehmann, 1989, Hülskamp *et al.*, 1989, Struhl, 1989]).

2. The next step in the segmentation process is the response of the zygotic genome to this maternal information, resulting in activation of the gap genes[1]. In the anterior half, the genes giant (gt) and and hunchback (hb) are activated as a response to Bcd and maternal Hb. In the posterior half, the gene knirps (kni) and gt are activated as a response to Cad. In the central region, the gene Krüppel (Kr) is activated. All these genes encode transcription factors that create cross-regulatory interaction between

[1] The name 'gap genes' refers to their mutant phenotype as mutations in these genes affect several contiguous segments [Nüsslein-Volhard & Wieschaus, 1980].

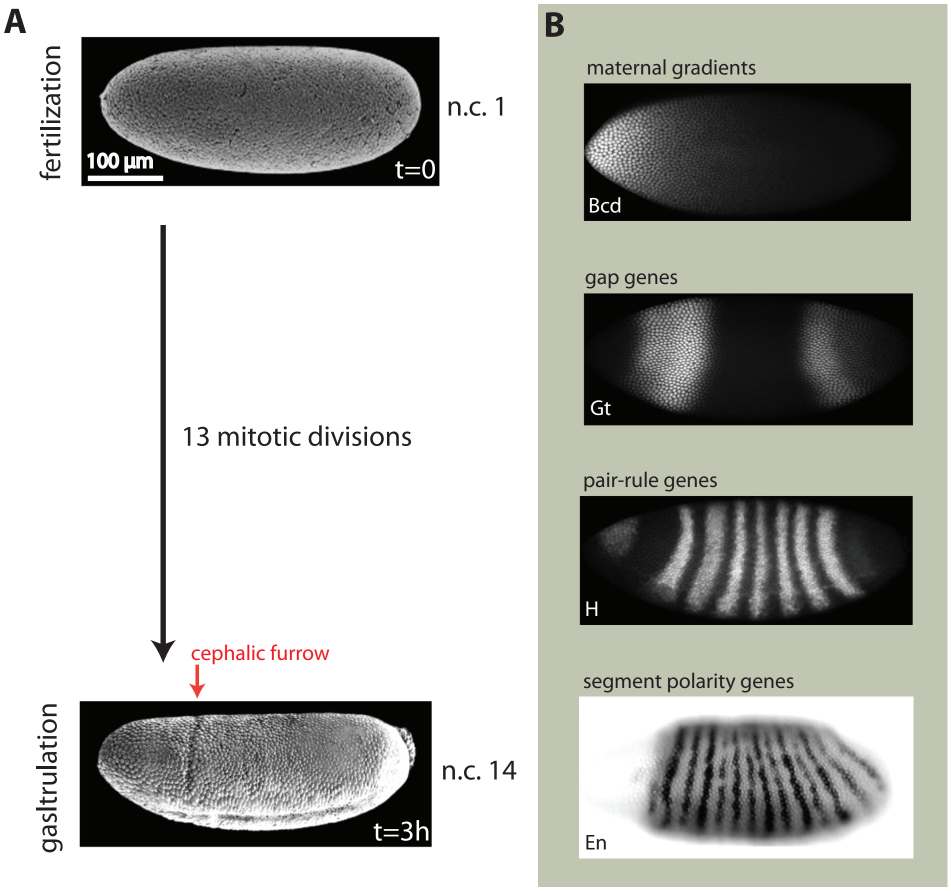

Figure 2: Schematic picture of the gene cascade, which regulates *Drosophila melanogaster* anteroposterior axis. **A)** In the first 3 hours of development, the embryo experiences 13 successive mitotic divisions and remains syncytial (without cell membranes between nuclei) until cellularization occurs during n.c 14. Gastrulation begins immediately after cellularization is complete approximately 3 hours after fertilization. The cephalic furrow is formed with a precision of a single row of cells. **B)** The regulatory hierarchy of the *Drosophila* segmentation gene network. First, maternal gradients are expressed across the whole embryo (Bicoid protein distribution shown as an example). They regulate the zygotic gap genes, expressed in broad overlapping domains (Giant protein shown as an example). Gap genes and pair-rule genes together regulate pair-rule genes, which are expressed in 7 stripes (Hairy protein shown as an example). Pair-rule genes in turn regulate segment-polarity genes, which are expressed in 14 stripes (Engrailed RNA shown as an example [Swantek & Gergen, 2004]). These stripes constitute the segmentation pre-pattern and correspond to the positions of parasegmental boundaries later in development. Embryos are shown with the anterior pole to the left and the dorsal side to the top.

them (see [Rivera-Pomar *et al.*, 1996, Sánchez & Thieffry, 2001] for a review).The result is the subdivision of the zygote space into subspaces defined by broad expression domains.

3. In turn, the gap genes transmit the morphogenetic information to the segmentation gene hierarchy with ever more refined spatial patterns of expression, namely the pair-rule genes, which are expressed in seven stripes. Some pair-rule genes are needed for the formation of even segments and others for the formation of odd segments. Finally, the pair-rule genes determine the activity of the segment polarity genes. Each of these is expressed as 14 stripes, one stripe corresponding to each segment.

Eventually the network generates distinct rows of nuclei that express the downstream genes in unique and distinguishable patterns [Gergen *et al.*, 1986].

Precision and reproducibility during embryonic development

The final developmental pattern of *Drosophila* shows two remarkable features that one shouldn't confuse:

1. The expression domains of the genes involved in the anteroposterior patterning are *precise* [Crauk & Dostatni, 2005, Gregor *et al.*, 2007a], meaning that, in a given embryo, neighboring cells can adopt different fates.

2. The patterns of gene expression are *reproducible* [Houchmandzadeh *et al.*, 2002, Gregor *et al.*, 2007a], meaning that the fate of the cells and their position are strongly correlated across embryos.

To illustrate this with a simple example, let us consider the formation of the cephalic furrow. The fact that its formation begins with a *single* row of initiator cells demonstrates the precision of the process. On the other hand, the fact that this row of initiator cells is systematically located at a fractional length of 33% of the anteroposterior axis demonstrates its reproducibility.

From the physics point of view, it is remarkable that such precision and reproducibility can be achieved in such a short amount of time using only a handful of genes. Gene expression is subject to intrinsic fluctuations that trace back to the randomness associated with regulatory interactions between molecules present at low absolute copy numbers. Moreover, there is random variability not only within, but also between, embryos, for instance in the strength of the morphogen sources. These biophysical limitations in the number of signaling molecules, available time for morphogen readout, and reproducibility of the initial and environmental conditions place severe constraints on the ability of the developmental system to generate reproducible gene expression patterns.

Although there is a consensus that particular genes carry positional information, much less is known quantitatively about how much information is being represented. Do the relatively broad, smooth expression profiles of the gap genes, for example, provide enough information to specify the exact pattern of development, cell by cell, along the anteroposterior axis? How much information does the whole embryo actually use in making this pattern? Answering these questions is important because we know that crucial molecules involved in the regulation of gene expression are present at low concentrations and even low absolute copy numbers, so that expression is noisy [Elowitz *et al.*, 2002, Ozbudak *et al.*, 2002, Blake *et al.*, 2003, Rosenfeld *et al.*, 2005, Golding *et al.*, 2005, Gregor *et al.*, 2007a, Tkacik *et al.*, 2008], and this noise must limit the transmission of information [Tkacik *et al.*, 2008, Ziv *et al.*, 2007, de Ronde *et al.*, 2010, Tkacik & Walczak, 2011]. Is it possible, as suggested theoretically [Tkacik *et al.*, 2008, Tkacik *et al.*, 2009, Walczak *et al.*, 2010, Tkacik & Walczak, 2011], that the information transmitted through these regulatory networks is close to the physical limits set by the bounded concentrations of the different transcription factors? To answer this and other questions, we need to measure positional information quantitatively, in bits. We do this here using the gap genes in *Drosophila* as an example.

Overview of this work

In chapter 1, we define the positional information carried by a set of morphogens following Shannon's original work [Shannon, 1948]. First, we start by asking which features of the morphogens (mean profile, variance and correlations) are informative to the cells from the positional point of view and we show, through schematic examples, how they influence positional information. Then, we make the case that the relevant metric to quantify positional information is the mutual information I (in its information-theoric definition) between the gene expression levels and the position. It is defined as a single number that tells us how much information cells can use overall from a set of morphogens. In particular, we claim that I is the only quantity that takes into account all the features described earlier. Finally, we establish the link between the mutual information and and the limits of local position decoding. In particular, we show how positional information puts mathematical limits on the precision up to which cells can infer their location and show how information can be 'mapped' along the anteroposterior axis of the embryo.

In chapter 2, we describe and discuss the protocols that we used to experimentally measure the positional information carried by the four gap genes, Hunchback (Hb), Krüppel (Kr), Giant (Gt) and Knirps (Kni). In particular, we present a set of careful control measurements that demonstrate clearly to what extent immunofluorescence protocols can be used for gene expression level quantification. We start by showing how the four genes can be simultaneously quantified in a single embryo. Then, we calibrate the developmental time point of fixed embryos with a precision of 1 - 2 min using live membrane furrow invagination as a calibrator. This allows us to sort out embryos precisely according to time and to limit experimental errors due to developmental time misidentification. Finally we identify the sources of experimental error which can perturb our measurement of the biologically relevant variance of profiles across the embryo. We show that antibody stainings provide an accurate way to measure mean expression profiles and that the vast majority of the observed variance is indeed biologically meaningful and cannot be only imputed to experimental biases.

Finally, in Chapter 3, we describe how the above experimental data are processed to quantify the positional information carried by the four gap genes. First, we start by covering the technical details related to the application of positional information formalism to real data sets. In particular, we show how to deal with a finite number of samples and introduce different approximation methods to estimate the information carried by one gene. Then, we move on to the case of multiple genes and use the experimental data described in Chapter 2 to compute the positional information carried by the gap genes hb, kr, gt, and kni. We then quantify the positional error up to which cells can estimate their location based on the readout of these four gap genes.

Chapter 1

Positional information in the context of embryonic development

1.1 Introduction

The concept of positional information has been widely used as a qualitative descriptor and has had enormous success in shaping our current understanding of spatial patterning in developing organisms [Wolpert, 1969, Wolpert, 1971, Tomlinson *et al.*, 1987, Fasano & Kerridge, 1988, Moses & Rubin, 1991, Reinitz *et al.*, 1995] (see [Wolpert, 1996] for a complete review). Mathematically, however, positional information has never been rigorously defined. Specific morphological features during early development have been studied in great detail and shown to occur reproducibly across wild-type embryos [Lecuit *et al.*, 1996, Jaeger & Reinitz, 2006, Jaeger & Irons, 2008], while perturbations to the morphogen system resulted in systematic shifts of these same features [Kraut & Levine, 1991, Capovilla *et al.*, 1992, Rivera-Pomar *et al.*, 1995]. This established a causal—but not quantitative—link between the positional information encoded in the morphogens and the resulting body plan. Building on previous as well as new results, we provide the missing quantitative link by proposing a mathematical formalism for positional information.

In *Drosophila*, the entire body plan of the future adult organism is established by a hierarchical network of interacting genes during the first three hours of embryonic development. The hierarchy is composed of three layers: long-range protein gradients that span the entire long axis of the egg, gap genes expressed in broad bands, and pair-rule genes that are expressed in a regular seven-striped pattern. The genes encode transcription factors, hence allowing the layers to interact. Positional information is provided to the system solely via the first layer, which is established from maternally supplied and highly-localized mRNA that act as protein sources for the maternal gradients. The network then uses these inputs to generate distinct rows of nuclei that express downstream genes in unique and distinguishable patterns.

It is remarkable that such precision can be achieved in such a short amount of time using only a handful of genes. Gene expression is subject to intrinsic fluctuations, which trace back to the randomness associated with regulatory interactions between molecules present at low absolute copy numbers. Moreover, there is also random variability across embryos: for instance, in the strength of the morphogen sources. These biophysical limitations—in the number of signaling molecules, in the time available for morphogen readout, in the reproducibility of the initial and environmental conditions—place severe constraints on the ability of the developmental system to generate reproducible gene expression patterns.

Traditionally, precision and reproducibility have been thought of as the ability to generate spatially sharp boundaries between broad bands of gene expression, where each gene would be either "on" or "off" in each of the expression domains. This view, which starts with the assumption that there are only two biologically meaningful levels of expression, has been challenged recently by showing that the morphogen *bicoid* transmits 1.5 bits of information to the gap gene *hunchback*; this would suffice for three (instead of two) distinguishable levels of *hunchback* response [Tkacik *et al.*, 2008]. Moreover, a total of, say, four gap genes, each of which can provide at most one bit ("on" or "off") of information, cannot suffice to specify $> 2^4$ nuclear rows uniquely, even in principle. Taken together, these considerations indicate that gap genes could be more than just "binary switches," encoding a single bit of positional information each in their expression domains. To address these conflicting interpretations rigorously, we need to be able to make *quantitative* statements about the positional information of the spatial gene expression profiles, without presupposing which features of the profile (e.g. sharpness of the boundary, size of the domains, position-dependent variability etc) actually encode this information.

Here we make the case that the relevant measure for positional information is the *mutual information* I — a central information-theoretic quantity defined by Shannon in the 40s [Shannon, 1948] — between expression profiles of the gap genes and position in the embryo. Briefly, this quantity, measured in bits, roughly counts (the binary logarithm of) the number of rows along the axis of the embryo that have distinguishable gene expression levels, given gene expression noise within an embryo and gene profile variability across embryos in a population. In other words, 2^I is an upper bound to the possible number of distinct cellular identities. We show how positional information puts mathematical limits to the ability with which cells in the developing embryo can infer their position if they respond to gap gene concentrations (and thus the morphogen gradient) alone. This allows us to quantitatively decide whether the picture of sharp "on"/"off" domains of gene expression is sufficient, and to ask how variability across embryos impedes the ability of the patterning system to transmit positional information.

In this chapter, we set out to achieve the following goals: first, we give positional information a formal definition within the context of Shannons information theory [Shannon, 1948] and, second, we identify features of developmental gene expression that increase or decrease the information. We start by establishing the information-theoretic framework for positional information of gene expression patterns. To develop an intuition, we begin with a one-dimensional toy example of a single gene, which we later generalize to a many-gene system. We present scenarios where positional information is stored in different qualitative features of gene expression patterns. To capture this intuition mathematically, we give a precise definition of positional information for one gene and for multiple genes. Finally, we show how a quantitative formulation of positional information is related to "decoding," i.e. the ability of the nuclei to infer their position in the embryo.

1.2 Features of spatial gene expression profiles that carry positional information

Let us consider the simplest possible example where the expression of a single gene varies with position x along the axis of a one-dimensional embryo. We choose units of length such that $x = 0$ corresponds to the anterior pole and $x = 1$ to the posterior pole. Suppose we are able to quantitatively measure the gene profile in $\mathcal{N}$ embryos, labeled with index $\mu = 1, \ldots, \mathcal{N}$. These quantitative experiments are providing us with $G^{(\mu)}(x)$, where G is the quantitative readout in embryo μ of the gene expression level, e.g. the light intensity profile of fluorescently labelled antibodies against a particular gene product. From a collection of embryos—after suitable data processing steps described later—we can then extract two statistics: the position-dependent "mean profile," capturing the prototypical gene expression pattern, and the position-dependent variance across embryos, which measures the degree of embryo-to-embryo variability or the reproducibility of the mean profile. We can transform the measurements $G(x)$ into profiles $g(x)$ with rescaled units, such that the mean profile $\bar{g}(x)$ is normalized to 1 at the maximum and to 0 at the minimum along x. After these steps, our description of the system consists of the mean profile, $\bar{g}(x)$, and the variance in the profile, $\sigma_g^2(x)$.

How much can a nucleus learn about its position if it expresses a gene at level g? We will compare three idealized cases, where we pick the shape of $\bar{g}(x)$ by hand and assume, for the start, that the variance is constant, $\sigma_g^2(x) = \sigma_g$. The first case is illustrated in Figure 1.1 A, where a step-like profile in $\bar{g}$ splits the embryo into two domains of gene expression: an anterior "on" domain, where $\bar{g}(x) = 1$ for $x < x_0$, and an "off" domain in the posterior, $x > x_0$, where $\bar{g}(x) = 0$. This arrangement has an extremely precise, indeed infinitely sharp, boundary at x_0; if we think that the precision of the boundary is the biologically relevant feature in this system, this arrangement would correspond to an ideal gene. But how much information can nuclei extract from such a profile? If x_0 were $1/2$, the boundary would reproducibly split the embryo into two equal domains: based on reading out the expression of g, the nucleus could decide whether it is in the anterior or posterior, a binary choice that is equally likely prior to reading out g. The positional information needed (and provided by

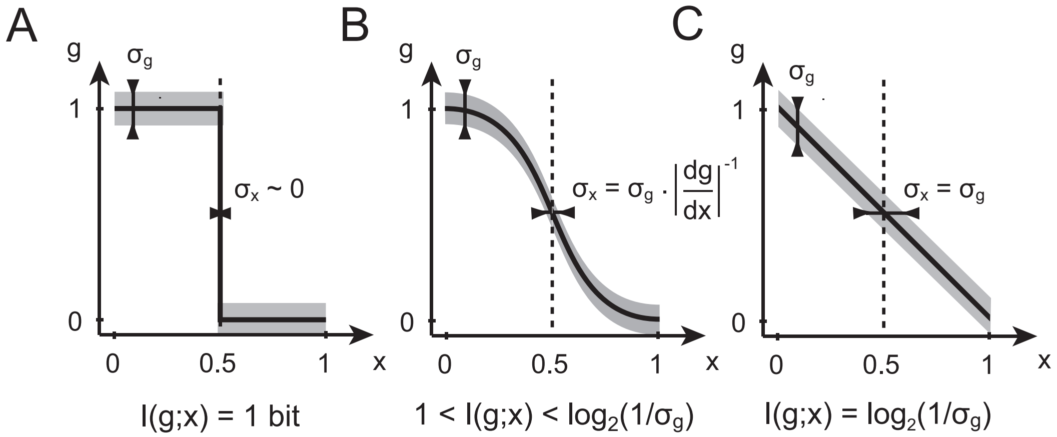

Figure 1.1: Positional information encoded by a single gene. Shown are three hypothetical mean gene expression profiles $\bar{g}(x)$ as a function of position x, with spatially constant variability σ_g. **A)** A step function carries (at most) 1 bit of positional information, by perfectly distinguishing between "off" (not induced, posterior) and "on" (fully induced, anterior) states. **B)** The boundary of the sigmoidal $g(x)$ function is now wider, but the total amount of encoded positional information can be higher than 1 bit because the transition region itself is distinguishable from the "on" and "off" domains. **C)** A linear gradient has no well-defined boundary, but nevertheless provides a further increase in information—if σ_g is low enough—by being equally sensitive to position at every x.

such a sharp profile!) to make a clear two-way choice between two *a priori* equally likely possibilities equals 1 bit.

Can a profile of a different shape do better? Figure 1.1 B shows a somewhat more realistic sigmoidal shape that has a steep, but not infinitely sharp, transition region. If the variance is small enough, $\sigma_g^2 \ll 1$, this profile can be more informative about the position. Nuclei far at the anterior still have $\bar{g} \approx 1$ (full induction or the "on" state), while nuclei at the posterior still have $\bar{g} \approx 0$ (the "off" state). But the graded response in the middle defines new expression levels in g that are significantly different from both 0 and 1. A nucleus with $g \approx 0.5$ will thus "know" that it is neither in the anterior nor in the posterior. This system will therefore be able to provide more positional information than the sharp boundary which is limited by 1 bit. Clearly, this conclusion is valid only insofar as the variance σ_g^2 is low enough; if it gets too big, the intermediate levels of expression in the transition region can no longer be distinguished and we are back to the 1 bit case.

The extreme contrast to the infinitely sharp gradient is the linear gradient, depicted in Figure 1.1 C. Wolpert already proposed that linear gradients might be efficient in encoding positional information [Wolpert, 1969], and indeed we can extend the argument for the sigmoidal case to convince ourselves that if σ_g^2 is not a function of position, the linear gradient is the best choice. Consider starting at the anterior and moving towards the posterior: as soon as we move far enough in x that the change in $\bar{g}(x)$ is above σ_g, we have created one more distinguishable level of expression in g, and thus a group of nuclei that, by measuring g, can differentiate themselves from their anteriorly-positioned neighbors. Finally, this reasoning gives us a hint about how to generalize to the case where the variance σ_g^2 depends on position, x. What is important is to count, as x covers the range from anterior to posterior, how much $\bar{g}(x)$ changes *in units of the local variability,* $\sigma_g(x)$—it will turn out that this is directly related to the mutual information between g and position.

Thus a sharp and reproducible boundary can correspond to a profile that does not encode a lot of positional information, and a linear profile where the boundary is not even well-defined can encode a high amount of positional information. Ultimately, whether or not there are intermediate distinguishable levels of gene expression depends on the variability in the profile. Therefore, any measure of positional information *must* be a function of both $\bar{g}(x)$ and $\sigma_g^2(x)$.

Does the ability of the nuclei to infer their position automatically improve if they can simultaneously read out the expression levels of more than one gene? Figure 1.2 A shows the case where two genes, g_1 and g_2, do not provide any more information than each one of them provides separately, because they are completely redundant. Redundant does not mean equal—indeed, in Figure 1.2 A the profiles are different at every x—but they are perfectly correlated (or dependent): knowing the expression level of g_1 one knows exactly the level of g_2, so g_2 cannot provide any additional new information about the position.

The situation is completely different if the two profiles are shifted relative to each other by, say, 25% embryo length. Note that none of the individual profile properties have changed: both are still infinitely sharp with two states of gene expression, and half of the nuclei express in each state. However, the two genes now partition the embryo into 4 different segments: the anterior-most domain which is combinatorially encoded by the gene expression pattern $\bar{g}_1 = \bar{g}_2 = 0$, the second domain with $\bar{g}_1 = 0,\ \bar{g}_2 = 1$, the third domain with $\bar{g}_1 = \bar{g}_2 = 1$ and the last domain with $\bar{g}_1 = 1,\ \bar{g}_2 = 0$. Upon reading out g_1 and g_2, a nucleus, a priori located in any of the four domains, can unambiguously decide on a single one out of the four possibilities. This is equivalent to making 2 binary decisions and, as we will later show, to 2 bits of positional information.

Finally, the most subtle case is depicted in Figure 1.2 C. Here, the mean profiles have exactly the same shapes as in Figure 1.2 B. What is different, however, is the correlation structure of the fluctuations. In certain areas of the embryo the two genes are strongly positively correlated, while in the others they are strongly negatively correlated. If these areas are overlaid appropriately on top of the domains defined by the mean expression patterns, an additional increase in positional information is possible. In the admittedly contrived but pedagogical example of Figure 1.2 C, the mean profiles and the correlations together define 8 distinguishable domains of expression, combinatorially encoded by 2 genes. Nuclei, having simultaneous access to the concentrations of the two genes, can compute which of the eight domains they reside in, although it might not be easy to implement such a computation in molecular hardware. Picking one of 8 choices corresponds to making 3 binary decisions, and thus to 3 bits of positional information. Note that in this case, each gene considered in isolation still carries 1 bit as before, so that the system of two genes carries more information than the sum of its parts—such a scheme is called synergistic encoding.

In sum, we have shown that the mean shapes of the profiles, as well as their variances *and* correlations, can carry positional information. Extrapolating to 3 or more genes, we see that the number of pairwise correlations increases and in addition higher-order correlation terms appear. Formally, positional information could be encoded in all of these features, but would become progressively harder to extract. Nevertheless, a principled and assumption-free measure should combine all statistical structure into a single number, a scalar quantity measured in bits, that can "count" the number of distinguishable expression states (and thus positions), as illustrated in the examples above.

1.3 Defining positional information

Using real data, determining the "distinguishable states" of gene expression is more complicated than in our toy models: the mean profiles have complex shapes and the (co)variability

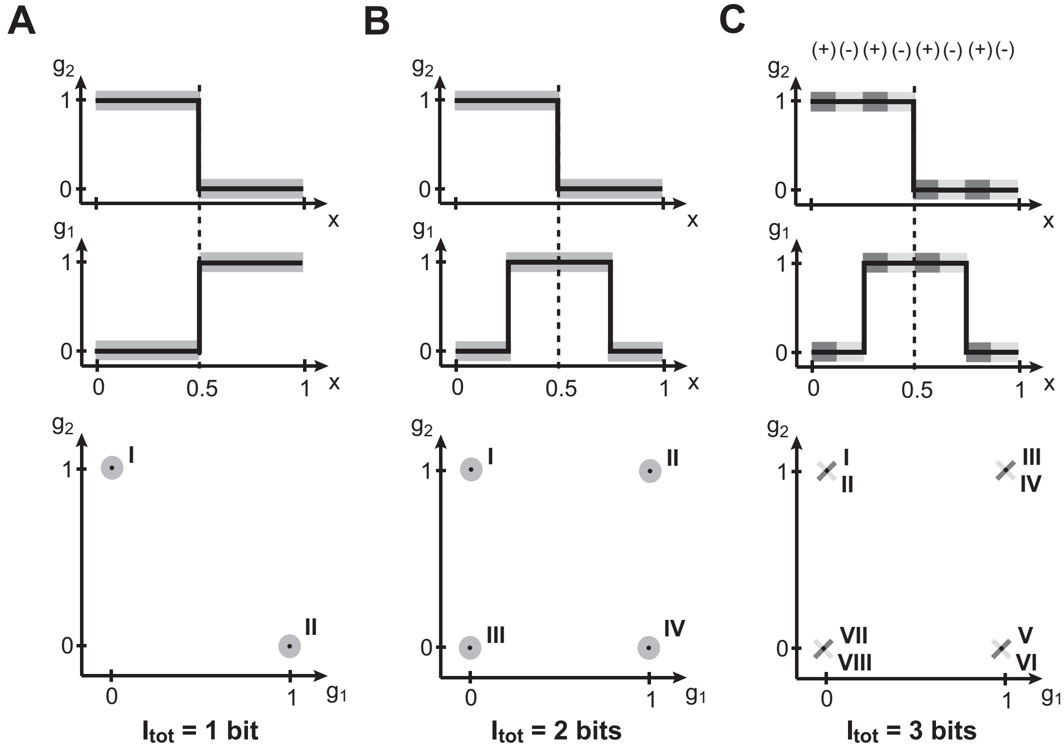

Figure 1.2: Positional information encoded by two genes. Spatial gene expression profiles are shown in the top row, while the bottom row schematically enumerates all distinguishable combinations of gene expression across the embryo. **A)** Genes g_1, g_2 are step functions, each encoding 1 bit of information. While the profiles are not the same, they are perfectly redundant, and the total number of jointly encoded “states” is only 2; this setup thus conveys only one bit of positional information, the same as each gene alone. **B)** The mean profiles of g_1, g_2 have been displaced such that the redundancy is broken and the total information encoded is 2 bits (4 distinct states of joint gene expression). **C)** If in addition to the mean profile shape the downstream layer can read out the (correlated) fluctuations of g_1, g_2, a further increase in information is possible. In this toy example, fluctuations are correlated (+) and anti-correlated (−) in various spatially separated regions, which allows the embryo to use this information along with the mean profile shape to distinguish 8 distinct regions, bringing the total encoded information to 3 bits.

depends on position. In the ideal case, where we could measure joint expression patterns of N genes $\{g_i\},\ i = 1, \ldots, N$ (for example $N = 4$ for four gap genes in *Drosophila*), in a large set of embryos, we could fully describe the position dependence of the expression levels with a conditional probability distribution $P(\{g_i\}|x)$. Concretely, for every position x in the embryo we would construct an N dimensional histogram of expression levels across all recorded embryos, which (when normalized) would yield the desired $P(\{g_i\}|x)$. This distribution contains all the information about how expression levels vary across embryos in a position-dependent fashion. For instance,

$$\begin{aligned}
\bar{g}_i(x) &= \int d^N\mathbf{g}\ g_i P(\{g_j\}|x), \qquad (1.1)\\
\sigma_i^2(x) &= \int d^N\mathbf{g}\ \left(g_i - \bar{g}_i(x)\right)^2 P(\{g_j\}|x), \qquad (1.2)\\
C_{ij}(x) &= \int d^N\mathbf{g}\ \left(g_i g_j - \bar{g}_i(x)\bar{g}_j(x)\right) P(\{g_j\}|x), \qquad (1.3)
\end{aligned}$$

are the mean profile of gene g_i, the variance across embryos of gene g_i, and the covariance between genes g_i and g_j, respectively. In principle, the conditional distribution contains also all higher-order moments that we could extract by integrating over appropriate sets of variables. Realistically, we are often limited in our ability to generate enough samples to construct $P(\{g_i\}|x)$ by histogram counts, especially when considering several genes simultaneously; the number of samples needed grows exponentially with the number of genes, N. However, estimating the mean profiles on the left-hand sides of Eqs. (1.1-1.3) can often be achieved from data directly. A reasonable first step (but one that has to be independently verified) is to assume that the joint distribution $P(\{g_i\}|x)$ of N expression levels $\{g_i\}$ at a given position x is Gaussian, and that it can be constructed using the measured mean values and covariances:

$$P(\{g_i\}|x) = (2\pi)^{-N/2}|C(x)|^{-1/2}\exp\left[-\frac{1}{2}\sum_{i,j=1}^{N}(g_i - \bar{g}_i(x))[C^{-1}(x)]_{ij}(g_j - \bar{g}_j(x))\right] \qquad (1.4)$$

We emphasize that this Gaussian approximation is not required to theoretically define positional information, but that it will turn out to be convenient when working with experimental data, as we will see in Chapter 3; in the cases of one or two genes it is often possible to proceed without making this approximation, which provides a convenient check for its validity.

