A statistical model relating transcription factor concentrations to positional information in the early *Drosophila* embryo

Garth Robert Ilsley

Peterhouse

A thesis submitted to the University of Cambridge
for the degree of Doctor of Philosophy

June 2010

DECLARATION

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

No parts of this dissertation have been submitted for any other qualification.

This dissertation does not exceed the specified limit of 60,000 words as set by the Biology Degree Committee.

14 June 2010 Garth Ilsley

A statistical model relating transcription factor concentrations to positional information in the early *Drosophila* embryo

Garth Robert Ilsley

The idea of morphogen gradients encoding positional information for a developing organism has long been discussed in the field of developmental biology, but only recently have quantitative models been proposed that relate measured transcription factor concentrations to enhancer activity. However, successful models are typically computationally time-consuming, thus limiting full exploration and interpretation of the data. This thesis addresses these problems using standard statistical techniques applied to a comprehensive data set with the *even skipped* (*eve*) locus as a test case.

The first part of the thesis introduces the data set. This is the precellular *Virtual Embryo* from the Berkeley Drosophila Transcription Network project. It comprises expression measurements of almost 100 genes in more than 6,000 individual nuclei at six time points. Different modelling approaches are evaluated in the context of this data set leading to a justification of logistic regression and the methods used to prepare the data set for further analysis.

The second part applies logistic regression to describe the response of the *eve* enhancers to known regulating transcription factors such as Hunchback. Predictions of behaviour under regulator perturbation are consistent with experimental results and the functional form is shown not to be arbitrarily flexible, both in terms of the regulators and regions of the embryo included.

The third part uses the framework developed above to find minimal explanatory models in the context of statistical model selection. It is found that the best scoring models depend on well-known regulators. The model selection techniques are then extended by directing the process using previous biological observations to analyse the *eve* 2 and *eve* 3+7 enhancers. The results are consistent with published research, but suggest specific additional hypotheses for the enhancers' regulation.

Finally, the thesis concludes by proposing a general model of positional information and discussing the biological implications of the results. Overall, the results show how transcriptional control can be allocated to discrete enhancers and that characterising their activity in relatively simple terms is sufficient to explain their precise spatially-defined response to transcription factor concentrations.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support of my wife, Tanja. I am hugely grateful to her for supporting me in giving up gainful employment to embark on this project, and throughout looking after our young children. Thank you Nathan, Vivian and Harriet for being yourselves. Thank you also to my parents for their loving interest.

I would like to thank Henning Hermjakob, who interviewed me, and Rolf Apweiler, my supervisor, for offering me a place at the European Molecular Biology Laboratory (EMBL). I am also a grateful recipient of the EMBL allowance for those with children; it has made this endeavour feasible.

I would like to thank Rolf for giving me the freedom to explore, so I could discover science in general and a topic that fascinates me. Thank you for supporting me in my interactions and discussions with other groups.

I thank Alvis Brazma at the European Bioinformatics Institute (EBI) for hosting me for a large part of a year and showing interest in my work, and Nils Gehlenborg for the good chats we had. I am also grateful for my time as an intern at Microsoft Research with Jasmin Fisher, where I had the opportunity to learn about the links between computer science and biology. I thank Jasmin for showing early interest in my work and believing in its potential.

I am indebted to the Machine Learning book reading group, organised by graduate students of the Inference Group at the Cavendish Laboratory, for introducing me to the joy of understanding. Thank you, in particular, to Philipp Hennig, Christian Steinruecken, Tamara Broderick and Carl Scheffler for excellent and enlightening discussions.

I would like to acknowledge my Thesis Advisory Committee: Michael Ashburner, Eileen Furlong, Wolfgang Huber, Jasmin Fisher, Nick Luscombe and Rolf Apweiler. In particular, I am grateful to Michael Ashburner for encouraging me to spend time in a fly lab and to Rolf in supporting this. I would also like to acknowledge my college, Peterhouse, for providing significant funding towards the consequent month-long visit to the lab of Angela DePace at Harvard Medical School. I thank Angela and her friendly group for welcoming me, and showing me the beauty of flies. In particular, I appreciate the help of Zeba Wunderlich, as well as Charless Fowlkes and Cris Luengo who were visiting there at the time, in understanding some of the trickier aspects of the *Virtual Embryo*. Tara Martin deserves special thanks for having confidence in my results and investing so much effort trying to verify them. May those flies be tamed yet!

I thank Wolfgang Huber for reading parts of chapter 3 and providing helpful criticism, and Nicolas Le Novère for reading an initial draft of chapter 2 and checking the mathematics of appendix A. Angela DePace read chapter 6 and helped improve its accessibility, for which I am grateful. Naturally, I am responsible for any errors that remain.

I am especially thankful to the Luscombe group at the EBI for making me welcome for a significant part of my PhD. Thank you to Aswin Seshasayee, Judith Zaugg, Kathi Zarnack, Annabel Todd, Florence Cavalli, Karthik Sivaraman, Filipe Cadete, Iñigo Martincorena, and especially Juanma Vaquerizas for sharing their knowledge with me. I am truly thankful to Nick Luscombe for the many discussions we have had and the advice he has given me. I find it remarkable that he was willing to read patiently through this thesis multiple times, giving detailed comments to improve its readability. But most of all, I thank Nick for his encouragement.

1

INTRODUCTION

The idea of morphogen gradients encoding positional information for a developing organism has long been discussed in the field of developmental biology, but only recently have quantitative models been proposed that relate measured transcription factor concentrations to enhancer activity. However, successful models are typically computationally time-consuming, thus limiting full exploration and interpretation of the data. This thesis addresses these problems using a comprehensive data set and the *even skipped* locus as a test case. The proposed model takes account of multiple enhancers and is able to show the sufficiency of actual, quantitative transcription factor gradients in accurately specifying position at cellular resolution. The thesis culminates in a model of positional information in the spirit of Wolpert's French flag problem.

This chapter introduces the concept of positional information and morphogen gradients. A particular tractable animal system for studying this problem is then outlined—that of the gene expression patterns in the *Drosophila melanogaster* embryo prior to gastrulation. The focus of the modelling work presented here is that of the *even skipped* gene, so a sketch is included of the allocation of its transcriptional control to different enhancers in the locus of the gene. Relevant aspects of previous modelling approaches are then described, and finally, the data set used in this work is introduced.

1.1 THE FRENCH FLAG PROBLEM

Lewis Wolpert first proposed the French flag problem in Wolpert (1968). The essential proposal is that in trying to understand patterning during development it is useful to separate questions regarding the establishment of positional information from those of how the cell interprets that information. Any process establishing positional information needs to take account of features at the level of the organism, such as boundaries or polarity, and it is important that the process is able to adjust to perturbations or to the scale of the organism. A separate question then, is how each cell converts the available positional information into a cellular parameter—the positional value—which is used according to that cell's lineage or history to make decisions that lead to spatial patterns in the developing embryo. If interpretation is at the level of a specific gene, then interpretation is equivalent to gene regulation: how is the gene regulated during development? The French flag problem asks how, at the cellular level, might the global (embryo-level) features of positional information be established and interpreted so that a pattern like the French flag is produced, with each region the same length and in the right order. This thesis is largely concerned with the interpretation of positional information, although, as will be shown in chapter 7, this has implications for its establishment.

Two main types, or models, of positional information were proposed (as refined in Wolpert, 1969): a *quantitative* variation that increases or decreases monotonically from a boundary, and a *qualitative* variation, such as a mechanism of cell counting from the boundary. It is only quantitative variation that is considered here. In this case, Lewis Wolpert envisaged that a simple threshold could be imposed on the available positional information to divide the embryo into two regions. He also pointed out that in order for positional information to scale according to the size of the embryo, it is necessary for the boundaries

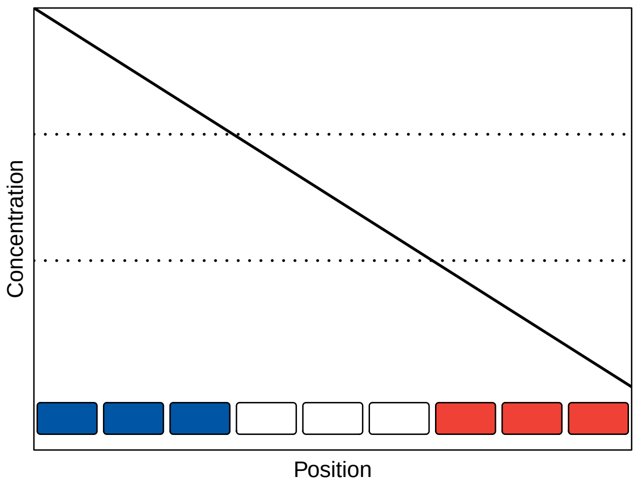


Figure 1.1: French flag model with a linear gradient of positional information. The dotted lines show two thresholds, used to specify the blue, white and red cells.

at both ends to be involved in the establishment of positional information. One way to do this is with a linear gradient that has a defined concentration at both ends. Since positional information is thought to act over a range of less than 1 mm and requires of the order of a few hours to establish (Wolpert, 1969, 1996), this led Francis Crick to propose that a diffusible morphogen (Crick, 1970) could set up the positional field in a developing embryo*. It is this solution to the French flag problem that is best known; for example, it is referred to as the *French flag model* in two recent reviews of positional information in *Drosophila* (Jaeger and Martinez-Arias, 2009; Papatsenko, 2009). Figure 1.1 shows the French flag model with a linear positional gradient, which could be established by a source of the morphogen on the left, and a sink of the morphogen on the right. Regardless of the specific form, Wolpert (1989) has stated that the terms positional signal (a signal to the cell imparting positional information), positional value (the cell's interpretation of the signal) and morphogen (a concentration gradient specifying position) should not be conflated. Thus, in this thesis, the

* The term morphogen was coined by Alan Turing to describe a diffusible chemical in a reaction-diffusion pattern-forming process (Turing, 1952).

term morphogen will refer to a factor that forms part of the positional information available to the cell; that is, it is used in calculating positional value. All the contours of equal positional value in turn define a *positional field* within which the organism develops. As will be shown, interpretation depends on multiple factors and therefore the resulting positional field cannot be derived from any one morphogen gradient.

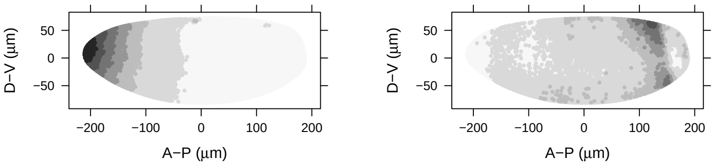
1.2 THE DROSOPHILA BLASTODERM

Drosophila melanogaster, a fruit fly, has played an important part in experimental investigations of positional information in a developing organism. In particular, it was host to the first molecular demonstration of a morphogen gradient—the Bicoid protein in the early embryo (Driever and Nüsslein-Volhard, 1988a,b). But *Drosophila* has also been central in investigations of how the striking gene expression patterns established before gastrulation give rise to later body structures, including the repeated segments typical of insects (reviewed by Akam, 1987; Ingham, 1988; Nasiadka et al., 2002).

This system is remarkably effective for understanding gene regulation for a number of reasons. The gene expression patterns are largely those of transcription factors—regulatory proteins that bind genomic DNA and control transcription of target genes. Since the expression patterns are formed progressively, with each class of transcription factors controlling the formation of the next, it is possible to assess directly how the concentrations of transcription factor inputs relate to their output. And because the patterns are spatial, the study of gene regulation can be approached in terms of positional information.

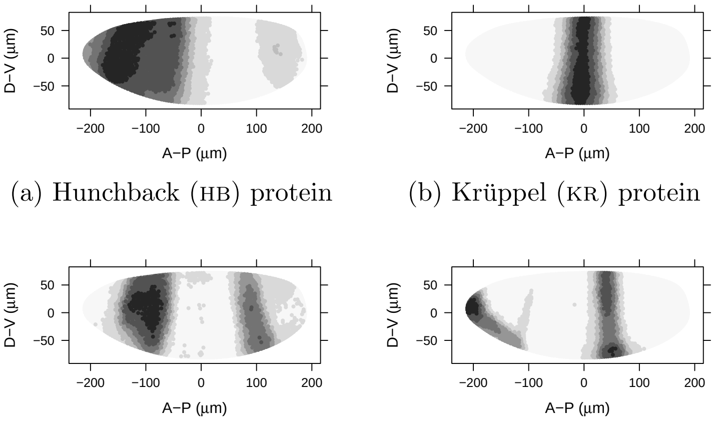
A further practical benefit for the measurement and observation of gene expression is that in the hour or so before gastrulation, the *Drosophila* embryo is a single-layered epithelium—the blastoderm—which has been formed by the migration of most of the nuclei to the periphery of the embryo (Campos-Ortega and Hartenstein, 1997). Fortuitously for the purposes of this thesis, the nuclei are not separated by cellular membranes. Rather they grow gradually down from the cortex during the stage leading up to gastrulation. The nuclei, therefore, cannot depend substantially on intra-cellular signalling for the specification of positional information, making the link between transcription factor concentrations and their measurable consequence relatively direct.

The patterning of gene expression progressively unfolds in the stages following fertilisation along the dorsoventral and anteroposterior axes. In this thesis, the focus is primarily the anteroposterior axis along which the future segments lie. This process begins with factors produced by the mother localised at the poles of the oocyte. Of these, *bicoid* (*bcd*), a transcription factor, is of crucial importance for the establishment of later patterns (Frohnhöfer and Nüsslein-Volhard, 1986). Initially the BCD protein forms a gradient decreasing towards the posterior of the embryo (figure 1.2). These maternally-supplied genes then regulate the next class of transcription factors: the gap genes (figure 1.3), such as *giant* (*gt*), *Krüppel* (*Kr*), *knirps* (*kni*) and *hunchback* (*hb*). These, together with the maternal factors, define the striped pattern of the pair rule genes such as *even skipped* (*eve*) and *fushi tarazu* (*ftz*); see figure 1.4. The pair rule genes are also transcription factors, and these then define the future parasegments, which later give rise to the segments of the adult. There are further classes of genes, such as those responsible for the polarity and identify of the segments, but these are not directly relevant to the matters of positional information discussed in this thesis.



(a) Bicoid (BCD) protein (b) *caudal* (*cad*) mRNA

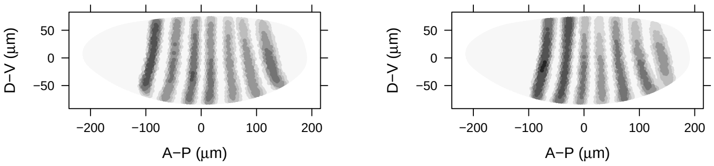
Figure 1.2: The expression of two important maternal genes. Darker shades signify higher expression.



(a) Hunchback (HB) protein (b) Krüppel (KR) protein

(c) Giant (GT) protein (d) *knirps* (*kni*) mRNA

Figure 1.3: The expression of a few important gap genes. Darker shades signify higher expression.



(a) *even skipped* (*eve*) mRNA (b) *fushi tarazu* (*ftz*) mRNA

Figure 1.4: The expression of two important pair rule genes. Darker shades signify higher expression.

1.2.1 *Positional information*

Each of these steps in the cascade has been studied with regard to positional information. One of the most influential has been a simple threshold model for the targets of BCD (reviewed in Ephrussi and Johnston, 2004). This has been based on the link between number and affinity of BCD binding sites in the regulatory region for the *hunchback* gene and the different consequent levels of *hunchback* activation (Driever and Nüsslein-Volhard, 1988a,b; Struhl et al., 1989). However, this view has become more complicated over time. For example, BCD alone is not sufficient to activate some targets (Simpson-Brose et al., 1994), and questions concerning precision in the read-out of the threshold have been raised (Houchmandzadeh et al., 2002).

Another fruitful area of research into positional information has been the gap genes and their cross-regulation. This has led to models of dynamically changing positional information (Jaeger et al., 2004) and a proposal regarding the importance of compensatory dynamic control (Manu et al., 2009b).

The precision and control of the boundaries between the gap genes is remarkable, and these are used in turn to define the location of the *eve* stripes, one of the best studied pair rule genes. It has been supposed that the interpretation of positional information in defining each of *eve*'s stripes is relatively simple (Papatsenko, 2009), and the prevalent model is that each stripe is specified by two repressor gradients converging from opposing directions to define each border. This has been proposed for *eve* stripe 2 (Stanojevic et al., 1991; Small et al., 1992; Andrioli et al., 2002), stripes 3 and 7 (Small et al., 1996; Clyde et al., 2003) and stripes 4 and 6 (Fujioka et al., 1999; Clyde et al., 2003). However, as will be demonstrated in this thesis, this is by no means certain.

1.3 ENHANCERS AND THE EVE LOCUS

The pair rule gene *even skipped* (*eve*) is interesting, but far from unique, in that its regulation is controlled by multiple enhancers. In contemporary usage, enhancers are discrete, contiguous regions of genomic DNA that are able to regulate the transcription of a gene[†]. This is achieved by recruiting combinations of transcription factors according to the enhancer's sequence composition, and the transcription factors in turn affect the rate of initiation of transcription of the target gene via co-regulators and the basal transcriptional apparatus. Many other layers of control are available to the eukaryotic cell, including post-translational modification of proteins via signalling cascades and other post-transcriptional regulatory events. The focus of this thesis is only one part of the regulatory mechanism: the information provided to enhancers by concentrations of proteins, specifically the transcription factors. As a result, these will also be referred to more generally as *regulators*.

Enhancers are modular in that they can be combined with a transgenic reporter gene and are able to drive the expression of a subset of the target gene's endogenous expression. They are therefore presumably combined additively in the endogenous locus to produce the overall pattern. Given the modular and regulatory nature of enhancers, they are also called *cis*-regulatory elements or modules (*cis* refers to the fact that the element is on the same molecule as the gene, the DNA).

Enhancers are therefore crucial to the temporal and spatial patterns observed in development, and some of the best studied are those near *eve* (see figure 1.5). Some enhancers drive a single stripe, such as *eve* 2 (Small et al., 1992); others pairs of stripes, such as *eve* 3+7 and *eve* 4+6 (Small et al., 1996; Clyde et al., 2003). Many of the

[†] For a general review of transcription factors and the control gene expression see Lemon and Tjian (2000). Arnosti (2003) reviews transcriptional regulation in the context of *Drosophila*.

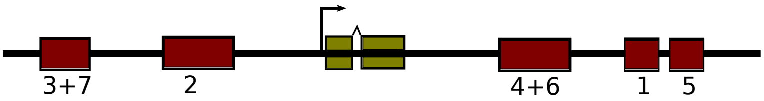


Figure 1.5: Schematic of the *eve* locus showing some of its enhancers (in red). Green shows the coding regions of *eve*. The arrow is the transcription start site.

associated transcription factors and their effects have been examined through mutational and misexpression studies and, as will be seen, this will be important in validating and extending the model of positional information described in this thesis. A further crucial benefit is that since the stripes are controlled by discrete enhancers, gene expression measurements of individual nuclei can be grouped according to the enhancer that controls their expression. Predictions of expression under different models can then easily be visualised and compared to actual expression. All of this will make it possible to discover and describe each enhancer's function accurately and reliably.