While the conditional distribution, $P(\{g_i\}|x)$, captures the behavior of gene expression levels *at a given* x, it is often useful to ask questions about the range of gene expression in the embryo as a whole. How often is the expression level larger than, say, 0.5? What is, in general, the correlation between the expression levels of gene 1 and gene 2? The global structure is encoded in the total distribution of expression levels, which can be obtained by averaging the conditional distribution over all positions:

$$P(\{g_i\}) = \langle P(\{g_i\}|x)\rangle_x = \int_0^1 dx\ P(\{g_i\}|x), \tag{1.5}$$

where $\langle\cdot\rangle_x$ denotes averaging over x. Note that we can think of Eq. (1.5) as a special case of averaging with a position-dependent weight,

$$P(\{g_i\}) = \int dx\ P(x) P(\{g_i\}|x), \tag{1.6}$$

where $P(x)$ is chosen to be uniform. As we shall see, in our case $P(x)$ will be the distribution of possible nuclear locations along the AP axis, which is roughly uniform.

When formulated in the language of probabilities, the relationship between the position x and the gene expression levels can be seen as a statistical dependency. If we knew this dependency were linear, we could measure it using, e.g., a linear correlation analysis between x and $\{g_i\}$. Shannon has shown [Shannon, 1948] that there is an alternative measure of total statistical dependence (not just of its linear component), called the *mutual information*, which is a functional of the probability distributions $P_g(\{g_i\})$ and $P(\{g_i\}|x)$, and is defined by

$$I(x \to \{g_i\}) = \int dx\ P_x(x) \int d^N\mathbf{g}\ P(\{g_i\}|x) \log_2 \frac{P(\{g_i\}|x)}{P(\{g_i\})}. \tag{1.7}$$

This positive quantity, measured in bits, tells us how much one can know about the gene expression pattern if one knows the position, x. It is not hard to convince oneself that the mutual information is symmetric, that is $I(\{g_i\} \to x) = I(x \to \{g_i\}) = I(\{g_i\}; x)$.

This is very attractive: we do the experiments by sampling the distribution of expression levels given position, while the nuclei in a developing embryo implicitly solve the inverse problem—knowing a set of gene expression levels, they need to infer their position. A fundamental result of information theory states that both problems are quantified by the same symmetric quantity, information $I(\{g_i\}; x)$. Furthermore, mutual information is not just one out of many possible ways of quantifying the total statistical dependency, but rather the unique way that satisfies a number of basic requirements, for example that information from independent sources is additive [Shannon, 1948, Cover & Thomas, 1991].

The definition of mutual information in Eq. (1.7) can be rewritten as a difference of two entropies (which are always nonnegative):

$$I(\{g_i\}; x) = S[P(\{g_i\})] - \langle S[P(\{g_i\}|x)]\rangle_x, \tag{1.8}$$

where $S[p(x)]$ is the standard entropy of the distribution $p(x)$ measured in bits (hence log base 2):

$$S[p(x)] = -\int dx\, p(x) \log_2 p(x). \tag{1.9}$$

Equation (1.8) provides an alternative interpretation of the mutual information $I(\{g_i\}; x)$ which is illustrated on the main panel of Figure 1.3. In the case of a single gene g, the "total entropy" $S[P(g)]$, represented on the left, measures the range of gene expression available across the whole embryo. This total entropy, or dynamic range, can be written as the sum of two contributions. One part is due to the systematic modulation of g with position x, and this is the useful part (the "signal"), or the mutual information $I(\{g_i\}; x)$. The other contribution is the variability in g that remains even at constant position x; this represents "noise" that carries no information about position, and is formally measured by the average entropy of the conditional distribution (the noise entropy), $\langle S[P(g|x)]\rangle_x$. Positional information carried by g is thus the difference between the total and noise entropies, as expressed in Eq. (1.8).

Mutual information is theoretically well founded, and is always non-negative, being 0 if and only if there is no statistical dependence of any kind between the position and the gene expression level. Conversely, if there are I bits of mutual information between the position and the expression level, there are $\sim 2^{I(\{g_i\};x)}$ distinguishable gene expression patterns that can be generated by moving along the anterior-posterior (AP) axis, from the head at $x = 0$ to the tail at $x = 1$. This is precisely the property we require from any suitable measure of positional information. We therefore suggest that, mathematically, positional information should be defined as the mutual information between expression level and position, $I(\{g_i\}; x)$.

1.4 Decoding position and positional error

Thus far we have discussed positional information in terms of the statistical dependency and the number of distinguishable levels of gene expression along the position coordinate. To present an alternative interpretation, we start by using the symmetry property of the mutual information and rewrite $I(\{g_i\}; x)$ as

$$I(\{g_i\}; x) = S[P(x)] - \langle S[P(x|\{g_i\})]\rangle_{P(\{g_i\})}, \tag{1.10}$$

i.e., the difference between the (uniform) distribution over all possible positions of a cell in the embryo, and the distribution of positions consistent with a given expression level. Here, $P(x|\{g_i\})$ can be obtained using Bayes' rule from the known quantities:

$$P(x|\{g_i\}) = \frac{P(\{g_i\}|x)P(x)}{P(\{g_i\})}. \tag{1.11}$$

The total entropy of all positions, $S[P(x)]$ in Eq. (1.10), is independent of the particular regulatory system – it simply measures the prior uncertainty about the location of the cells in the absence of knowing any gene expression level. If, however, the cell has access to the expression levels of a particular set of genes, this uncertainty is reduced, and it is hence possible to localize the cell much more precisely; the reduction in uncertainty is captured by the second term in Eq. (1.10). This form of positional information emphasizes the *decoding view,* that is, that cells can infer their positions by simultaneously reading out protein concentrations of various genes (Figure 1.3).

Positional information is a single number: it is a global measure of the reproducibility in

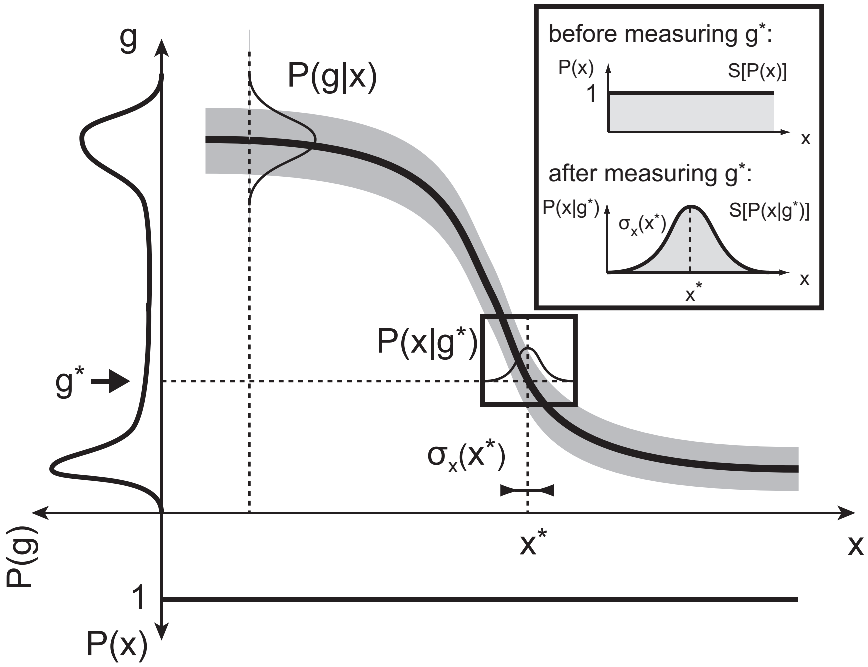

Figure 1.3: The mathematics of positional information and positional error for one gene. A schematic representation of the mean profile of gene g (thick black line) and its variability (shaded envelope) across embryos. Nuclei are distributed uniformly along the AP axis, which is mathematically equivalent to saying that the prior distribution of nuclear positions $P(x)$ is uniform (shown at the bottom). For each position x, the gap gene expression levels $P(g|x)$ are, in this example, Gaussian. The total distribution of expression levels across the embryo, $P(g)$, is determined by averaging $P(g|x)$ over all positions, and is shown on the left. The positional information $I(g;x)$ can be computed by averaging the difference of entropies of $P(g)$ and $P(g|x)$ over all positions x, as in Eq. (1.8). **Inset**: Decoding, or estimating the position of the nucleus, from a measured expression level g^*. Prior to the "measurement," all positions are equally likely. After observing the value g^*, the positions consistent with this measured value are drawn from $P(x|g^*)$. The best estimate of the true position, x^*, is at the peak of this distribution, and the positional error, $\sigma_x(x)$, is the distribution's width. Due to the symmetry of mutual information, positional information $I(g;x)$ is also equal to the average difference between the entropy of the uniform distribution $P(x)$ and the entropy of $P(x|g)$.

the patterning system. Is there a local quantity that would tell us, position by position, how well cells can read out their gene expression levels and infer their location? Is positional information "distributed equally" along the AP axis, or is it very non-uniform, such that cells in some regions of the embryo are much better at reproducibly assuming their roles?

Estimation theory tells us that the optimal estimator of the true location x of a cell, once we (or the cell) measure the gene expression levels $\{g_i^*\}$, is the maximum a posteriori (MAP) estimate, $x^*(\{g_i^*\}) = \text{argmax } P(x|\{g_i^*\})$. In cases like ours, where the prior distribution $P(x)$ is uniform, this equals the maximum likelihood (ML) estimate,

$$x^*(\{g_i^*\}) = \text{argmax } P(\{g_i^*\}|x); \tag{1.12}$$

thus, for each expression level readout, this "decoding rule" gives us the most likely position of the cell, x^*. The inset of Figure 1.3 illustrates the decoding in the case of one gene.

How well can this (optimal) rule perform? The expected error of the estimated x^* is given by $\sigma_x^2(x^*) = \langle (x - x^*)^2 \rangle$, where brackets denote averaging over $P(x|\{g_i^*\})$. This error is a function of the gene expression levels; however, we can also evaluate it for every x, since we know the mean gene expression profiles, $\bar{g}_i(x)$, for every x. Thus, we define a new quantity, the *positional error* $\sigma_x(x)$, which measures how well cells at a true position x are able to estimate their position based on the gene expression levels alone. This is the local measure of positional information that we were aiming for.

Independently of how cells actually read out the concentrations mechanistically, it can be shown that $\sigma_x(x)$ cannot be lower than the limit set by the Cramer-Rao bound [Cover & Thomas, 1991]:

$$\sigma_x^2(x) \geq \frac{1}{\mathcal{I}(x)}, \tag{1.13}$$

where $\mathcal{I}(x)$ is the Fisher information given by

$$\mathcal{I}(x) = -\left\langle \frac{\partial^2 \log P(\{g_i\}|x)}{\partial^2 x} \right\rangle_{P(\{g_i\}|x)}. \tag{1.14}$$

Despite its name, the Fisher information $\mathcal{I}$ is not an information-theoretic quantity and unlike the mutual information I, the Fisher information depends on position. Is there a connection between the positional error, $\sigma_x(x)$, and the mutual information $I(\{g_i\};x)$? Below, we sketch the derivation, following [Brunel & Nadal, 1998], demonstrating the link for the case of one gene, g.

Let's assume that the Gaussian approximation of Eq. (1.4) holds and that the distribution of the levels of a single gene at a given position is

$$P(g|x) = \frac{1}{\sqrt{2\pi\sigma_g^2(x)}} \exp\left\{-\frac{1}{2}\frac{(g-\bar{g}(x))^2}{\sigma_g^2(x)}\right\}. \tag{1.15}$$

We can use the Gaussian distribution to compute the Fisher information in Eq (1.14); if the noise is small, $\sigma_g \ll \bar{g}$, we find that

$$\sigma_x^2(x) \geq \frac{1}{\mathcal{I}(x)} = \left(\frac{d\bar{g}}{dx}\right)^{-2} \sigma_g^2(x). \tag{1.16}$$

This result is intuitively straightforward: it is simply the transformation of the variability in gene expression, $\sigma_g^2(x)$, into an effective variance in the position estimate, $\sigma_x^2(x)$, and the two are related by slope of the input/output relation, $\bar{g}(x)$.

A crucial next step is to think of x as determining gene expressions g_i probabilistically, and the $\hat{x}$ as being a function of these gene expression levels—that is, when computing x^* neither we nor the nuclei have access to the true position. This forms a dependency chain, $x \to \{g_i\} \to x^*$. Since each of these steps is probabilistic, it can only lose information, such that by information processing inequality [Cover & Thomas, 1991] we must have $I(\{g_i\};x) \geq I(x^*;x)$. The mutual information between the true location and its estimate is given by

$$I(\hat{x};x) = S[P(x^*)] - \langle S[P(x^*|x)]\rangle_{P(x)}. \tag{1.17}$$

Under weak assumptions, the first term in our case is approximately the entropy of a uniform distribution. While we don't know the full distribution $P(x^*|x)$ and thus cannot compute its entropy directly, we know its variance, which is just the square of the positional error, $\sigma_x^2(x)$. Regardless of what the full distribution is, its entropy must be less or equal to the entropy of the Gaussian distribution of the same variance, which is $S[P(x^*|x)] = \log_2\sqrt{2\pi e\sigma_x^2(x)}$

bits. Putting everything together, we find that:

$$I(\{g_i\};x) \geq I(x^*;x) \geq -\frac{1}{2}\langle \log_2 \left(2\pi e \sigma_x^2(x)\right)\rangle_x. \tag{1.18}$$

Therefore, positional information $I(\{g_i\};x)$ puts an upper bound to the average ability of the cells to infer their locations, that is, to the smallness of the positional error $\sigma_x(x)$. In a straightforward generalization of a single gene case, the positional error for the multiple genes case is given by:

$$\sigma_x^2(x) \geq \left(\sum_{i,j=1}^{N} \frac{d\bar{g}_i}{dx}[C(x)^{-1}]_{ij}\frac{d\bar{g}_j}{dx}\right)^{-1}, \tag{1.19}$$

where C_{ij} is the covariance matrix of the profiles, as defined in Eq. (1.3). This extends the fundamental connection, Eq. (1.18), between the positional information and positional error, to the case of multiple genes. Importantly, all quantities—the mean profiles and their covariance—in Eq. (1.19) can be obtained from experimental data, so $\sigma_x(x)$ is a quantity that can be estimated.

In sum, we have shown that a rigorous mathematical framework of positional information can quantify the reproducibility of gene expression profiles in a global manner. By framing the cells' problem of finding their location in the embryo in terms of an estimation problem, we have shown that the same mathematical framework of positional information places precise constraints on how well the cells can infer their positions by reading out a set of genes. These constraints are universal: regardless of how complex the mechanistic details of the cells' readout of the gene concentrations levels are, the expression level variability prevents the cells from decreasing the positional error below σ_x. The concept of positional error easily generalizes to the case of multiple genes, and it clearly shows how positional information is distributed along the AP axis.

Chapter 2

Quantification of the gap genes dynamics and reproducibility

2.1 Introduction

The final macroscopic outcome of developmental processes in multicellular organisms results in structures that are remarkably similar between individuals of a given species. This reproducibility has its origins in the reproducible spatial patterns of morphogen concentrations in the early embryo [Lawrence, 1992]. To fully understand the source and maintenance of this reproducibility, a number of conceptual and technical challenges need to be overcome. Conceptually, we are seeking to uncover novel quantitative traits in developmental regulatory networks, such as the time dependence of simultaneously measured expression levels, and their position-dependent variances and cross-correlations across populations of comparable embryos. Technically, such quantities must be measured in a way that is not biased by systematic experimental errors.

Quantitative data in this field is largely based on immunofluorescence staining [Surkova *et al.*, 2008, Fowlkes *et al.*, 2008, Ay *et al.*, 2008, Pisarev *et al.*, 2009], a technique that is restricted to fixed tissue and induces inherent experimental error due to variable conditions. These fluctuations are thought to have only a minor effect on the measurement of the mean expression profiles [Myasnikova *et al.*, 2009]. If one is interested in quantifying the variations that naturally arise from embryo to embryo, however, these experimental errors are bound to distort the signal. Moreover, the fact that this technique uses fixed tissues makes gene expression dynamics difficult to reconstruct, and because the number of simultaneously stainable genes is limited, comparisons of different genes at any given time point are restricted.

Here we address these issues in the context of the gap gene network in the early fly embryo and show that precise measurements in conjunction with proper quantification of experimental errors enables (i) simultaneous spatial and temporal monitoring of gene expression levels; (ii) quantification of reproducibility at an unprecedented level of precision; and (iii) hitherto unmatched access to questions about the interactions between the genes.

In embryos of the fruit fly *Drosophila melanogaster*, the earliest sign of a reproducible pattern has been found in the gradient of Bicoid (Bcd), which is the primary anterior determinant of patterning along the major body axis, and which is established by maternally deposited mRNA at the anterior pole of the egg [Ephrussi & St Johnston, 2004]. In its main region of activity between 10 and 60% egg length (EL), the gradient has been shown to be reproducible to ∼10% in Bcd concentration across different embryos, which translates into a positional accuracy of $1 - 2\%$ EL [Gregor *et al.*, 2007a].

Bcd, acting as a transcription factor, triggers the expression of Hunchback (Hb) and other downstream zygotic genes such as the gap, pair-rule, and segment polarity genes that interact with each other according to a complex network [von Dassow *et al.*, 2000, Tomancak *et al.*, 2007, Fakhouri *et al.*, 2010, Jaeger, 2011] and, three hours after fertilization, form precise patterns in which neighboring cells have readily distinguishable levels of gene expression [Gergen & Wieschaus, 1986], and which are reproducible from embryo to embryo [Houchmandzadeh *et al.*, 2002, Gregor *et al.*, 2007a]. In order to understand the strategies that nature uses to make embryonic development so reproducible, it is crucial to precisely monitor the variations over time in gene expression levels across embryos [Gregor *et al.*, 2007b]. In particular, what is the final reproducibility at gastrulation and how does it compare to the reproducibility of the Bcd gradient [Gregor *et al.*, 2007a]? How does the gap gene network responds to the Bcd input [Ochoa-Espinosa *et al.*, 2009] and transmit reproducibility from Bcd to the pair-rule genes [Spirov & Holloway, 2003, Manu *et al.*, 2009]?

In this paper, we present a set of methods and a number of careful control measurements that clearly demonstrate to what extent antibody stainings can be used for quantitative analysis. First, we developed a new set of antibodies for the four major gap genes, Hunchback, Krüppel (Kr), Giant (Gt) and Knirps (Kni), which allows us to stain individual embryos with all four antibodies at once, and hence to extract causal intergenic relationships of these genes at the particular developmental time point of the embryo's fixation. Second, we calibrate the developmental time point of fixed embryos with a precision of $1-2$ min using live membrane furrow invagination as a calibrator. Third, we identify 8 sources of experimental measurement errors, and show that the combined sum of these errors for any gene is less than 20% of the total variance in our data set, leaving ∼80% of the variance open for biological interpretation.

To demonstrate the potential of this approach, we connect the dynamics and reproducibility of the gap gene expression patterns with the positional precision of various position markers. Concretely, we compute the positional error of 20 individual position markers (such as boundary positions and the positions of the peak concentrations of all four major gap genes) in a set of 163 embryos. We show that at any time during nuclear cycle (n.c.) 14 all markers of all genes are reproducible above one internuclear distance. We further see a dynamic increase of this reproducibility (averaged over all markers) from one internuclear distance at the beginning of n.c. 14 to half an internuclear distance $40-45$ min later. Further analysis of the gap gene dynamics reveals that this reduction coincides with an overall zenith of the gap gene network, when all gap genes simultaneously peak in their maximum concentration as well as their maximum boundary slopes. These results lead us to postulate that collective network dynamics increase positional accuracy from the one internuclear distance level inherent to the Bcd gradient to half a internuclear distance observed in the pair-rule genes along the central 80% of the length of the embryo.

2.2 Quadruple antibody stainings in individual embryos

When labeling different proteins with fluorescent probes in the same tissue, there are two main limitations. The first limitation is biochemical: antibodies are raised in specific

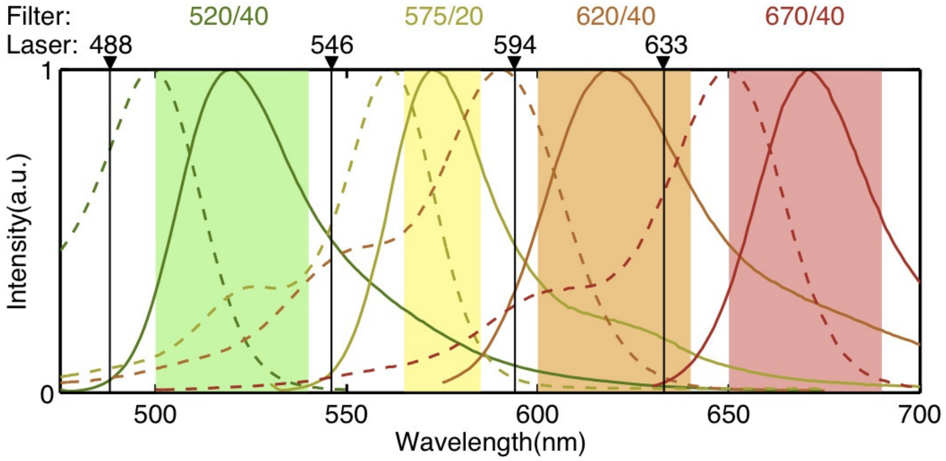

Figure 2.1: Spectral imaging settings for simultaneous measurements of four fluorescent dyes. Absorption (dashed lines) and emission (solid lines) spectra of the dyes used for 4 simultaneous immunostainings; laser excitation wavelengths (black arrows) and bandwidths of the emission filters for each detection channel (light color patches) are indicated.

host animals, so each antibody used to label a given tissue had to be made in a different animal, of which only a limited number exist commercially. The second limitation is optical: secondary antibodies are conjugated with fluorophores whose absorption and emission spectra can significantly overlap, which leads to crosstalk between the different optical channels, making quantification of co-labelled proteins challenging. Some attempts have been made in the past to try to overcome those limitations (direct labeling, use of quantum dots as fluorescent probes [Choi *et al.*, 2009], spectral imaging [Dickinson *et al.*, 2001], blind source separation [Neher *et al.*, 2009]) but there has been to our knowledge no attempt to use the classic immunofluorescence staining protocol to simultaneously quantify more than three co-labelled proteins with high precision.

To overcome the biochemical limitation, we developed a new set of primary polyclonal antibodies against three of the four main gap genes [Kosman *et al.*, 1998]. All antibodies were raised in a different host species (rat anti-Kni, guinea pig anti-Gt, and mouse anti-Hb), allowing us to use them in combination with a rabbit anti-Kr antibody (gift of C. Rushlow) to simultaneously stain of all four major gap genes in the same embryo without the risk of cross-reactivity between reagents.

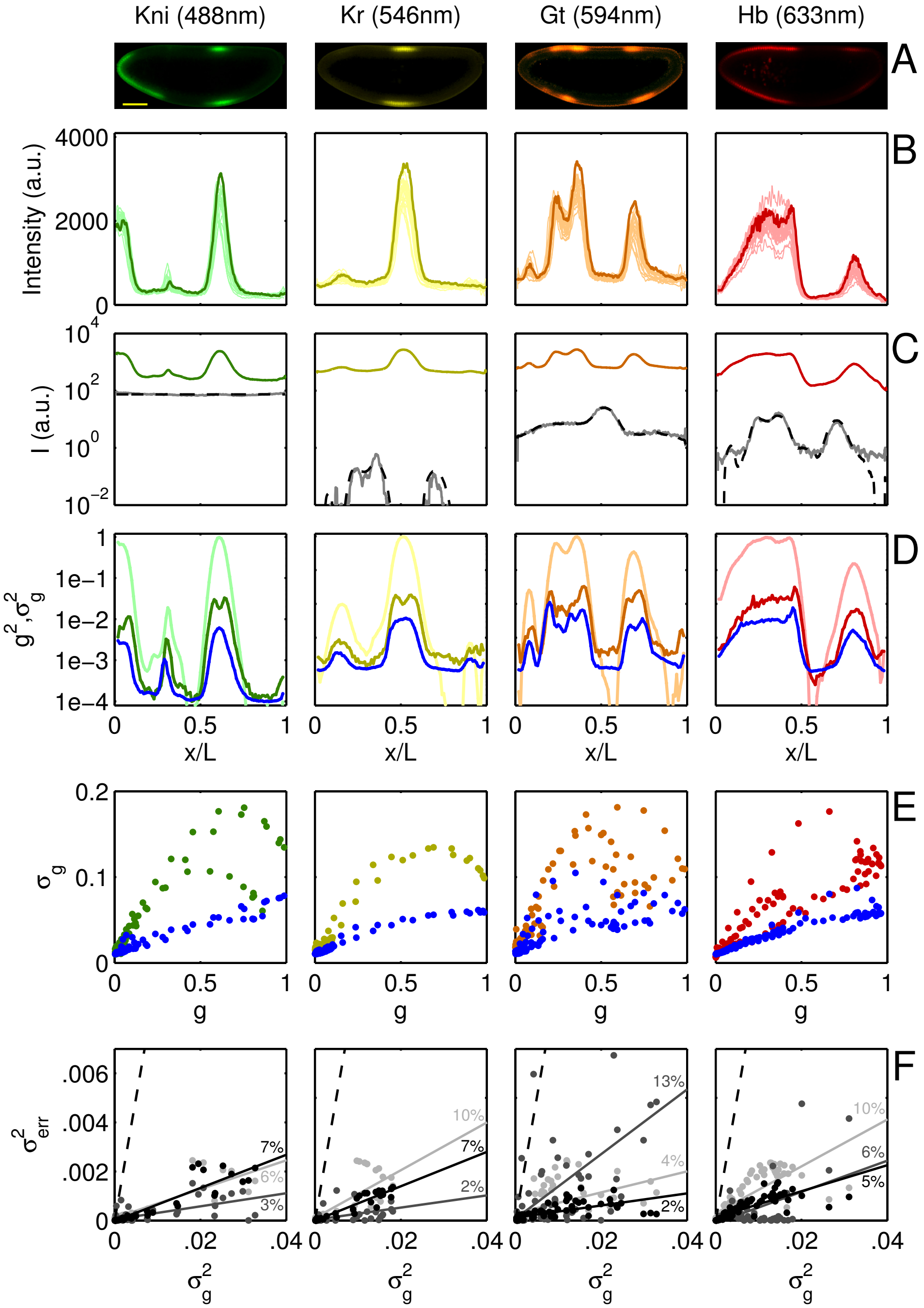

Figure 2.2: Systematic errors of gap gene expression profiles. **A)** Optical sections through the midsagittal plane of a single *Drosophila* embryo with co-immunofuorescence staining against the four gap genes Kni (green), Kr (yellow), Gt (orange) and Hb (red); scalebar 100 μm.**B)** Raw intensity profiles (dorsal side) of 23 selected embryos (light colors); embryo depicted above is highlighted in darker color. **C)** Quantification of spectral crosstalk and fluorophore bleed-through. For each channel, the average intensity profile I of 10 embryos immunofluorescently labeled with three antibodies lacking the specific antibody corresponding to that optical channel is shown in gray. Black dashed line shows a cross-talk estimate using a reconstruction algorithm (see text). Average profiles from B are shown in color. **D)**

Squared mean dorsal profiles (light color) and corresponding time-corrected variances (dark color) measured across 23 embryos ($\delta_{\mathrm{FC}} = 10 - 20\,\mu$m) as a function of fractional egg length x/L. Estimated summed total variance from major sources of measured systematic errors (staining, imaging, orientation, and time) are shown in blue. **E)** Standard deviation of gene expression levels in time-corrected normalized profiles (color) and standard deviation due to systematic error (blue) as a function of gene expression level g (for 100 equally spaced bins along the AP-axis). **F)** For each gap gene, we show a scatter plot of the variances due to the major sources of systematic error versus the total variance measured across embryos: imaging and staining (light grey), age (dark grey), and orientation (black). For each source of systematic error, data points were fitted with a straight line; slopes represent estimated average contributions to the overall variance. For reference, dashed line represents the case $\sigma_{\mathrm{g}}^2 = \sigma_{\mathrm{err}}^2$.