1.4 MODEL TYPES

Many aspects of the anteroposterior (A-P) patterning of the *Drosophila* embryo have been modelled over the past few decades. (For a recent review see Jaeger, 2009.) This section will describe the distinguishing features of the main modelling approaches relevant to the work of this thesis. Models of *Drosophila* A-P patterning can be placed into three groups, depending on the level of abstraction they provide.

1.4.1 *Differential equation models*

The most mechanistic are differential equation models that deal in actual rates of transcription with the aim of modelling the dynamic behaviour of the system. One of the early influential models for developmental biology was that of Turing (1952), who demonstrated using a system of differential equations that chemicals with initially small spatial inhomogeneities could form complex and stable patterns given appropriate reaction and diffusion rates. Thus, differential equation models in development are sometimes referred to as reaction-diffusion models, although *reaction* might mean gene transcription and translation, and *diffusion* might refer to any transport mechanism.

As quantitative data have become available for *Drosophila* development, these models have been applied in their analysis. von Dassow et al. (2000) and Jaeger et al. (2004) are good examples. The key feature of this method is that it provides a basis for simulation. In other words, given initial conditions for each of the variables in the system, the model determines how the system changes over time. The behaviour of the system can then be compared to the available experimental data. The model can be improved incrementally by adding or removing variables and connections or tweaking parameters of the model and considering whether this reduces the discrepancy between the model and the data. Thus, the goal is to construct ultimately a plausible model that explains the data.

In the case of *Drosophila* A-P patterning, extensive dynamic data are not yet available so this amounts to ensuring that the model simulation fits the data qualitatively, or in the case of quantitative data, to snapshots in time. It is also difficult to model movements of nuclei and changing morphology, so these simulations usually restrict themselves to a short phase of development. The *Drosophila* blastoderm is useful in this regard because the shape is relatively stable and in the last 50 minutes or so before gastrulation there are no nuclear cleavages. It is often assumed that the nuclei do not move, although this assumption has been questioned (Keränen et al., 2006).

The selection of parameters can be automated, as in the gene circuit approach (Mjolsness et al., 1991; Reinitz et al., 2003; Jaeger et al., 2004). Here an algorithm is used to minimise the discrepancy between the data and the model simulation. Since each change in parameters requires the simulation to be re-run, the parameter selection process can take a long time: earlier approaches took months, but more recently this had been reduced to days (Perkins et al., 2006). One result of this approach is that the data are restricted to measurements of gene expression along the lateral midline of the embryo—a one dimensional data set.

However, although parameter selection can be automated, the topology of the network needs to be specified to some extent[‡]. Also, the range of parameter values that are consistent with the model can vary over orders of magnitude (von Dassow et al., 2000), which together with the lack of quantitative data, suggests a more abstract approach, that of logical, or qualitative models.

1.4.2 *Qualitative models*

Rather than being based on purely pragmatic reasons such as the lack of quantitative data, according to Bolouri (2008), qualitative models were originally inspired by the work of Jacob and Monod who emphasised the ON/OFF nature of gene regulation. In these models, variables (genes or other actors) are given a small number of discrete states. Different formalisms of qualitative models then provide different ways of specifying the rules for how the variables change state over time.

The work of Sánchez and Thieffry (2001, 2003) provides examples of *Drosophila* A-P patterning models in the formalism of Thomas (1973). Here, the input variables are discretised (assigned to a few levels) based on thresholds, which are then combined linearly (a weighted sum) to give the resulting output (e.g., rate of transcription). Since the inputs can only have a few different levels, the output is correspondingly lim-

[‡] Although certain parameters with a value near zero can suggest the removal of a network connection.

ited in the number of values it can take, making the analysis of large networks and multiple alternative topologies tractable. Albert and Othmer (2003) is an example of another formalism. Again, each input is assigned a discrete set of values, two in this case. The output is calculated by a Boolean (logical) function of the inputs, with each update occurring synchronously across the whole network.

These models, and many others, have provided intriguing insights into the nature of robustness of different networks (i.e., the ability to tolerate different parameter values or initial conditions). However, they are strictly qualitative and cannot be used to explain variations in data which are not amenable to discretisation. Additionally, and rather importantly, the models are used after thresholding. In other words, they presume that positional information is implemented by thresholding each regulator separately. The proposition of this thesis is that this is a flawed assumption.

1.4.3 *Steady state models*

A further class of models avoids both discretisation of the input variables (i.e., the imposition of thresholds) and computationally time-consuming simulations. This is done by assuming steady state. This is explained more fully in appendix A, but the essential aspect is that the parameters of the model can be fitted to data directly since relative levels of steady state measurements reveal the underlying relative transcriptional rates.

These types of models have been particularly useful in relating occupancy of DNA binding sites by transcription factors and the consequent effect on transcription rates. Janssens et al. (2006) and Segal et al. (2008) are good examples for patterning along the anteroposterior axis of *Drosophila*. Zinzen et al. (2006) is a further example with enhancers regulated by a gradient along the dorsoventral axis. A similar approach is taken by Papatsenko and Levine (2008) where dual regulation by Hunchback (HB) is explored.

Given that developmental systems are by nature not stable, it is surprising that these models have been relatively successful (Bolouri and Davidson, 2003). Therefore, these methods have often been used for unicellular organisms, such as yeast and bacteria; and it is in this context that thermodynamic models based on the relative energies of different transcription factor binding configurations have been developed, for example (e.g., Ackers et al., 1982; Hill, 1985; Bintu et al., 2005). It is thus of interest to assess to what extent a multinuclear dynamic system such as the *Drosophila* blastoderm supports the steady state assumption. This is explored more fully in chapter 2.

1.4.4 *The approach of this thesis*

The approach of this thesis also avoids discretisation of the input variables, but like the logical models, discretises the response (into ON or OFF). As a result, it is able to handle uncertainties in the data and yet retain sufficient information to make quantitative predictions. Thus far, most success in quantitative modelling of *eve* has been for stripe 2 and 7 (e.g. Janssens et al., 2006). Papatsenko and Levine (2008) have reproduced stripe 3, but not stripe 7 since the model lacks a repressor for the posterior border of stripe 7. Segal et al. (2008) have reasonable success for the gap genes, but their predictions for the pair rule genes (including *eve*) are not very reliable (see Levine, 2008).

The work presented in this thesis begins by modelling stripe 2 successfully and continues from there to develop a framework within which alternative models and regulators can be evaluated. Thus, the primary regulators of the *eve* 3+7, *eve* 4+6 and *eve* 2 enhancers are recovered (chapter 5). Unlike the models above, which are along only one axis of the embryo (sometimes truncated), this work can reproduce both wild-type and mutant gene expression patterns along the full length of both axes of the embryo (for *eve* 3+7 and *eve* 2 in chapter 6). Further, it enables the importance of dual regulation by some of *eve*'s regulators to be evaluated (chapters 4 and 6).

The analysis depends critically, though, on the data set underlying the models. The next section introduces the *Virtual Embryo*.

1.5 THE BDTNP AND THE VIRTUAL EMBRYO

The Berkeley Drosophila Transcription Network Project (BDTNP) is a multidisciplinary team at Lawrence Berkeley National Laboratory, the University of California, Berkeley, and the University of California, Davis. According to their website (BDTNP, 2007), their goal is

> to decipher the transcriptional information contained in the extensive cis-acting DNA sequences that direct the patterns of gene expression that underlie animal development. Using the early embryo of the fruitfly *Drosophila melanogaster* as a model, we are developing experimental and computational methods to systematically characterize and dissect the complex expression patterns and regulatory interactions already present prior to gastrulation.

In April 2008 the BDTNP released the *Virtual Embryo* data set (Fowlkes et al., 2008) which is the basis for the analysis presented in this thesis. It is built from in situ gene expression measurements in 2180 embryos for 95 genes at six time cohorts during developmental stage 5, the stage leading up to gastrulation (Campos-Ortega and Hartenstein, 1997). This section describes the key features of this data set.

1.5.1 *Data acquisition*

As name implies, the *Virtual Embryo* does not contain the expression level of any particular embryo, but rather the combined data of many embryos. The data acquisition process is described in Luengo Hendriks et al. (2006). In essence, embryos were demembranated and fixed, prepared for imaging, and fluorescently stained for the mRNA of two different genes as well as the location of nuclear DNA. Imaging was done with a two-photon scanning microscope serially for each of the three fluorescent channels. Then, using various image analysis methods, these were converted into quantitative measurements for each nucleus of the expression levels for the two genes. Additionally, the 3D coordinates of every nucleus were recorded to generate a *PointCloud*. It was these *PointClouds* that were used in creating the *Virtual Embryo*.

Of the 95 genes imaged, 23 are early acting transcription factors with at least 25 embryos imaged each. The remaining 72 genes are known or putative targets of these early transcription factors, and were primarily imaged in the period leading up to gastrulation. The imaging of mRNA was always done in pairs with one of the pair being *eve* or *ftz*. The data set also contains protein levels from 215 embryos for four proteins (GT, KR, BCD and HB).

1.5.2 *Assignment of embryos into temporal cohorts*

During stage 5 the nuclei do not divide and the number are thought to remain stable (Keränen et al., 2006). Most have migrated to the embryo surface (with the exception of the yolk nuclei which are not included in the *Virtual Embryo*). At the beginning of this stage, the nuclei are not separated by cell membranes but over the next 50 minutes or so, the membrane extends from the surface towards the centre of the embryo eventually separating them. This provides the basis for assigning the embryos to different temporal cohorts. Since the rate of ingrowth differs between the dorsal and ventral sides, the six cohorts were defined depending on how far the membrane on the ventral side had invaginated: 0-3%, 4-8%, 9-25%, 26-50%, 51-75% and 76-100%. This corresponds to times after egg laying spanning approximately 70 to 120 minutes with about 10 minutes between each cohort.

1.5.3 *Scaling of fluorescence*

Although it is plausible that fluorescence is proportional to actual mRNA count, the actual scale can vary between embryos owing to variations in labelling efficiency, microscope performance and other experimental artefacts (Luengo Hendriks et al., 2006). The BDTNP project fixed and hybridised embryos in batches, which were mixed in developmental stage. Fluorescence varied across batch, embryo and nuclei, but since the goal of the project was to capture relative variation between nuclei within an embryo and across time, the variation across batch and embryo was reduced before averaging to create the *Virtual Embryo*. The process described here was done per probe.

The first step was to find the relative change in intensity of the 99th percentile expression level across temporal cohorts, with the maximum of these scaled to 1. Using embryos from different developmental time points that were within a single batch, various partial time series of the 99th percentile could be estimated by averaging the relevant embryos (the coloured lines in figure 1.6). Different batches were scaled by minimising the squared error relative to the mean, and finally, using Gaussian process regression (Rasmussen and Williams, 2006), a smooth curve was fitted to the whole time course. The smoothed curve then, provided the maximum for each temporal cohort used to scale the relative fluorescence levels of each embryo.

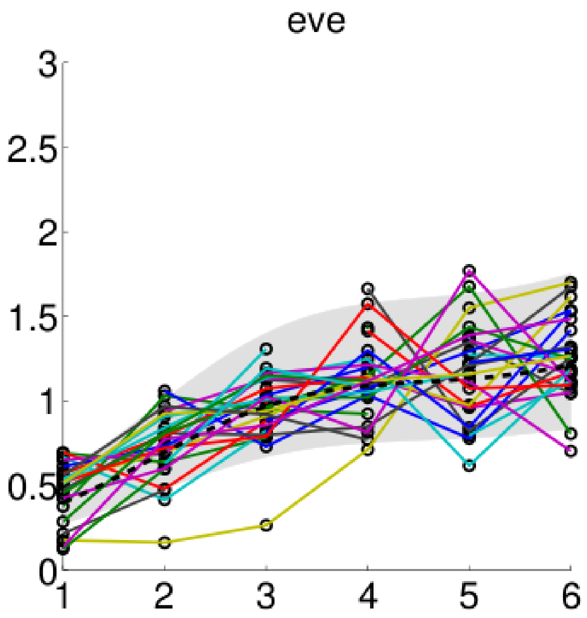


Figure 1.6: Gaussian Process Regression applied to *eve* levels in different batches and embryos. Individual coloured curves show estimates of the 99th percentile expression level using embryos from a given hybridisation batch. A separate gain parameter is estimated in order to align batches and remove the effects of variable reaction efficiency or choice of fluorophore. The black dashed curve shows a fit to log expression levels given by Gaussian process regression using a squared exponential covariance function with characteristic length of 3 (in the graph units, roughly 30 minutes) and independent noise model with standard deviation of 0.3. Figure from Fowlkes et al. (2008), used with permission from Elsevier.

After this, a separate offset and gain was chosen for each embryo to minimise the variability within the cohort, but still matching the maximum for that cohort. Implicit in this process is that biological variation can also be scaled in this manner to produce meaningful relative measurements. This then, provided the basis for the values that were averaged for all corresponding nuclei to produce the final value in the *Virtual Embryo*. Calculating the correspondence between nuclei is known as *registration*. There were two parts to this: temporal and spatial.

1.5.4 *Temporal registration*

The beginnings of the *Virtual Embryo* is a model of the positions of 6078 'virtual nuclei' for each of the 6 time points corresponding to the 6 temporal cohorts of the *Virtual Embryo*. As described in Keränen et al. (2006) the density and positions of the nuclei change over time, although the number does not. They observed that some nuclei travelled as far as three cell diameters during stage 5. The basis for the positions of the 'virtual nuclei' is a model of nuclear flow. Using the average embryo shape and nuclear density pattern from each cohort, a model was fitted to ensure that the nuclei would match the density at these time points as closely as possible with as little movement over time as possible. The resulting model then provides the necessary temporal correspondence between nuclei and specifies the nuclear positions in the 'morphological template'. A schematic of the process is shown in figure 1.7. Further details are described in Keränen et al. (2006) and Fowlkes and Malik (2006).

1.5.5 *Spatial registration*

The second component is the correspondence between the nuclei of each *PointCloud* and that of the morphological template—this is *spatial registration*. This consisted of two steps. The first was called *course registration* and involved scaling the *PointCloud* to match the average egg length for its cohort. The second step, *fine registration*, was iterative: the nuclei in each *PointCloud* were assigned to the nearest nucleus in the morphological template, enabling the average expression of the marker genes, *eve* and *ftz* to be calculated for each 'virtual nucleus' . Using the boundaries of these genes, each *PointCloud* was then warped onto the morphological template giving a more precise match to the borders of these stripes. A new average was then be computed and the

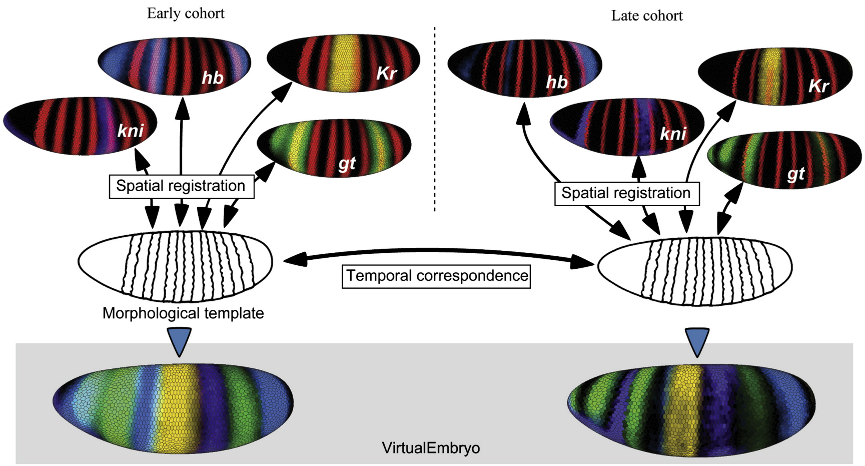


Figure 1.7: Schematic of the process used to create the *Virtual Embryo*. On the top panel, each individual embryo is stained for nuclei, a common marker gene (red) and a gene of interest (second colour). In the centre panel, within each temporal cohort, the marker gene is used to guide spatial registration on to a morphological template; temporal correspondences between cohorts are provided by a model of typical nuclear movements. On the bottom panel, once correspondences across embryos have been established, expression measurements are averaged and composited to create a model *Virtual Embryo* in which the expression of many genes can be analysed. Figure from Fowlkes et al. (2008), used with permission from Elsevier.

step repeated until there was no significant change in correspondence between the nuclei, thus ending the fine registration process. Finally, all genes were averaged, and together with the dynamic morphological template, this made the *Virtual Embryo*.

It should be pointed out that the mRNA probes were co-stained with *eve* or *ftz*, but the protein stains were done either individually, or with another protein. This means that the protein measurements could not be finely registered. Their location is only adjusted with the course scaling of embryo size (Zeba Wunderlich, personal communication).

1.5.6 *Regulator discovery*

As part of the *Virtual Embryo* paper (Fowlkes et al., 2008), regulatory relationships were inferred from the data. 17 regulators were considered

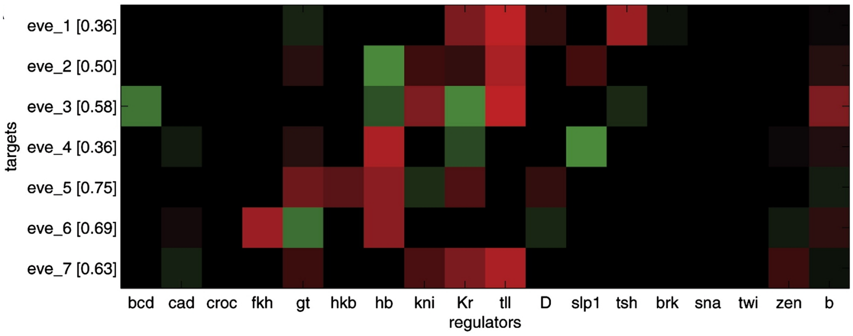


Figure 1.8: Regulator discovery figure from Fowlkes et al. (2008), used with permission from Elsevier. Coefficients for each of 17 regulators (columns) determined by fitting each target *eve* stripe (rows). Green indicates activation, red indicates repression, black indicates no interaction. The right-most column indicates the constant offset, b. Quality of fit (R^2) values are shown in brackets. Figure from Fowlkes et al. (2008), used with permission from Elsevier.

for each target gene (using protein measurements when available). Of these the best 6 were selected to fit a model. This was done for each stripe of a target gene. Figure 1.8 shows the results for the *eve* gene. Given that *Kr*, *gt*, *hb* and *bcd* are the main regulating genes for stripe 2 (Small et al., 1992; Andrioli et al., 2002), one of the best characterised stripes, their success was limited, especially compared with the results of section 5.4.

1.5.7 *Candidate genes and their abbreviations*

There are 95 genes available in the *Virtual Embryo*, but only a subset were used for the analysis of this thesis. One of the difficulties of searching for regulators that can describe the position of stripes in the embryo is that other pair rule genes can provide excellent information. For example, *ftz* is expressed in a complementary pattern between the stripes of *eve* (Frasch and Levine, 1987) and so can be used in a model as a repressor of *eve*, regardless of whether it does or does not regulate *eve*. Some of these relationships might reflect biological reality, but

eve	*even skipped*
ftz	*fushi tarazu*
h	*hairy*
odd	*odd-skipped*
opa	*odd-paired*
prd	*paired*
run	*runt*

Table 1.1: The symbols and names of pair rule genes excluded as potential regulators.

they do obscure the question as to whether the products of other genes (including the gap genes) provide sufficient spatial information for *eve*. Therefore, the pair rule genes with the clearest stripe patterns were excluded as candidates. This list is shown in table 1.1.

Since the measurements underlying the *Virtual Embryo* were not taken for all time points for all probes, the final set of candidate genes were those that had measurements for the appropriate temporal cohort, namely cohort 3 (as justified in chapter 2). The full set of candidate genes are given in table 1.2. It is worth reiterating that protein measurements were used when available, namely for GT, KR, BCD and HB.

1.5.8 *Notation*

The availability of protein measurements makes it important to identify these in any model. Thus, in this thesis, the following convention will be used. When the gene symbol refers to an mRNA measurement or to the gene more generally, italics will be used, for example, *eve* or *even skipped*. For protein measurement, small caps will be used for the abbreviated name, such as GT. The full name for the protein product will contain an initial capital, for example, Giant.

Maternal	
bcd	*bicoid*
cad	*caudal*
Gap	
gt	*giant*
hb	*hunchback*
Kr	*Krüppel*
kni	*knirps*
tll	*tailless*
croc	*crocodile*
hkb	*huckebein*
knrl	*knirps-like*
fkh	*forkhead*
cnc	*cap-n-collar*
oc	*ocelliless*
Pair rule or pair rule associated.	
slp1	*sloppy paired 1*
slp2	*sloppy paired 2*
D	*Dichaete*
sob	*sister of odd and bowl*
tsh	*teashirt*

Dorsoventral	
brk	*brinker*
sna	*snail*
twi	*twist*
zen	*zerknüllt*
Other	
Dfd	*Deformed*
Doc2	*Dorsocross2*
Traf1	*TNF-receptor-associated factor 1*
bun	*bunched*
emc	*extra macrochaetae*
fj	*four-jointed*
path	*pathetic*
rho	*rhomboid*
sala	*spalt-adjacent*
srp	*serpent*
trn	*tartan*
term	*terminus*
Cyp310a1	*Cyp310a1*
CG4702	
CG10924	
CG17786	

Table 1.2: The symbols and names of the 38 candidate regulators from the *Virtual Embryo*. Categorised following Brody (2010) and BDTNP (2007).

1.6 THESIS OUTLINE

Positional information within a developing organism can be described in different ways. In this thesis the focus is how transcriptional activity varies according to position within the embryo and how this is related to varying transcription factor concentrations. The first question to be considered is whether a relatively simple function can describe how an enhancer might use transcription factors concentrations to determine the correct transcriptional response. Secondly, if a simple function can be found, does this enable positional information to be studied abstractly and generally in the spirit of the French flag problem?

In order to model the enhancer function suitably it is necessary to settle on what the inputs to the function will be, and what output the function will produce. This is examined in chapter 2. Chapters 3 and 4 are then concerned with the form of the function. Chapters 5 and 6 use this functional form to find the primary regulators of the *eve* stripes and to demonstrate that it can describe the enhancers' behaviour under different experimental conditions. Finally, chapter 7 builds on the earlier results to introduce a model of positional information, which is used to analyse positional information in the *Drosophila* embryo at the level of the gap genes and the *eve* stripes.

2

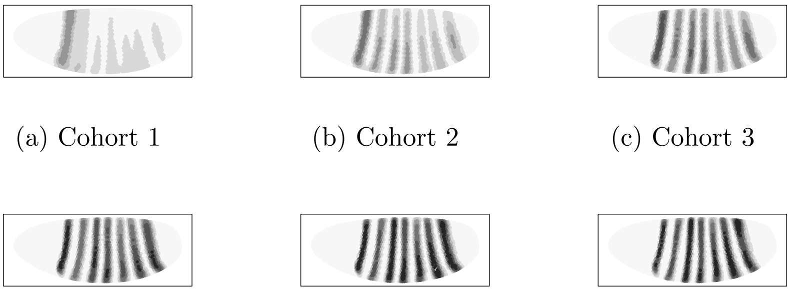
INITIAL ANALYSIS AND DATA PREPARATION

This chapter is concerned with the inputs and outputs that will be used subsequently to study positional information and *eve*'s transcriptional response in the *Virtual Embryo*. The *Virtual Embryo* provides a rich source of quantitative, continuous measurements at six time points for *eve* as well as many other factors, and yet, the models of this thesis will be less expressive: each nucleus of the embryo will be classified according to whether *eve* is ON or OFF—a binary classification model. Further, when transcription factor protein measurements are not available, mRNA measurements will be substituted, not from earlier in the data set as might be expected, but from the same time point as the *eve* response being considered.

This chapter justifies these decisions. It also serves as a detailed introduction to the *Virtual Embryo* data and its characteristics, which will be important in the analysis of subsequent *eve* models.

2.1 CHANGING PATTERNS IN THE VIRTUAL EMBRYO

Before beginning the analysis of the transcription of *eve* in the *Virtual Embryo*, it is worth remembering that this occurs in a dynamic context: the expression levels of *eve* and its regulators are changing over time. The expression levels of *eve* are plotted in figure 2.1. From this it can be seen that, to begin with, *eve* is more strongly expressed around stripe 1. Over time, *eve* expression increases and becomes spatially restricted resulting in clear stripes.



(a) Cohort 1

(b) Cohort 2

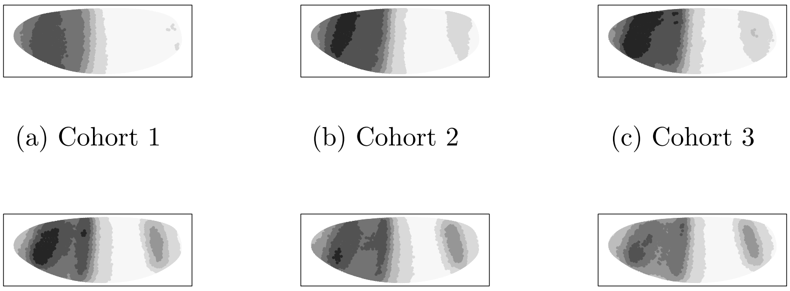
(c) Cohort 3

(d) Cohort 4

(e) Cohort 5

(f) Cohort 6

Figure 2.1: *eve* expression in the *Virtual Embryo* across six time points. Darker shades signify higher expression.



(a) Cohort 1

(b) Cohort 2

(c) Cohort 3

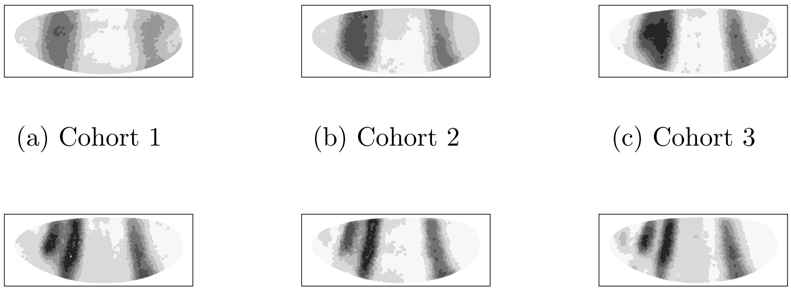
(d) Cohort 4

(e) Cohort 5

(f) Cohort 6

Figure 2.2: hb expression in the *Virtual Embryo* across six time points. Darker shades signify higher expression.

There are a large number of potential regulators in the data set, but two will be plotted here for illustration: the protein measurements for *hunchback* (*hb*) and *giant* (*gt*). Clearly, they too are changing over time (figures 2.2 and 2.3).



(a) Cohort 1

(b) Cohort 2

(c) Cohort 3

(d) Cohort 4

(e) Cohort 5

(f) Cohort 6

Figure 2.3: GT expression in the *Virtual Embryo* across six time points. Darker shades signify higher expression.

2.2 THE BASIC DIFFERENTIAL EQUATION

The justification for a classification model will be approached by starting with a more general description of the relationship between the rates of transcription and the concentrations of transcription factors. Differential equations are the standard way of specifying this (Alon, 2006; Bolouri, 2008). As is usual in modelling, these rely on a number of simplifying assumptions, the most basic of which is that there are enough molecules (of mRNA and protein) to approximate their numbers with continuous variables. This is a reasonable assumption for the *Virtual Embryo*: there are so many different measured values (scaled to lie between zero and one) that no discrete levels are detectable.

Additional assumptions might differ between different differential equation models and their validity can be difficult to determine a priori. Thus, it is reasonable to begin with the simplest assumptions and adjust these when the model is unable to explain the data.

Over the time period considered, the following are assumed:

- The changes in volume of the embryo or nuclei are negligible. Thus, relative concentration and relative amount are equivalent.

- The overall *eve* mRNA degradation rate is proportional to the amount of *eve* mRNA.
- The proportion of each transcription factor in the active form is constant per transcription factor. So, if the measured concentration of one factor is twice that in another nucleus, then the concentration of the active form is also double.
- Diffusion is not significant.

The relevant differential equation for a particular nucleus is then:

$$\frac{dy}{dt} = f(\mathbf{x}) - \beta y. \tag{2.1}$$

The left-hand side refers to the rate of change of y with respect to time t, where y is the amount of *eve* mRNA. $\mathbf{x}$ is the vector of concentrations of transcription factors, $f(\mathbf{x})$ is the transcription rate of *eve*, which is dependent on the transcription factor concentrations, and β is the rate of decay per mRNA transcript, the decay constant.

The goal of this thesis is to find a simple function f per enhancer. However, the concentrations of transcription factors $\mathbf{x}$ are changing over the relevant time period—figures 2.2 and 2.3 and Jaeger et al. (2004). This suggests that the transcription rate $f(\mathbf{x})$ is also changing, which provides one justification for a classification model. By dispensing with a fully dynamic model and considering $f(\mathbf{x})$ to be either ON or OFF, variability in transcription rate becomes less important. It can then be assumed that the values of $\mathbf{x}$ that lead to a high transcription rate will be similar to the values that lead to a medium rate (both ON), but quite different from the values that result in no transcription (OFF). A classification model has the further benefit of avoiding simulation and computational complexity so that different forms of f can be tested quickly and directly.

However, there are still some issues to be resolved. Firstly, how should mRNA measurements be used when protein measurements are not available? And secondly, how should the response of *eve* be classified? Should it be based on the change of *eve* mRNA from one time point to another, or would a simple threshold on the current level of *eve* suffice? Again, the decisions made here will depend on the characteristics of the data, including the reliability of temporal registration. As will be demonstrated, it is useful to test whether $f(\mathbf{x})$ is indeed constant or not. Doing so will reveal a potential flaw in the *Virtual Embryo*'s temporal registration process and raise an intriguing question regarding how *Drosophila* might make use of positional information.

2.3 CONSTANT TRANSCRIPTION RATE

If *eve*'s transcription in a particular nucleus is constant, its concentration over time should follow one of two forms depending on whether it is ON or OFF.

2.3.1 *Transcription is off*

If β is constant and there is no transcription, the differential equation 2.1 has a solution of:

$$y = y_0 e^{-\beta t} \tag{2.2}$$

where y_0 is the concentration of y at the start of the time period ($t = 0$). In this case, nuclei with no transcription should follow a curve similar to that in figure 2.4. Examining the profiles of nuclei with decreasing transcription in the *Virtual Embryo* will reveal to what extent this is correct.

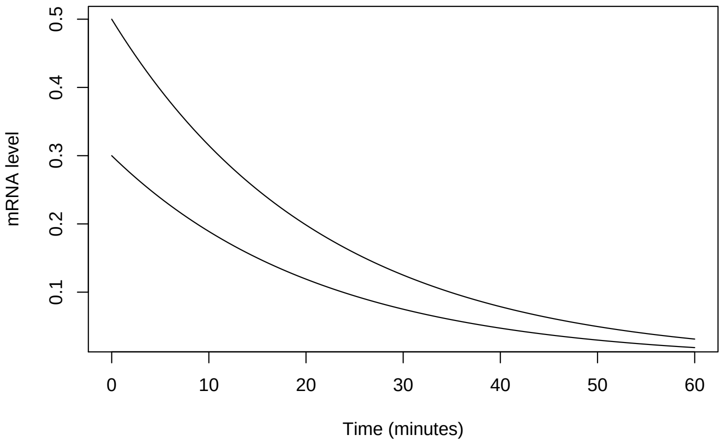


Figure 2.4: Illustration of exponential decay with a half-life of 15 minutes, and two different starting values.

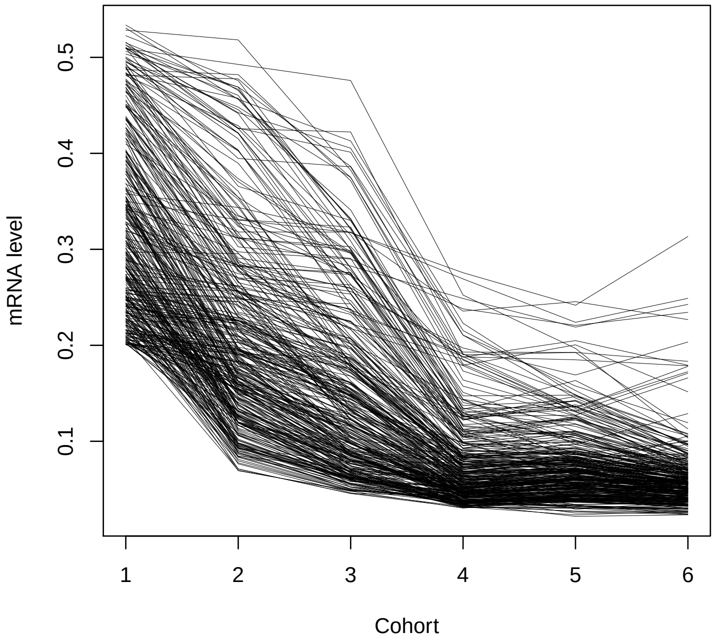
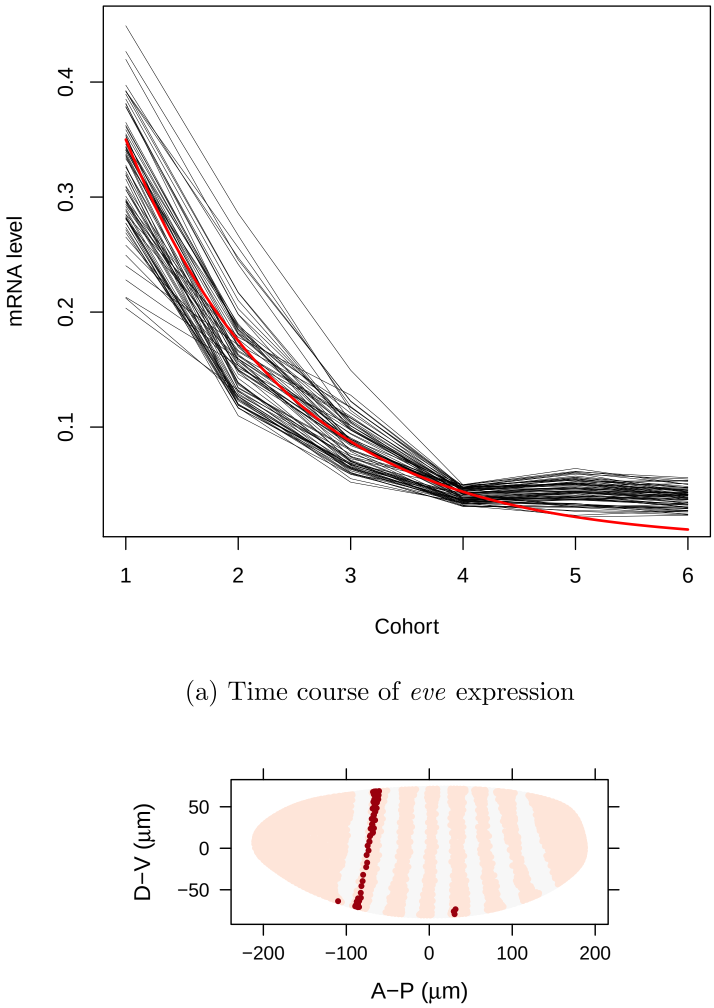


Figure 2.5: Time course of *eve* mRNA levels in all nuclei of the *Virtual Embryo* where the level of *eve* is decreasing over the whole period from cohort 1 to 4.

Figure 2.5 shows all 340 nuclei in the *Virtual Embryo* that are decreasing over the whole period from time points 1 to 4. As can be seen, many nuclei do not follow simple exponential decay. This is especially evident from some nuclei that are unchanged (horizontal lines) between time points 2 and 3. The horizontal lines are difficult to explain biologically, thus suggesting an experimental or data processing artefact. One possibility is that the time periods between cohorts are not constant since these are determined by the extent of cellularisation (see section 1.5.2). It could be that the time period between 2 and 3 is much shorter than for the other periods. However, as the horizontal lines affect a minority of nuclei, this is unlikely to be the cause. Another possibility is that it is the product of incorrect temporal or spatial registration, which will be considered in more detail later in this chapter.

If those nuclei where *eve* does not change substantially[*] between cohorts 1 and 2 or cohorts 2 and 3 are excluded, and the time course of the remaining nuclei plotted (figure 2.6a), it can be seen that these appear to follow simple exponential decay, with a half-life of about 10 minutes (the red curve). Interestingly, the nuclei of the subset are found mostly in a narrow strip posterior to stripe 1 (figure 2.6b). Comparing this with figure 2.1 shows that it corresponds to the refinement of stripe 1. This might be the only part of the embryo where simple exponential decay is present, in which case the relative levels of mRNA across cohorts might be reasonable (at least up to cohort 4). Other more complicated patterns in figure 2.5 most likely result instead from sharpening of the borders of the *eve* stripes and potential shifts in their location (Keränen et al., 2006).

[*] Where the change is less than 0.05.



(a) Time course of *eve* expression

(b) Location of the nuclei

Figure 2.6: A subset of nuclei from figure 2.5. The red curve in (a) shows exponential decay, with a half-life of 10 minutes. The location of the nuclei are shown in red in (b) where the light grey regions indicate the positions of the *eve* stripes.

2.3.2 *Transcription is on*

When transcription is constant (i.e., $f(\mathbf{x}) = \alpha$), the differential equation 2.1 has a solution of

$$y = \frac{\alpha}{\beta} - e^{-\beta t}(\frac{\alpha}{\beta} - y_0), \tag{2.3}$$

where y_0 is the initial level of y at time $t = 0$. As before, β is the decay constant. At steady state, decay is balanced by transcription, that is $0 = \alpha - \beta y$ or $y = \alpha/\beta$. Thus, the equation shows that the difference between the steady state, α/β, and the current level y decays exponentially. When this is the case, the time course of *eve* should look similar to the curves plotted in figure 2.7.