To overcome the optical limitation, we chose fluorophores that maximized the span of the employable laser excitation wavelengths to limit simultaneous excitation of multiple fluorophores, and we adjusted the bandwidths of the emission channels in order to minimize simultaneous emission from multiple fluorophores in the same imaging channel (see Figure 2.1).

We used this setup to stain and image a batch of 163 *Drosophila* embryos at the blastoderm stage. A typical embryo imaged in all four channels is depicted in Figure 2.2 A. Typical intensity profiles extracted from these images are shown in Figure 2.2 B, and cross-talk between channels is quantified in Figure 2.2 C.

2.3 Developmental time measurements in fixed embryos

2.3.1 Finding a 'clock' for the embryos

The *Drosophila* gap genes are endogenously transcribed for a time span of $1 - 2\,h$ ending with gastrulation, $3\,h$ after the onset of embryonic development. During this entire time, the patterns of gene expression change significantly. When quantifying the reproducibility or the dynamics of these patterns in fixed embryos, it is crucial to determine each individual embryo's age with high fidelity. To this end, we limit our analysis to n.c. 14, the $1\,h$ time window before gastrulation. During this time, cell membranes form across the entire embryo surface, and we can use the progression of this cellularization process (by measuring the depth δ_{FC} of the membrane furrow canal (FC)) as a 'clock' to estimate the age of the embryo (see Figure 2.3) [Lecuit *et al.*, 2002].

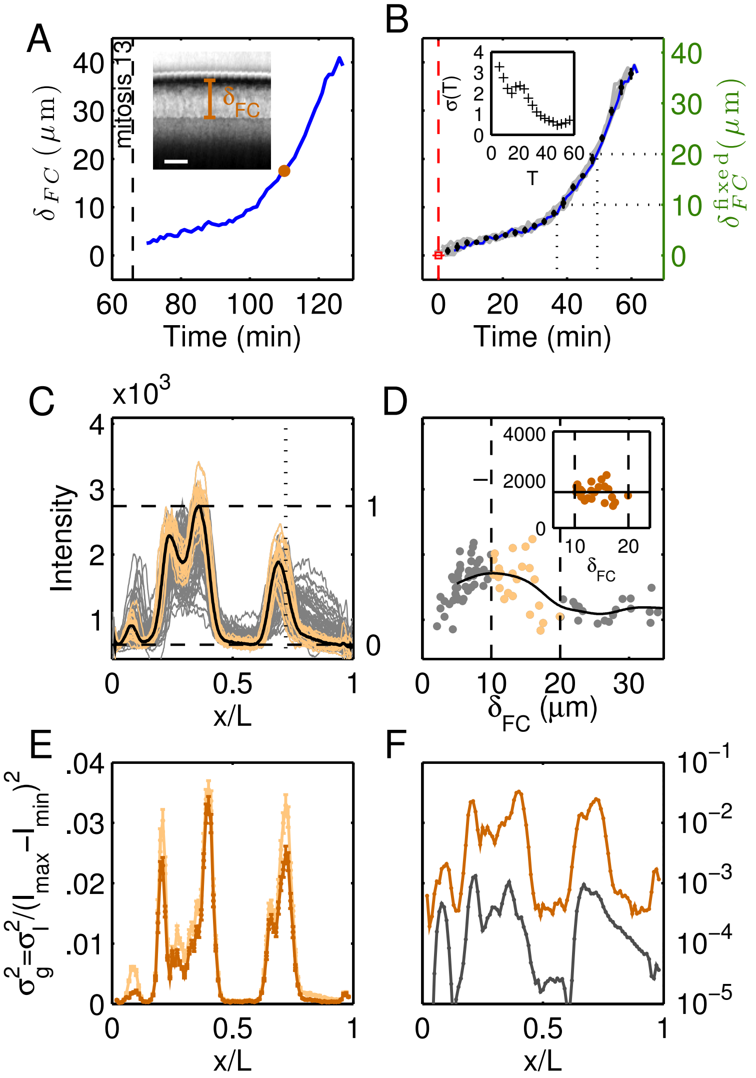

Figure 2.3: Timing of gene expression profiles in fixed embryos. **A)** Depth of FC (δ_{FC}) during blastoderm cellularization measured along the dorsal side of a live-imaged wild-type embryo during n.c. 14. Dashed line corresponds to the onset of n.c. 14. Inset shows a bright field microscopy image of a fraction (40-50%EL) of the dorsal side of the embryo at time point indicated by orange dot; scale bar 10 μm. δ_{FC} was measured as indicated by orange bar. **B)** Invagination of the membrane for 8 embryos (gray lines) with binned means and standard deviations in black. The mean of the onsets of n.c. 14 is indicated as a red dashed line and their standard deviation is represented by the error bar of the red square. Profile in A is shown in blue. Scale on the right shows the adjusted length measured in fixed embryos (on average 5%EL shrinkage w.r.t living embryos, see Online Methods). Inset shows the measurement error of the age estimation of the embryo as a function of time. **C)** Raw dorsal Gt intensity profiles of 87 embryos imaged in their midsagittal plane (DV orientation) with $\delta_{\text{FC}} = 0-40\,\mu$m (gray) and a subset of 23 embryos with $\delta_{\text{FC}} = 10-20\,\mu$m (light orange). The mean intensity profile of the 23 embryos is shown in black and its minimum and maximum in the 10-90%EL region are shown as dashed lines (defining the minimum (0) and maximum (1) gene expression levels, respectively). **D)** Intensity measured at x/L=0.72 (gray dotted line

in C) as a function of δ_{FC}, each point representing a different embryo. The 23 points of the $10 - 20\,\mu$m batch are plotted in light orange. The black line represents a nearest neighbor averaging with a Gaussian filter ($\sigma = 5\,\mu$m). Insert shows light orange data points with weighted average subtracted, i.e. time-corrected. **E)** Variance of Gt gene expression levels computed on the same 23 embryos shown in panel C before (light orange) and after (dark orange) time correction, respectively . Error bars obtained by bootstrapping. **F)** Estimation of residual variance in age determination (gray) due to measurement uncertainty of δ_{FC} after time-correction of the profiles (see Online Methods); for comparison, the variance of the time-corrected normalized profiles is plotted in dark orange. For a similar analysis for other gap genes see Figures 2.4 and 2.6.

To evaluate the precision of this method, we compare δ_{FC} across live-imaged embryos as shown in Figure 2.3 A. After profile alignment and correction for tissue shrinkage due to the fixation process, we find that the standard deviation of δ_{FC} across embryos never exceeds $1\,\mu$m, which translates into a temporal precision of ~ 2 min during the first 30 min of n.c. 14 and into less than 1 min during the next 30 min as shown in Figure 2.3 B (inset).

Hence, using this clock we can precisely select embryos for any given developmental time point during n.c. 14. Figure 2.3 C shows qualitatively how time selection reduces the variability of intensity profiles across embryos.

The precision of $1 - 2$ min with which we can determine embryo age during n.c. 14 allows us to correct embryo-to-embryo intensity profile fluctuations that are due to the underlying pattern dynamics. Hence for each position we compute the average time dependence of the expression level for each gene and use that average to apply a zero-order correction to the variances of the gene expression profiles (Figure 2.3 D). For profiles of embryos pooled together in the particular time window of $\delta_{\mathrm{FC}} = 10 - 20\mu$m (corresponding to $37 - 49$ min into n.c. 14) we find that in the case of Gt this correction is on average 15% of the original variance (Figure 2.3 E). For the other genes this time correction similarly reduces the profile variances by 10-20% on average, as shown in Figure 2.4.

2.3.2 Correction for profile normalization and time dependence

The 87 embryos co-immunostained against Kni, Kr, Gt and Hb and imaged in DV orientation were sorted w.r.t. increasing FC depth δ_{FC}. For each optical channel k and fractional length x/L, the intensity cloud $I_{k,x/L}(\delta_{\mathrm{FC}})$ was fitted using a Gaussian-weighted average of standard deviation $\sigma = 2.5\,\mu$m (see Figure 2.3 D for Gt at $x/L = 0.72$). For each time bin these

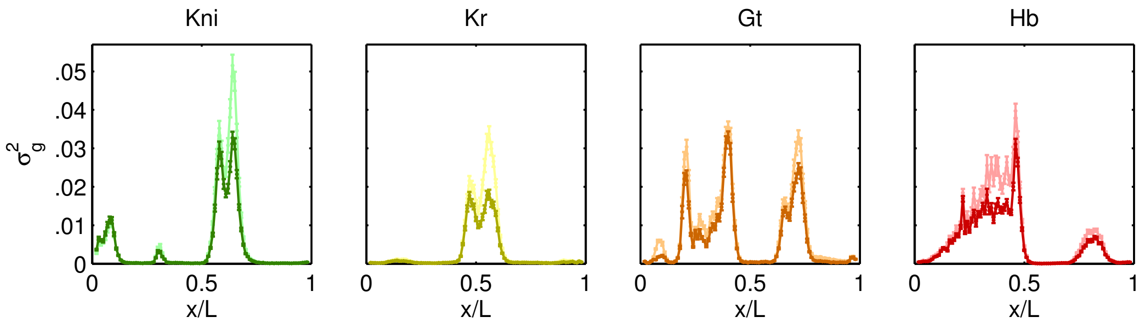

Figure 2.4: Reduction of gene expression profile variance due to time correction. Comparison of the time-corrected (dark colors) and non-time-corrected (light colors) variances in gene expression for each gap gene computed over 23 embryos (DV orientation only, $\delta_{\mathrm{FC}} = 10 - 20\ \mu$m). Error bars have been obtained by exhaustively bootstrapping over all subsets of $n - 1$ embryos.

fits were subtracted from the raw intensity profiles. Time-corrected intensity profiles were converted into gene expression profiles as follows: we determined the background intensity $I_{\min}$ (resp. maximum intensity $I_{\max}$) of the whole batch as the minimum (resp. maximum) intensity of any mean of 10 age-consecutive profiles; $I_{\min}$ was subtracted from each intensity profile and the resulting curve was rescaled by $1/(I_{\max} - I_{\min})$, so that the maximum average gene expression level over the course of n.c. 14 is 1. (Note: to avoid confusion when double-labeling axes, we sometimes call the maximum gene expression level G_0 as in Figure 2.13 and Figure 2.14.)

2.4 Systematic and statistical error quantification

When quantifying fluctuations of developmental processes such as the reproducibility across embryos, it is crucial to establish the contributions of the various systematic errors to the measured levels of gene expression. These errors induce extra variance in the levels across a population of embryos, masking the true statistical variance that naturally arises from the intrinsic variability of embryonic development. For an analysis of gap gene reproducibility to be meaningful, it is crucial to measure systematic error contributions as precisely as possible in order to be convinced that the sum of all systematic variances is only a small fraction of the overall variance σ_g^2.

To this end, we first verified the linearity between immunofluorescence intensity and protein concentration (Figure 2.5), and then we assessed the contributions of 6 possible sources of systematic errors in our experiments susceptible to perturb our measurement of σ_g.

1. The gap gene expression profiles change during the course of n.c. 14 such that the uncertainty in our estimation of an embryo's exact age can induce inconsistencies within a set of randomly selected embryos (Figure 2.3 F and Figure 2.4).
2. Laser fluctuations and photon detection induce noise in the imaging process (Figure 2.8).
3. Primary and secondary antibodies have non-specific binding properties (Figure 2.8 F) that affect the overall staining quality (Figures 2.7 B-C and 2.8).
4. There is potential cross-talk between neighboring optical channels (Figure 2.1).
5. Errors on the determination of the focal plane affect the profile extraction.
6. The profiles depend on the azimuthal embryo orientation which is hard to precisely control when mounting the samples (Figures 2.9-2.11). Figure 2.2 D shows in blue our estimate of the summed total variance σ_{err}^2 induced by these sources of systematic errors.

For any of the gap genes, it represents on average less than 20% of the gene expression variance σ_{g}^2 measured across embryos (Figure 2.2 E). The major sources of systematic errors stem from embryo orientation, age determination, and the imaging process. They are highly gene dependent but none is larger than 13% of σ_{g}^2 (see Figure 2.2 F).

We conclude that the largest contribution to σ_{g}^2 ($\sim$80%) is statistical and is hence due to biological fluctuations across embryos, and that our experimental method is suitable to quantify the gap gene expression patterns and analyze their reproducibility. Note that whenever the measured total variance of gene expression levels across embryos is used, one should be aware that one is working with an overestimate of the underlying biological variance and that measures of reproducibility are lower bounds on what nature can potentially achieve.

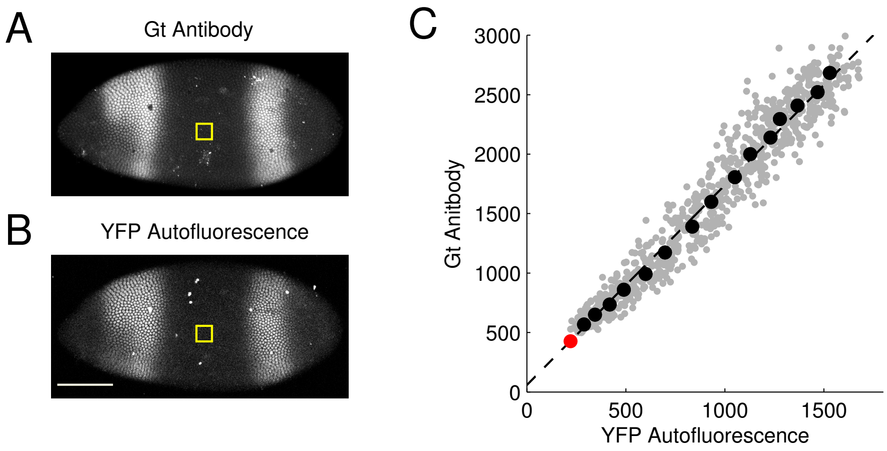

Figure 2.5: Linearity of gap gene antibody stainings. An embryo containing a Gt-YFP fusion protein was formaldehyde fixed at approximately 40 min into n.c. 14, stained with a guinea pig anti-Gt primary antibody and an Alexa-647 conjugated secondary antibody, and imaged via confocal microscopy. **A)** Image of YFP fluorescence; scale bar is 100 μm.) **B)** Image of immunofluorescence of same embryo as in A. Yellow square in A and B corresponds to a 50×50 pixel window used to determine background intensity. **C)** Scatter plot of raw nuclear intensities, in gray, and their binned average (black) computed over 15 equally populated bins (see Online Methods). Red dot shows mean intensity measured in yellow square in A and B. Standard deviations for each bin are smaller than the width of the black and red dots.

2.4.1 Linearity of antibody stainings

A fixed embryo of a Gt-YFP transgenic fly (gift of Michael Ludwig [Ludwig *et al.*, 2011]) was co-labeled with a fluorescent antibody against Gt (Figure 2.5 A) and nuclear fluorescence was compared to endogenous YFP fluorescence (Figure 2.5 B). To control for the difference in backgrounds, intensities were also compared over a 50 × 50 pixel window in a region were *gt* is not expressed. As shown on the scatter plot of Figure 2.5 B, the immunostaining and autofluorescence intensities are linearly related to each other. This result is all the more convincing because the linear interpolation of the means of nuclear intensities, computed across equally populated bins, also fits the background intensity (symbolized by a red point).

2.4.2 Embryo age determination

Despite all our efforts to precisely time embryo age, a fraction of the variance of the time-corrected profiles is still due to our systematic error in estimating the age of the embryo; due both to the limited reproducibility of $\delta_{\mathrm{FC}}(t)$ in the live imaging (Figure 2.3 B) and our

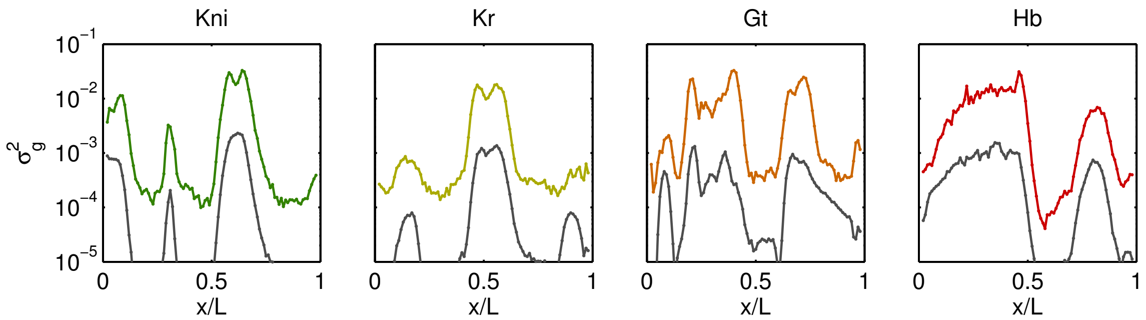

Figure 2.6: Residual gene expression profile variance due to measurement error in δ_{FC}. For each gap gene, the estimate of the residual time variance due to the uncertainty (± 1 μm) on the measure of δ_{FC} is shown in gray (see also Online Methods). For reference, the variances of the time-corrected normalized profiles is shown in dark colors.

measurement error of δ_{FC} in fixed tissues. The live imaging reproducibility error is estimated to $0.6\,\mu$m by averaging the standard deviation of the 8 gray curves over the whole n.c. 14 on the main panel of Figure 2.3 B . The measurement error was estimated to $0.7\,\mu$m by manually measuring δ_{FC} 5 times for all 163 embryos presented randomly and computing the average standard deviation over embryos. Hence the resulting total uncertainty on δ_{FC} is $\sqrt{0.6^2 + 0.7^2} \sim 1\,\mu$m. To find the contribution of this error to the observed variance, we added a Gaussian $1\,\mu$m noise to the original δ_{FC}'s of the 80 profiles shown in Figure 2.3 D-E and computed the variance of the newly time-corrected profiles $\sigma'_g(x)$. The variance $\sigma_{\mathrm{age}}(x)$ due to the error in age determination was then estimated as the extra variance introduced by the perturbation in δ_{FC} compared to the original data set, $\sigma_{\mathrm{age}}(x) = \sigma'_g(x) - \sigma_g(x)$ (see gray curve in Figure 2.3 F). By plotting σ_{age} vs σ_g and then fitting it with a straight line (see Figure 2.2 F), we find that the systematic error for embryo age determination contributes for Giant on average 13% (slope of the fit) to the total variance measured across the batch in the $\delta_{\mathrm{FC}} = 10 - 20\,\mu$m window (for other gap genes, see Figure 2.6).

2.4.3 Fluorophore cross-talk

A series of four control experiments was performed. In each experiment embryos were labeled as before with the omission of one of the four secondary antibodies, i.e. the one for which fluorophore cross-talk was assessed in that experiment. Embryos were imaged with the same experimental conditions as the quadruple staining batch, and dorsal intensity profiles for each channel were extracted from the images as described above (images were selected for a FC depth of $\delta_{\mathrm{FC}} = 10 - 20\,\mu\mathrm{m}$). Average intensity profiles measured in each channel are shown in Fig. 1C on a log-linear scale with the corresponding secondary antibody present (reference signal in dark color) and absent (control signal in gray). For any experiment, the mean intensity profile $\tilde{I}_i(x)$ measured in channel i can be written as a sum over 4 channels k as $\tilde{I}_i(x) = \sum_{k=1}^{4} C_{k\rightarrow i} I_k(x) + b_k$, where C is a matrix of unit diagonal terms and where the off-diagonal terms represent the crosstalk contribution of channel k to channel i; b_k is a positive constant for the contribution of the autofluorescence of heat fixed wild-type embryos to the signal in channel k, and $I_k(x)$ is the intensity of channel k without any cross-talk or autofluorescence contributions. To determine the off-diagonal coefficients of C, $\tilde{I}_i^{\mathrm{cont}}(x) = \sum_{\substack{k=1\\k\neq i}}^{4} C_{k\rightarrow i} I_k(x) + b_k$ of the ith control experiment was computed, where the mean intensities $I_k(x)$ were approximated to first order by the measured mean intensities $\tilde{I}_k(x)$, with all secondary antibodies present. Coefficients $c_{k\rightarrow i}$ and b_k were obtained by minimizing the Euclidean distance between $\tilde{I}_i^{\mathrm{cont}}(x)$ and $\sum_k c_{k\rightarrow i} I_k(x)$ for any channel i. The final $C_{k\rightarrow i}$'s and the b_k's are given by

$$C = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0.0085 & 0 \\ 0 & 0.0001 & 1 & 0.0029 \\ 0 & 0 & 0.0006 & 1 \end{pmatrix},$$

and $B = \begin{pmatrix} 73.7 & 0 & 0 & 0 \end{pmatrix}$. The corresponding estimated mean control profiles are plotted in black dashed lines on the bottom panel of Figure 2.3 C.

2.4.4 Imaging process

A single wild-type embryo was fixed and incubated simultaneously with rat anti-Hb and guinea pig anti-Hb antibodies. After excessive rinsing it was incubated in Alexa conjugated anti-rat(488nm), anti-guinea pig(546nm), and anti-guinea pig(647nm) antibodies and imaged twice on each of the three optical channels. For each channel, the nuclear intensities of the two successive images are scatter plotted Figure 2.8 A-C. After correction for the systematic error due to photobleaching and change in laser power (change in the slope of a linear

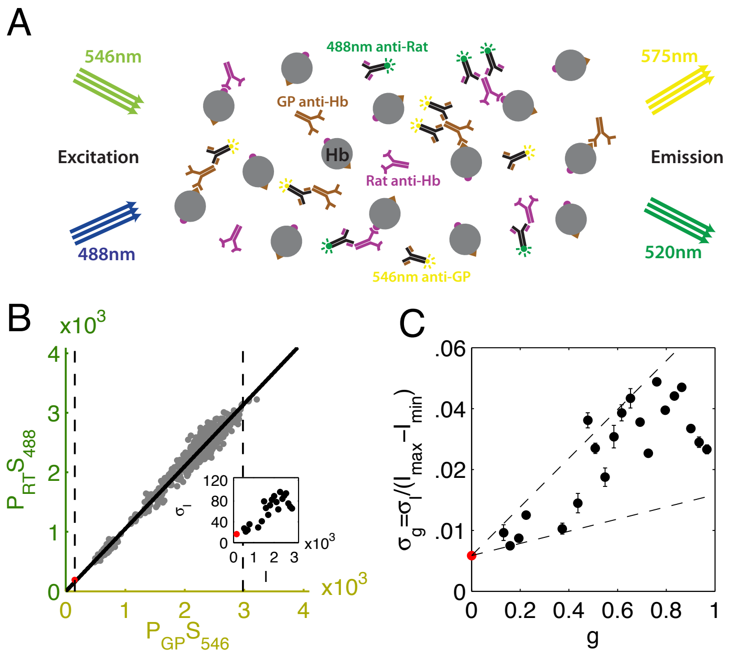

Figure 2.7: Imaging noise and antibody non-specificities. **A)** Schematic representation of a simultaneous staining process by two distinct primary anti-Hb antibodies and their respective secondary antibodies conjugated with differently colored dyes. The systematic error σ_I in nuclear intensity measured in each optical channel comes from imaging noise (mainly due to photon counting and laser intensity fluctuations) as well as the random binding processes for primary and secondary antibodies. **B)** Scatter plot of the nuclear intensities from two channels with different anti-Hb antibodies and Alexa conjugated dyes (anti-rat(488) and anti-guinea pig(546)). Two dashed lines define the minimum (red dot) and the maximum gene expression levels, $I_{\min}$ and $I_{\max}$, respectively. Inset shows dependence of measurement noise σ_I (standard deviation of the distance of the points to the diagonal divided by $\sqrt{2}$) as a function of the mean intensity I, computed over 40 equally spaced bins along the whole intensity range (only data points with five or more nuclei are shown). Standard deviation of zero expression level region shown as a red dot. **C)** Rescaled measurement noise $\sigma_g = \sigma_I/(I_{\max} - I_{\min})$ (computed from the data shown in inset of panel B) as a function of the gene expression level g. Red dot shows the background noise, and two dashed lines represent the 1% and 4% measurement noise levels, respectively. Error bars computed by bootstrapping.

fit to these data is less than 2%), the standard deviation of the intensity σ_I is plotted as a function of its mean I for 40 bins equally spaced in intensity, as shown in insets of Figure 2.8 A-C (for consistency, only bins with 5 or more nuclei are shown). For each optical channel, the background intensity I_{min} is determined as the average intensity of ten randomly selected nuclei in the $60 - 70\%$ EL region, and the maximum gene expression intensity I_{max} by averaging over the 10 brightest nuclei in the image. The standard deviation of the background is less than 0.5% of the total amplitude of the signal $I_{\text{max}} - I_{\text{min}}$ (background noise), and the standard deviation within each bin is on average 1% of the gene expression level after background subtraction and normalization by $I_{\text{max}} - I_{\text{min}}$ (measurement noise).

2.4.5 Antibody non-specificities

Secondary antibodies. Using the same data as described above in Figure 2.8, panel E compares the nuclear intensities for the two stainings with identical primary antibodies (guinea pig anti-Hb) but with secondary antibodies of different colors (Alexa-546 and Alexa-647 guinea pig). Applying the same noise analysis as in the previous section, we find a background noise of 1.7% of the total amplitude of the signal and the measurement noise to be equal to 0.7% of the gene expression level after background subtraction and normalization by $I_{\text{max}} - I_{\text{min}}$. After subtracting the imaging noise, we find that the sole secondaries generate a background noise equal to $\sqrt{0.7^2 - 0.5^2} = 0.5\%$ of the maximum level of gene expression and a measurement noise less than $\sqrt{1.7^2 - 1^2} = 1.4\%$ of the gene expression level. All error bars of standard deviations have been obtained by bootstrapping.

Primary antibodies. Using the same data as described above in Figure 2.8, panel D compares the nuclear intensities for the two stainings with primary antibodies raised in two different animals and with secondary antibodies that are Alexa conjugated with two different colors, anti-rat(488nm) and anti-guinea pig(546nm), respectively. A background noise equal to $\sqrt{1^2 - 0.7^2} = 0.7\%$ of the maximum level of gene expression is detected, and a measurement noise less than $\sqrt{2.7^2 - 1.7^2} = 2.1\%$ of the maximum gene expression level. All error bars of standard deviations have been obtained by bootstrapping.