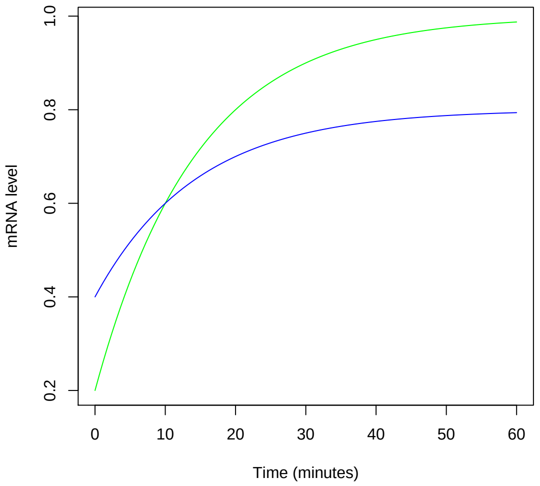
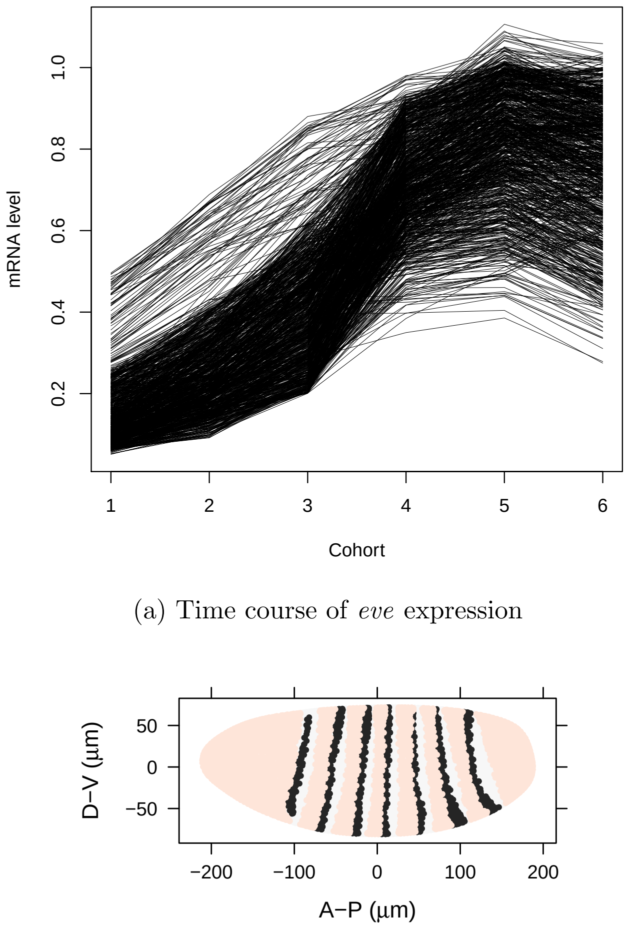


Figure 2.7: Changing mRNA levels with a constant transcription rate and constant decay (half-life of 10 minutes). Blue has a transcription rate 80% that of green. The absolute concentration is scaled to a maximum of 1.

Figure 2.8 shows the time series of *eve* expression and location of all nuclei whose expression increases at each time point from 2 to 5 (900 out of 6,078 nuclei). The time series does not follow the standard curve



(a) Time course of *eve* expression

(b) Location of the nuclei

Figure 2.8: mRNA levels in nuclei for which *eve* is always increasing from cohorts 2 to 5. The location of the nuclei in (b) are indicated in black, within the *eve* stripes, the light grey regions.

of equation 2.3 and figure 2.7, but instead looks sigmoidal from time points 1 to 5, with expression decreasing after that. This sigmoidal curve indicates that constant transcription is not valid. However, it is intriguing that these nuclei are all on the anterior borders of the *eve* stripes. As the next two sections will show, this is most likely an artefact of the nuclear flow model underlying temporal registration.

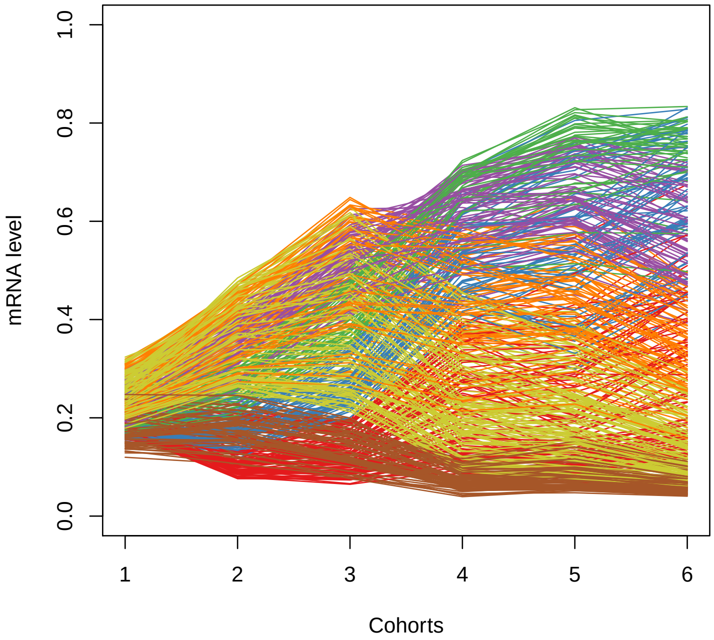


Figure 2.9: Time series of *eve* expression in the nuclei of *eve* stripe 2. The nuclei have the same colours as in figure 2.11.

2.4 TIME COURSE CHARACTERISATION

In order to characterise the dynamics, or time course, of *eve*, it is useful to consider the nuclei of a single stripe. The *eve* 2 enhancer has been well studied, and *eve* stripe 2 will form the basis for the model of the next chapter.

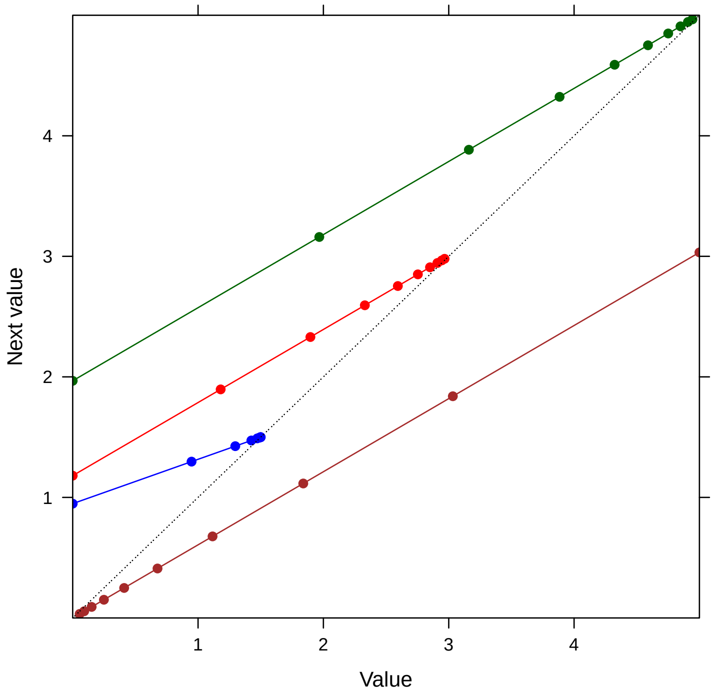
It is thus fitting to focus on the dynamic behaviour of this stripe. As can be seen from figure 2.9, the 508 nuclei of *eve* stripe 2[†] do not follow a single time course. In particular there is no clear tendency towards steady state. However, the time course does appear to have some organisation or pattern. The goal of this section is to characterise this in more detail.

[†] The stripes are defined as where expression is above 0.2 for cohort 3 plus bordering nuclei.

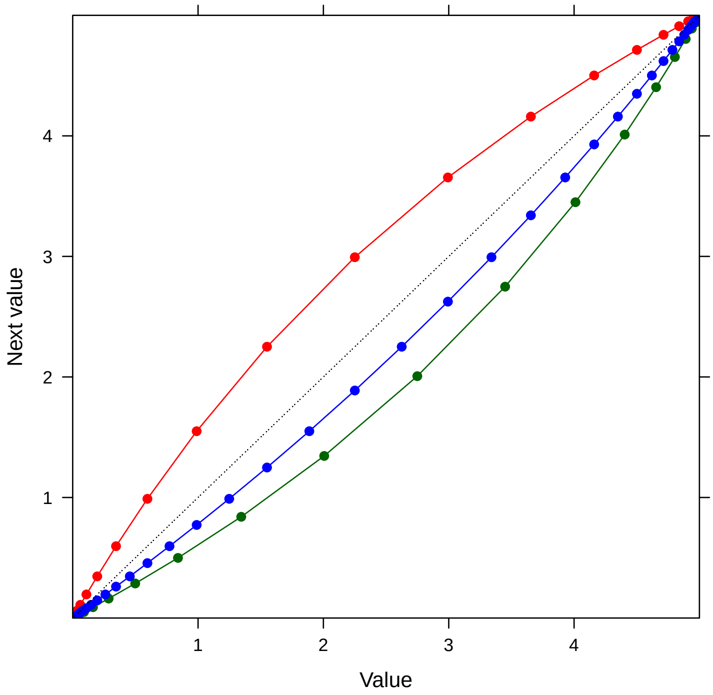
This will be done by comparing each expression level with the next in the time series. By definition, when a time series is following equation 2.2 (exponential decay) or equation 2.3 (towards steady state), the plot will show the second value is proportional to the first. Figure 2.10 illustrates this. In figure 2.10a the red, blue and green curves correspond to hypothetical, increasing exponential functions tending towards steady state. The brown curve is exponential decay. Blue has a higher decay constant than the other three, and green has a higher transcription rate than red. This can be discerned from the figure as the parameters affect the slope of the line and where it reaches steady state (the dashed line). Figure 2.10b shows the same type of plot for a sigmoid function—to be specific, a logistic function. Red is an increasing logistic function, green has parameters of exactly the same magnitude except it is decreasing, while blue is a more slowly decreasing logistic function.

Applying this approach to *eve* stripe 2 is quite revealing. The results are plotted in figure 2.11a for each cohort. Firstly, it is evident that the nuclei cover a full range of values within each cohort, and this helps reveal the pattern. As expected the pattern is not exponential, but rather from a logistic function. This is most obvious in cohort 3: the underlying pattern appears to be one of logistic growth and logistic decrease.

Further details are revealed in figure 2.11 by grouping and colouring the nuclei, and then plotting their physical location within stripe 2 (figure 2.11b). It is remarkable that they arrange themselves in narrow strips along the anteroposterior axis parallel to the border of the stripe. This shows clearly that the nuclei's expression dynamics are correlated with their location in the stripe. It can also be seen that the transition along the pattern (in figure 2.11) from one nuclei to the next is smooth and that the strips next to each other are more similar than strips further apart. There are two possible interpretations.

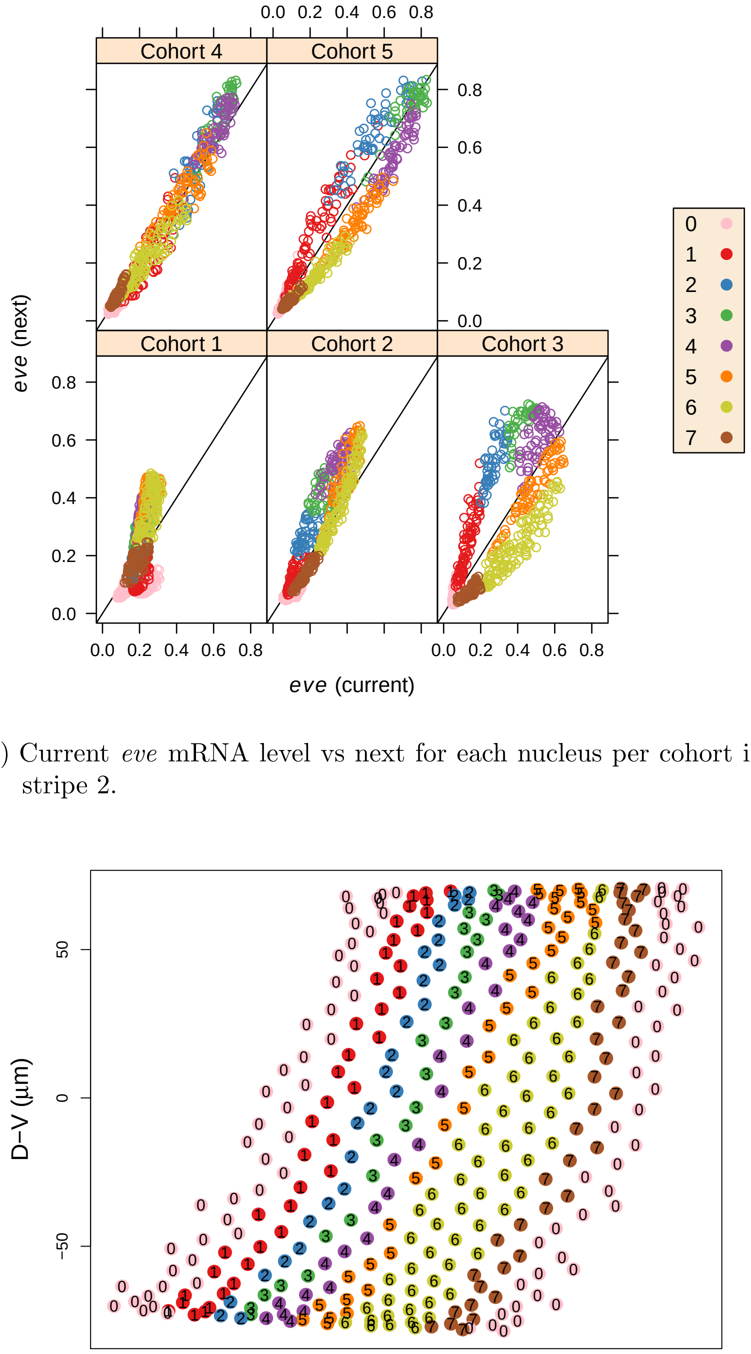


(a) Exponential functions



(b) Logistic functions

Figure 2.10: All values against their next value for a time series following exponential and logistic functions. The dashed line shows equal values. See text for details.



(a) Current *eve* mRNA level vs next for each nucleus per cohort in stripe 2.

(b) Position of nuclei in stripe 2, numbered and coloured with the same groups used in (a).

Figure 2.11: Exploring expression dynamics of nuclei by position within *eve* stripe 2.

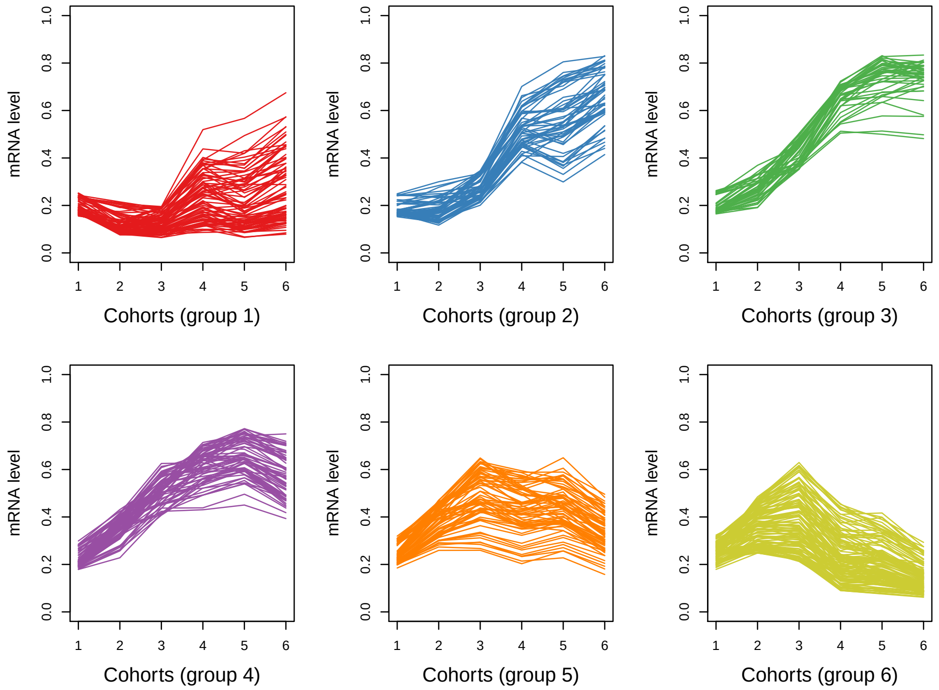


Figure 2.12: Time course per group of nuclei in *eve* stripe 2 (excluding the bordering nuclei); coloured as in figure 2.11.

The first explanation is that each nucleus is following the same course in time, but the point at which it started along this trajectory depends on its location in the stripe. To see this, consider the time courses for each group plotted in figure 2.12. It is possible to view each subfigure as a sliding window on an overall curve of logistic growth followed by some sort of decay—perhaps logistic if the behaviour of the later cohorts is interpreted loosely. This can also be viewed by observing that there is an apparent progression over time of the coloured groups around the 'oval' in figure 2.11a. It is difficult to see what process could give rise to this exquisite control.

A second, simpler, explanation is that there are two processes underlying the pattern, each one operating uniformly and smoothly: the first process varying along the A-P axis and the second related to mRNA level. As will be seen in the next section, if the former corresponds to an artefact of the nuclear flow model, then the latter can be explained by simple transcriptional dynamics.

2.5 VIRTUAL EMBRYO TEMPORAL CORRESPONDENCES

The approach used to link measurements of nuclei between temporal cohorts in the *Virtual Embryo* is relatively indirect compared to the approach of linking nuclei spatially (spatial registration). It depends on two models: Gaussian process regression (Rasmussen and Williams, 2006), a type of curve fitting, and the nuclear flow model. Either of these might have introduced artefacts into the dynamics of *eve*, and it is thus worth considering this issue in more detail.

2.5.1 *Gaussian Process Regression*

The Gaussian process regression step (see section 1.5.3) was designed to find a smooth fitting curve to describe the progression of the 99th percentile expression level of the relevant gene across each cohort. As can be seen from figure 1.6 on page 17, the 99th percentile is expected to increase across each cohort, but by a reducing amount over time. This curve is used to scale each individual embryo accordingly before being averaged to produce the measurements for that gene in the *Virtual Embryo*. It is thus informative to examine the 99th percentile of expression levels in the final averaged embryo to see if it matches the smoothed curve from Gaussian process regression.

Figure 2.13 shows that the 99th percentile does increase over time, although more linearly than expected with a drop in the final cohort. This suggests that the scaling of fluorescence across time is potentially plausible at least for the *eve* gene which has many embryos underlying its average, but that this scaling is by no means certain, particularly for the last cohort. Further, other transcription factor concentrations might not be as reliable. Nevertheless, it will be shown that the other model underlying the temporal correspondences is far more likely to have introduced difficult to understand dynamics into the data set.

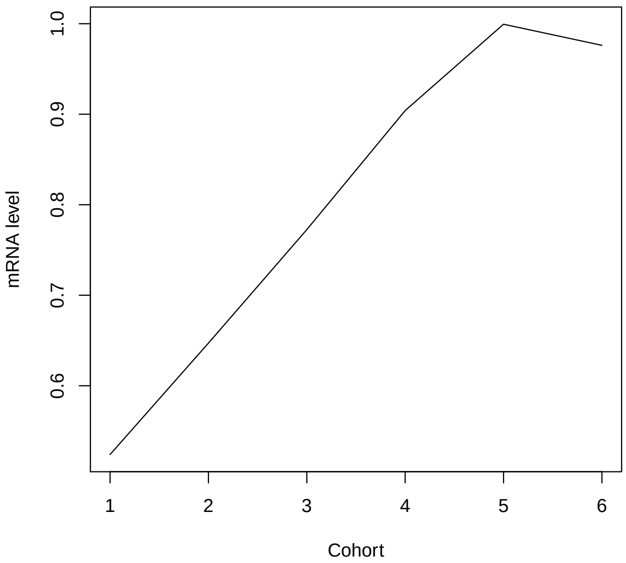


Figure 2.13: 99th percentile expression level for each cohort of the *Virtual Embryo*.

2.5.2 *Nuclear flow model*

The *Virtual Embryo* is constructed from fixed embryos and so there is no direct means of assessing how nuclei relate to each other over time. The correspondence method used depends on a model which interpolates the positions of the nuclei based on fixed material so that that the resulting nuclear density matches that of the embryos (see section 1.5.4 and references within). Using this, an earlier analysis of the dynamics of the PointClouds (Keränen et al., 2006), proposed that the spatial movement of a stripe over time (pattern flow) could be decomposed into expression flow and nuclear flow. Expression flow results from regulatory interactions and changes in transcription, whereas nuclear flow independently contributes to stripe movement. Since the *Virtual Embryo* calculates a correspondence for each nucleus across each cohort, nuclear flow movement should not be present in the *Virtual Embryo* if expression is plotted according to nucleus. Considering this, the movement of the *eve* stripe 2 pattern over time can thus be examined. Figure

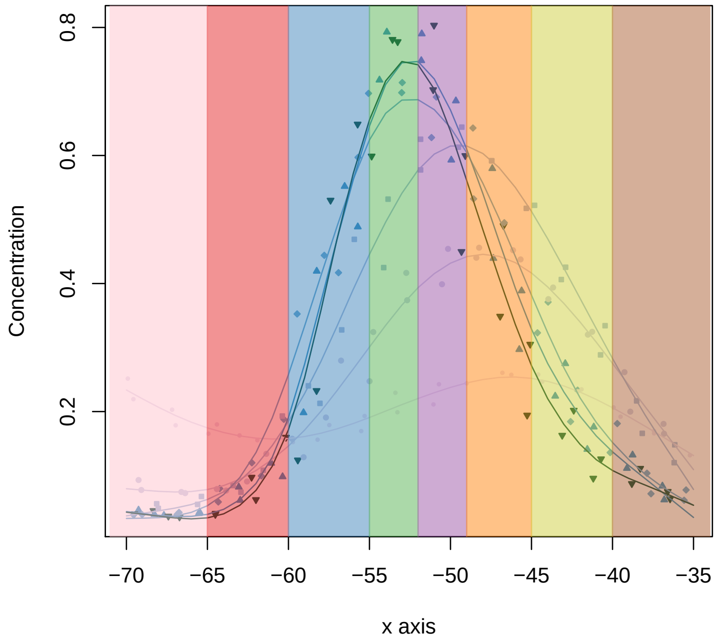


Figure 2.14: Expression level of *eve* stripe 2 nuclei within a narrow lateral strip, with smooth curves fitted for each cohort. The background colours correspond to the groups in figure 2.15. The curves are coloured with increasing darkness according to cohort.

2.14 shows the expression level of individual nuclei in the region of the second *eve* stripe, but restricted to a narrow lateral stripe on one side of the embryo. Using penalised regression splines (with the R function `gam` in the package *mgcv*), a smooth curve is fitted for each cohort. This makes it easy to visualise expression flow over time: the stripe is increasing in expression, becoming more refined, but also moving to the anterior of the embryo. Further, and importantly, the smoothed data maintains the same overall dynamics as before (compare figure 2.11 to figure 2.15, and figure 2.9 to figure 2.16) meaning that deductions made from the smoothed data will be applicable to the actual data. This is helpful since it is easier to analyse the smoothed data.

As can be seen from the time series of the smoothed data (figure 2.16) and the colour code shown in figure 2.14, there is a gradual shift

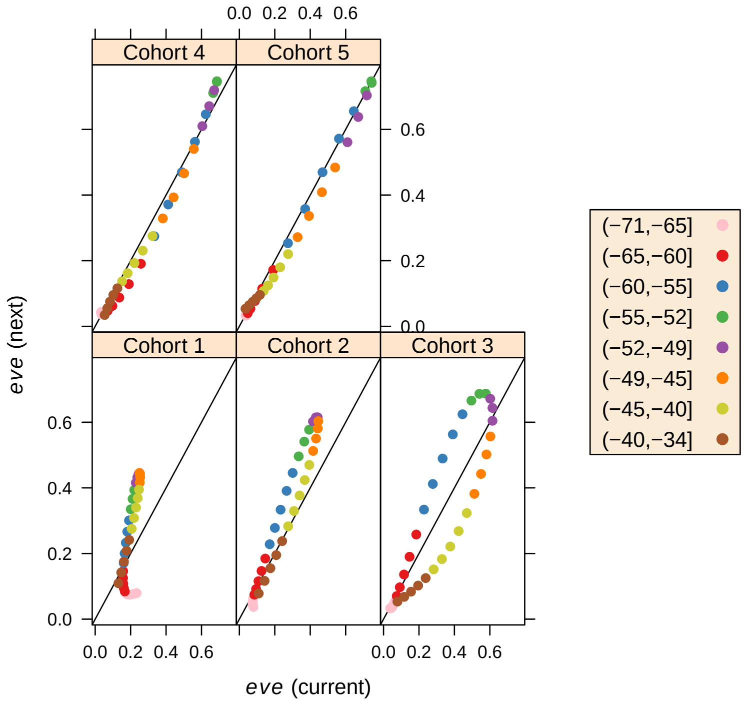


Figure 2.15: Current eve mRNA level vs next for each nucleus per cohort for the smoothed data. The nuclei are divided into groups corresponding to their location on the x-axis. This corresponds to the background colours in figure 2.14.

in the dynamics as one moves along the A-P axis. The interpretation of this shift depends to a large degree on reliability of the nuclear flow model. The relative composition of *eve* stripe 2 movement according to the nuclear flow model can be seen in figure 2.17. In particular, stripe 2 is not expected to move overall, but the nuclear flow model predicts that the nuclei move towards the posterior, apparently compensated by an expression flow towards the anterior. But as shown above, the dynamics underlying expression flow are complicated.

Another explanation for *eve* 2 is that it is not subject to nuclear flow or expression flow. It is thus instructive to consider whether the dynamics in the case of no expression flow are any simpler. And in fact, they are clearly so as shown in figure 2.19. This time series results from centring the smoothed stripes as in figure 2.18. It is now easier to see that the dynamics of *eve* stripe 2 can be explained by a curve

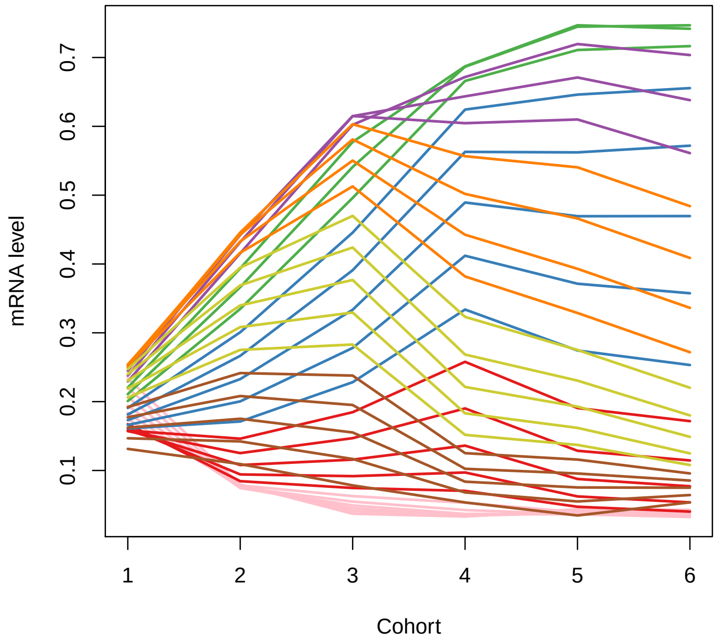


Figure 2.16: Time series of smoothed data.

following equation 2.3 and figure 2.7, plus refinement or sharpening of the borders of the stripe. By adding anterior movement, the dynamics of figure 2.12 will be obtained. This is confirmed with figure 2.20 that shows the dynamics are now more like an exponential function, with the exception of the borders of the stripe where refinement is taking place.

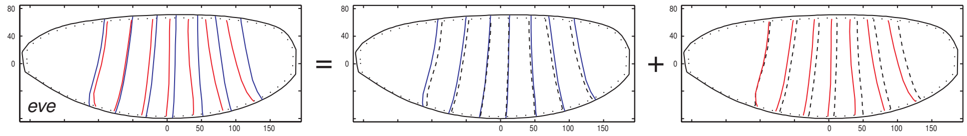


Figure 2.17: Pattern flow = nuclear flow + expression flow. Solid lines are the location of the actual stripe borders (blue is early and red is late). The dashed black lines show the location of the late stripe border in the absence of expression flow (calculated using the nuclear flow model). From Keränen et al. (2006); used in accordance with the Creative Commons Attribution License (`http://creativecommons.org/licenses/by/2.0`).

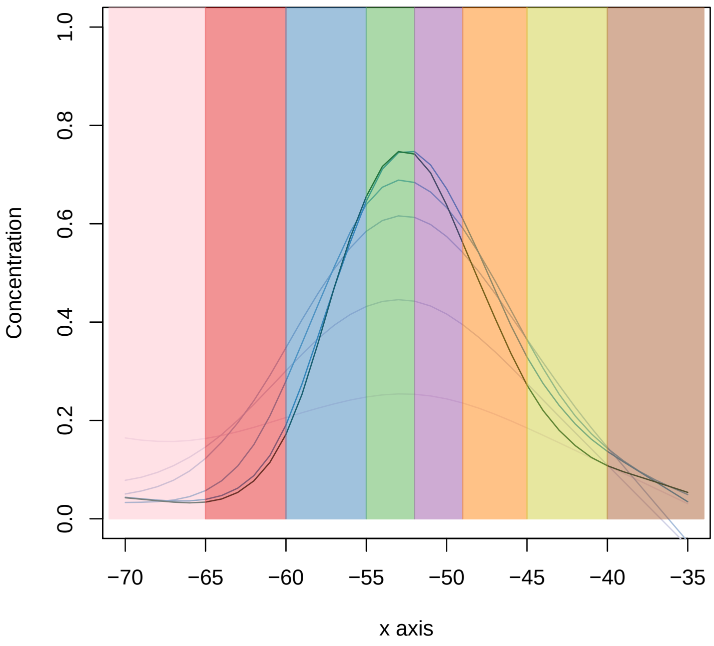


Figure 2.18: Smoothed stripes, centred.

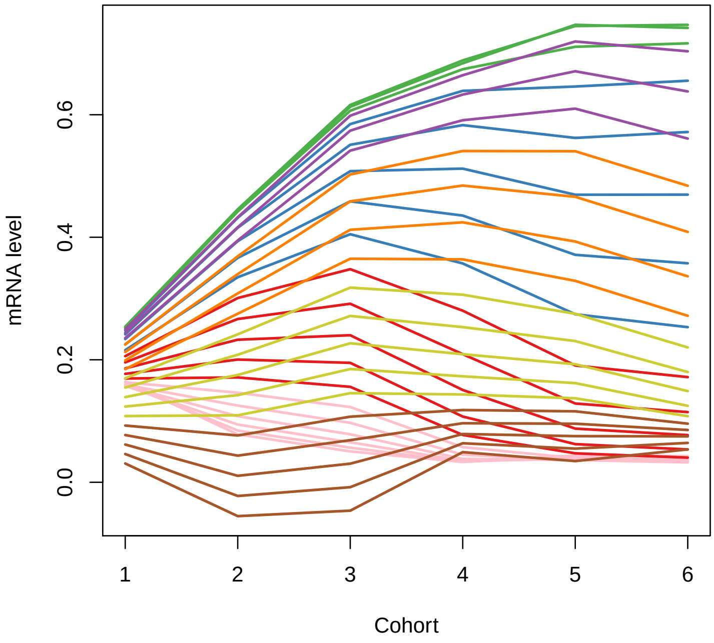


Figure 2.19: Time series of centred, smoothed data.

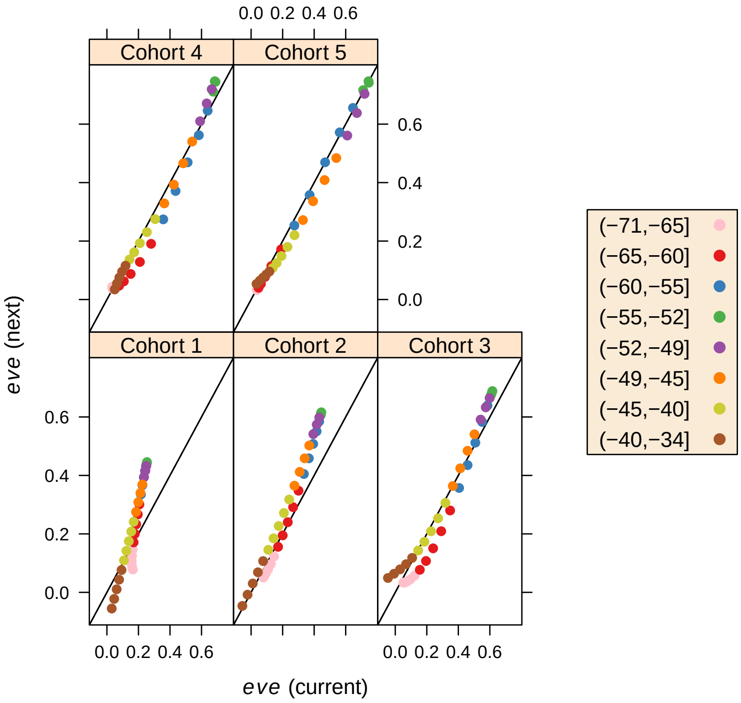


Figure 2.20: Figure 2.15 replotted with centred data.

2.6 MODEL INPUTS AND OUTPUTS

The main findings of this chapter are clear. The dynamics of *eve* stripe 2 in the *Virtual Embryo* can be explained by a combination of the following:

- constant transcription and proportional mRNA degradation
- expression flow to the anterior (whether real or artefactual)
- refinement and narrowing of the stripe

Since the gap gene concentrations are apparently changing, it is remarkable that the transcription rate of *eve* might be constant. This matter is not resolved in this thesis, but section 8.4 offers a few suggestions. The most interesting of these suggestions, though, depends on the general model of positional information described in chapter 7, and this in turn, depends on the enhancer models developed in this thesis.

This chapter now concludes with the set-up of the inputs and outputs for these models.

2.6.1 *Model inputs*

There is no obvious explanation from the data for the movement of stripe 2, and this suggests that caution should be exercised in the use of the temporal correspondences of the nuclei in the *Virtual Embryo*. It is thus sensible to use expression measurements from the same cohort. Considering figure 2.1, cohort 3 is a good choice: it is the earliest cohort where the *eve* stripes are clear. This cohort has the added benefits of being before the significant stripe movement between cohort 3 and 4 (see figure 2.14) and being the last cohort that includes measurements for an important regulator, BCD. It is not ideal to use mRNA in place of protein measurements, but as will be seen, the models perform well in spite of this.

2.6.2 *Model outputs*

The discretisation of *eve*'s response avoids uncertainties in transcription rate, but it is also important because of the use of mRNA measurements as a proxy for protein concentrations. It is possible that there are movements in the various expression patterns (for example, moving gap gene patterns are analysed in Jaeger et al., 2004), and therefore, mRNA concentrations from the same time point will be shifted slightly relative to the actual protein concentrations.

Discretisation is reasonable since the *eve* stripes are narrow and the change in concentration from maximum to low occurs over only a few nuclei. Thus, the nuclei are divided into two classes: ON or OFF. The decision is based on a threshold expression level of 0.2, which was chosen from figure 2.14 with the aim of including the majority of the nuclei in the stripe. These discretised values will form the basis for the models of the next chapter.

Before proceeding, however, it is interesting to look at the discretised response of the *eve* gene in *regulator space*, which is where each nucleus is positioned according to the concentrations of its regulators. Figures 2.22, 2.23 and 2.24 show a slice through regulator space (i.e., only two regulators are considered at a time). Each nucleus is coloured according to which stripe it belongs to using the discretisation approach described above (see figure 2.21). From this it is clear that the nuclei of each stripe cluster together in regulator space, which strongly suggests a classification model, the topic of the next chapter.

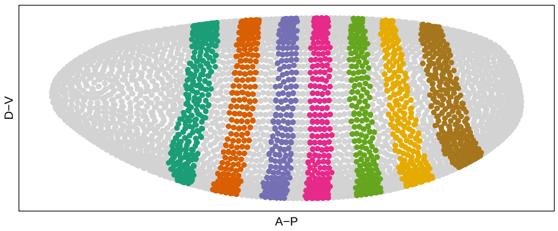


Figure 2.21: Nuclei colour-coding for figures 2.22, 2.23 and 2.24

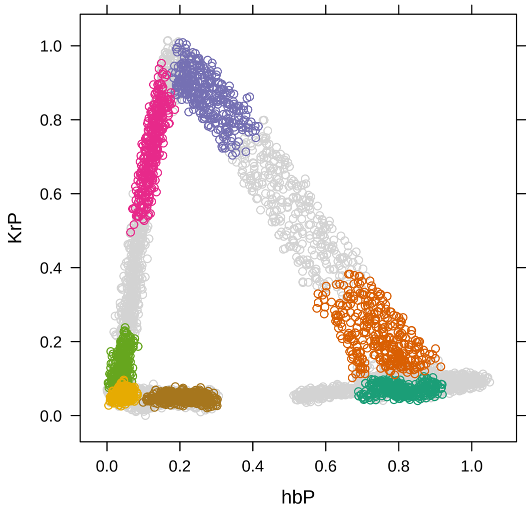


Figure 2.22: Nuclei of the *Virtual Embryo* for cohort 3 plotted according to the amount of HB and KR. Nuclei coloured as in figure 2.21

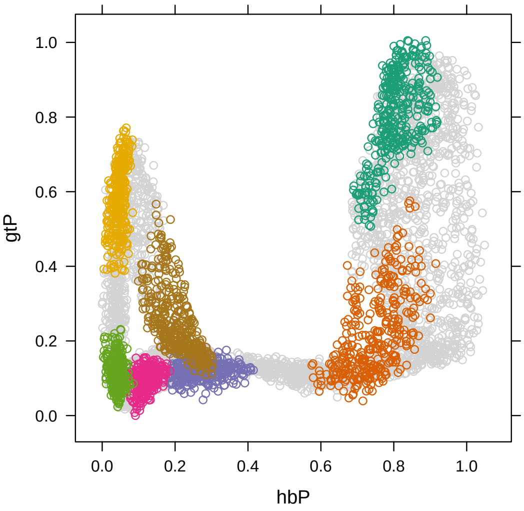


Figure 2.23: Nuclei of the *Virtual Embryo* for cohort 3 plotted according to the amount of HB and GT. Nuclei coloured as in figure 2.21

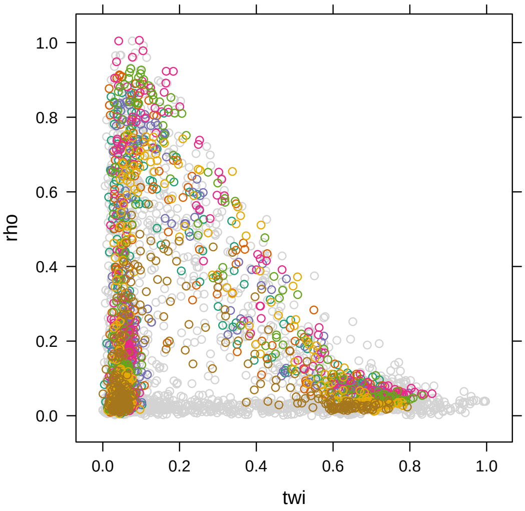


Figure 2.24: Nuclei of the *Virtual Embryo* for cohort 3 plotted according to the amount of *rho* and *twi*. Nuclei coloured as in figure 2.21

3

MODEL BEGINNINGS: A LINEAR CLASSIFIER

The previous chapter justified the use of a classification model for the *Virtual Embryo*: the response of the gene belongs to one of two classes, ON or OFF. The regulator concentrations define the position of the nucleus in regulator space (see section 2.6) and it is the goal of a classification model to separate the classes in this space.