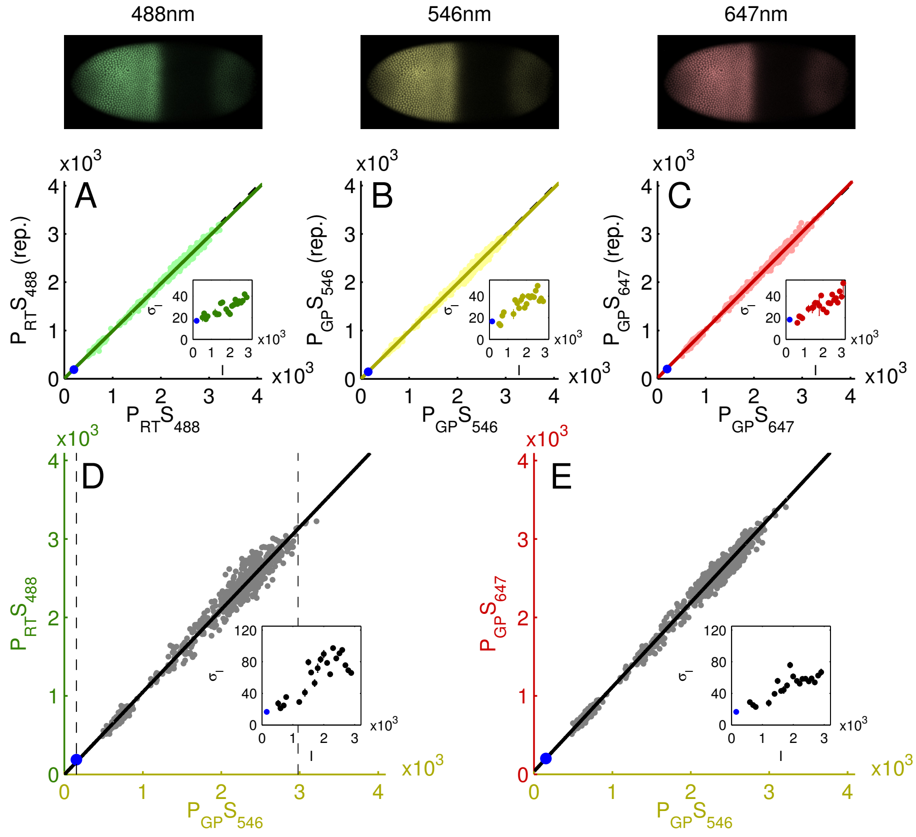

Figure 2.8: Systematic error contributions due to noise in the imaging process, and due to non-specificities of primary and secondary antibodies. An embryo, heat fixed approximately 40 min into n.c. 14, was stained with two different primary antibodies (rat and guinea pig anti-Hb) and three different secondary antibodies (Alexa conjugated 488 anti-rat, 546 anti-guinea pig, 647 anti-guinea pig). Embryo imaged in three different channels via confocal microscopy can be seen at top (green-488, yellow-546, red-647). **A-C)** Scatter plots of the raw nuclear intensities (712 nuclei) from two independent successive imaging sequences, taken in the same channel (light colored dots, repeat sequence on the y-axis). Blue dots represent the measurement of the background, performed by averaging the intensity over 10 nuclear size masks randomly selected in the 60-70%EL region of the embryo. The dark color lines are linear fits to the raw nuclear intensities, and the black dashed line corresponds to equal intensities on the x- and y-axis (slope=1). Insets show intensity measurement noise σ_I as a function of mean intensity computed across 40 equally spaced bins along the full intensity spectrum. For consistency only bins with 10 or more points are shown. **D)** Scatter plot of the nuclear intensities from channels with different primary and secondary antibodies (anti-rat(488) and anti-guinea pig(546)). Two dashed lines show were the zero and one levels of gene expression are determined. Inset shows the measurement noise of the intensity as a function of the mean intensity. **E)** Scatter plot of the nuclear intensities from channels with different secondary antibodies (anti-guinea pig(546) and anti-guinea pig(647)). Inset as in D.

2.4.6 Focal plane determination

For each individual embryo the focal plane has to be hand adjusted before image acquisition. Using the bright field channel, we adjust the focal plane to be at the midsagittal plane of the embryo but estimate our uncertainty to be $6\,\mu$m, or one nuclear diameter. The resulting error is estimated by computing the variances of nuclear intensities across 7 images taken at consecutive heights spaced by $1\,\mu$m in a single embryo.

2.4.7 Embryo azimuthal orientation

The inability to control the azimuthal angle around the embryo's AP-axis when acquiring two-dimensional images is the biggest contribution to the overall systematic error. For this reason we present two independent methods to estimate this error.

First method. 46 embryos with $10 < \delta_{\mathrm{FC}} < 20\,\mu$m were manually separated into two equally likely orientations: DV, when imaged closer to the midsagittal plane (23 embryos $\langle|\phi_1|\rangle \simeq 22.5^\circ$) and LR, when imaged closer to the coronal plane (23 embryos $\langle|\phi_2|\rangle \simeq 67.5^\circ$) (see Figure 2.9 A-B). The dependence on azimuthal angle of each gene in the dorsal region was estimated as $dI(x)/d\phi = (I_{\mathrm{DV}}(x) - I_{\mathrm{LR}}(x))/45$. The error on the intensity measurement induced by this azimuthal dependence was then estimated by propagating the corresponding variance $\sigma_I = \sigma_\phi \cdot dI/d\phi$ (see Figure 2.9 C), assuming that embryos in the DV orientation have their absolute azimuthal angles uniformly distributed between 0° and 45°. The error in the intensity values was converted into an error in gene expression levels σ_g by dividing σ_I by the squared amplitude of the measured intensities.

Second method. We imaged the dorsal surface and the coronal plane of a single embryo flattened and co-immunostained against the gap genes. The embryo was chosen such that its FC depth ($12\,\mu$m) falls within the same depth window as above ($10 < \delta_{\mathrm{FC}} < 20\ \mu$m). For each gap gene, the average intensity of all identifiable nuclei was extracted, and the azimuthal angle with the midsagittal plane was computed by $\phi = 90 \times y/y_{\mathrm{max}}$, where y is the distance from the nuclei to the dorsal line represented as a dashed line in Figure 2.10, and y_{max} is the the distance from the outer edge of the embryo to the AP-axis. Nuclei were sorted in 10° wide bins ranging from -55° to 55°, represented by colors ranging from

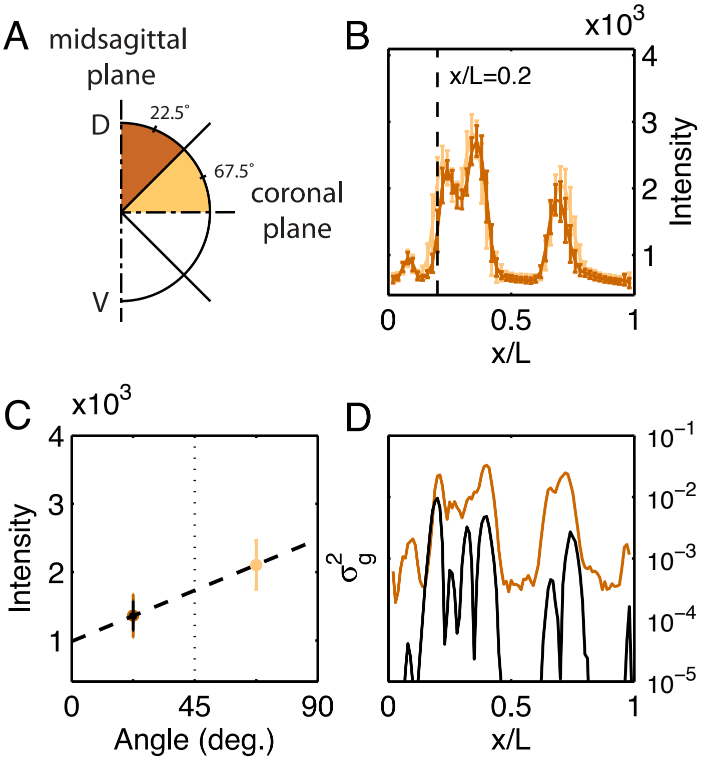

Figure 2.9: Contribution of azimuthal embryo orientation uncertainty to Gt gene expression profile reproducibility I. **A)** Schematic representation of the absolute azimuthal angle distribution of the embryos depending on their imaging plane. **B)** Mean and standard deviation of dorsal time-corrected Gt intensity profiles of embryos whose imaging plane is closer to the midsagittal plane (dark orange, DV orientation, 23 samples, $\delta_{\mathrm{FC}} = 10 - 20\ \mu$m) or closer to the coronal plane (light orange, LR orientation, 23 samples, $\delta_{\mathrm{FC}} = 10 - 20\ \mu$m) **C)** Linear estimate of the DV dependence of raw dorsal Gt intensity at the fractional egg length $x/L = 0.2$. The systematic error in intensity due to embryo orientation uncertainty of the dark brown profiles σ_I (shown in black) is estimated by propagating the uncertainty on the azimuthal angle σ_ϕ (uniform distribution $\phi = 0 - 45°$), $\sigma_I = \sigma_\phi \cdot dI/d\phi$, where $dI/d\phi$ is the slope of the dashed line (see text). **D)** Variance of Gt gene expression profiles due to embryo orientation uncertainty (in black), computed as $\sigma_g^2 = \sigma_I^2/(I_{\max} - I_{\min})^2$; total variance of gene expression profiles of the time corrected dorsal profiles is shown in dark orange for comparison.

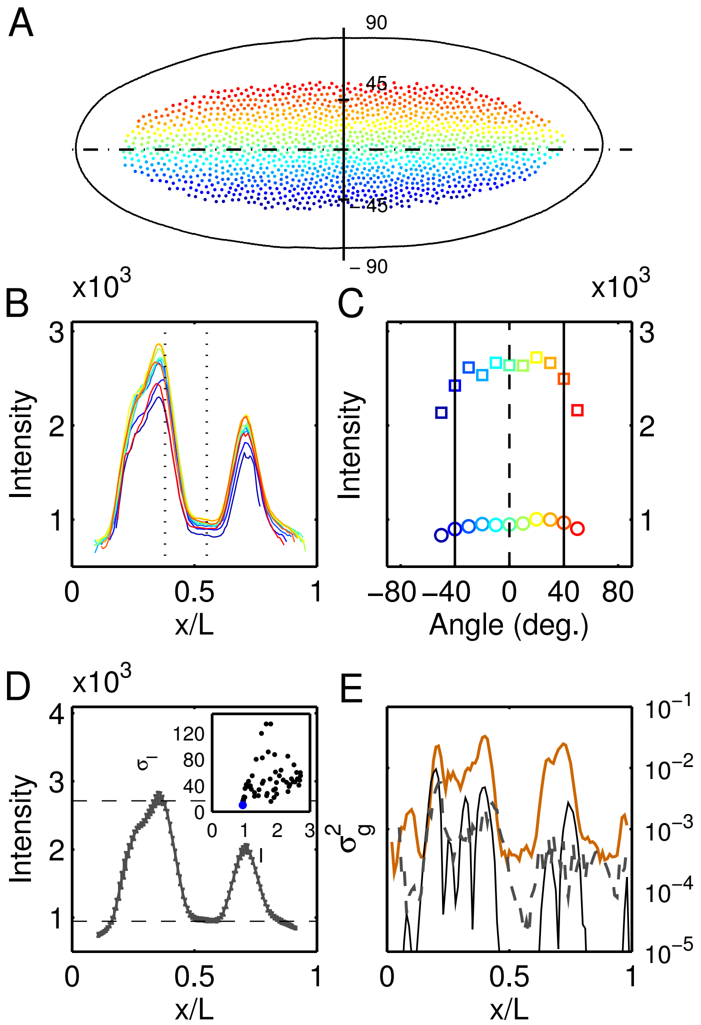

Figure 2.10: Contribution of azimuthal embryo orientation uncertainty to Gt gene expression profile reproducibility II. The dorsal side of a flattened embryo (fixed 40 min after the onset of n.c. 14 (FC depth = 12μm) and immunostained against *kni*, *kr*, *gt* and *hb*) was imaged. **A)** Nuclear mask of the embryo. Each nucleus was assigned an 'angle' ϕ defined as $y/y_{\max}$, where y is the y-coordinate of the nucleus and $y_{\max}$ the y-coordinate of the outline (in black) at the same AP position (x-coordinate). **B)** Gt profiles for each bin. **C)** Dependence of intensity on angle. Intensity measured in each bin is plotted for $x/L = 0.38$ (□) and $x/L = 0.55$ (○). Typical eye-selection allows allows for discrimination of embryos for which the midsaggital plane lies within $-40° < \phi < 40°$ (black lines). **D)** Mean and standard deviation of Gt intensity profiles computed across the the 7 profiles in the $-40° < \phi < 40°$ bin. Inset shows the standard deviation as a function of the mean for 100 equally spaced points along the AP-axis, with the blue dot representing the standard deviation of the background intensity. **E)** Variance of the 7 previous profiles plotted as a function of the fractional length (dashed grey line); total variance of gene expression profiles of the time corrected dorsal profiles is shown in dark orange for comparison; plain black line shows an estimation of the variance induced by the embryo orientation uncertainty by the alternative method of Fig. S6.

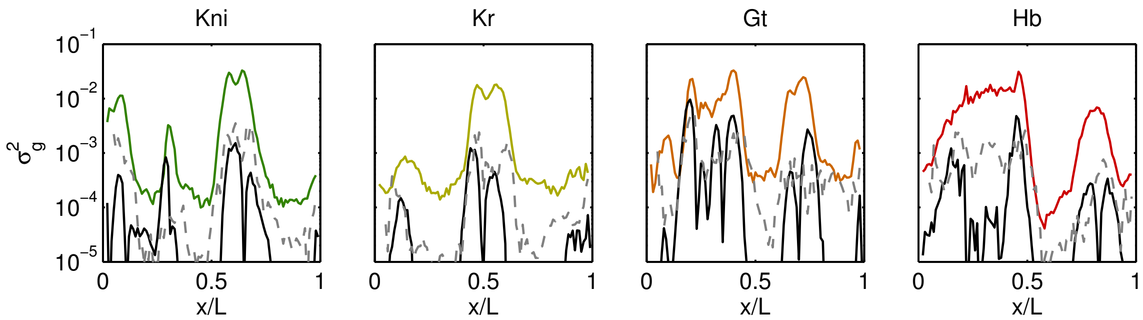

Figure 2.11: Contribution of azimuthal embryo orientation uncertainty to gap gene expression profile reproducibility. For each gap gene, the estimate of the variance due to the uncertainty on the measure of the orientation of the embryo (logarithmic scale) is shown as a function of x/L; estimated as in Fig. 2.9 (black plain line) and as in Fig. 2.10 (gray dashed line). The variances of the time-corrected normalized profiles are shown as a reference in dark colors.

blue to red. For each bin a mean nuclear intensity profile along the AP-axis was extracted, illustrated in Figure 2.10 B-C for Gt intensity profile $I(x/L)$ as a function of the angle ϕ. The dependence on ϕ is shown at the position of highest and lowest levels of gene expression ($x/L = 0.38$ and $x/L = 0.55$) . For each AP position x/L, the mean level of gene expression $g(x/L)$ was computed and its standard deviation across angle bins $\sigma^2(x/L)$ is shown for angles varying from $-40°$ to $40°$ (which we estimate to be our range of misidentification of the dorsal line). The mean profile and corresponding standard deviation for Gt is represented as a function of x/L in Figure 2.10 D. Results are in good agreement with the first method (see Figure 2.10 E for Gt and Figure 2.11 for the other gap genes).

2.5 Dynamics of gap gene expression profiles

The results in the preceding three sections allow us to fully reconstruct the dynamics of the four major gap gene expression profiles from fixed tissue: simultaneous labeling of the four genes in single embryos preserves temporal causality; embryos can be staged with a temporal precision of $1 - 2\,\text{min}$; and systematic errors are low and controllable. We merged and age-classified the dorsal intensity profiles of Hb, Kr, Gt, and Kni simultaneously extracted from 80 embryos during n.c. 14. Profiles were rescaled such that their maximum level of gene expression over the course of n.c. 14 is 1 for each gap gene. Chronologically ordered 1 min snapshots were generated by time-averaging intensities at each position with a 3 min

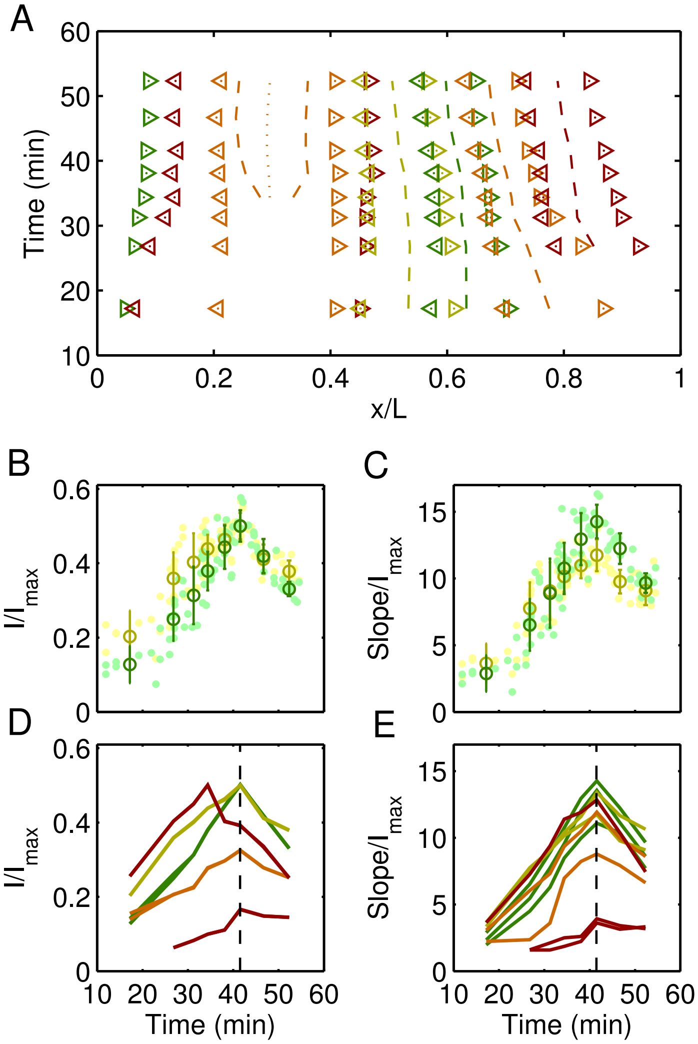

Figure 2.12: Dynamics of the main boundaries of the gap gene expression profiles. **A)** Positions of anterior (◁) and posterior (▷) borders (inflection points) of the gap gene expression profiles as a function of time. Color code for different genes as above. As a guide-to-the-eye, the peaks of the major stripes are plotted in dashed lines and the interstripe local minimum of Gt is plotted in dotted line. **B)** Developmental progression of the intensity at the posterior Kr border (yellow) and the anterior Kni border (green). For each embryo and each gene, intensity at the inflection point is plotted as a function of embryo age in n.c. 14 (dim dots). Darker circles represent the averages and standard deviations of the intensities of these data points over 8 equally populated time-bins (10 embryos per bin). Data have been normalized by $I_{\max}$, the maximum intensity reached over all embryos, positions, and bins during n.c. 14. **C)** Developmental progression of the absolute value of the slope at the Kr (yellow) and Kni (green) crossing as in B. **D)** Summary of the time dependences of the inflection point intensities for all gap gene borders (green and yellow lines correspond to the dark circles in B (for all stripes, anterior and posterior intensities almost fall on the top of each other). **E)** Summary of the time dependences of the inflection point slopes of all gap gene borders (green and yellow lines correspond to dark circles in C).

wide gaussian distribution to produce a movie of the gap gene time-dependence during the entire course of n.c. 14.

To analyze the statistical properties of gap gene dynamics, embryos were sorted according to their age and divided in 8 equally populated time bins $T_1 - T_8$ with 10 samples each and whose average ages range from 15 to 55 min after the onset of n.c. 14. For each embryo and each gene, we tracked the position of the peaks in gene expression as well as the position of the inflection points (i.e. pattern boundaries) in the region of transition between high and low levels of gene expression. The locations of these markers shift during n.c. 14, as shown in Figure 2.12 A, some by almost 10%EL. Notice that gene expression patterns tend to be more dynamic near the edges of the embryo, where the terminal group genes are active. Overall our analysis of these shifts is in good agreement with previous reports [Jaeger *et al.*, 2004].

To understand how the average levels of gene expression change with time, we monitor the immunofluorescence intensity of the boundary inflection points as well as the absolute value of the slope at these points as a function of embryo age. In Figure 2.12 B andFigure 2.12 C we analyze the intensity of two of these points at $x/L \simeq 58\%$, where the expression boundaries of Kr and Kni intersect. We find that the intensities and the slopes of the borders simultaneously reaches their maximum around 42 min after the onset of n.c. 14. Remarkably, this trend is conserved for the borders of all other gap genes (Figure 2.12 D and Figure 2.12 E), except for the intensity of the main Hb border, which peaks at 34 min. These results suggest that the collective intrinsic dynamics of the gap gene network has a culmination point at 42 min, where one should expect the network to hold special properties.

2.6 Reproducibility of gap gene expression profiles

Next, we examined how the reproducibility of the profiles evolves over time during n.c. 14, using the distributions of the positional markers described in Figure 2.12 in eight different time bins $T_1 - T_8$ for increased statistical significance. We define the marker reproducibility $\sigma_{x/L}$ as the standard deviation across embryos of the markers' fractional positions x/L. In Figure 2.13 A these standard deviations are plotted as a function of their respective mean fractional position for time bins T_1, T_6, and T_8 (blue data; all $T_1 - T_8$ inFigure 2.14).

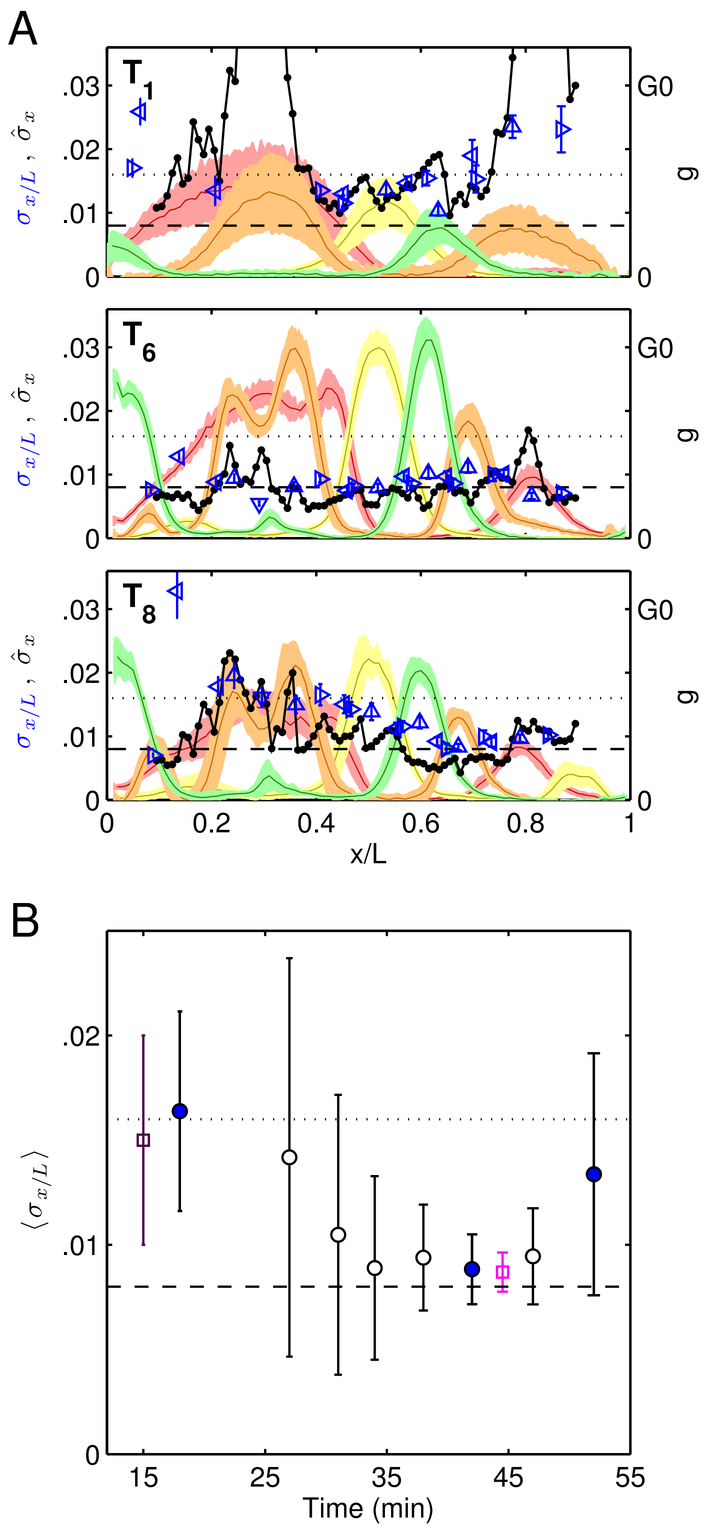

Figure 2.13: Temporal evolution of gene expression profile reproducibility. **A)** Reproducibility (standard deviations $\sigma_{\mathrm{x/L}}$) of the marker positions in Fig. 4A as a function of fractional egg length x/L (left border $\triangleleft$, right broder $\triangleright$, local max Δ, and local min ∇) for 3 specific time windows ($T_1 = 10 - 25$ min, $T_6 = 41 - 45$ min, and $T_8 = 51 - 55$ min) are shown in blue. The

positional error $\hat{\sigma}_x$ achieved by simultaneously decoding the four gap genes is shown in black for 100 AP positions (see Online Materials). For reference, mean and standard deviation of the gap genes profiles are shown in the background in dark and light colors, respectively (gap gene color code as above), scaled between zero and their maximum expression level G_0. **B)** Time dependences of the average reproducibility $\langle \sigma_{\mathrm{x/L}} \rangle$ of the markers in Fig. 4A (80 dorsal profiles were sorted according to age and then divided into 8 bins of 10 profiles). For each time bin (see Fig. S10) the average and standard deviation across the blue markers is shown as a function of time. The three time bins in A are depicted with filled blue circles. For comparison, the maximum positional reproducibility of Bcd profiles 15 min into n.c. 14 (average between $10 - 60\%$ EL) [Gregor *et al.*, 2007a] is shown with a violet square, and the average positional reproducibility of the pair rule genes *rnt*, *eve* and *prd* 45 min into n.c. 14 is shown with a magenta square [Dubuis *et al.*, 2011]. For reference, the internuclear distance and the half-internuclear distance are shown in dotted and dashed lines, respectively.

For the time class T_6 ($41 - 45$ min), we notice three interesting features. First, for all genes, $\sigma_{x/L}$ is independent of position, which is only observed in this particular time window if the embryos have been sorted out by age and orientation (see Figure 2.15). In particular, the constancy of $\sigma_{x/L}$ across the Hb/Kr, Kr/Kni, Kni/Gt, and Gt/Hb borders hints at a mutual regulation of these genes, rather than a uni-directional repression, in which case it would have a smaller variance than the gene it regulates. Second, the value of $\sigma_{x/L}$ is on average approximately half the inter-nuclear distance in n.c. 14 (i.e. $4\,\mu$m or 0.85%EL, see Figure 2.14 B). Hence, at this level of reproducibility, nuclei located on the gap gene boundaries have already 15 min prior to gastrulation access to a measure of their position with an error smaller than half the distance to their closest neighbors. Although this might not be enough to distinguish themselves from their neighbors with 100% accuracy, it suggests that the level of reproducibility necessary to form the cephalic furrow at the time of gastrulation (~1%EL) is already achieved earlier during n.c. 14 in several regions of the embryo.

The third feature concerns how the markers' reproducibility changes over the course of n.c. 14. For each time window we averaged the $\sigma_{x/L}$ across all markers and mapped the result $\langle \sigma_{x/L} \rangle$ as a function of time, shown in Figure 2.13 B. The variance of the gap genes decreases during the first 35 min, forms a bottle neck that has its minimum at 42 min, and finally increases again until gastrulation. The dynamics of gap gene reproducibility are remarkably correlated with the overall gap gene dynamics uncovered in Figure 2.12, giving us a hint of a functional significance for the collective network dynamics. The reproducibility evolves in time and increases synchronously from one nuclear distance to half a nuclear distance, and thus these dynamics can be seen as the driving element in a relay process where the gap gene network obtains a reproducibility from the Bcd input of one nuclear distance [Gregor *et al.*, 2007a], and subsequently passes a two-fold higher reproducibility on to the the pair-rule genes toward the end of n.c. 14.