Two modelling approaches will be considered here: logistic regression and the Support Vector Machine (SVM), both well-established and common methods for binary classification (Bishop, 2007; Ripley, 2008). These are introduced and then applied to *eve* stripe 2 where it is shown that, in this case, a linear decision boundary is sufficient.

3.1 NOTATION AND TERMINOLOGY

This section introduces the terminology and notation used in comparing the Support Vector Machine (SVM) and logistic regression models.

The data set has $\mathcal{D}$ explanatory variables (the regulator concentrations) and one response variable—the transcriptional state of the *eve* gene. The response variable can have one of two values: ON or OFF. The data set consists of N observations, one for each nucleus[*], where each observation includes one value for each of the regulators, as well as the related response of the *eve* gene. In order to indicate all the values for the explanatory variables of observation i, $\mathbf{x}_i$ will be used, with bold to indicate a vector. $\mathbf{x}$ will be used to refer to the values for any observation, for example, in a function. If the value of the jth

[*] In cohort 3—see the previous chapter.

explanatory variable for the ith observation is required, this will be indicated by x_{ji} (note the lack of bold).

The usual goal of a classification model is to find a function that can classify a future observation into one of two classes, in this case ON or OFF. To do this, the explanatory or input variables might first be transformed. In the models considered here, any transformation is fixed in advance of fitting the model and the transformation can be written as a vector of M functions, each of which operates on the input vector $\mathbf{x}$ producing one of M values in *feature space*. In essence, the dimension of the input space has changed from $\mathcal{D}$ dimensions to M dimensions. The vector of functions can be written as $\boldsymbol{\phi}(\mathbf{x})$, and each of these M values can be referred to by $\phi_m(\mathbf{x})$. Using this notation, the relevant classification models may be written as:

$$y = f(\boldsymbol{\phi}(\mathbf{x}))$$

where the class decision is made based on the value of y. Since, the classification models considered here are linear, they can be described as:

$$y = \beta_0 + \beta_1\phi_1(\mathbf{x}) + \cdots + \beta_m\phi_m(\mathbf{x}) \tag{3.1}$$

which are linear in the feature space. In this case, if the decision is based on a specific threshold for y it gives rise to a linear decision boundary, a hyperplane in the feature space. This will be non-linear in the original input space if a non-linear transformation has been applied. Figure 3.1 shows an example of a linear decision boundary in a two-dimensional ($M = 2$) feature space.

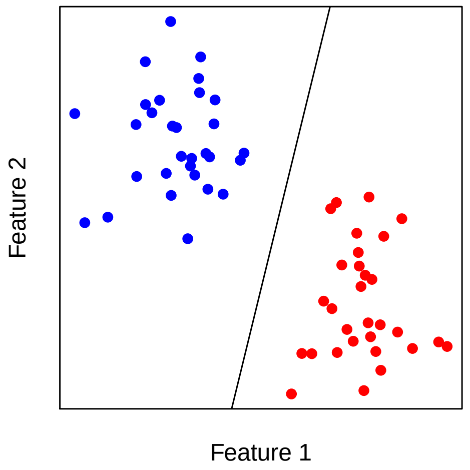


Figure 3.1: Two classes (red and blue) separated by a linear classification boundary.

3.2 SUPPORT VECTOR MACHINE

The first classification model to be considered is the Support Vector Machine (SVM), a flexible method that includes the ability to transform the data in many different ways. There are two relevant conceptual parts to an SVM: it is a maximum margin classifier and it makes use of kernel methods to implicitly transform the data. Each of these will now be described.

3.2.1 *Maximum margin classification*

The SVM finds a linear decision boundary in feature space by maximising the distance of the margin. In linearly separable data, the margin is the distance to the nearest points (the *support vectors*) on either side of the decision boundary. Since not every data set can be linearly separable, an error function is introduced that penalises points that are on the wrong side of the relevant margin in the training data according to how far they are from it. By choosing the error function to be linear in the distance from the margin, the algorithm can be viewed as a quadratic programming problem (the optimisation of a quadratic function subject to some inequality constraints). In this case, a global optimum can be found (Bishop, 2007).

It is worth noting that the output of the SVM is a number that, depending on whether it is positive or negative, classifies the observation into one of two classes. It has no natural probabilistic interpretation.

3.2.2 *Kernel methods*

If the SVM's maximum margin method was used directly to find the optimal expression for equation 3.1, it would be quadratic programming problem in M variables, where M is the dimension of the feature space. But a dual formulation allows it to be expressed in terms of N variables instead, where N is the number of observations. This is not much use on its own, especially if $N > M$. However, in the dual formulation, the feature vectors occur only in the form $\boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}')$ (an inner product) for each pair of data points $(\mathbf{x}, \mathbf{x}')$. Kernel methods allow this expression to be replaced by a general kernel function $k(\mathbf{x}, \mathbf{x}')$ as long as it meets certain conditions (e.g., positive definite). Now instead of transforming the input space up front and then finding a linear classifier, it is possible to define the transformation implicitly in terms of a kernel function using the dual formulation. Since this formulation depends on the number of original data points N rather than the dimensionality of the transformed input space, it can be used even when the kernel results in a high (or even infinite) dimensional feature space. A second benefit is that different kernels can be tried without changing the algorithm.

In this work, the following kernels are relevant:

- Linear kernel (no non-linear transformation of the input space)

- Polynomial kernel (a polynomial transformation)
- "Gaussian" or radial basis kernel (introduced later)

3.3 LOGISTIC REGRESSION FOR CLASSIFICATION

Logistic regression is a standard statistical technique that is used for modelling binary or binomial responses to explanatory variables (see Collett, 2002 and Dobson and Barnett, 2008). The explanatory variables can be discrete or continuous, and the response variable is the proportion of successes in n independent trials distributed according to the binomial distribution. Using the notation of a generalised linear model (Nelder and Wedderburn, 1972), the underlying probability p_i for the ith observation can be written:

$$\log\left(\frac{p_i}{1-p_i}\right) = \eta_i, \tag{3.2}$$

η_i is the *linear predictor*, the linear combination of the explanatory variables; that is,

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

where x_{ji} is the value of the jth explanatory variable for the ith observation, and the β are to be estimated.

The left-hand side of equation 3.2 is the logarithm of the odds, or *log-odds*. Thus logistic regression models the log-odds of the response for observation i as a linear function of the explanatory variables for that observation, or equivalently, how the log-odds change in terms of a linear combination of the changes in the explanatory variables.

Logistic regression can be used for the *classification* of two classes (Ripley, 2008; Bishop, 2007). In this case, $n = 1$ (the response is distributed according to the Bernoulli distribution) and each observation belongs to a specified class (e.g., ON) with probability p_i, which is determined by the values of the explanatory variables. Although the actual transcriptional response of *eve* to its regulator concentrations is continuous, as shown in the previous chapter, there are uncertainties over its value and so it is prudent to assign it to one of two classes (ON or OFF). Therefore, p_i refers to the probability of *eve*'s rate of transcription being greater than the discretisation threshold and it is this probability which is being modelled. This interpretation is fairly standard in connection with ordinal logistic regression (Dobson and Barnett, 2008). For simplicity, this probability will be referred to as the probability that *eve* is ON.

Although the relation between the response variable and explanatory variables is non-linear, the explanatory variables are first combined linearly (the η) before being transformed by the non-linear function of equation 3.2. Thus, if one classifies by picking a threshold for p and classifying a new observation according to whether it is above or below the threshold, this will relate to a fixed threshold for η. As a result, the decision boundary will be linear (i.e., a hyperplane) just as in the case of the SVM model (see figure 3.1). The difference is that the input variables have not been transformed by the use of a kernel and the feature space is identical to regulator space.

The non-linear transformation for p and the fixed decision boundary is most easily visualised in one-dimension, which is illustrated in figure 3.2. This was generated with artificial data of two classes with a single transcription factor as the explanatory variable. In this case, the classes are not perfectly separable by a decision boundary. The boundary drawn follows from choosing a boundary at $p = 0.5$, which results in a specific value for the (single) explanatory factor. In the more general case, this would correspond to a linear decision boundary. Thus, as shown in the figure, when the linear combination of explanatory

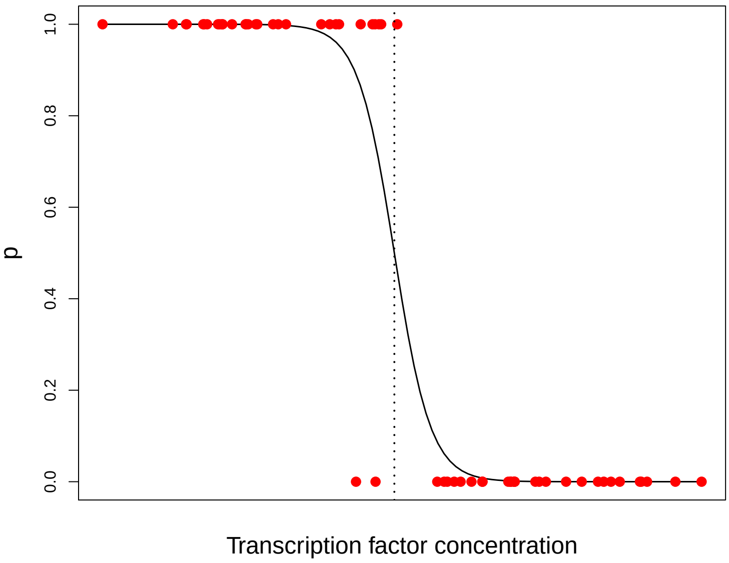


Figure 3.2: Linear logistic regression separating two classes. Artificial data of nuclei with different concentrations of a transcription factor are plotted as red points according to their measured concentration and class, ON (1) or OFF (0). Logistic regression applied to these data produces a prediction (p) as a function of transcription factor concentration, which is plotted as the solid curve. The dashed line indicates a transcription factor concentration boundary corresponding to $p = 0.5$.

factors η is close to the boundary, p is close to 0.5, and when it is further away, its classification becomes more certain (tending towards 0 or 1).

3.4 FITTING AND EVALUATION

In both classification models, the classifier contains unknown parameters, the βs of equation 3.1. These are fitted (or 'estimated', or 'trained') using the data (which is therefore called the *training data*). A common concern is that the parameters might be good at describing the training data, but are not good at generalising to other contexts, such as future predictions. In other words, the training data are in some sense not representative of other relevant data. This is called *over-fitting.*

3.4.1 *Model assumptions and over-fitting*

For statistical models, the question of over-fitting can be addressed partially by considering the suitability of the assumptions the model makes of the data. In the case of logistic regression, it is assumed that the explanatory variables are fixed and only the variation of the response is modelled directly. It is therefore worth considering to what extent the *Virtual Embryo* satisfies this assumption. As described in section 1.5, the measurements for each nuclei of the *Virtual Embryo* are averaged measurements. This should reduce variability in the final averaged measurement assuming the measurement values of each individual embryo follow the same distribution and each embryo measurement is independent. This makes the assumption of fixed values for the explanatory variables reasonable, but it is difficult to assess this fully since each of the many steps that produced the *Virtual Embryo* could affect its suitability.

Fortunately, this is not necessary for the work of this thesis. Firstly, Fowlkes et al. (2008) explored the issue of measurement error and bias in the *Virtual Embryo.* They showed that for various pairs of regulators, their relationship in the averaged embryo was similar to that in the individual embryos where they were co-stained. This makes it plausible that averaging will mainly remove obscuring variation. Secondly, in the context of the analysis presented here, the exact parameter values are not important. Merely being able to classify the response of the *eve* gene based on regulator concentrations from the *Virtual Embryo* would be of interest as it would demonstrate that positional information is available to the *eve* enhancers. And finally, as discussed by Ripley (2008), logistic regression is able to perform well as a classifier even in the cases where specific distributional assumptions are not met[†]. In fact, these considerations are not even relevant to an SVM since it is not a probabilistic model. For the SVM, it is usual to assess the model's suitability on a case by case basis, often by holding back some of the data for testing (Ripley, 2008; Hsu et al., 2010). The same approach could be used for logistic regression.

However, for the work of this thesis, this technique is of limited value. The regulators are expressed in spatial regions or domains (see section 1.2). In other words, the expression levels of the nuclei are not independent, but are similar to their neighbours. This means that nuclei with similar expression levels for the relevant regulators are classified together and therefore, it is not a particularly stringent requirement to generalise from a random subset of the training data to the excluded nuclei.

If classification is possible, a more interesting consideration is whether the classifier, in some way, actually reflects the function of the enhancer. In other words, is an enhancer a classifier? In this context, over-fitting means that the classifier is too flexible or insensitive to biological constraints and is unlikely to represent the functional form of the enhancer. If, however, the classifier can be shown to generalise well to other experimental situations, then perhaps the parameters of the classifier might correspond to biological parameters such as binding sites for transcription factors. Therefore, the following aspects will be tested, which are more relevant biologically:

- How much freedom do the set of regulators have? For example, can the regulators of *eve* 2 be used to fit other stripes? If they can, it suggests that the choice of regulators does not matter

[†] Logistic regression can be justified via distributional assumptions of the explanatory variables. As explained in Ripley (2008) and Bishop (2007), if the probability distribution of explanatory variables conditional on their class is a member of the exponential family of distributions (which includes the normal distribution), and the scaling parameter (or covariance matrix for the normal distribution) is the same for both classes, then the log-odds of an observation belonging to a particular class can be expressed as a linear combination of the explanatory variables.

particularly and that the model is unlikely to reflect biological reality. A related consideration is whether the best fitting models contain the known regulators of that enhancer. This latter issue is explored in chapter 5.

- What is the behaviour after perturbing the regulator inputs, for example, setting one of them to zero? Does the resulting prediction appear plausible in terms of null mutants? Of course, this test is limited since the perturbed gene can affect other regulators, and this will not be captured in the classification model. Nevertheless, it can provide a degree of confidence that the model is behaving reasonably.
- Model flexibility can also be related to enhancer structure in the *eve* locus. For example, which pairs of stripes can be trained together, and does this reflect known enhancer structure, such as pairs of stripes that are regulated by a single *eve* enhancer?

3.4.2 *Implementation details*

The Support Vector Machine models were trained using the R package *e1071*, which uses the LIBSVM library (Chang and Lin, 2001). The default method (C-classification) was used and the data were scaled to zero mean and unit variance. With the exception of the kernel, which is explicitly mentioned where relevant, the only other parameters supplied were $\gamma = 1$ and cost $= 1$. The predictions made by SVM are binary: ON or OFF.

The logistic regression models of this chapter and the next were fitted using the R function `glm` from the *stats* package, which uses *Iteratively Re-weighted Least Squares*, the standard algorithm for fitting a generalised linear model (Nelder and Wedderburn, 1972; Dobson and Barnett, 2008). For many of the models of this thesis, `glm` issued a warning message. This was because most of the logistic models that classify the *eve* stripes successfully have some fitted probabilities very near 0 or 1. (The nuclei on the borders of the stripes have intermediate values.) Although this can suggest problems in certain situations, here, in agreement with Ripley (2008), it is viewed as a desirable outcome of classification. The exact parameter values are not crucial and it will not be necessary to perform hypothesis tests or to estimate confidence intervals for these. Rather, it is of interest whether individual regulators function as activators or repressors (positive or negative coefficients), particularly in predictions of mutant embryos. Also, as long as the nuclei with intermediate fitted values occur along the borders of the stripes, their accuracy is not crucial as it is with other applications of logistic regression (such as modelling drug dose and response). Thus, it is important that model success is ultimately evaluated visually.

3.4.3 *Visual fits*

The *Virtual Embryo* contains three dimensional coordinates for each nucleus, which is measured in μm from the embryo's centre of mass. The visual fits used in this thesis are orthographic views of the embryo. Specifically, each nucleus is plotted at its (x, z) coordinate, ignoring the y coordinate. Since the x- and z-axes are oriented with respect to the anteroposterior and dorsoventral axes respectively, this corresponds to a lateral view, with anterior to the left, and dorsal at the top. In order to plot both sides, the embryo is split where $y = 0$. This means the viewpoint for both sides is from the left; the right lateral view is as if one is looking through the embryo, ignoring the left side. For the sake of compactness, when the predictions for both sides are much the same, the embryo is not split and all the nuclei are plotted in one composite view.

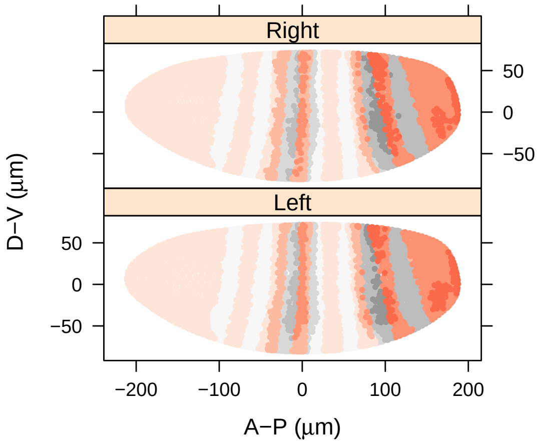


Figure 3.3: An example of a visual fit, an orthographic projection of the embryo. A-P and D-V are the anteroposterior and dorsoventral axes respectively. The scale is from the embryo centre of mass. The nuclei are coloured as described in the text.