Can we understand from our data how this observed rise in reproducibility is achieved? We define a new quantity, the positional error $\hat{\sigma}_x$ resulting from the simultaneous position decoding of the four gap genes, and look at its temporal progression during n.c. 14 (black curve in Figure 2.13 A and Figure 2.14). Initially, transcriptional inputs and gap gene interactions prior to n.c. 14 have already generated profiles whose decoding leads to a positional error of less than one internuclear distance in the 40-80%EL region (T_1). Next, secondary stripes emerge from the interactions between Gt and Hb and between Gt and Kr, reducing the positional error to less than one internuclear distance also in the 10-40%EL and 90-80%EL regions (T_4). The subsequent dynamics of the gap gene network lead to more reproducible profiles overall and positional error eventually reaches its minimum of half a nuclear distance everywhere in the $41 - 45\,\mathrm{min}$ window (T_6). Reproducibility eventually deteriorates during the last 10 min of n.c. 14, just before gastrulation ($T_7 - T_8$).

2.7 Discussion

Simultaneous quantification of gene expression profiles and their variability over the course of embryonic development is crucial for understanding the mechanisms that lead to the formation of organisms with a reproducible body plan. Not only do they provide an estimate of the physical limits reached by the system, but they also provide insights on how these limits are achieved. Here we have shown that immunofluorescent techniques, handled properly, are suitable to make such measurements with high accuracy and precision. We have illustrated this by monitoring, for the first time, the simultaneous expression dynamics of the four gap genes Hb, Kr, Gt, and Kni during n.c. 14, and by quantifying the origin of the positional accuracy of reproducible profiles.

A methodical analysis of the different sources of experimental errors shows that 20%

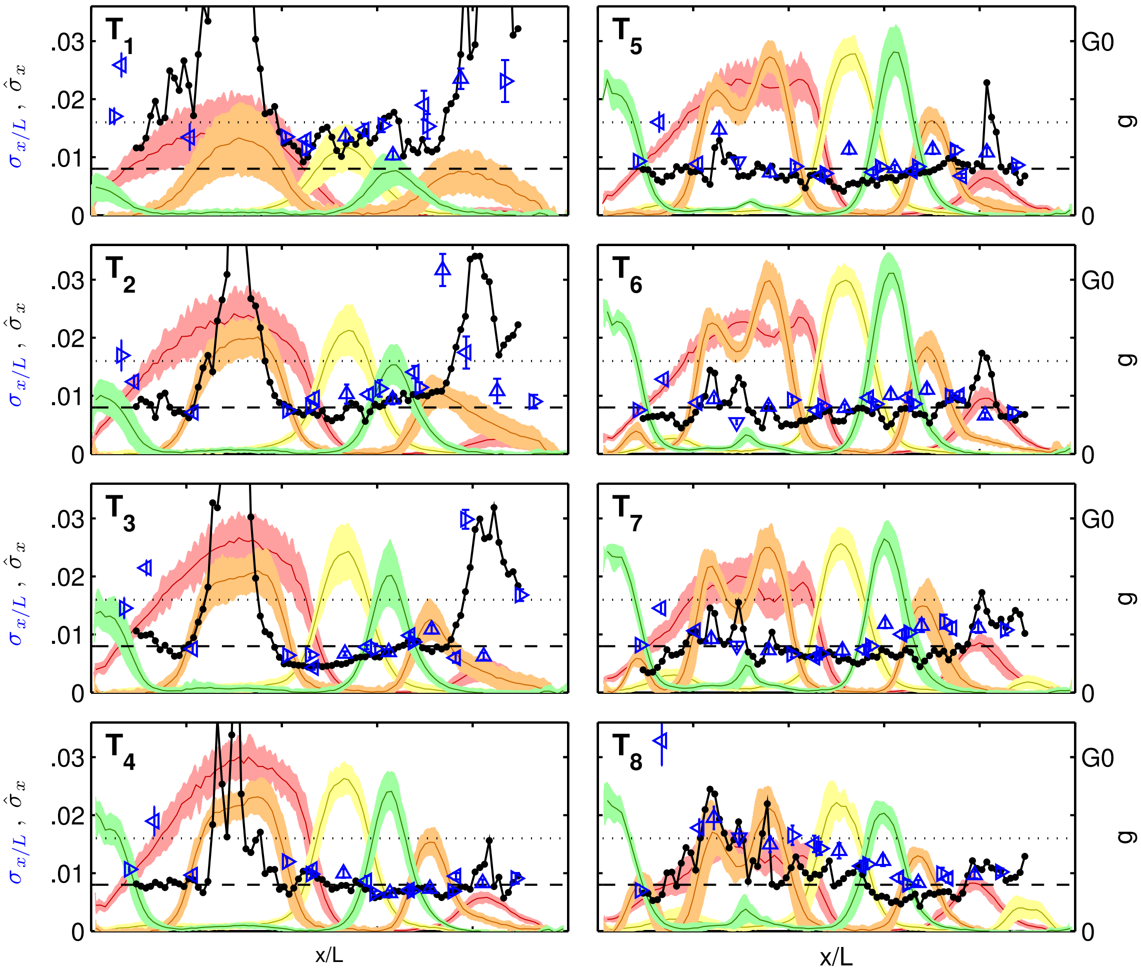

Figure 2.14: Developmental progression of gap gene expression profile reproducibility. Position dependence of the reproducibility of the markers, computed as the standard deviation of their positions across embryos (left border ◁, right boder ▷, local max △, and local min ▽) for 8 time windows covering 90% of n.c. 14 ($T_1 = 10-25$ min, $T_2 = 26-29$ min, $T_3 = 30-33$ min, $T_4 = 34-36$ min, $T_5 = 37-40$ min, $T_6 = 41-45$ min, $T_7 = 46-50$ min, and $T_8 = 51-55$ min) are shown in blue. The positional error $\hat{\sigma}_x$ obtained by simultaneously decoding all four gap genes (four-dimensional extension of Fig. S10F) is shown in black for 100 AP positions. For reference, the mean and standard deviation of the gap gene profiles are shown in the background in dark and light colors, respectively. Profiles have been scaled such that the maximum level of gene expression of their mean over the 8 time windows is G_0. The internuclear distance and half-internuclear distances are shown as dotted and dashed lines, respectively.

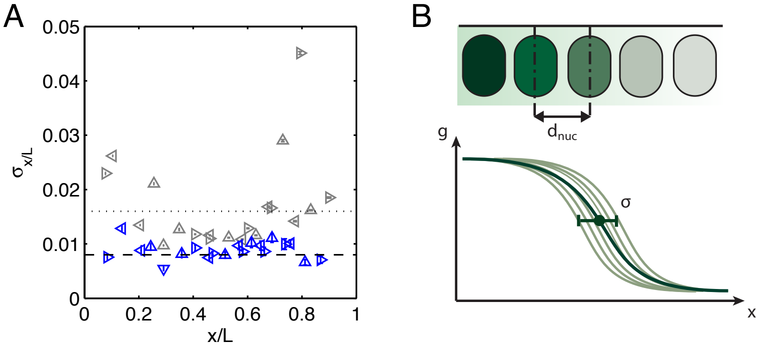

Figure 2.15: Influence of the age and orientation window on the reproducibility of the markers. **A)** Reproducibility of various dorsal markers depending on their position along the anteroposterior axis for 163 embryos without prior selection for age and orientation (grey) and 10 embryos selected to be 41 to 45 min into n.c. 14 (time class T6), and oriented in the midsagittal plane (blue). The average standard deviation of the positions across markers corresponds to half the internuclear distance (1.6 μm), shown as a dashed line as depicted on panel A. **B)** Schematic representation of the reproducibility of a border across profiles and how it relates to the internuclear distance.

of the total variability of our gene expression profiles is due to systematic errors that are inherent to our protocols. This leaves 80% of the observed embryo-to-embryo variability as potentially biologically meaningful statistical fluctuations. In particular, this result makes our experimental method suitable to study the limits by which nuclei can read out transcription factor concentrations [Tkacik *et al.*, 2008, Tkacik *et al.*, 2009], and it will enable us to measure the precision with which individual nuclei can infer their position based on the simultaneous measurement of multiple input transcription factor concentrations.

The analysis of the gap genes dynamics and the evolution of their variability during n.c. 14 suggests that nature uses a 'relay-type' strategy to refine the knowledge that nuclei have about their position from fertilization to gastrulation: Reproducibility is inherent in Bcd, then passed on to the gap genes early n.c. 14, and relayed further to the pair-rule genes late n.c. 14. At the Bcd level, the physical mechanisms that establish the maternal gradient already allow nuclei to achieve position decoding up to one nuclear distance. This reproducibility increases further to a half-nuclear distance at the peak of gap gene expression, 15 min before gastrulation.

This cartoon-like reconstruction of the gap gene dynamics should stimulate the development of more straightforward experimental methods to quantify the time evolution of morphogens expression levels. In particular, it would be crucial to simultaneously monitor the concentration of several morphogens in living embryos. Such methods, which rely on live imaging of embryos expressing molecular fusions of fluorescent proteins are known to be perturbed by photobleaching and the folding and maturation times of the fluorescent proteins [Little *et al.*, 2011]. Therefore, the method presented here will serve as a fiducial mark to assess the validity of those future versions of transgenic flies expressing fluorescent proteins.

2.8 Materials and Methods

Antibody preparation. To allow simultaneous imaging of proteins encoded by all four gap genes, we generated polyclonal antibodies in mice, rats and guinea pigs to His-Trx tagged full-length Hb, Kni and Gt fusion proteins [Kosman *et al.*, 1998]. To image Kr protein, we use a Rabbit anti-Kr antibody generated by Chris Rushlow.

Antibody staining and confocal microscopy. All embryos were collected at 25°C and dechorionated in 100% bleach for 2 minutes, heat fixed in a saline solution (NaCl,Triton X-100) and vortexed in a vial containing 5 mL of heptane and 5 mL of methanol for one minute to remove the vitelline membrane. They were then rinsed and stored in methanol at -20 C. Embryos were then labelled with fluorescent probes. We used rat anti-Kni, guinea pig anti-Gt, rabbit anti-Kr, and mouse anti-Hb. Secondary antibodies were respectively conjugated with Alexa-488 (rat), Alexa-568 (rabbit), Alexa-594 (guinea pig) and Alexa-647 (mouse) from Invitrogen, Grand Island, NY. Embryos were mounted in Aqua-Poly/Mount (Polysciences Inc., Warringyon, PA). Samples were imaged on a Leica SP5 laser-scanning confocal microscope and image analysis routines were implemented in Matlab software (MathWorks, Natick, MA). Images were taken with a Leica 20× HC PL APO NA 0.7 oil immersion objective, and with sequential excitation wavelengths of 488, 546, 594 and 633 nm. The bandwidth of the detection filters were set up as shown in Figure 2.1 to minimize fluorophore cross-talk while still allowing good detection in each optical channel. For each embryo, three high-resolution images (1024 × 1024 pixels, with 12 bits and at 100 Hz) were taken along the anterior-posterior axis (focused at the midsagittal plane) at 1.7× magnified zoom and averaged together. With these settings, the linear pixel dimension corresponds to $0.44 \pm 0.01\,\mu$m. All embryos were prepared, and images were taken, under the same conditions: (i) all embryos were heat fixed together, (ii) embryos were stained and washed together in the same tube, and (iii) all images were taken with the same microscope settings in a single acquisition cycle.

Measurements of the invagination depth of the membrane furrow canals. We used the depth of the furrow canal (FC) during blastoderm cellularization as a 'clock' to estimate the age of the embryo (origin taken at the onset of nuclear n.c. 14). To assess the precision of this clock, movies of the invagination process on the dorsal side of 8 wild-type embryos (one frame per minute) were taken using bright field microscopy. For each embryo

i, the onset of n.c. 14 was identified as the frame where the nuclear membrane is the most dissolved during mitosis 13. The depth $\delta_{\mathrm{FC}}^{(i)}$ of the FC was monitored as a function of absolute time until gastrulation as shown on 2.3 A,. To compare the growth of the cellularization membrane across embryos, the time traces $\delta_{\mathrm{FC}}^{(i)}(t)$ $(i = 1 \ldots 8)$ were aligned frame by frame, allowing an arbitrary shift in time ΔT_i (to compensate for inaccuracies in determining the onset of n.c. 14) and a shift in depth $\Delta\delta_{\mathrm{FC}}^{(i)}$ (to compensate for inaccuracies in determining the apical membrane; constrained by a maximum amplitude of 1μm). Shifts were optimized by minimizing the Euclidean distance $\chi^2 = \int_t \left| \tilde{\delta}_{\mathrm{FC}}^{(i)}(t) - \left\langle \tilde{\delta}_{\mathrm{FC}}^{(i)}(t) \right\rangle \right|^2 dt$, where each $\tilde{\delta}_{\mathrm{FC}}^{(i)}(t)$ is given by $\tilde{\delta}_{\mathrm{FC}}^{(i)}(t) = \delta_{\mathrm{FC}}^{(i)}(t + \Delta T_i) + \Delta\delta_{\mathrm{FC}}^{(i)}$, and $\langle\rangle$ denotes averaging across embryos. To estimate the temporal error σ_t on an embryo's age, we propagated the positional error σ_δ (error bars in 2.3 B) of the measurement of the furrow canal depth as $\sigma_t = \sigma_\delta \cdot |d\delta/dt|^{-1}$ (see inset in 2.3 B). Correction for embryo deformation due to fixation procedure was evaluated by comparing the AP and DV lengths of live with fixed embryos. Both dimensions shrink by $\sim$5%, leading to a corrected FC depth $\delta_{\mathrm{FC}}^{\mathrm{fixed}}$ shown in green on the right axis of Figure 2.3 B.

Expression profile quantification. Profiles were extracted by sliding, in Matlab software, a disk of the size of a nucleus along the edge of the embryo in the midsagittal plane and computing the average intensity of its pixels [Houchmandzadeh *et al.*, 2002]. The coordinates of the disk centers were projected on the anteroposterior axes of the embryo. The dorsal and ventral profiles were separately extracted. For consistency, only dorsal profiles are used in our analysis.

Identification of the Nuclei. For experiments requiring nuclear intensity quantification (see Figures 2.5, 2.7, 2.8, and 2.10), we used the average intensity of the different optical channels to identify nulcei. We first examine each pixel $\mathbf{x}$ in the context of its 11×11 pixel neighborhood; let the mean intensity in this neighborhood be $\bar{I}(\mathbf{x})$ and the variance be $\sigma^2(\mathbf{x})$. We construct a normalized image, $\psi(\mathbf{x}) = [I(\mathbf{x}) - \bar{I}(\mathbf{x})]/\sigma(\mathbf{x})$, which is smoothed with a Gaussian filter (standard deviation 2 pixels) and thresholded with a threshold optimized by eye-inspection. Locations of nuclei were assigned as the center of mass in the connected regions above threshold. Each nuclear mask was manually corrected to avoid misidentifications.

Determination of boundary positions and peaks in gene expression.
Semi-automatic identification of expression peaks. Each raw intensity profile was binned (1000 points) and linearly filtered by computing, for each bin, the average intensity taken over the 30 nearest neighbor bins (corresponding to 3%EL, or approximately two internuclear distances). The effect of this operation is to limit intensity variations due to the presence of the nuclei and thus to compute an average nuclear/cytoplasmic intensity. A region of N bins around the presumed boundary was manually identified (e.g. 60-100%E.L, or $N = 400$ bins for the posterior stripe of Hb). To avoid identifying local extrema solely due to higher nuclear concentrations, we randomly picked up $N/3$ uniformly distributed bins in that region and computed the peak intensities and positions over the 10% intensity of highest (resp. lowest) bins. We similarly defined for each embryo and each profile the maximum intensity $I_{\max}$ as well as the background intensity $I_{\min}$.
Boundary identification. A region of N bins was manually identified at the presumed location of the half-maximum intensity point of the boundary (e.g. 30-70%EL, or $N = 400$ bins, for the main border of *hb*). The code required the finding of two neighboring data points in

that region with higher and lower intensities than $(I_{\max} + I_{\min})/2$ (given the sharp slope of the border, false positives were extremely rare and manually corrected). The intensity and position of the half-maximum amplitude bin was then identified as the mean intensity and position of the two neighboring points. The slope was determined by fitting the points in the 3%EL neighboring region with a straight line (which corresponds to approximately two internuclear distances).

Influence of the age and orientation window on the reproducibility of the markers To address the contribution of our sorting method to the value and constancy of $\sigma_{x/L}$ we compared our results to the reproducibility that is achieved across 163 quadruple stained embryos, without proper age and orientation classification (gray data points in 2.15. In this case, we find that the standard deviation across embryos is on average more than twice as large as in our original measurement, which can be explained by the fact that $\sigma(x/L)$ now includes the dependence of time in the profiles along the entire n.c. 14, as well as the dependence on embryo orientation w.r.t the dorsal side. The resulting reproducibility of the markers as a function of x/L tends to be higher at the anterior and posterior ends of the embryo, where the orientation greatly influences the patterns of gene expression and where the time shift of the markers is most important.

Reproducibility quantification of Bcd, and the pair pair-rule genes. The reproducibility of Bcd was quantified by averaging the effective rms error $\sigma(x)$ in positional readout over the 20-60% EL region (15 embryos live-imaged 15 min after entry into mitosis 13, [Gregor *et al.*, 2007a]. The reproducibility of the pair rule genes was quantified by measuring the standard deviation of the position of the peaks of the 7 stripes of Even-Skipped and Runt across 12 embryos selected in a 40 to 50 min time window after the beginning of n.c. 14 [Dubuis *et al.*, 2011].

Computation of σ_x. For a single gene g, we compute the positional error for 100 equally spaced positions along the AP axis, using $\sigma_x(x) = \sigma_g(x)\,|d\bar{g}/dx|^{-1}$. At each point the error bars on the positional error are estimated by bootstrapping over 10 samples of size $\mathcal{N}/2$, where $\mathcal{N}$ is the number of profiles collected for gene g. The positional error at a given x can be visualized as the AP distance between the intersection points of $\bar{g}(x)$ and a rectangle of width $\sigma_g(x)$. The positional error is smaller in the regions of high profile slope, where the variations in gene expression are reliably translated into variations in position. In a straightforward generalization of the single gene case the positional error for the multiple gene case is given by:

$$\sigma_x^2(x) = \left(\sum_{i,j=1}^{N} \frac{d\bar{g}_i}{dx} [C(x)^{-1}]_{ij} \frac{d\bar{g}_j}{dx} \right)^{-1},$$

where C_{ij} is the covariance matrix of the profiles that is directly estimated from the quadruple staining at each position x. The positional error at a given x can be visualized as the AP distance between the positions of the intersection points of $\bar{g}(x)$ and a cylinder whose base is the ellipsoid in the $\{hb, kr\}$ plane such that $\sum_{i,j}(g_i(x) - \bar{g}_i)[C^{-1}(x)]_{ij}(g_j(x) - \bar{g}_j) \leq 1/4$. Note that the optimal positional decoding performed with several genes (e.g., $N = 2$) at a given x does not correspond to the positional error carried by the gene with the least individual positional error at that position, i.e. the combined error can be smaller than the individual errors due to the noise averaging by the N readouts and the correlation structure.

Chapter 3

Positional information carried by the four main gap genes

3.1 Introduction

So far, we have developed the conceptual tools that are necessary to rigorously define positional information and we have shown that all features required for the computation can be measured experimentally. Now we wish to actually compute the quantities defined in Chapter 1 to estimate how much information is carried by the four gap genes of *Drosophila*. Is this information comparable to the information that the embryo uses at the time of gastrulation? Is it enough to specify the position of each cell along the anteroposterior axis? These questions are all the more interesting since the processes that govern pattern formation in the embryo are driven by molecules that are present at low concentrations, such that noise inevitably limits transmission of information through the different regulatory layers of the network [Elowitz *et al.*, 2002, Tkacik *et al.*, 2009, Tkacik & Walczak, 2011]. Some theoretical studies have suggested that the transmitted information is close to the maximum allowed by the physical limits on the system [Tkacik *et al.*, 2008]. Armed with the tools developed in Chapter 1, we can now test these ideas on experimental data and gain more insights about the strategies that nature uses to make embryonic development so reproducible.

In this chapter, we cover the technical details of applying positional information and positional error formalism to real data sets, using the four major gap genes in early *Drosophila* development as a test case. We start by presenting the necessary statistical techniques to consistently merge data from separate pairwise gap gene immunostaining experiments into a single dataset. We then estimate mutual information directly from data for one and for two gap genes. To generalize the method to higher dimensions, we introduce the Gaussian noise approximation and an adaptive Monte Carlo integration scheme and we extract the information jointly carried by all four gap genes. This analysis leads us to some insights on how information is coneyed through the layers of the *Drosophila* regulatory network.

3.2 Inference methods

3.2.1 Estimating information with limited amounts of data

Measuring positional information from a finite number $\mathcal{N}$ of embryos is challenging due to estimation biases. Good estimators are thus often more complicated than the naive approach, which consists of applying Eq. (1.7) directly to an estimate of $P(\{g_i\},x)$ obtained by simply histogramming the data. Here we start by introducing this naive estimator, which we subsequently develop into a series of better estimators, based on an idea termed *direct estimation* [Strong *et al.*, 1998, Slonim *et al.*, 2005].

The easiest way to obtain an estimate for $P(\{g_i\},x)$ is to convert the range of continuous values for g_i and x into b discrete adaptive bins; we can then create a histogram of the joint distribution of $\{g_i\}$ and x from $\mathcal{N}$ embryos, and normalize it to obtain $\tilde{P}_{b,\mathcal{N}}(\{g_i\},x)$, the estimate of the joint distribution. We can compute a *naive estimate* of the positional information, $I^{\rm DIR}_{b,\mathcal{N}}(\{g_i\};x)$ by using $\tilde{P}$ directly in Eq. (1.7):

$$I^{\rm DIR}_{b,\mathcal{N}}(\{g_i\};x) = -\sum_{\{g_i\}} \tilde{P}_{b,\mathcal{N}}(\{g_i\},x)\log_2\left[\frac{\tilde{P}_{b,\mathcal{N}}(\{g_i\},x)}{\tilde{P}_{b,\mathcal{N}}(\{g_i\})\tilde{P}_{b,\mathcal{N}}(x)}\right] \tag{3.1}$$

where the subscripts indicate the explicit dependence on sample size and number of adaptive bins.$\tilde{P}_{b,\mathcal{N}}(\{g_i\})$ and $\tilde{P}_{b,\mathcal{N}}(x)$ are obtained by summing $\tilde{P}_{b,\mathcal{N}}(\{g_i\},x)$ along the appropriate dimensions. It is known that naive estimators suffer from estimation biases that scale as $1/\mathcal{N}$ and $1/b^{N+1}$ (as we are partitioning not only the N genes but also the x axis). Using techniques discussed in [Strong *et al.*, 1998, Slonim *et al.*, 2005], we can obtain a *direct estimate* of the mutual information by first computing a series of naive estimates for a fixed value of b and for fractions of the whole data set. Concretely, we pick fractions $n = [0.95\ 0.9\ 0.85\ 0.8\ 0.75\ 05] \times \mathcal{N}$ of the total number of samples $\mathcal{N}$. At each fraction, we randomly pick n embryos 100 times and compute $\langle I^{\mathrm{DIR}}_{b,n}(\{g_i\};x)\rangle$ (where averages are taken across 100 random data subsets). This gives us a series of data points that can be extrapolated to an infinite data set by fitting a linear model for $\langle I^{\mathrm{DIR}}_{b,n}(\{g_i\};x)\rangle$ vs $1/n$ (Fig. 3.1 A). The intercept of this linear model yields $I^{\mathrm{DIR}}_{b,n\to\infty}(\{g_i\};x)$ and we can repeat this procedure for a set of ever more adaptive bins. To avoid overpartioning the data, we first need to define a bin number limit b^* (depending on $\mathcal{N}$) for which the 1/n extrapolation is valid. This is done by analyzing the same data set but now shuffled and finding the critical number of bins for which the extrapolated information $I^{\mathrm{shuffled}}_{b,\infty}(\{g_i\};x)$ is zero within error bars. The inset of Fig. 3.1 B shows that for the analyzed *hb* profiles ($\mathcal{N}$=23), $b^* = 50$. Extrapolating $I^{\mathrm{DIR}}_{b,\infty}(\{g_i\};x)$ to $1/b^2 \to 0$ for $b < b^*$, we obtain the final estimate $I^{\mathrm{DIR}}_{b\to\infty,n\to\infty}(\{g_i\};x)$, called *direct estimate (DIR)* of positional information (red dot in Fig. 3.1 B). No prior knowledge about the shape of the distribution $P(\{g_i\}|x)$ is assumed by the direct estimation method. A potential disadvantage of this method is that it still requires a substantial amount of data (large $\mathcal{N}$) because de-biasing extrapolations can only work when under-sampling is in the assumed scaling regime. In practice, our current data sets suffice for the direct estimation of positional information carried by one gap gene.

In order to extend this method to more than two genes, one needs to resort to approximations for $P(\{g_i\}|x)$, the simplest of which is the Gaussian approximation, shown in Eq. (1.4). In this case, we can write down the entropy of $P(\{g_i\}|x)$ analytically. For a single gene we get

$$S[P(g|x)] = \frac{1}{2}\log_2\left(2\pi e\sigma_g^2(x)\right), \tag{3.2}$$

while the straightforward generalization to the case of N genes is given by

$$S[P(\{g_i\}|x)] = \frac{1}{2}\log_2\left((2\pi e)^N|\mathbf{C}(x)|\right), \tag{3.3}$$

where $|\mathbf{C}(x)|$ is the determinant of the covariance matrix $C_{ij}(x)$. From Eq. (1.8) we know that positional information $I(g;x) = S[P(g)] - \langle S[P(g|x)]\rangle_x$. Here the second term ("noise entropy") is therefore easily computable from Eq. (3.2). The first term ("total entropy") can be estimated as above by the direct method, i.e. histogramming $P(g)$ for various sample sizes n and bin sizes Δ, and extrapolating $n \to \infty$ and $\Delta \to 0$ (Fig. 3.1 C). This combined procedure, where we evaluate one term in the Gaussian approximation and the other one directly, has two important properties. First, the total entropy is usually much better sampled than the noise entropy, because it is based on the values of g pooled together over every value of x; it can therefore be estimated in a direct (assumption-free) way even when the noise entropy cannot be. Second, by making the Gaussian approximation for the second term, we are always *overestimating* the noise entropy and thus *underestimating* the total positional information, because the Gaussian distribution is the maximum entropy distribution with a given mean and variance (other, non-Gaussian distributions with the same mean and variance can only have smaller entropies). Therefore, in a scenario where the first term in Eq. (3.2) is estimated directly, while the second term is computed analytically from the Gaussian ansatz, we always obtain a lower bound on the true positional information (Fig. 3.1 D). We call this bound the *first Gaussian approximation* (FGA).