The nuclei are coloured as shown in figure 3.3 according to the relevant model's prediction, which is made for every nucleus regardless of the size of the training set. To make comparisons between models possible, the colour scale is consistent and does not depend on the maximum or minimum prediction from any one model. For logistic regression, the colour scale is from 0 (light) to 1 (dark). The same colour scale is used for an SVM model, but in this case, only two values are possible: 1 for ON and 0 for OFF. The colour scale for predictions within stripes[‡] is grey-scale, ranging from white to black. Predictions outside of stripes are on a red scale, with peach for values near 0. Thus, in figure 3.3, the result of a logistic regression prediction, two main regions of expression are predicted: one near stripe 3 (counting from the anterior tip) and another around stripes 6 and 7 extending to the posterior tip. There are patches of higher predictions, including some within stripe 6, but on the whole, only predictions near zero are clear-cut.

It is important to check convergence after each model fit. Thus the function producing the visual fits uses a blue and orange colour

[‡] Corresponding to *eve* expression greater than 0.2

scheme if the underlying model has not converged—this is of particular importance for the custom algorithm introduced in section 5.3.

3.4.4 *Selection of the training data*

Using the entire embryo as training data requires the classifier to operate at the level of the *eve* gene. Since it is known that the transcription of *eve* is controlled by discrete enhancers (see section 1.3), it is reasonable instead to select the subset of the embryo that is relevant to the enhancer under consideration. This is helpful in that each resulting model can be related directly to a discrete regulatory region of the *eve* locus and thus to potential mechanism and the experimental results of transgenic reporters. Further, the models considered in this thesis do not classify well at the level of the gene. Figure 3.4 shows the model predictions of a logistic regression model and an SVM with a Gaussian kernel[§] after using the whole embryo as training data, with all candidate regulators.

In order to select the nuclei to construct the appropriate training data, it is, however, necessary to make an assumption of enhancer function. An hypothesis of this thesis is that transcription factor concentrations are sufficient to determine enhancer function across the whole embryo. In other words, higher-level repression of enhancers is not relevant. Thus, it is assumed that every *eve* stripe enhancer is available to transcription factors in every nucleus, but that the regulator concentrations are such that no expression ensues in the regions outside of the stripes. The overall expression of *eve* is then simply the sum of each enhancer's contribution.

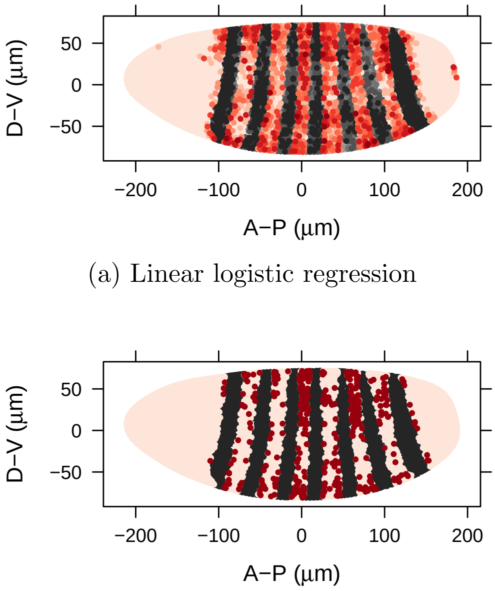
To construct the training data it is first necessary to choose which stripes the enhancer is responsible for. For example, the *eve* 2 enhancer is responsible for *eve* stripe 2. Therefore, *eve* stripe 2 should be

[§] The Gaussian kernel performs better than the polynomial kernel in this case.

included in the training data. Also, all the nuclei outside of the other stripes should be added as negative examples.

There are two variations of this that are used in the following models. In one approach, the training data consists of every nucleus, with the *eve* response set to ON in the relevant stripe(s), and OFF elsewhere (including in the other stripes). In situ hybridisation data of transgenic reporters lends support to this perspective, and therefore this approach will be taken for discovering regulators and modelling enhancer function in detail as in chapters 5 and 6.

However, this chapter and the next consider the suitability of different modelling approaches more generally. It is thus useful to suspend judgement on which stripes are regulated together. In the second approach, one or more stripes will still be added as positive examples to the training data, but rather than include the other stripes as negative examples, they are excluded. The regions between the stripes remain in the training data as before, except that to avoid possible inaccuracies in where the stripe borders are drawn, the immediate neighbouring nuclei of the other stripes are also left out. The advantage of this approach is that it makes it possible to see if certain stripes could be co-regulated.



(a) Linear logistic regression

(b) SVM with a Gaussian kernel

Figure 3.4: Model predictions for all *eve* stripes, using the 38 regulators from table 1.2.

3.5 APPLICATION TO *eve* STRIPE 2

The two main classification models that have been introduced will now be applied to *eve* stripe 2. The expression of this stripe is controlled by a single, well-studied enhancer with relatively well-known regulators. For the remainder of the chapter the primary regulators from Andrioli et al. (2002) will be used: BCD, HB, KR, GT and *slp1*. KR and GT are thought to define the borders of the stripe through repression, HB and BCD are broadly distributed activators (Small et al., 1992), and they propose that SLP1 acts as a repressor in the anterior region (Andrioli et al., 2002).

3.5.1 *SVM with a Gaussian kernel*

As shown in section 2.6, the nuclei of the stripes are found together in regulator space. This is to be expected if the concentrations of regulators vary in gradients across the anteroposterior axis. Therefore, a natural classification model to consider is an SVM with a Gaussian kernel. The Gaussian kernel effectively measures the distance points are away from a central location (such as the middle of a cluster of points), and is therefore good at separating a cluster of nuclei from the rest of those in regulator space. It is thus not surprising that this model, after training, can successfully predict stripe 2 (figure 3.5). It is, though, of some importance: it indicates that sufficient positional information is available for the *eve* 2 enhancer using only the concentrations of these primary regulators. However, as described in section 3.4.1, the next question to ask is whether the classifier function is biologically plausible.

First though, by way of illustration of the discussion in section 3.4.1, it is instructive to test over-fitting by holding out some of the training data. Figure 3.6 shows a model trained on a random subset of the data

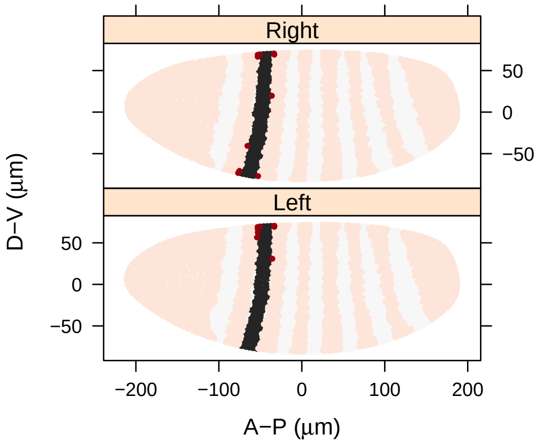


Figure 3.5: Prediction for *eve* stripe 2 using an SVM with a Gaussian kernel.

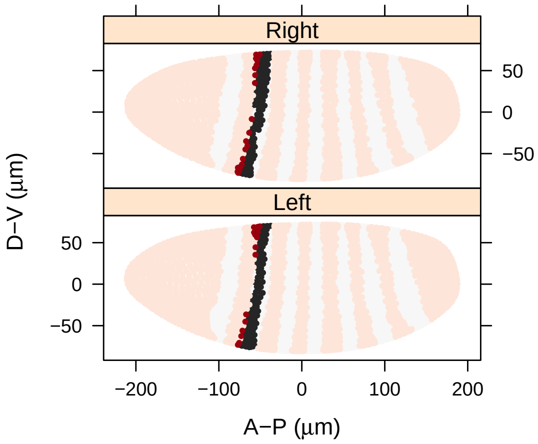


Figure 3.6: Prediction for *eve* stripe 2 using an SVM with a Gaussian kernel trained on a random subset.

consisting of 10% of the positive and negative examples (i.e., keeping the same ratio). It is indeed able to successfully predict across the whole embryo. In some contexts this would be taken as evidence that the model generalises well, but the prediction for a null mutant of the repressor *Kr* shows otherwise. Rather than predicting expanded expression resulting from derepression, the SVM model predicts no stripe (figure 3.7). This is because a concentration of zero for KR is too far away from the values found in the nuclei of *eve* stripe 2 (refer to figure 2.22). As a result, the SVM classifier predicts that any nucleus with no KR could not be part of the stripe. More worringly, as shown

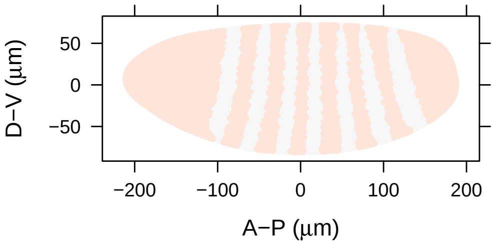


Figure 3.7: *Kr* null mutant prediction for *eve* stripe 2 using an SVM with a Gaussian kernel.

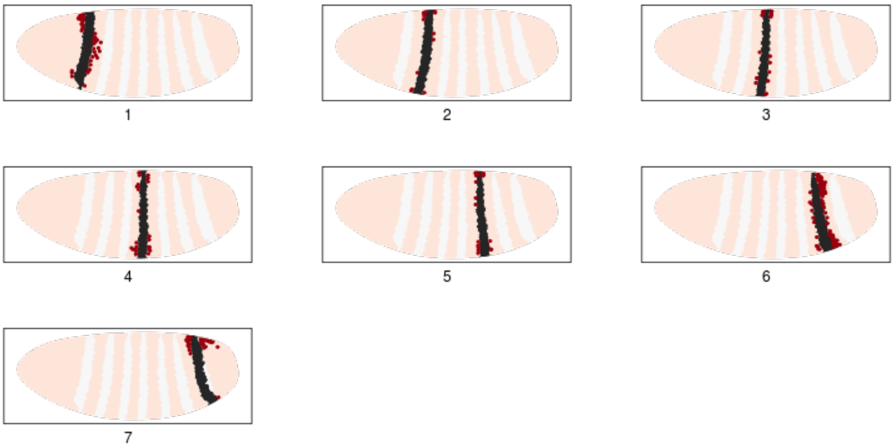


Figure 3.8: Predictions for each stripe using an SVM with a Gaussian kernel and the *eve* stripe 2 regulators.

in figure 3.8, an SVM with a Gaussian kernel is able to fit almost every stripe with the same set of limited regulators. Note in particular that this set of regulators does not include *kni*, which is thought to be crucial for stripes 4+6 and 3+7 (Small et al., 1996; Clyde et al., 2003).

One further complication that has been glossed over is that the SVM requires two parameters to be supplied in advance of the training procedure¶. In other words, these parameters are not learnt directly, but need to be chosen through higher-level methods such as cross-validation. The ability of the SVM to classify each stripe can be worsened by changing these parameters, but for all values tested, stripes 4 and 5 still classified very well, and the others quite well. In these cases, the *Kr*

¶ The parameters are cost, a scaling for the misclassification penalty and γ, a parameter for each kernel mutant still did not produce an extension of stripe 2, but rather had reduced or no expression. The essential difficulty of the SVM is that if these parameters are set through methods such as cross-validation, then it will by design give rise to excellent classification. Further, the biological intepretability of these extra parameters is limited. Therefore, in the analysis presented here, the parameters used are those that are optimal for classifying the training set[‖].

3.5.2 *A linear classifier is sufficient and reasonable*

Given the biologically implausible flexibility of the Gaussian kernel, it is necessary to consider whether simpler kernels for the SVM might generalise better. The simplest kernel one could consider is the linear kernel, which amounts to no transformation of the input (or regulator) space. In this case the decision boundary between the two classes is linear. So, for example, if a repressor is important in defining the border of the two classes of nuclei, then once the decision border is crossed, further increases in the concentration of the repressor will not cause the boundary to be crossed again.

The results for an SVM with a linear kernel and that of logistic regression are quite similar. However, in the case of a linear kernel, an SVM is of limited benefit. As described in section 3.2.2, its attractiveness is the ability to transform the input space in many ways, whereas, as will be demonstrated in the remainder of this thesis, the logistic regression model can be extended in straightforward ways. Therefore, in the case of untransformed regulator space, the logistic regression model will be used instead. However, in the next chapter, as the model is extended to other stripes, it will be important to consider further transformations of the input space, and the SVM will feature once more.

[‖] A value of 1 for each works well.

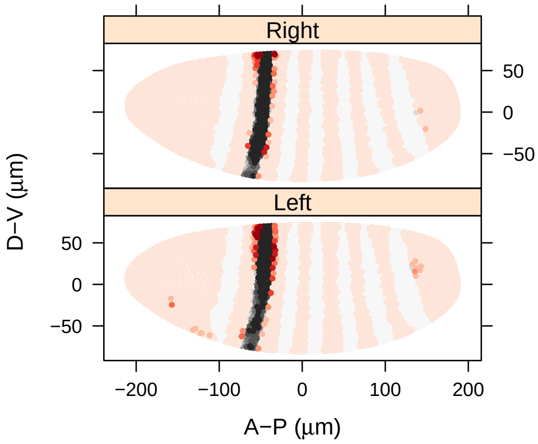


Figure 3.9: Prediction for *eve* stripe 2 using logistic regression model.

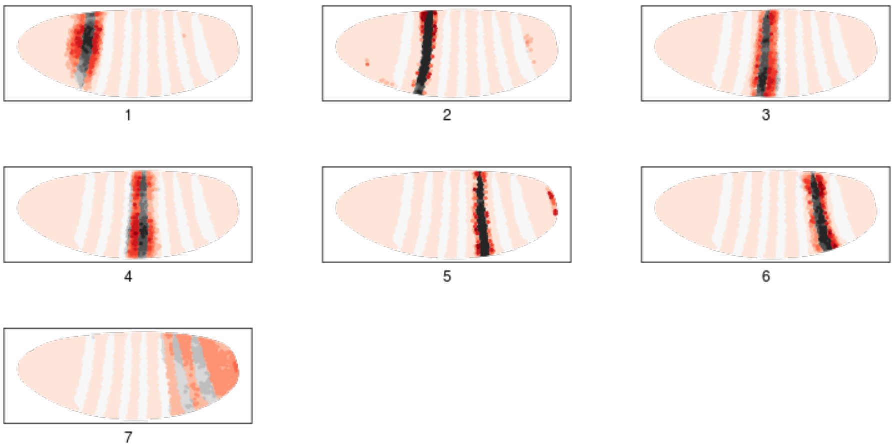


Figure 3.10: Predictions for each stripe using logistic regression and the *eve* stripe 2 regulators.

For stripe 2, a linear classifier performs well and responds reasonably to perturbation, although not perfectly. Figure 3.9 shows the results of logistic regression on *eve* stripe 2—it can be successfully classified. Further, the linear classifier does not have arbitrary flexibility to fit any stripe precisely using the *eve* stripe 2 regulators (see figure 3.10). It does better on stripe 5 than the others, but it is thought that many of the regulators of *eve* 2 overlap with those of *eve* 5 (Fujioka et al., 1999), so this is not a concern. It is also interesting to examine the predictions of perturbed regulator inputs (figures 3.11, 3.12, 3.13 and 3.14). They generally agree with expectation: KR and GT are repressors and HB is an activator. BCD, however, does not function as an activator. The reasons for this, and the null mutant predictions in general, are discussed in chapter 6 where a modified model is proposed. But before that, it is necessary to introduce an extension to the model, which is the topic of the next chapter.

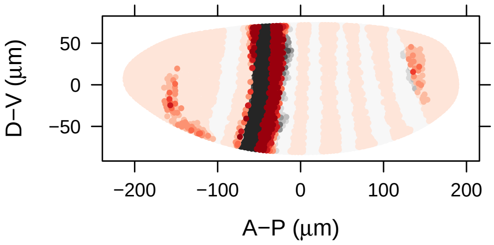


Figure 3.11: *Kr* null mutant for *eve* stripe 2 using logistic regression.

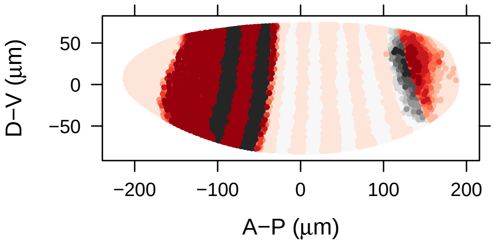


Figure 3.12: *gt* null mutant for *eve* stripe 2 using logistic regression.

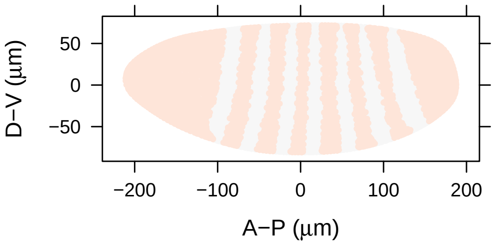


Figure 3.13: *hb* null mutant for *eve* stripe 2 using logistic regression.

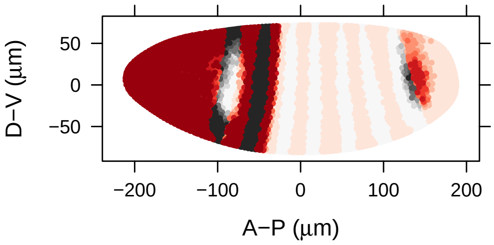


Figure 3.14: *bcd* null mutant for *eve* stripe 2 using logistic regression.

4

MODEL EXTENSION: DUAL REGULATION

The previous chapter demonstrated that a linear classifier can model the functionality of *eve* 2, a well-characterised enhancer. However, it has been suggested in connection with an enhancer that controls two stripes, *eve* 3+7, that dual regulation might play an important role (Papatsenko and Levine, 2008). Dual regulation refers to the capability of a regulator to act as an activator or a repressor depending on its concentration. If this is relevant, then a linear classifer will not be able to describe the functionality of *eve* 3+7 fully. This chapter demonstrates that this is the case, but shows how the model can be extended relatively simply to address this shortcoming.

4.1 LINEAR LOGISTIC REGRESSION IS NOT SUFFICIENT

In order to test the suitability of a linear classifier, it is necessary to choose an appropriate set of regulators, but the choice of regulators is less straightforward for *eve* 3+7 than for *eve* 2. Two gap genes, *knirps* (*kni*) and *hunchback* (*hb*), are thought to be important for defining the borders of the *eve* stripes (Small et al., 1996; Clyde et al., 2003). However, the role of other regulators is less well known, and as shown in figure 4.1, *kni* and HB are not sufficient for predicting stripes 3 and 7. Unfortunately, including the full set of (38) candidate regulators from table 1.2 (page 22) produces an incoherent prediction, which is best observed from the *visual fit*—figure 4.2*.

* Adding more regulators only improves the flexibility of the model in fitting the data, and so inevitably improves the likelihood of the model.