For three or more gap genes the amount of data can be insufficient to reliably apply either the direct estimate or FGA, and one needs to resort to yet another approximation, called *second Gaussian approximation* (SGA). As in the FGA, for the second Gaussian approximation we also assume that $P(\{g_i\}|x) \approx \mathcal{G}(\{g_i\}; \bar{g}_i(x), C_{ij}(x))$ is Gaussian, but we make another assumption in that $P(\{g_i\})$ obtained by integrating over these Gaussian conditional distributions is a good approximation to the true $P(\{g_i\})$. The total distribution we use for the estimation is therefore:

$$P(\{g_i\}) = \int_0^1 dx\; P(\{g_i\}|x). \tag{3.4}$$

For each position, the noise entropy in Eq. (3.3) is proportional to the logarithm of the determinant of the of the covariance matrix, which scales as $1/n$ if estimated from a limited number n of samples. We therefore estimate the information $I_n^{\mathrm{SGA}}(\{g_i\}; x)$ for fractions

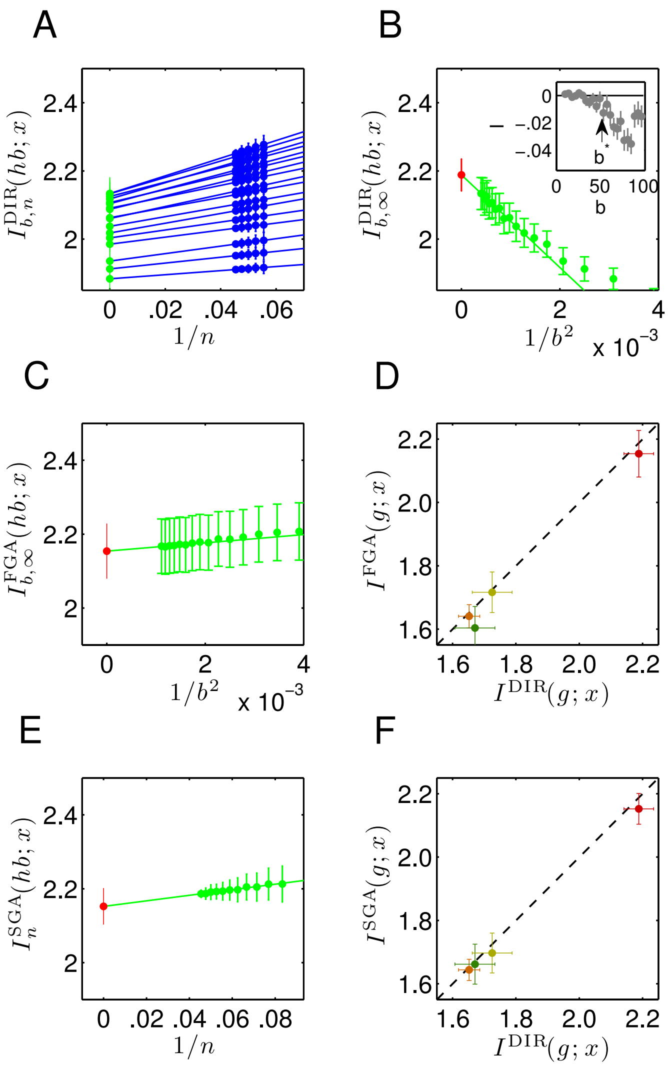

Figure 3.1: Estimates of positional information (in bits) carried by the gap genes in the 10-90% egg length region, using various approximations. **A,B)** *Direct estimate* of positional information carried by *hb* (dorsal profiles, $\delta_{\mathrm{FC}} = 10-20$ μm, $\mathcal{N} = 23$). Blue points show the dependence of $I^{\mathrm{DIR}}_{b,n}(hb;x)$ on n for different choices of b (each line corresponds to a different number of adaptive bins). By extrapolating $I^{\mathrm{DIR}}_{b,n}(hb;x)$ with respect to $1/n$ we obtain $I^{\mathrm{DIR}}_{b,\infty}(hb;x)$ for various bin sizes b. In B, each green square is an extrapolation for an infinite number of samples estimated in A, for the chosen value of b. The red dot represents the final estimated information, obtained by extrapolating $b \to 0$; here $I^{\mathrm{DIR}}(hb;x) = 2.19 \pm 0.05$ bits. **C)** *First Gaussian approximation.* Green dots show the dependence of the information estimate on the bin size; each point is an extrapolation to infinite number of samples. We find $I^{\mathrm{FGA}}(hb;x) = 2.15 \pm 0.07$ bits. **D)** Comparison of the *first Gaussian approximation* with the *direct estimate*, computed from the dorsal profiles of *hb* (red), *kr* (yellow), *gt* (orange), and *kni* (green) with $\delta_{\mathrm{FC}} = 10-20$ μm. The two approximations give the same result within the error bars. **E)** *Second Gaussian approximation.* Green dots show the dependence of the information estimate on the bin size. In the infinite sample limit, we find $I^{\mathrm{SGA}}(hb;x) = 2.15 \pm 0.05$ bits. **F)** Comparison of the *second Gaussian approximation* with the *direct estimate*, computed from the dorsal profiles of the gap genes with $\delta_{\mathrm{FC}} = 10-20$ μm. Color as in D.

of the whole data set and then extrapolate for $n \to \infty$ (Fig. 3.1E). For single genes, the *second Gaussian approximation* is in good agreement with the *direct estimate*, as shown in Fig. 3.1F, where most points fall onto the diagonal.

The second Gaussian approximation relies on our ability to (i) estimate well the covariance matrix $C_{ij}(x)$ between gap genes i and j at every position x, and (ii) carry out the non-trivial integration in Eq. (3.4) when distributions are high dimensional. Both issues are discussed in detail below.

3.2.2 Inferring consistent covariance matrices from partial measurements

To compute positional information from our data using the second Gaussian approximation, we need to measure N mean profiles, $\bar{g}_i(x)$, and the $N \times N$ covariance matrix, $C_{ij}(x)$. Measuring individual gap gene expression profiles using, e.g., immunostaining is a standard experimental technique in developmental biology. In contrast, estimating the covariance matrix, $C_{ij}(x)$, would require simultaneously labeling all N gap genes in each embryo using fluorescent probes of different colors. While simultaneous stainings of two genes are not unusual, it is not easy to scale the method up to more genes while maintaining a precise and quantitative readout. Hence, in this section we present a technique that allows estimation of a consistent $N \times N$ covariance matrix based on a collection of embryos stained for different pairs of gap genes. To test the validity of our approach, we also performed a simultaneous recording of all 4 gap genes in *Drosophila* embryos which allows us to quantitatively compare the two alternative methods.

Estimating a joint covariance matrix from pairwise staining experiments is a non-trivial problem for two reasons. First, each diagonal element of the covariance matrix, i.e. the variance of an individual gap gene, is measured in multiple experiments, but the obtained values might vary due to statistical and systematic measurement errors. Second, true covariance matrices are positive definite, i.e. $\det(C(x)) > 0$, a property that is not guaranteed by naively filling in different terms of the matrix by computing them across sets of embryos collected in different experiments. This is a consequence of small sampling errors that can strongly influence the determinant of the matrix. We therefore need a principled way to find a single best and valid covariance matrix from multiple partial observations, a problem that has had considerable history in statistics and finance [Ait-Sahalia & Kimmel, 2007, Phillips & Yu, 2009, Lee & Liu, 2012].

We start by considering a number $\mathcal{N}_{ij}$ of embryos that have been co-stained for the pair of gap genes (i, j). Let the full dataset consist of all such pairwise stainings: for N gap genes, this is a total of $\binom{N}{2}$ pairwise experiments, where $i, j = 1, \ldots, N$ and $i < j$. Thus, in the case of the four major gap genes in *Drosophila* embryos, *hb*, *kr*, *gt* and *kni* (in that order), the total number of recorded embryos is $\mathcal{N} = \sum_{(i,j)} \mathcal{N}_{ij}$, where the sum is across all six pairwise measurements: $(1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (3, 4)$. These pairwise measurements give us estimates of the mean profiles and the 2×2 covariance matrices for each pair (i, j):

$$\hat{g}_i(x) = \frac{1}{\mathcal{N}_{ij}} \sum_{\mu=1}^{\mathcal{N}_{ij}} g_i^{(\mu)}(x), \tag{3.5}$$

$$\hat{C}_{ij}(x) = \frac{1}{\mathcal{N}_{ij}} \sum_{\mu=1}^{\mathcal{N}_{ij}} (g_i^{(\mu)}(x) - \hat{g}_i)(g_j^{(\mu)}(x) - \hat{g}_j). \tag{3.6}$$

The index μ enumerates all the embryos recorded in a pairwise experiment (i, j). For four gap genes and six pairwise experiments, we will get six 2×2 partial covariance matrices $\hat{C}$, and 6×2 estimates of the mean profile $\hat{g}$. Our task is to find a single set of 4 mean profiles $\bar{g}_i(x)$, and a single 4×4 covariance matrix $C_{ij}(x)$, that fits all pairwise experiments best. In the next paragraphs, we will show how this can be computed for the arbitrary case of N gap genes.

To infer a single set of mean profiles $\bar{g}_i(x)$ and a single consistent $N \times N$ covariance matrix $C_{ij}(x)$, we use maximum likelihood inference. We assume that at each position x our data is generated by a single N-dimensional Gaussian distribution of Eq. (1.4) with unknown mean values and an unknown covariance matrix, which we would like to find, but we can only observe two of the mean values and a partial covariance in each experiment; the other variables are integrated over in the likelihood.

Following this reasoning, the log likelihood of the data for pairwise staining (i, j) at position x is

$$\begin{aligned}
\mathcal{L}_{ij} &= \frac{1}{\mathcal{N}_{ij}} \log \int \prod_{k\neq i,j} dg_k \; P(\{g_i\}|x) \\
&= \frac{1}{\mathcal{N}_{ij}} \ln \prod_{\mu=1}^{\mathcal{N}_{ij}} \frac{1}{2\pi\sqrt{C_{ii}C_{jj}-C_{ij}^2}} \\
&\times \; \exp\left[-\frac{1}{2}\frac{C_{jj}(g_i^{(\mu)}-\bar{g}_i)^2+C_{ii}(g_j^{(\mu)}-\bar{g}_j)^2-2C_{ij}g_i^{(\mu)}g_j^{(\mu)}}{C_{ii}C_{jj}-C_{ij}^2}\right],
\end{aligned} \tag{3.7}$$

where the log likelihood $\mathcal{L}$, as well as the mean profiles, covariance elements and the measurements all depend on x.

Since all pairwise experiments are independent measurements, the total likelihood $\mathcal{L}_{\text{tot}}(x)$ at a given position x is the sum of the individual likelihoods

$$\mathcal{L}_{\text{tot}}(x) = \sum_{(i,j)} \mathcal{L}_{ij}(x) \tag{3.8}$$

After some algebraic manipulation, the total log likelihood can thus be written as:

$$\begin{aligned}
\mathcal{L}_{\text{tot}}(x) = -\sum_{(i,j)} & \ln(2\pi)+\ln\left(C_{ii}(x)C_{jj}(x)-C_{ij}^2(x)\right) \\
& +C_{jj}(x)\frac{\hat{C}_{ij}(x)-2\bar{g}_i(x)\hat{g}_i(x)+\bar{g}_i^2(x)}{C_{ii}(x)C_{jj}(x)-C_{ij}^2(x)} \\
& +C_{ii}(x)\frac{\hat{C}_{jj}(x)-2\bar{g}_j(x)\hat{g}_j(x)+\bar{g}_j^2(x)}{C_{ii}(x)C_{jj}(x)-C_{ij}^2(x)} \\
& +2C_{ij}(x)\frac{\hat{C}_{ij}(x)-\hat{g}_i(x)\bar{g}_j(x)-\hat{g}_j(x)\bar{g}_i(x)+\bar{g}_i(x)\bar{g}_j(x)}{C_{ii}(x)C_{jj}(x)-C_{ij}^2(x)},
\end{aligned} \tag{3.9}$$

where $\hat{g}_i(x)$ and $\hat{C}_{ij}(x)$ are the experimentally determined profiles and covariance elements defined in Eqs. (3.5,3.6). For each position x, we search for $\bar{g}_i(x)$ and $C_{ij}(x)$ that maximize $\mathcal{L}_{\text{tot}}(x)$. Before proceeding, however, we have to guarantee that the search can only take place in the space of positive semi-definite matrices $C_{ij}(x)$ (i.e. $\det \mathbf{C} \geq 0$). We enforce this constraint by spectrally decomposing C_{ij} and parametrizing it in its eigensystem [Pinheiro

& Bates, 1996]. To this end, we write

$$\mathbf{C}(x) = \mathbf{PDP}^{\mathrm{T}}, \tag{3.10}$$

where $\mathbf{D}$ is a diagonal matrix parametrized with the variables $\alpha_1, \ldots, \alpha_N$ that determine the diagonal elements in $\mathbf{D}$, i.e. $D_{ii} = \exp(\alpha_i)$. The orthonormal matrix $\mathbf{P}$ is decomposed as a product of the $N(N-1)/2$ rotation matrices in N dimensions:

$$\mathbf{P} = \prod_{k=1}^{N(N-1)/2} \mathbf{R}_k(\varphi_k), \tag{3.11}$$

where $\mathbf{R}_k(\varphi_k)$ is a rotation matrix that can be written as:

$$\mathbf{R}_k(\varphi_k) = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos(\varphi_k) & 0 & \cdots & 0 & -\sin(\varphi_k) & \\ & & & 0 & 1 & & & 0 & \\ & & & \vdots & & \ddots & & \vdots & \\ & & & 0 & & & 1 & 0 & \\ & & & \sin(\varphi_k) & 0 & \cdots & 0 & \cos(\varphi_k) & \\ & & & & & & & & 1 \end{pmatrix}. \tag{3.12}$$

In the representation given by Eq (3.10), the likelihood is a function of N values $\bar{g}_i$, N parameters α_i, and $N(N-1)/2$ angles φ_k at each x. In case of 4 gap genes, this is a total of 14 parameters that need to be computed by maximizing $\mathcal{L}_{\mathrm{tot}}(x)$ at each location x.

There is no guarantee that there is a unique minimum for the log likelihood; moreover, there could exist sets of covariance matrices that all lead to essentially the same value for $\mathcal{L}_{\mathrm{tot}}(x)$. To break these degeneracies, we regularize the likelihood by penalizing the matrices that have (almost) singular determinants. This is equivalent to picking, out of all the matrices that fit the data equally well, the covariance matrix corresponding to the most random (maximum entropy, or largest determinant) distribution. Thus, we maximize:

$$\tilde{\mathcal{L}}_{\text{tot}}(x) = \mathcal{L}_{\text{tot}}(x) - \lambda \left|\mathbf{C}(x)\right| \tag{3.13}$$

where λ is a regularization parameter: larger values of λ will favor more "spherical" distributions, while small values will allow distributions that can be very squeezed in some directions. The value of λ depends on the dataset; one way to determine it is to find the maximum log likelihood $\mathcal{L}$ first without regularizer and assess the error bar on this estimate by bootstrapping, and then select the regularizer λ with the largest possible value such that $\tilde{\mathcal{L}}(x)$ remains within the statistical error of $\mathcal{L}$. Using this criterion, we found that the value $\lambda = 0.05$ worked well for our dataset.

To compute $\bar{g}_i(x)$ and $C_{ij}(x)$, we initialize $\bar{g}_i(x)$ to the mean profile across all pairwise experiments (i, j); we initialize α_i to the mean of the log of the diagonal terms $\hat{C}_{ii}$; and we initialize all rotation angles $\varphi_k = 0$. Nelder-Mead simplex method is used to maximize $\tilde{\mathcal{L}}_{\text{tot}}(x)$ [Lagarias *et al.*, 1998]. Finally, we compute $\mathbf{C}(x)$ from α_i and φ_k using Eqs (3.10,3.11,3.12).

To check the validity of this algorithm, we performed an experiment in which we stained embryos simultaneously for all four gap genes g_i (*hb*, *kr*, *gt*, and *kni*, respectively). We split the dataset obtained from this staining into six "virtual" pairwise experiments, i.e. into experiments that recorded simultaneously gap genes (1,2), (1,3), (1,4), (2,3), (2,4) and (3,4). We used the maximum likelihood merging method outlined above to compute $\bar{g}_i(x)$ and $C_{ij}(x)$ (denoted by a subscript ML) for these virtual datasets. The same quantities can also be evaluated directly from the quadruple staining (denoted by a subscript 4G). Figure 3.2 shows that both the inferred mean profiles and the covariance matrix elements match the directly measured values, validating the correctness of our inference method.

Qualitatively the two methods show a near perfect overlap. We assess the similarity ξ between these two sets of curves by comparing the Euclidean distance between the estimated and the true result with the Euclidean distance between the expected result and its mean over x:

$$\xi = \frac{1}{1 + \sqrt{\frac{\int_0^1 dx |f_{\text{ML}}(x) - f_{\text{4G}}(x)|^2}{\int_0^1 dx |\bar{f}_{\text{4G}} - f_{\text{4G}}(x)|^2}}} \tag{3.14}$$

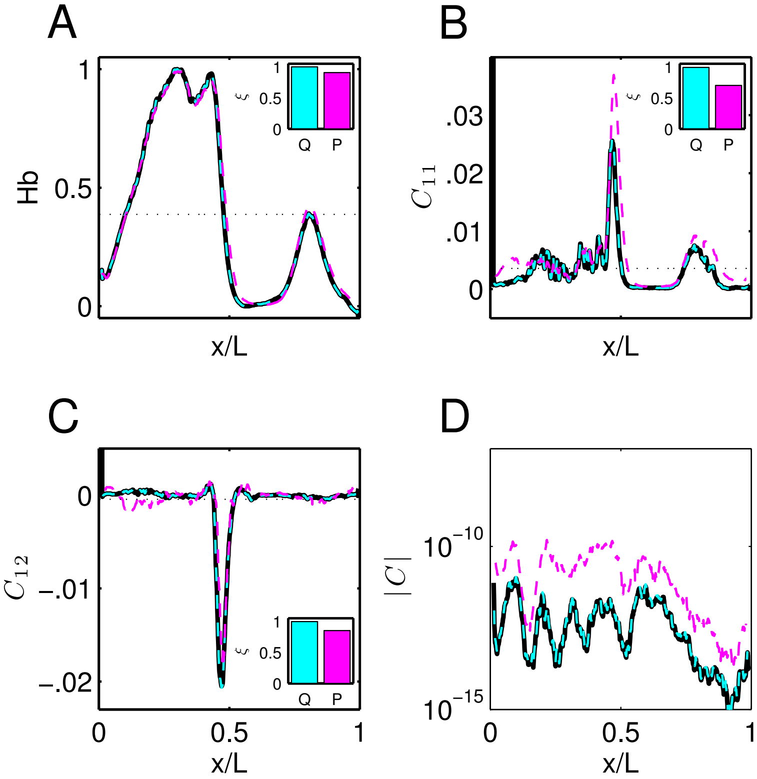

Figure 3.2: Merging data sets using maximum likelihood reconstruction. **A)** Mean *hb* profiles as a function of relative egg length extracted from quadruply-stained embryos (black), from quadruply-stained embryos after application of maximum likelihood method (see text, cyan), and from pairwise-stained in independent experiments after application of maximum likelihood method (magenta). *Inset*: Similarity ξ (see text), comparing either the cyan or magenta curve to the black reference. $\xi = 0.5$ corresponds to the similarity between the black curve and its average over x (dotted line in main panel). **B)** Same comparison as in A for *hb* profile variances. **C)** Same comparison as in A for the cross-correlation between *hb* and *kr* profiles. **D)** Same comparison as in A for the determinant of the covariance matrix.

where $\xi = 1$ is a perfect match, $\xi = 0.5$ the similarity between $f_{4\mathrm{G}}$ and its mean $\bar{f}_{4\mathrm{G}}$ and $\xi \to 0$ the similarity between two very different curves, or even cases where the inferred curve has divergences along the AP axis. This metric is particularly relevant for assessing curves that reproduce the details of position dependence better than a model that is simply a constant along AP, equal to the mean value. The mean profiles (Fig. 3.2 A) and the elements of the consistent 4×4 covariance matrix (Fig. 3.2B and 3.2 C) have similarities greater than 0.99, validating the convergence of the algorithm.

The method was then applied to six pairwise immunostaining experiments that have been carried out on different embryo collections and imaged on different days. Qualitatively the matchup with the quadruple staining is still astoundingly good: similarities for the various quantities in Fig. 3.2 range from 0.70 to 0.91. In particular, we see a bit more noise in the merged pairwise experiments compared to the quadruplet staining, which is to be expected, because the pairwise experiments also include experiment-to-experiment variance which the quadruplet does not. Clearly, however, most of the mean and covariance behavior is consistent between different experiments. The maximum likelihood method can therefore be used to reconstruct the N mean profiles and the $N \times N$ covariance matrix of a set of N genes by simply using $N(N-1)/2$ pairwise experiments of co-stained embryos. This represents an important advance and a powerful tool in connecting theoretically relevant quantities to what can be practically measured.

3.2.3 Monte Carlo integration of total entropy

To apply the second Gaussian approximation to three or more gap genes, we need not only a consistent estimate of the mean profiles and the covariance, but also have to compute the total entropy, $S[P(\{g_i\})]$, by integrating $P(\{g_i\}|x)$ over all x as prescribed by Eq. (3.4). The straightforward way to numerically evaluate this integral would be to partition the integration domain $[0,1]^N$ into a grid with grid spacing Δg in each dimension, evaluate the conditional distribution $P(\{g_i\}|x)$ on the grid for each x, and average the results to get $P(\{g\})$. Unfortunately, for 3 or 4 gap genes this is numerically infeasible because of the curse of dimensionality for any reasonably fine-grained partition.

To address this problem we make use of the fact that on most of the integration domain

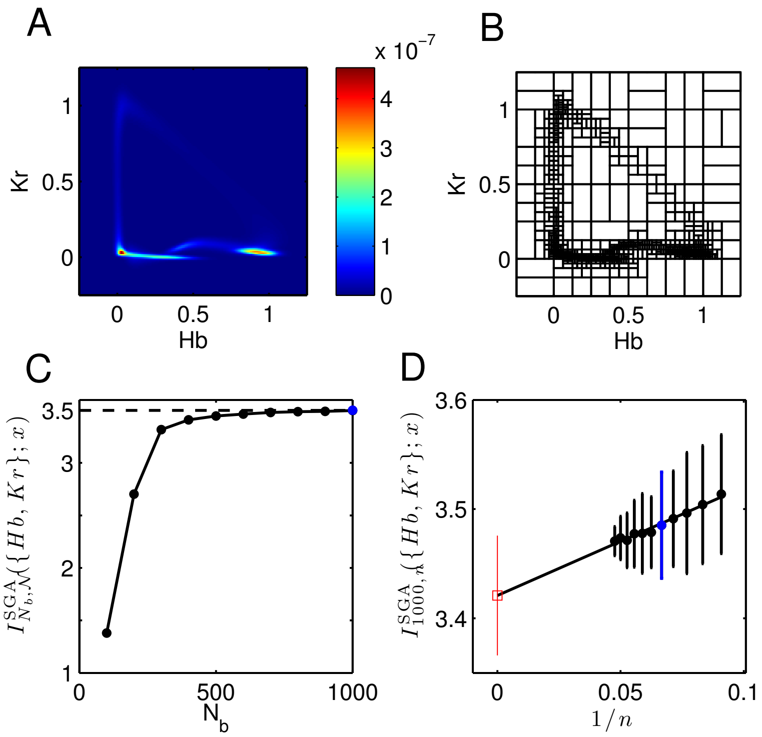

Figure 3.3: Monte Carlo integration of the total information for two genes. **A)** Heat map of the total probability distribution $P(\{hb, kr\})$ computed by using the sampled means and covariance matrix, and assuming a Gaussian distribution for $P(\{hb, kr\}|x)$ at each position. **B)** Adaptive partitioning of the integration domain for the $\{hb, kr\}$ distribution generated by the Monte Carlo algorithm for $N_{\mathrm{boxes}} = 1000$. **C)** Convergence of the estimated positional information I^{SGA} carried by hb and kr, computed for $\mathcal{N} = 15$ embryos as the number of boxes increases. **D)** The dependence of positional information on the number of samples, $\mathcal{N}$. Blue data point corresponds to the converged information computed in C. Error bars are computed by bootstrapping over ten subsets of size $n/2$. We find $I^{\mathrm{SGA}}(\{Hb, Kr\}; x) = 3.53 \pm 0.05$ bits.

$P(\{g_i\})$ is zero (Fig. 3.3 A). This is due to the fact that the variability among embryos is small, which means that most of the probability weight is concentrated in the small volume around the path traced out by $\bar{g}_i(x)$ as x changes from 0 to 1. We thus designed a method that partitions the whole integration domain adaptively into volume elements such that the total probability weight in every box is approximately the same.

The following algorithm was used to compute the total entropy, $S[P(\{g_i\})]$:

1. The whole domain for $\{g_i\}$ is recursively divided into boxes such that no box contains more than 1 percent of the total volume.

2. For each box i with volume ω_i, we use Monte Carlo sampling to randomly select

$t = 1, \ldots, T$ points $\mathbf{g}^t$ in the box and approximate the weight of the box i as $\pi(i|x) = \omega_i T^{-1} \sum_{t=1}^{T} P(\mathbf{g}^t|x)$; we use $T = 100$.

3. Analogously, we evaluate the approximate total weight of each box $\pi(i)$, by pooling Monte Carlo sampled points across all x.
4. $\pi(i|x)$ and $\pi(i)$ are renormalized to ensure $\sum_i \pi(i|x) = 1$ for every x and $\sum_i \pi(i) = 1$.
5. The conditional and total entropies are computed as $S_{\text{noise}} = -\langle \sum_i \pi(i|x) \log_2 \pi(i|x) \rangle_x$ and $S_{\text{tot}} = -\sum_i \pi(i) \log_2 \pi(i)$.
6. The positional information is estimated as $I(\{g_i\}; x) = S_{\text{tot}} - S_{\text{noise}}$.
7. The box $i^* = \text{argmax}_i \pi(i)$ with the highest probability weight $\pi(i^*)$ is split into two smaller boxes of equal volume $\omega(i^*)/2$, and the estimation procedure is repeated by returning to step 2. Additional Monte Carlo sampling only needs to be done within the newly split box; for the other boxes old samples can be reused.
8. The algorithm terminates when the positional information achieves desired convergence.

Fig. 3.3 A and 3.3 B show how the partitioning works in a case of two genes ($\{hb, kr\}$) which is easy to visualize. This partition is spatially highly refined where the distribution $P(\{hb, kr\})$ has a lot of weight and remains coarse elsewhere. This numerical integration of the information converges to a limit value as the number of boxes increases (Fig. 3.3 C).

To correct for the dependancy on the finite sample size, we randomly select fractions of the data, $n = [0.95\ 0.9\ 0.85\ 0.8\ 0.75\ 0.5] \times \mathcal{N}$, where the total number of recorded embryos is $\mathcal{N}$. As in direct estimation, we run the above algorithm using the means and covariance matrix computed from 10 random subsets at each data fraction, and extrapolate towards $n \to \infty$, as shown in Fig. 3.3 D.

3.2.4 Decoding positional information from experimental data

To see how the availability of positional information changes along the AP axis, we use Eq. (1.19) to compute and plot the positional error, first for a single gene, and then for a gene pair, using experimentally measured mean profiles and the profile covariance matrix.

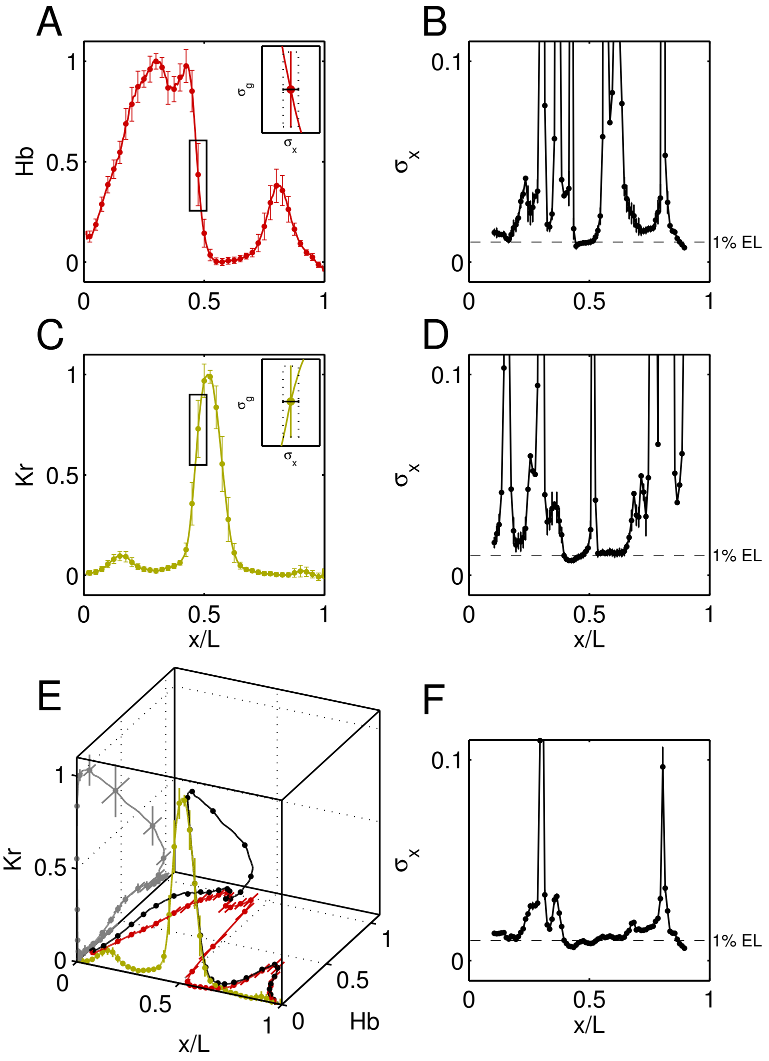

Figure 3.4: Positional error for a single gene and a pair of genes. **A)** Mean dorsal *hb* profile and standard deviation across 24 embryos (age: 35 to 45 min into nuclear cycle14) as a function of fractional egg length. The inset shows a close-up of the transition region with the positional error σ_x geometrically determined from the mean $\bar{g}(x)$ and the standard deviation σ_g of the profiles. **B)** Positional error for *hb* as a function of position, computed using Eq. (3.15), for 100 bins along the AP axis. The dashed line is a reference for $\sigma_x = 0.01$ or 1% EL. Error bars are obtained by bootstrapping 10 times over $\mathcal{N}/2$ embryo subsets. **C)** Mean profile and standard deviation for *kr* from the same data set as in A. **D)** Positional error for *kr* as a function of position (as in B). **E)** Three-dimensional representation of *hb* and *kr* profiles from A and C as a function of relative position. Gray curve shows the mean and standard deviations in the *hb*-*kr* plane; c.f. their joint distribution in Fig. 3.3A. Black curve shows the joint *hb*-*kr* mean expression as a function of position x. **F)** Positional error computed using Eq. (1.19) for a pair of genes, *hb* and *kr*, using their mean expression behavior and covariance shown in E.

For a single gene g, we compute the positional error for 100 equally spaced positions along the AP axis, using

$$\sigma_x(x) = \sigma_g(x)\,|d\bar{g}/dx|^{-1}. \tag{3.15}$$

At each point the error bars on the positional error are estimated by bootstrapping over 10 samples of size $\mathcal{N}/2$, where $\mathcal{N}$ is the number of profiles collected for gene g. Fig. 3.4 illustrates this procedure geometrically for the case of *hb* (Fig. 3.4A) and *kr* (Fig. 3.4 C). The positional error at a given x can be visualized as the AP distance between the intersection points of $\bar{g}(x)$ and a rectangle of width $\sigma_g(x)$. The positional error is smaller in the regions of high profile slope, where the variations in gene expression are reliably translated into variations in position (Fig. 3.4 B and 3.4 D).

A similar analysis for a pair of genes is shown in Figs. 3.4 E and F. Note that the optimal positional decoding performed with several genes (e.g., $N = 2$) at a given x does not correspond to the positional error carried by the most informative gene at that position; the combined error can be smaller than the individual errors due to the noise averaging by the N readouts and the correlation structure of the profiles. Fig. 3.4 E and 3.4 F illustrate how σ_x is constructed geometrically for the case of $N = 2$ genes (here *hb* and *kr* measured simultaneously). The positional error at a given x can be visualized as the AP distance between the positions of the intersection points of $\bar{g}(x)$ and a cylinder whose base is the ellipsoid in the $\{hb, kr\}$ plane such that $\sum_{i,j}(g_i(x) - \bar{g}_i)[C^{-1}(x)]_{ij}(g_j(x) - \bar{g}_j) \leq 1/4$.

3.3 Results

3.3.1 Information carried by individual genes and pairs

We applied the different information estimation methods described in the previous section to measure the positional information carried by the individual gap genes, *kni*, *kr*, *gt* and *hb*, as well as the six corresponding pairs (dorsal profiles, 35-45 minutes into cycle 14). The results are summarized in Table 3.1, where we see that various estimation methods agree to within 4% on average. The positional information carried by individual genes is always in the range of 1.5-2.5 bits. This indicates that each gap gene provides more information about

	kni	kr	gt	hb
DIR	$1.67 \pm .05$	$1.72 \pm .05$	$1.65 \pm .05$	$2.19 \pm .04$
FGA	$1.60 \pm .05$	$1.71 \pm .06$	$1.64 \pm .06$	$2.15 \pm .05$
SGA	$1.66 \pm .05$	$1.70 \pm .05$	$1.64 \pm .04$	$2.15 \pm .04$

	kni/kr	kni/gt	kni/hb	kr/gt	kr/hb	gt/hb
SGA	$3.22 + 0.06$	$3.25 + 0.07$	$3.25 + 0.07$	$3.21 + 0.05$	$3.53 + 0.06$	$3.37 + 0.01$

Table 3.1: Information (in bits) carried by single genes and gene pairs, using different methods. For single genes we computed the information and the error bar using the *direct estimate* (DIR), the *first Gaussian approximation* (FGA) and the *second Gaussian approximation* (SGA), as described in the Materials and Methods section. As expected, the *first Gaussian approximation* is a lower bound of the *direct estimate*. For the pairs, we only used the *second Gaussian approximation*.

position than the traditional "decision threshold" that would separate the "on" and "off" domains of expression. We will explore this result in detail later.

By comparing the information carried by pairs of genes with the sum of the informations carried by each gene separately, we notice that pairs are redundant but that the redundancy is small (15-20% on average). This is expected because gap genes are often expressed in complementary regions of the embryo and will thus mostly convey new information about position. Unlike our toy examples in Chapter 1, real gene expression levels are continuous and noisy, and some degree of redundancy might be useful in mitigating the effects of noise, as has been shown in other biological information processing systems. Taken together, we see that positional information increases with the addition of the second gene by much more than 0.5 bits, as would be expected for fully redundant case on theoretical grounds, and also faster than for the theoretical case of non-redundant (but non-interacting) genes, where the increase for the second gap gene would be limited to 1 bit [Tkacik *et al.*, 2009]. The ability to generate non-monotonic profiles of gene expression (e.g. bumps) is therefore crucial for high information transmissions achieved in *Drosophila* gap gene network [Walczak *et al.*, 2010].

3.3.2 How much information does the embryo use?

If the expression profile of each gap gene were described by on/off domains with sharp boundaries, not only would a single gene carry at most one bit of information, but four genes taken together could carry at most four bits - and this would happen only if the

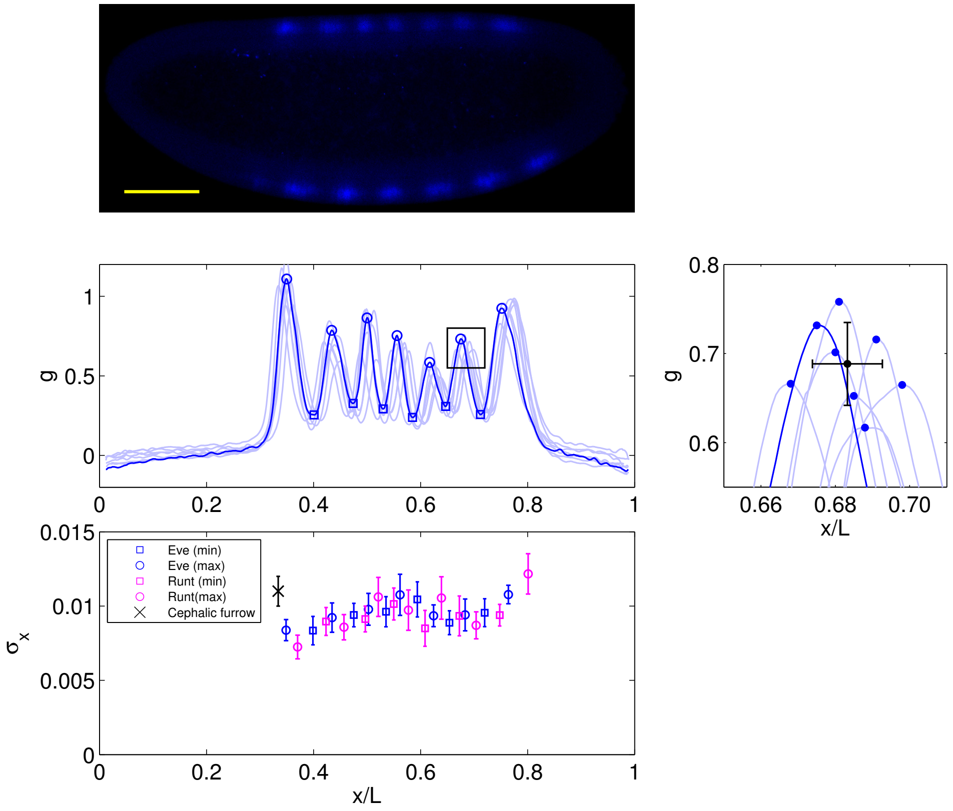

Figure 3.5: Reproducibility of multiple pattern elements along the anteroposterior axis. Top panel: optical section through the midsagittal plane of a Drosophila embryo with immunofuorescence staining against Eve protein; scale bar is 100 μm. Middle panels: normalized dorsal profiles of fluorescence in- tensity from 12 embryos selected in a 40 to 50 min time window after the beginning of nuclear cycle 14 (light blue lines); dorsal profile of top panel embryo is in darker blue. Zooming in on a single peak (as shown at right), we can measure the standard deviation of both the expression level and position of this element in the pattern. Bottom panel summarizes results from such measurements on Even-skipped (blue) and Runt (magenta), plotting the standard deviation of the position σ_x as a function of the mean position x, together with a similar measurement on the reproducibility of the cephalic furrow. Note that all of the elements are positioned with 1% accuracy or better.

spatial arrangement of the different expression domains was carefully aligned to minimize redundancy. Four bits of information corresponds to, at most, 16 reliably distinguishable states encoded by these genes, which seems insufficient to account for the complex final pattern. But how much information does the embryo really need, or use? At best, every nucleus could be labelled with a unique identity, so that with N nuclei the embryo could make use of $\log_2 N$ bits. Along the anteriorposterior axis, we can count nuclei in a single midsagittal slice through the embryo, and in the middle 80% of the embryo where the images are clearest we have $N = 58 \pm 4$ along the dorsal side and $N = 59 \pm 4$ along the ventral side, where the error bars represent standard deviations across a population of 57 embryos in nuclear cycle 14, corresponding to 5.9 ± 0.1 bits of information. But do individual cells in fact know their identity? More precisely, are the elements of the anteroposterior pattern specified with single cell resolution?

There are several experiments suggesting that elements of the final body plan of the maggot can be traced to identifiable rows of cells along the anteroposterior axis [Gergen & Wieschaus, 1986], which is consistent with the idea that each row of cells has a reproducible identity. More quantitatively, we can ask about the reproducibility of various pattern elements in early development, elements that appear not long after the expression patterns of the gap genes are established. A classic case is the cephalic furrow, which can be observed in live embryos and is known to have a position along the anteroposterior axis that is reproducible with $\sim 1\%$ accuracy (see, for example, [Gregor *et al.*, 2007a]). Is the cephalic furrow special, or can the embryo more generally position pattern elements with $\sim 1\%$ accuracy? The striped patterns of pair rule gene expression allow us to ask about the position of multiple pattern elements, seven peaks and six troughs of expression along the anteroposterior axis. As shown in Fig. 3.5, all of these elements have positions that are reproducible to within 1% of the embryo length. This strongly suggests that all cells "know" their position along the anteroposterior axis with $\sim 1\%$ precision.

3.3.3 Decoding the positional information carried by the four gap genes

Do the four gap genes, taken together, carry enough information to specify position with $\sim 1\%$ accuracy suggested in the previous section? To answer this, we start by looking

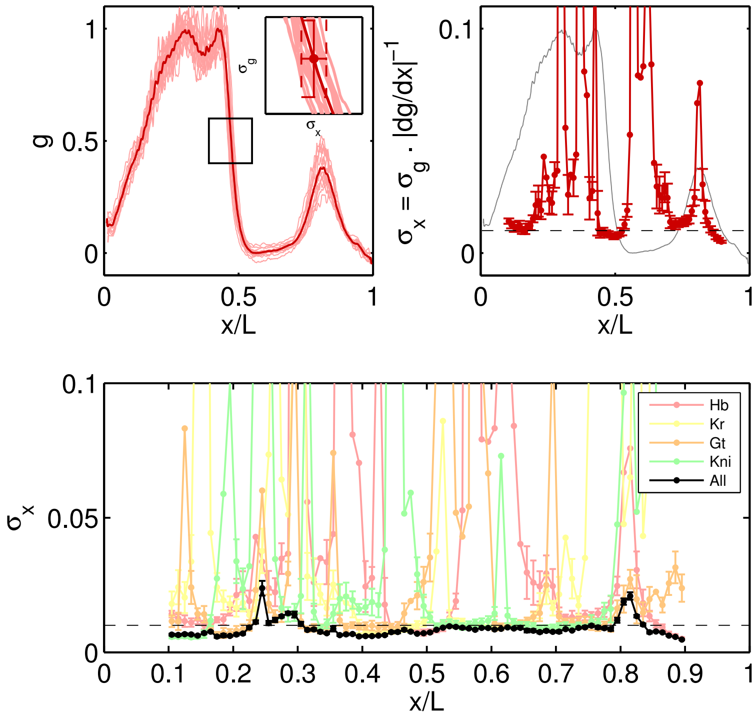

Figure 3.6: Positional error as a function of position. Upper left panel: geometrical interpretation of the positional error of a single gene at a given position. $\sigma_x(x)$ is proportional to the reproducibility of the profiles and is inversely proportional to the derivative of the mean profile. The upper right panel summarizes the positional error carried by Hunchback for a hundred points along the anteroposterior axis. Lower panel: total positional error carried by the four gap genes (in black), from Eq. 1.19; Error bars are from bootstrapping. For reference, the individual positional errors are plotted in lighter colors in the background. Note that the total positional error is nearly constant and equal to 1% of the total egg length.

more directly at how the information is encoded and to use the positional error σ_x defined in Eq. 1.19. Measurements of σ_x are summarized in Fig. 3.6. Remarkably, the reliability of position estimates based on the four gap genes is $\sim 1\%$, almost precisely equal to the observed reproducibility with which pattern elements are positioned along the AP axis. This is strong evidence that the gap genes, taken together, carry the information needed to specify the full pattern. Further, this positional accuracy is almost constant along the length of the embryo, which again is consistent with what we see in Fig. 3.5. This constancy emerges in a nontrivial way from the expression profiles, the noise levels, and the correlation structure of the noise. If we try to make estimates based on one gene, we can reach $\sim 1\%$ accuracy in a very limited region of the embryo, and estimates from the different genes have their optimal precision in different places. The detailed structure of the spatial profiles insures that these signals can be combined to give nearly constant accuracy.

Figure 3.7 gives us some clue on how the gap genes "tile" the AP axis and how this constancy is achieved. By reading out single genes the nuclei can already achieve positional errors of less than 1.5% egg length in specific regions, but can be very bad in determining position in other regions of the embryo. Looking more closely, we note that the lowest positional errors are consistently achieved in transition regions, where the expression level changes steeply with position, and not in regions of stable response; in a way, the transition regions can be thought of as mini-gradients that carry useful information. As we pointed out in Chapter 1, the ability to use the transition regions is by no means obvious, but is a consequence of the overall high levels of reproducibility achieved by the gap gene system. We can easily imagine a fly that has the same mean profiles, but a different structure of variability encoded in the covariance matrix $C_{ij}(x)$, such that transition regions contribute no discriminatory ability and all positional information comes from properly decoding the domains where the expression levels plateau.

Let's now focus on the information itself. If the errors in estimating position really are Gaussian, then according to Eq. 1.18 $I = \langle \log_2[L/\sigma_x\sqrt{2\pi e}]\rangle$, where L is the length of the embryo and $\langle \dots \rangle$ denotes an average over the possible position dependence of the error σ_x. With $\sigma_x/L \sim 0.01$, we have $I \sim 4.57 \pm 0.02$ bits. On the other hand, we can use our measurements of the means and the covariance matrix as well as the Monte Carlo

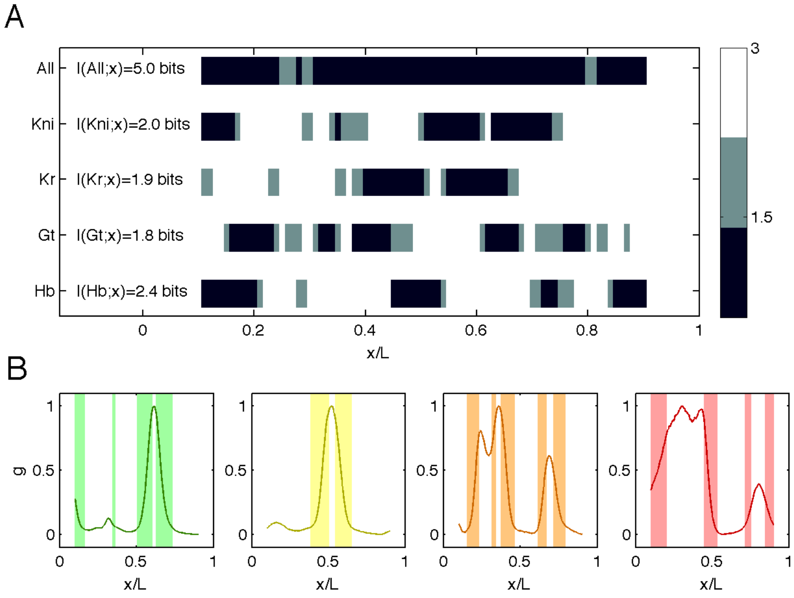

Figure 3.7: Positional error along the AP axis of the embryo. **A)** Map of the local positional error as a function of position, for a segment in the embryo between 10-90% egg length. Each row represents positional error from a single gap gene, except for the top row which shows decoding for all 4 gap genes simultaneously. Information estimates using SGA are shown next to each row. Shade of gray (colorbar) indicates the positional error ($\sigma_x < 1.5\% =$ black, $\sigma_x < 3\% =$ grey, $\sigma_x > 3\% =$ white). **B)** Comparison of the regions where $\sigma_x < 1.5\%$ EL (shaded regions) for each gene with the corresponding mean profile (lines), for *kn* (green), *kr* (yellow), *gt* (orange), and *hb* (red). Consistently, the regions providing lowest positional errors are the regions where gap gene expression changes quickly with position (transition regions).

integration of the total entropy described in section3.2.3, and we find $I = 5.0 \pm 0.23$ bits. These estimates supports our approximations, and gives us confidence that the measurement of σ_x on Fig. 3.6 really does characterize the encoding of positional information by the gap genes.

We notice that our estimate of the positional information carried by the four gap genes doesn't exactly match the $I = 5.9$ bits that one would need to undoubtedly distinguish the ~ 60 rows of cells along the dorsal side of the embryo (as explained in the previous section). So, "where is the missing bit?". What our results show is actually that the information carried by the gap genes allows to decode $\sim 2^5$ states which is roughly half of the number of nuclei along the dorsal side of the embryo. However the distribution of σ_x suggests that the information carried is uniformly distributed along these nuclei. So, the gap genes are not specifying the position of every other nucleus but instead are providing just enough information to each nucleus to decode its position with half an internuclear distance precision. This result is all the more convincing that this is actually the positional accuracy that is achieved a few minutes later by the pair-rule genes.

3.3.4 How much can the embryo achieve with on/off decoding

To better understand the previous results and contrast them with the picture where each gap gene profile can individually be either "off" or "on", but no information can be encoded in the transition regions, we built a model as follows. We assumed that there exists a separate threshold in the expression level of each of the four gap genes which separates the "on" and "off" states for each individual profile of that gene. We can look for these threshold levels computationally (by exhaustively trying all the settings) and find such a combination of four thresholds that, were gap gene expressions really binary, the gap genes would provide the largest amount of positional information. We argue that since this is the theoretical maximum given the measured mean profiles and variances of gap gene expression, the fly could not do better with binary expression levels and "on" / "off" domains alone.

In detail, we quantize each gap gene g_i with an individual threshold θ_i that discriminates the gene expression levels of each individual profile $g_i^{(j)}$ into two ("on" if $g_i^{(j)} > \theta_i$, "off" if $g_i^{(j)} < \theta_i$) and then search for the set of thresholds $\{\theta_i\}$ that maximize the information carried altogether by the four quantized gap genes, defined as in Eq. (1.8), where the $\{g\}$'s are now 4-digit binary patterns (with 0 corresponding to "off" and 1 corresponding to "on").

To compute the first term of the formula (total entropy), we quantize the mean profile of each gap gene and define $p(\{g\})$ as the occurrence frequency of the pattern $\{g\}$ in the 10-90% EL region. To compute the second term of the formula (noise entropy), we notice that $p(\{g\}|x)$ is 0 or 1 "almost everywhere", except at each border k of the $\{g\}$ domain, which fluctuates with a standard deviation $\sigma_x(k) = \sigma_g(x_k) \cdot |d\bar{g}/dx|^{-1}_{x=x_k} \ll 1$. If we assume these fluctuations to be Gaussian (as previously), then each border contributes by $2.6\,\sigma_x(k)$ bits to the noise entropy (see Appendix). Now, we allow each individual threshold θ_i to be $10, 20, \ldots 90\%$ of the maximum gene expression level and we exhaustively search for the set of thresholds that maximizes the total information carried by the quantized profiles. Figure 3.8 shows the binary gene expression domains, the corresponding 4-digit code, and, the fluctuations of the borders of the binary patterns.

In this optimal binary representation, all gap genes carry less than one bit of information. Furthermore, the total information does not exceed 2.8 bits, 30% smaller than 4 bits, the maximum amount of information available in theory to 4 binary genes. Even in this best case scenario, the nature loses more than one bit of information (and thus half of the distinguishable states) compared to the available total capacity. The reason for this is the uneven use of different expression combinations. Clearly, the information carried by the quantized profiles is maximized when each of the 2^4 binary combinations is used with equal probability (in this case, along $\sim 6\%$ of AP length for each pattern) along the AP axis of the embryo. What we see, however, is that patterns are used in region sizes that range from 1% EL (e.g. 0111) to 20%EL (e.g. 0011). This encoding not only decreases the positional information available, but also means that the positional error is strongly nonuniform along the AP axis: the uncertainty about the true position is very large in larger regions encoded by the single combination.

In sum, even if we (or the fly) were to make an optimal partition of the expression profiles into binary "on" / "off" domains, between a third and a half of total positional information would be lost, as well as the ability to decode the position precisely anywhere along the AP axis. This drop in the ability of the "binarized" patterning system to encode positional

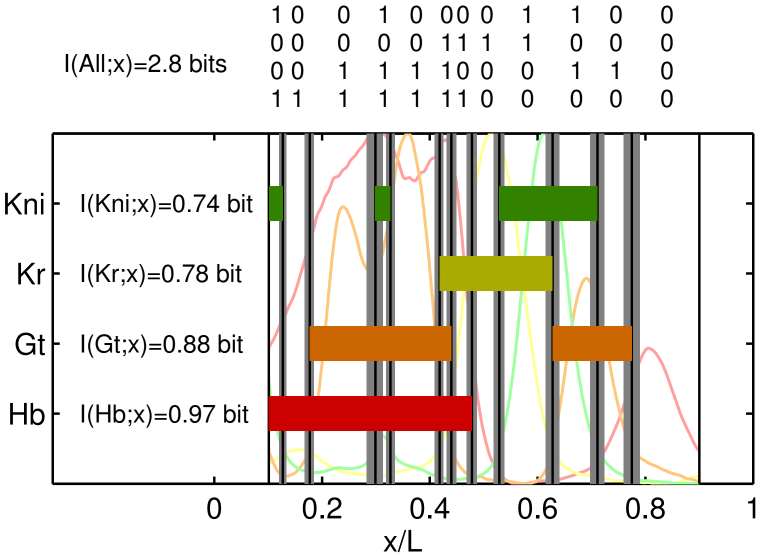

Figure 3.8: The binary view of the gap gene system. For each gap gene i, we quantize each of the profiles $g_i^{(j)}$ previously used such that the level of gene expression is "on" (or 1) if g_i is greater than a threshold θ_i, and 0 otherwise. Here we show the resulting domains of gene expression (dark color bars) as well as their fluctuations (in grey) for a set of thresholds that maximize the total information in the $10 - 90\%$ EL region ($\theta_{kni} = 0.1$, $\theta_{kr} = 0.1$, $\theta_{gt} = 0.1$, $\theta_{hb} = 0.4$). For reference, the mean profiles are plotted in dim color in the background. Information carried by the quantized profiles of the individual genes is shown on the left. The joint pattern of gap gene activity at each position is represented by a four-digit binary code (shown above), and the total information encoded jointly by the "on"/"off" gap gene expression domains is computed as explained above.

information is in part caused by the fact that single genes really carry much more than 1 bit of information and thresholding destroys this excess capacity. But even taking this into account, we note that thresholding into domains does not decrease the ability of each single gene to much below 1 bit; indeed, the genes on average encode still > 0.7 bits each (out of 1 bit maximum). What happens is that the resulting *joint* combination of patterns and domains, if binary as assumed here, is far from optimal for encoding positional information and reaching a low and stable positional error.

3.3.5 A signature of optimization?

The discussion thus far concerns the amount of information that actually is transmitted by the levels of gap gene expression. But we know that the capacity to transmit information is strictly limited by the available numbers of molecules, and that significant increases in information capacity would require vastly more than proportional increases in these numbers [Tkacik *et al.*, 2009]. Given these limitations, however, cells can still make more or less efficient use of the available capacity. To maximize efficiency, the input/output relations and noise characteristics of the regulatory network must be matched to the distribution of input transcription factor concentrations [Tkacik *et al.*, 2008]. This matching principle has a long history in the analysis of neural coding [Barlow, 1959, Laughlin, 1981, Brenner *et al.*, 2000], and in [Tkacik *et al.*, 2008] it was suggested that the regulation of Hunchback by Bicoid might provide an example of this principle. Here we consider the generalization of this argument to the gap gene network as a whole. If we imagine that there is a single primary morphogen, then the expression levels of the different gap genes, taken together, can be thought of as encoding the concentration c of this morphogen. By analogy with Eq. 1.19, these expression levels can be decoded with some accuracy $\sigma_{\text{eff}}(c)$, which itself depends on the mean local concentration. The key result of [Tkacik *et al.*, 2008] is that, when noise levels are small, all the symbols in the code should be used in proportion to their reliability, or in inverse proportion to their variability. Thus, if we point to a cell at random, we should see that the concentration of the primary morphogen is drawn from a distribution

$$P_{\text{input}}(c) = \frac{1}{Z} \cdot \frac{1}{\sigma_c^{\text{eff}}(c)} \tag{3.16}$$

where the constant Z is chosen to normalize the distribution. But the input is a morphogen, so its variation is connected with the physical position x of cells along the embryo: we should have $c = c(x)$. Then if the cells are distributed uniformly along the length of the embryo, the probability that we find a cell at x is just $P(x) = 1/L$, and hence we must have

$$P_{\text{input}}(c)\, dc = P(x)\, dx = \frac{dx}{L} \tag{3.17}$$

which implies that

$$P_{\text{input}}(c) = \frac{1}{L} \left| \frac{dc(x)}{dx} \right|^{-1}. \tag{3.18}$$

We have two expressions for the distribution of input transcription factor concentrations: Eq 3.18, which expresses the role of the input as morphogen, encoding position x, and Eq 3.16, which expresses the solution to the problem of optimizing information transmission through the network of genes that respond to the input. Putting these expressions together, we have

$$\frac{Z}{L} = \frac{1}{\sigma_c^{\text{eff}}(c)} \left| \frac{dc(x)}{dx} \right| = \sigma_x(x), \tag{3.19}$$

where in the last step we recognize the equivalent positional noise $\sigma_x(x)$ by analogy with Eq 1.19. Thus, optimizing information transmission predicts that the positional uncertainty $\sigma_x(x)$ will be constant along the length of the embryo, as observed in Fig. 3.6. A more detailed version of this argument is given in the Appendix.

Conclusion

The final result of embryonic development appears precise and reproducible. Less is known quantitatively about the degree of this precision and about the stages of development at which precision first becomes apparent. Our central result is that, in the early *Drosophila* embryo, the patterns of gap gene expression provide enough information to specify the positions of individual cells with a precision of $\sim 1\%$ along the anteroposterior axis. This is the same precision with which subsequent pattern elements are specified, from the pair rule expression stripes through the cephalic furrow, so that all the required information is available from a local readout of the gap genes. The precise value of the information that we observe is also interesting. It corresponds to any nucleus being able to locate it self with an error bar that is smaller than the distance to its neighbor, but it not quite large enough to specify the position of every cell uniquely. The difference between these statements is that when we make an estimate with error bars, the estimate comes from a distribution with tails, and the (small) overlap of the tails of these distributions means that one cannot quite identify every cell. It is possible that cells do not quite have unique identities or that the missing information is hiding in correlations among the errors at different points. Although the gap genes encode position with an error bar, the difference in positions coded by expression levels in neighboring cells could have a much smaller error bar. While further experiments are required to settle this issue, we find it remarkable that the gap gene expression levels carry so much information such that an enormously precise pattern is available very early in development. The information that gene expression levels can carry about position is limited by noise. Because the concentrations of transcription factors are low, and because the absolute copy numbers of the output proteins are small, there are physical sources of noise that cannot be reduced without the embryo investing more resources in making these molecules. Given these limits, it still is possible to transmit more information through the gap gene network by “matching” the distribution of input signals to the noise characteristics of the network. Although this matching condition is in general complicated, in the limits that the noise is small it can be expressed very simply: the density of cells along the anteroposterior axis should be inversely proportional to the precision with which we can infer position by decoding the signals carried in the gap gene expression levels. Since cells are almost uniformly distributed at this stage of development, this predicts that an optimal network would have a uniform precision, and this is what we find. This uniformity emerges despite the complex spatial dependence of all the ingredients, and thus seems likely to be a signature of selection for optimal information transmission.

Appendix A

Methods

A.1 Integration of the noise entropy for the gap gene quantization

Here we show how to integrate the noise entropy part of the information in the case of the gap gene quantization. For a given binary word $\{g\}$, $p(\{g\}|x)\log_2 p(\{g\}|x)$ is zero for almost all x except in the regions where $p(\{g\}|x)$ quickly transitions between 0 and 1. At a given border k located in x_k, we assume that the fluctuations of the transition positions of the individual quantized profiles follow a Gaussian distribution of standard deviation $\sigma_x(k) \ll 1$ so that, near the border k, the conditional probability $p(\{g\}|x)$ can be written as an error function:

$$p(\{g\}|x) = \int_{-\infty}^{x} dx' \frac{1}{\sqrt{2\pi\sigma_x^2(k)}} \exp\left[-\frac{(x'-x_k)^2}{2\sigma_x^2(k)}\right]$$

(or the corresponding complementary function). We can then numerically integrate the corresponding term in the noise entropy:

$$-\int dx\, p(\{g\}|x)\log_2 p(\{g\}|x) \simeq 1.3\,\sigma_x(k). \tag{A.1}$$

At a given border, the two neighboring words each contribute by $1.3\,\sigma_x(k)$ bits so that the border contributes in total by $2.6\,\sigma_x(k)$ bits to the noise entropy. Note that as the

$\sigma_x(k)$'s go to zeros, the noise entropy term cancels, as expected.

A.2 Maximizing the flow of information through the network

A.2.1 Optimal input distribution

During the early developmental process, the positional information is orignally encoded in some maternal gradients, namely Bicoid, Nanos and Torso, and flows through the gene network. In this section we show that the constancy of the positional error $\sigma_x(x)$ observed previously is the consequence of maximizing this flow, suggesting that the gene network is operating close to optimal capacity.

For the sake of simplicity, we consider the case where the information flows from a single input transcription factor c to a set of N genes $g_1, \dots, g_N$. The information transmitted by the input to the output is the mutual information between the distributions of c and $\{g\}$ and can be written as

$$\begin{aligned} I\left(c;\{g\}\right) &= \int \mathrm{d}c \int \mathrm{d}g\, P(c,\{g\}) \log_2 \left[\frac{P(c,\{g\})}{P(c)P(\{g\})}\right] \\ &= -\int \mathrm{d}c\, P_{\text{input}}(c) \log_2 P_{\text{input}}(c) \\ &\quad -\int \mathrm{d}g\, P_{\text{output}}(\{g\}) \left[-\int \mathrm{d}c\, P(c|\{g\}) \log_2 P(c|\{g\})\right] \end{aligned} \tag{A.2}$$

with

$$P_{\text{output}}(\{g\}) = \int \mathrm{d}c\, P_{\text{input}}(c) P(\{g\}|c). \tag{A.3}$$

Using Bayes' rule we can rewrite

$$P(c|\{g\}) = P(\{g\}|c) P_{\text{input}}(c) / \int \mathrm{d}c\, P_{\text{input}}(c) P(\{g\}|c) \tag{A.4}$$

and it becomes clear that the transmitted information depends only on two factors: the distribution of inputs $P_{\text{input}}(c)$ and the structure of the gene network encoded in $P(\{g\}|c)$.

If this the first is properly adjusted according to the second, then the gene network can maximize the positional information transmission from the input to the output.

From now, we are going to assume that the noise in the system is small so that the information can be written

$$\begin{aligned} I\left(c;\{g\}\right) &= -\int \mathrm{d}c\, P_{\text{input}}(c) \log_2 P_{\text{input}}(c) \\ &\quad -\int \mathrm{d}c\, P_{\text{input}}(c) \left[-\int \mathrm{d}c\, P(c|\{\overline{g}(c)\}) \log_2 P(c|\{\overline{g}(c)\}) \right]. \end{aligned} \tag{A.5}$$

To find the distribution of inputs $P_{\text{input}}(c)$ that maximizes the information flow satisfies $\int \mathrm{d}c\, P_{\text{input}}(c) = 1$, we introduce a Lagrange multiplier and solve

$$\frac{\delta}{\delta P_{\text{input}}(c)} \left[I\left(c;\{g\}\right) - \lambda \int \mathrm{d}c\, P_{\text{input}}(c) \right] = 0, \tag{A.6}$$

which leads to

$$P_{\text{opt}}(c) = \frac{1}{Z} \exp\left[\ln(2) \int \mathrm{d}c\, P(c|\{\overline{g}(c)\}) \log_2 P(c|\{\overline{g}(c)\}) \right], \tag{A.7}$$

where Z is a normalization constant such that $\int \mathrm{d}c\, P_{\text{input}}(c) = 1$. Now if we make the same assumption than in the rest of this article, which is that the noise is Gaussian

$$P(c|\{g\}) = \frac{1}{\sqrt{2\pi\sigma_c^2(g)}} \exp\left[-\frac{(c - \overline{c}(g))^2}{2\sigma_c^2(g)} \right], \tag{A.8}$$

then we have that $\int \mathrm{d}c\, P(c|\{\overline{g}(c)\}) \log_2 P(c|\{\overline{g}(c)\}) = 1/2 \log_2[2\pi\sigma_c^2(g)]$ and equation A.7 becomes

$$P_{\text{opt}}(c) = \frac{1}{Z'} \cdot \frac{1}{\sigma_c^2(g)}, \tag{A.9}$$

where Z' is another renormalization constant. As discussed in [Tkacik *et al.*, 2008], this tells us that the system can optimize information transmission by using the symbols c in proportion to their reliability.

A.2.2 Constancy of the positional error

If the maternal input is an "image" of the position and that the nuclei are uniformly distributed along the anteroposterior axis, we must have:

$$P(c)\,\mathrm{d}c = P(x)\,\mathrm{d}x = \frac{\mathrm{d}x}{L}. \tag{A.10}$$

By plugging this in equation A.9, we obtain

$$\sigma_x(x) = \frac{Z'}{L} = \mathrm{Cte}. \tag{A.11}$$

Thus, optimizing the information flow through the network predicts that the positional error should be constant along the embryo, as observed in figure 3.6.

Bibliography

[Ait-Sahalia & Kimmel, 2007] Ait-Sahalia, Y. & Kimmel, R. (2007). Maximum likelihood estimation of stochastic volatility models. *Journal of Financial Economics* **83**(2), 413–452.

[Alberts *et al.*, 2002] Alberts, B., Johnson, A., Lewis, J., Raff, M., & Roberts, K. (2002). *Molecular Biology of the Cell.* Garland Science, New York.

[Ay *et al.*, 2008] Ay, A., Fakhouri, W. D., Chiu, C., & Arnosti, D. N. (2008). Image processing and analysis for quantifying gene expression from early Drosophila embryos. *Tissue engineering. Part A* **14**(9), 1517–1526.

[Barlow, 1959] Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. in *Proceedings of the Symposium on the Mechanization of Thought Processes* (Uttley, D. V. & Blake, A. M., eds.) pp. 537–574 London. The mechanisation of thought processes.

[Blake *et al.*, 2003] Blake, W. J., Kaern, M., Cantor, C. R., & Collins, J. J. (2003). Noise in eukaryotic gene expression. *Nature* **422**(6932), 633–637.

[Boveri, 1901] Boveri, T. (1901). Die Polarität von Ovocyte, Ei und Larve des Stronglyocentrotus lividus. *Zool. Jahrb. Abt. Anat. Ontog. Tiere* **14**, 630–653.

[Brenner *et al.*, 2000] Brenner, N., Bialek, W., & de Ruyter van Steveninck, R. (2000). Adaptive rescaling maximizes information transmission. *Neuron* **26**(3), 695–702.

[Brunel & Nadal, 1998] Brunel, N. & Nadal, J. P. (1998). Mutual information, Fisher information, and population coding. *Neural Computation* **10**(7), 1731–1757.

[Capovilla *et al.*, 1992] Capovilla, M., Eldon, E. D., & Pirrotta, V. (1992). The giant gene of Drosophila encodes a b-ZIP DNA-binding protein that regulates the expression of other segmentation gap genes. *Development* **114**(1), 99–112.

[Chen & Schier, 2001] Chen, Y. & Schier, A. F. (2001). The zebrafish Nodal signal Squint functions as a morphogen. *Nature* **411**(6837), 607–610.

[Choi *et al.*, 2009] Choi, Y., Kim, H. P., Hong, S. M., Ryu, J. Y., Han, S. J., & Song, R. (2009). In situ visualization of gene expression using polymer-coated quantum-dot-DNA conjugates. *Small* **5**(18), 2085–2091.

[Conklin, 1905] Conklin, E. G. (1905). Organ-forming substances in the eggs of ascidians. *Bio. Bull.* **8**, 205–230.

[Cover & Thomas, 1991] Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory.* J. Wiley & Sons, New York.

[Crauk & Dostatni, 2005] Crauk, O. & Dostatni, N. (2005). Bicoid determines sharp and precise target gene expression in the Drosophila embryo. *Current Biology* **15**(21), 1888–1898.

[de Ronde *et al.*, 2010] de Ronde, W. H., Tostevin, F., & ten Wolde, P. R. (2010). Effect of feedback on the fidelity of information transmission of time-varying signals. *Physical Review E* **82**, 031914.

[Dickinson *et al.*, 2001] Dickinson, M. E., Bearman, G., Tille, S., Lansford, R., & Fraser, S. E. (2001). Multi-spectral imaging and linear unmixing add a whole new dimension to laser scanning fluorescence microscopy. *BioTechniques* **31**(6), 1272–1278.

[Driever & Nüsslein-Volhard, 1988a] Driever, W. & Nüsslein-Volhard, C. (1988a). A gradient of Bicoid protein in Drosophila embryos. *Cell* **54**(1), 83–93.

[Driever & Nüsslein-Volhard, 1988b] Driever, W. & Nüsslein-Volhard, C. (1988b). The Bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner. *Cell* **54**(1), 95–104.

[Driever & Nüsslein-Volhard, 1989] Driever, W. & Nüsslein-Volhard, C. (1989). The Bicoid protein is a positive regulator of hunchback transcription in the early Drosophila embryo. *Nature* **337**(6203), 138–143.

[Driever *et al.*, 1989] Driever, W., Thoma, G., & Nüsslein-Volhard, C. (1989). Determination of spatial domains of zygotic gene expression in the Drosophila embryo by the affinity of binding sites for the bicoid morphogen. *Nature* **340**(6232), 363–367.

[Dubnau & Struhl, 1996] Dubnau, J. & Struhl, G. (1996). RNA recognition and translational regulation by a homeodomain protein. *Nature* **379**(6567), 694–699.

[Dubuis *et al.*, 2011] Dubuis, J. O., Tkacik, G., Wieschaus, E. F., Gregor, T., & Bialek, W. (2011). Positional information, in bits. *arXiv.org* **q-bio.MN**.

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

[Ephrussi & St Johnston, 2004] Ephrussi, A. & St Johnston, D. (2004). Seeing is believing: the bicoid morphogen gradient matures. *Cell* **116**(2), 143–152.

[Fakhouri *et al.*, 2010] Fakhouri, W. D., Ay, A., Sayal, R., Dresch, J., Dayringer, E., & Arnosti, D. N. (2010). Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo. *Molecular Systems Biology* **6**, 341.

[Fasano & Kerridge, 1988] Fasano, L. & Kerridge, S. (1988). Monitoring positional information during oogenesis in adult Drosophila. *Development* **104**, 245–253.

[Fowlkes *et al.*, 2008] Fowlkes, C. C., Hendriks, C. L. L., Keränen, S. V. E., Weber, G. H., Rübel, O., Huang, M.-Y., Chatoor, S., DePace, A. H., Simirenko, L., Henriquez, C., Beaton, A., Weiszmann, R., Celniker, S., Hamann, B., Knowles, D. W., Biggin, M. D., Eisen, M. B., & Malik, J. (2008). A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. *Cell* **133**(2), 364–374.

[Gergen *et al.*, 1986] Gergen, J. P., Coulter, D., & Wieschaus, E. F. (1986). Segmental pattern and blastoderm cell identities. in *Gametogenesis and The Early Embryo* (Gall, J. G., ed.) pp. 195–220 New York. Gametogenesis and the Early Embryo.

[Gergen & Wieschaus, 1986] Gergen, J. P. & Wieschaus, E. F. (1986). Localized requirements for gene activity in segmentation of Drosophila embryos - analysis of armadillo, fused, giant and unpaired mutations in mosaic embryos. *Wilhelm Rouxs Archives of Developmental Biology* **195**(1), 49–62.

[Gierer & Meinhardt, 1972] Gierer, A. & Meinhardt, H. (1972). A theory of biological pattern formation. *Kybernetik* **12**(1), 30–39.

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

[Gregor *et al.*, 2007a] Gregor, T., Tank, D. W., Wieschaus, E. F., & Bialek, W. (2007a). Probing the limits to positional information. *Cell* **130**(1), 153–164.

[Gregor *et al.*, 2007b] Gregor, T., Wieschaus, E. F., McGregor, A. P., Bialek, W., & Tank, D. W. (2007b). Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell* **130**(1), 141–152.

[Houchmandzadeh *et al.*, 2002] Houchmandzadeh, B., Wieschaus, E., & Leibler, S. (2002). Establishment of developmental precision and proportions in the early Drosophila embryo. *Nature* **415**(6873), 798–802.

[Huang *et al.*, 2005] Huang, S., Eichler, G., Bar-Yam, Y., & Ingber, D. E. (2005). Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Physical Review Letters* **94**(12), 128701.

[Hülskamp *et al.*, 1989] Hülskamp, M., Schröder, C., Pfeifle, C., Jäckle, H., & Tautz, D. (1989). Posterior segmentation of the Drosophila embryo in the absence of a maternal posterior organizer gene. *Nature* **338**(6217), 629–632.

[Illmensee & Mahowald, 1974] Illmensee, K. & Mahowald, A. P. (1974). Transplantation of posterior polar plasm in Drosophila. Induction of germ cells at the anterior pole of the egg. *Proc Natl Acad Sci USA* **71**(4), 1016–1020.

[Irish & Lehmann, 1989] Irish, V. & Lehmann, R. (1989). The Drosophila posterior-group gene nanos functions by repressing hunchback activity. *Nature* **338**, 646–648.

[Jaeger, 2011] Jaeger, J. (2011). The gap gene network. *Cellular and Molecular Life Sciences* **68**(2), 243–274.

[Jaeger & Irons, 2008] Jaeger, J. & Irons, D. (2008). Regulative feedback in pattern formation: towards a general relativistic theory of positional information. *Development* **135**, 3175–3183.

[Jaeger & Reinitz, 2006] Jaeger, J. & Reinitz, J. (2006). On the dynamic nature of positional information. *BioEssays* **28**(11), 1102–1111.

[Jaeger *et al.*, 2004] Jaeger, J., Surkova, S., Blagov, M., Janssens, H., Kosman, D., Kozlov, K. N., Manu, Myasnikova, E., Vanario-Alonso, C. E., Samsonova, M., Sharp, D. H., & Reinitz, J. (2004). Dynamic control of positional information in the early Drosophila embryo. *Nature* **430**(6997), 368–371.

[Kosman *et al.*, 1998] Kosman, D., Small, S., & Reinitz, J. (1998). Rapid preparation of a panel of polyclonal antibodies to Drosophila segmentation proteins. *Development Genes and Evolution* **208**(5), 290–294.

[Kraut & Levine, 1991] Kraut, R. & Levine, M. (1991). Spatial regulation of the gap gene giant during Drosophila development.. *Development* **111**(2), 601–609.

[Lagarias *et al.*, 1998] Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. *Siam Journal on Optimization* **9**(1), 112–147.

[Laughlin, 1981] Laughlin, S. B. (1981). A simple coding procedure enhances a neuron's information capacity. *Z Naturforsch.* **36**, 910–912.

[Lawrence, 1966] Lawrence, P. A. (1966). Gradients in insect segment - orientation of hairs in milkweed bug oncopeltus fasciatus. *Journal of Experimental Biology* **44**(3), 607–610.

[Lawrence, 1992] Lawrence, P. A. (1992). The making of a fly: the genetics of animal design. Blackwell Scientific.

[Lecuit *et al.*, 1996] Lecuit, T., Brook, W. J., Ng, M., Calleja, M., & Sun, H. (1996). Two distinct mechanisms for long-range patterning by Decapentaplegic in the Drosophila wing. *Nature* **381**, 387–393.

[Lecuit *et al.*, 2002] Lecuit, T., Samanta, R., & Wieschaus, E. (2002). Slam encodes a developmental regulator of polarized membrane growth during cleavage of the Drosophila embryo.. *Developmental Cell* **2**(4), 425–436.

[Lee & Liu, 2012] Lee, W. & Liu, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *Journal of multivariate analysis* **111**, 241–255.

[Little *et al.*, 2011] Little, S. C., Tkačik, G., Kneeland, T. B., Wieschaus, E. F., & Gregor, T. (2011). The formation of the Bicoid morphogen gradient requires protein movement from anteriorly localized mRNA. *PLoS Biology* **9**(3), e1000596.

[Ludwig *et al.*, 2011] Ludwig, M. Z., Manu, Kittler, R., White, K. P., & Kreitman, M. (2011). Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. *PLoS Genetics* **7**(11), e1002364.

[Manu *et al.*, 2009] Manu, Surkova, S., Spirov, A. V., Gursky, V. V., Janssens, H., Kim, A.-R., Radulescu, O., Vanario-Alonso, C. E., Sharp, D. H., Samsonova, M., & Reinitz, J. (2009). Canalization of gene expression in the Drosophila blastoderm by gap gene cross regulation. *PLoS Biology* **7**(3), e1000049.

[McMahon *et al.*, 2003] McMahon, A. P., Ingham, P. W., & Tabin, C. J. (2003). Developmental roles and clinical significance of hedgehog signaling. *Current Topics in Developmental Biology* **53**, 1–114.

[Meinhardt, 1989] Meinhardt, H. (1989). Models for positional signalling with application to the dorsoventral patterning of insects and segregation into different cell types. *Development* **107**, 169–180.

[Morgan, 1904] Morgan, T. H. (1904). An attempt to analyze the phenomena of polarity in Tubularia. *Journal of Experimental Zoology* **1**(4), 587–591.

[Morgan, 1905] Morgan, T. M. (1905). Polarity considered as a phenomenon of gradation of materials. *Journal of Experimental Zoology* **2**(4), 495–506.

[Moses & Rubin, 1991] Moses, K. & Rubin, G. M. (1991). Glass encodes a site-specific DNA-binding protein that is regulated in response to positional signals in the developing Drosophila eye. *Genes & Development* **5**(4), 583–593.

[Murata & Wharton, 1995] Murata, Y. & Wharton, R. P. (1995). Binding of pumilio to maternal hunchback mRNA is required for posterior patterning in Drosophila embryos. *Cell* **80**(5), 747–756.

[Myasnikova *et al.*, 2009] Myasnikova, E., Surkova, S., Panok, L., Samsonova, M., & Reinitz, J. (2009). Estimation of errors introduced by confocal imaging into the data on segmentation gene expression in Drosophila. *Bioinformatics* **25**(3), 346–352.

[Neher *et al.*, 2009] Neher, R. A., Mitkovski, M., Kirchhoff, F., Neher, E., Theis, F. J., & Zeug, A. (2009). Blind source separation techniques for the decomposition of multiply labeled fluorescence images. *Biophysical Journal* **96**(9), 3791–3800.

[Neumann & Cohen, 1997] Neumann, C. & Cohen, S. (1997). Morphogens and pattern formation. *BioEssays* **19**(8), 721–729.

[Nüsslein-Volhard, 1977] Nüsslein-Volhard, C. (1977). Genetic analysis of pattern formation in embryo of Drosophila melanogaster. Characterization of maternal effect mutant Bicaudal. *Wilhelm Rouxs Archives of Developmental Biology* **183**(3), 249–268.

[Nüsslein-Volhard & Wieschaus, 1980] Nüsslein-Volhard, C. & Wieschaus, E. (1980). Mutations affecting segment number and polarity in Drosophila. *Nature* **287**(5785), 795–801.

[Ochoa-Espinosa *et al.*, 2009] Ochoa-Espinosa, A., Yu, D., Tsirigos, A., Struffi, P., & Small, S. (2009). Anterior-posterior positional information in the absence of a strong Bicoid gradient. *Proc Natl Acad Sci USA* **106**(10), 3823–3828.

[Ozbudak *et al.*, 2002] Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D., & van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. *Nature Genetics* **31**(1), 69–73.

[Painter *et al.*, 1999] Painter, K. J., Maini, P. K., & Othmer, H. G. (1999). Stripe formation in juvenile Pomacanthus explained by a generalized turing mechanism with chemotaxis. *Proc Natl Acad Sci USA* **96**(10), 5549–5554.

[Phillips & Yu, 2009] Phillips, P. C. B. & Yu, J. (2009). *Handbook of Financial Time Series*. Springer, Berlin.

[Pinheiro & Bates, 1996] Pinheiro, J. C. & Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing* **6**(3), 289–296.

[Pisarev *et al.*, 2009] Pisarev, A., Poustelnikova, E., Samsonova, M., & Reinitz, J. (2009). FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic Acids Research* **37**(Database issue), D560–6.

[Reinitz *et al.*, 1995] Reinitz, J., Mjolsness, E., & Sharp, D. H. (1995). Model for cooperative control of positional information in Drosophila by bicoid and maternal hunchback. *Journal of Experimental Zoology* **271**(1), 47–56.

[Rivera-Pomar *et al.*, 1995] Rivera-Pomar, R., Lu, X., Perrimon, N., Taubert, H., & Jäckle, H. (1995). Activation of posterior gap gene expression in the Drosophila blastoderm. *Nature* **376**(6537), 253–256.

[Rivera-Pomar *et al.*, 1996] Rivera-Pomar, R., Niessing, D., Schmidt-Ott, U., Gehring, W. J., & Jäckle, H. (1996). RNA binding and translational suppression by bicoid. *Nature* **379**(6567), 746–749.

[Rosenfeld *et al.*, 2005] Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S., & Elowitz, M. B. (2005). Gene regulation at the single-cell level. *Science* **307**(5717), 1962–1965.

[Sánchez & Thieffry, 2001] Sánchez, L. & Thieffry, D. (2001). A logical analysis of the Drosophila gap-gene system. *Journal of Theoretical Biology* **211**(2), 115–141.

[Shannon, 1948] Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* **27**(3), 379–423.

[Slack, 2001] Slack, J. (2001). Essential Developmental Biology. Blackwell Publishing.

[Slonim *et al.*, 2005] Slonim, N., Atwal, G. S., Tkačik, G., & Bialek, W. (2005). Estimating mutual information and multi-information in large networks. 2005. *Arxiv preprint cs/0502017*.

[Spirov & Holloway, 2003] Spirov, A. V. & Holloway, D. M. (2003). Making the body plan: precision in the genetic hierarchy of Drosophila embryo segmentation. *In Silico Biology* **3**(1-2), 89–100.

[Strong *et al.*, 1998] Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R., & Bialek, W. (1998). Entropy and information in neural spike trains. *Physical Review Letters* **80**(1), 197–200.

[Struhl, 1989] Struhl, G. (1989). Differing strategies for organizing anterior and posterior body pattern in Drosophila embryos. *Nature* **338**(6218), 741–744.

[Stumpf, 1968] Stumpf, H. F. (1968). Further studies on gradient-dependent diversification in pupal cuticle of Galleria mellonella. *Journal of Experimental Biology* **49**(1), 49–&.

[Surkova *et al.*, 2008] Surkova, S., Kosman, D., Kozlov, K., Manu, Myasnikova, E., Samsonova, A. A., Spirov, A., Vanario-Alonso, C. E., Samsonova, M., & Reinitz, J. (2008). Characterization of the Drosophila segment determination morphome. *Developmental Biology* **313**(2), 844–862.

[Swantek & Gergen, 2004] Swantek, D. & Gergen, J. P. (2004). Ftz modulates Runt-dependent activation and repression of segment-polarity gene transcription. *Development* **131**(10), 2281–2290.

[Tkacik *et al.*, 2008] Tkacik, G., Callan, C. G., & Bialek, W. (2008). Information flow and optimization in transcriptional regulation. *Proc Natl Acad Sci USA* **105**(34), 12265–12270.

[Tkacik & Walczak, 2011] Tkacik, G. & Walczak, A. M. (2011). Information transmission in genetic regulatory networks: a review. *Journal of Physics: Condensed Matter* **23**(15), 153102.

[Tkacik *et al.*, 2009] Tkacik, G., Walczak, A. M., & Bialek, W. (2009). Optimizing information flow in small genetic networks. *Physical Review E* **80**(3 Pt 1), 031920.

[Tomancak *et al.*, 2007] Tomancak, P., Berman, B. P., Beaton, A., Weiszmann, R., Kwan, E., Hartenstein, V., Celniker, S. E., & Rubin, G. M. (2007). Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biology* **8**(7), R145.

[Tomlinson *et al.*, 1987] Tomlinson, A., Bowtell, D., & Hafen, E. (1987). Localization of the Sevenless protein, a putative receptor for positional information, in the eye imaginal disc of Drosophila. *Cell* **51**, 143–150.

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

[von Dassow *et al.*, 2000] von Dassow, G., Meir, E., Munro, E. M., & Odell, G. M. (2000). The segment polarity network is a robust developmental module. *Nature* **406**(6792), 188–192.

[Walczak *et al.*, 2010] Walczak, A. M., Tkacik, G., & Bialek, W. (2010). Optimizing information flow in small genetic networks. II. Feed-forward interactions. *Physical Review E* **81**, 041905.

[Wilson, 1904] Wilson, E. B. (1904). Experimental studies on germinal localization. *Journal of Experimental Zoology* **1**(1), 1–72.

[Wolpert, 1969] Wolpert, L. (1969). Positional information and the spatial pattern of cellular differentiation.. *Journal of Theoretical Biology* **25**(1), 1–47.

[Wolpert, 1971] Wolpert, L. (1971). Positional information and pattern formation. *Current Topics in Developmental Biology* **6**, 183–224.

[Wolpert, 1996] Wolpert, L. (1996). One hundred years of positional information. *Trends in Genetics* **12**, 359–369.

[Wolpert *et al.*, 2002] Wolpert, L., Beddington, R., Brockets, J., & Jessel, T. (2002). Principles of Development. Oxford University Press.

[Ziv *et al.*, 2007] Ziv, E., Nemenman, I., & Wiggins, C. H. (2007). Optimal Signal Processing in Small Stochastic Biochemical Networks. *PloS One* **2**(10), e1077.