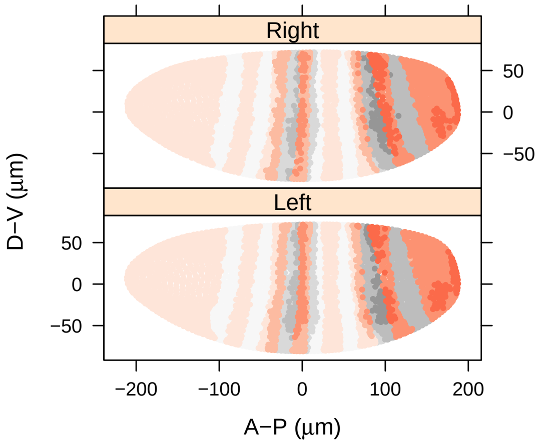


Figure 4.1: Linear logistic regression with HB and KNI applied to *eve* 3+7.

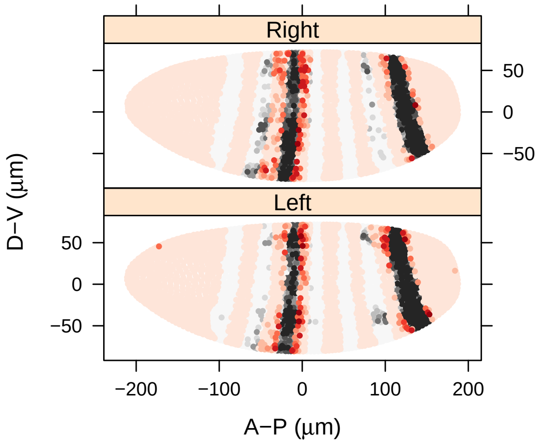


Figure 4.2: Linear logistic regression with 38 regulators applied to *eve* 3+7.

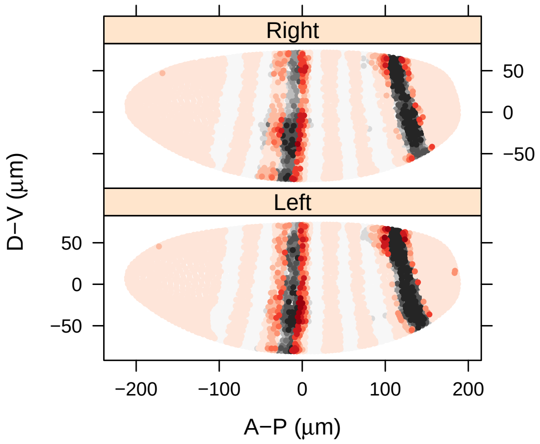


Figure 4.3: Linear logistic regression with 11 regulators applied to *eve* 3+7.

This can be remedied by choosing a smaller set of regulators. For the purposes of this section and the next, a set of common gap and maternal transcription factors were selected from the fuller set, namely: BCD, *cad*, GT, HB, KR, *kni*, *tll*, *hkb*, *fkh*, *slp1* and *D*. (This selection process was fairly ad hoc; model selection is explored more rigorously in chapters 5 and 6.) The resulting prediction from the smaller set can be seen in figure 4.3. The fit from the linear logistic regression model is reasonable in the sense of broadly predicting expression around stripes 3+7, but it lacks precision in specifying the borders of the stripes, particularly stripe 3.

This could be a result of a missing regulator or lack of precision in the data. A lack of precision could, for example, be a consequence of inaccurate registration of various embryos which would lead to blurring near the stripe borders after averaging. This would apply particularly to protein measurements since these are not finely registered (see section 1.5.5). Also, mRNA measurements might approximate protein concentrations poorly, for example if the *eve* stripes are shifting or the relevant proteins are subject to post-translational regulation. However, an SVM with a polynomial kernel can fit the stripes successfully (figure 4.4). This demonstrates that a slightly more complicated function is able to describe the enhancer's function, and it thus important to understand what is missing from the linear classifier.

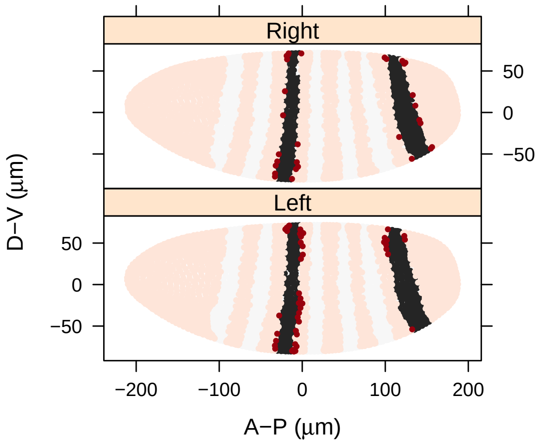


Figure 4.4: SVM with a polynomial kernel with 11 regulators applied to *eve* 3+7.

4.2 INTERACTIONS AND DUAL REGULATION

The simplest non-linear polynomial kernel is of degree 2[†], and this is sufficient to produce a good fit—that of figure 4.4. This effectively transforms the input space to include each explanatory variable, the square of each explanatory variable, all pairwise interactions (the multiplication of each possible pair) and an offset term.

These aspects can be separated into interactions and dual regulation, as will be done in the remainder of this section, but exploring this with an SVM is difficult. In an SVM, all the input variables are transformed in the same way, and yet it would be useful to see if only a few require dual regulatory capabilities. Also, it will be helpful to assess the relative importance of interactions with respect to dual regulation. Further, in the case of the polynomial kernel, the dimension of the feature space is not much higher than the original input space, thus limiting the advantage of an algorithm operating in the size of the observations (nuclei) rather than the number of input variables (regulators). Logistic regression does not have these problems, but additionally has the benefit that

[†] This kernel is $(\mathbf{x}^T\mathbf{x}' + c)^2$, where $\mathbf{x}$ and $\mathbf{x}'$ are the two input vectors as described in section 3.2.2.

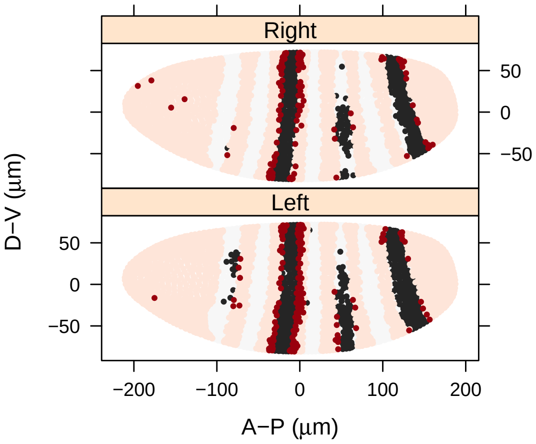


Figure 4.5: Logistic regression with interactions and 11 regulators applied to *eve* 3+7.

its output is a probability. This makes it easy to identify 'grey areas' between classes, particularly visually. It also supports model selection methods, which will be discussed in chapter 5. Thus, rather than exploring alternatives within the framework of the SVM, it is much more straightforward and useful for the purposes of this thesis to modify the linear logistic model. The extended logistic model will be used from here on.

4.2.1 *Interactions*

One explanation given for imprecise predictions is that molecular interactions between transcription factors (such as cooperative binding to DNA) might be missing from the model (e.g., as proposed in Segal et al., 2008 for their model). It is thus of interest to see the effect of adding pairwise interaction terms to the model; namely, the pairwise multiplication of all relevant regulator concentrations.

The model prediction, however, is not particularly easy to interpret. When training on 3 and 7 together, the interaction model predicts expression in stripe 5, thus suggesting that it might be co-regulated with stripes 3 and 7 (figure 4.5). One might argue that this interaction is biologically interesting, but the goal of this thesis is to capture enhancer functionality, rather than higher-level functions such as enhancer interaction. Thus, in this context, the function is not capturing the behaviour being considered. One option is to restrict which interactions are in the model and to consider those that behave most reasonably. However, there is limited biological information to guide the interpretation and, as the next section will show, dual regulation is a simpler and more robust explanation.

4.2.2 *Dual regulation*

A linear logistic regression model effectively means that the enhancer is only imposing a single threshold. Increasing (or decreasing) the concentration of a transcription factor can cause the threshold to be crossed, but further increases (or decreases) in concentration will have no further effect as far as the binary response of the enhancer is concerned. Yet, it has been proposed that HB might have dual regulation capabilities, which has been modelled in the context of the *eve* 3+7 enhancer (Papatsenko and Levine, 2008).

Dual regulation can be modelled by adding a quadratic term for the relevant regulator to the model. Thus, for example, if the quadratic term is given a negative coefficient in the trained model (so that it is a concave downwards function), then the explanatory variable is able to contribute positively at lower concentrations and negatively at higher concentrations (like the dual regulation model described in Papatsenko and Levine, 2008).

And indeed, the quadratic logistic regression model is able to predict as expected for stripes 3 and 7 (figure 4.6). This is a significant result since it suggests that dual regulation alone is sufficient to explain the function of the *eve* 3+7 enhancer; interactions between different regulators are not necessary in this case. Nevertheless, as was done for

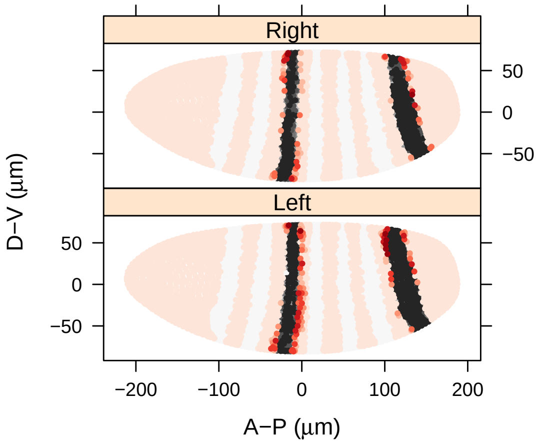


Figure 4.6: Quadratic logistic regression with 11 regulators applied to *eve* 3+7.

eve 2 in section 3.5.2, it is important to test the biological plausibility of the model.

4.3 VALIDATION OF DUAL REGULATION

As it turns out, the model underlying figure 4.6 does not generalise well. The most apparent flaw for the biological plausibility of the model is the *kni* null mutant prediction (figure 4.7), which suggests that *kni* is an activator, rather than a repressor as expected. In particular, there should be increased expression between stripes 3 and 7 (Clyde et al., 2003), but instead decreased expression is observed. It is worth mentioning (although the results are not shown) that *kni* still acts as a repressor if its expression is increased—both stripes are removed.

One can see how this arises by considering figure 4.8, the *regulator function* of KNI in the model. The regulator function[‡] is simply the contribution of that regulator to the linear predictor of the model (η from equation 3.2 on page 55) as a function of its concentration. At low concentrations, KNI is an activator, even though its overall function is a

[‡] Defined in section 5.3.2 by equation 5.2 on page 97.

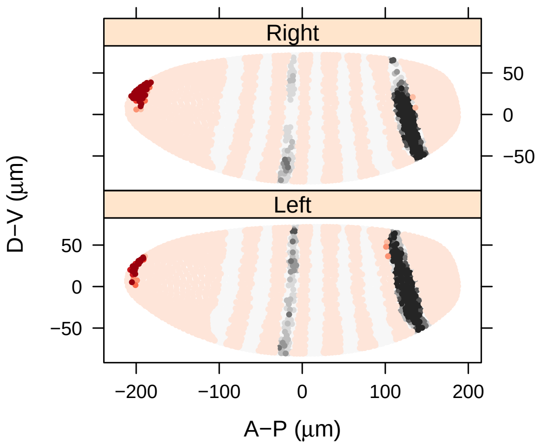


Figure 4.7: *kni* null mutant prediction using the quadratic logistic regression model with 11 regulators applied to *eve* 3+7.

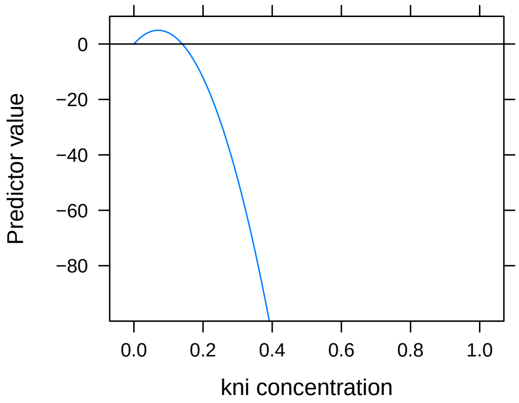


Figure 4.8: The regulatory function for *kni* in the full quadratic logistic regression model of *eve* 3+7.

repressor as expected. The supposed dual regulatory activity is rather extreme, with activation quickly switching to very strong repression.

This problem demonstrates that although the prediction on the available data is good, adding too many terms can lead to unreasonable fits that do not generalise well to situations outside the training data—and this was with a smaller set of only 11 regulators. (The models are, however, able to generalise quite well from subsets of the training data to the whole embryo; see section 3.4.1.)

There are two ways to address this. One approach is to restrict the number of terms (particularly quadratic) that are added to the model. Another is to modify the fitting algorithm. The advantage of the logistic regression model is that both of these are relatively straightforward. The purpose of this chapter is to demonstrate the validity of dual regulation, and for this it will suffice to demonstrate that a minimal model with dual regulation generalises well. A modification to the fitting algorithm is introduced in section 5.3.

4.3.1 *A minimal model helps*

It is in fact possible to get a much improved prediction with a small number of regulators (in this section, HB, *kni* and *tll*[§]), simply by including a quadratic term for Hunchback (HB). The fit is shown in figure 4.9. This is strongly suggestive that the model is capturing actual biological behaviour especially since, as mentioned previously, Hunchback has previously been implicated as a dual regulator in the context of *eve* 3+7 (Papatsenko and Levine, 2008). Further, the added flexibility of the model is still biologically plausible.

Firstly, the *kni* mutant prediction (figure 4.10) is much closer to expectation as KNI functions as a repressor between the stripes. It is possible to explain the remaining discrepancy, which is described in section 6.3 where a fuller treatment is given in the context of model selection.

Secondly, this set of regulators is not able to fit any arbitrary pair of stripes. Figure 4.11 shows the results of training and predicting for all possible stripes using these regulators. Stripes 3 and 7 clearly stand out. Interestingly, the stripe pairs 3+4 and 6+7 predict complementary expression in whichever of stripe 3 or 7 was not included in the training

[§] Section 6.3.2 on page 134 shows the need for additional repressors posterior of stripe 7.

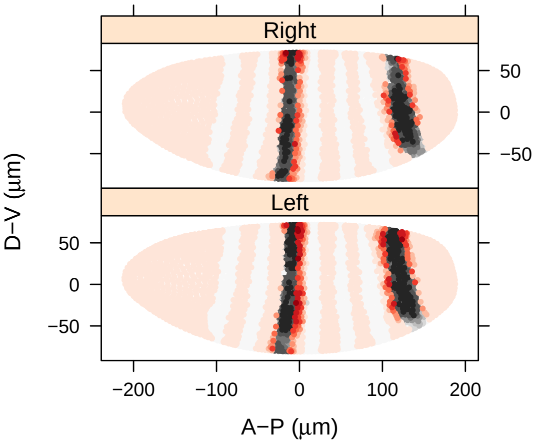


Figure 4.9: Logistic regression model for a minimal model including a quadratic HB applied to *eve* 3+7.

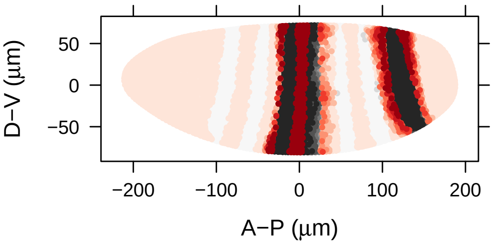


Figure 4.10: *kni* mutant prediction after logistic regression with a minimal model including a quadratic HB applied to *eve* 3+7.

data. This further underlines the compatibility of the model to the data in the vicinity of stripes 3 and 7. It also suggests, however, that the data might not be sufficiently precise to distinguish the expression in the vicinity of stripes 3 and 7 from that firmly in the stripe. For instance, the prediction for stripes 3 and 7 does blur a little into the neighbouring nuclei. More positively though, it does indicate that the discretisation approach is indeed robust to slight expression shifts that might occur over the relevant time period.

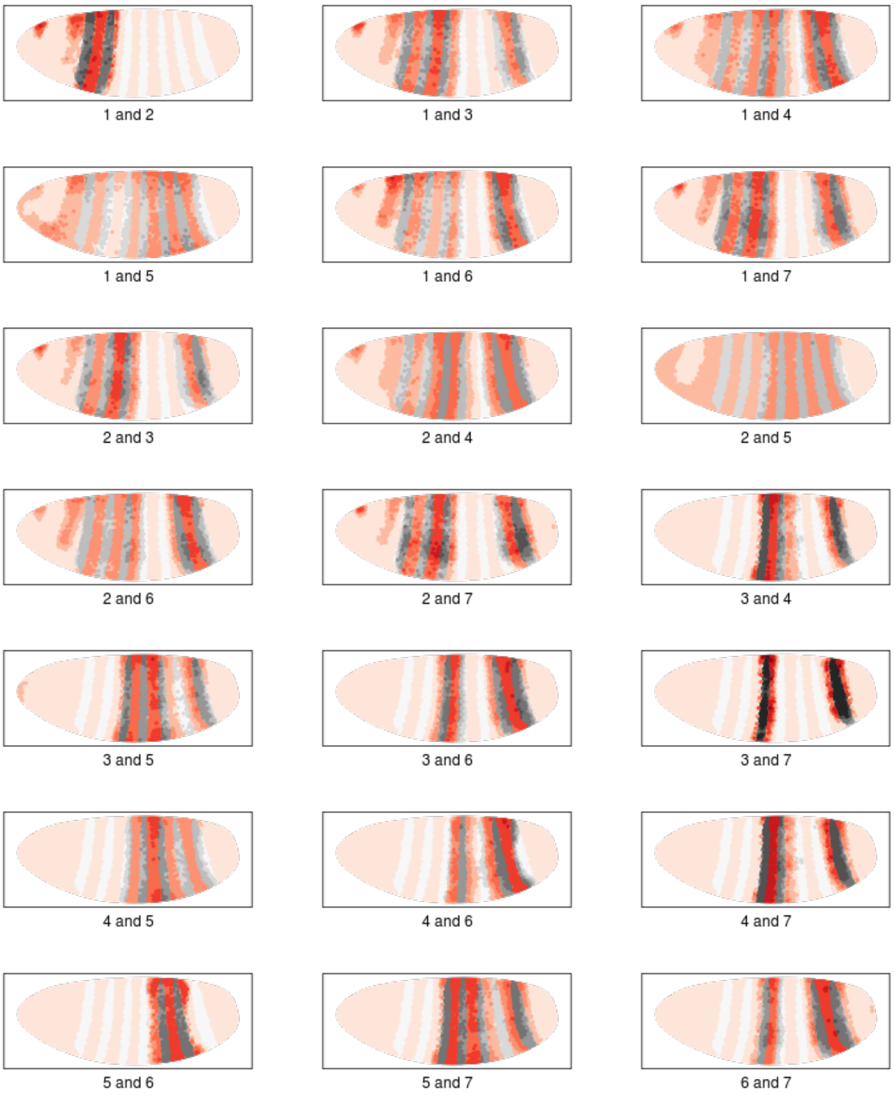
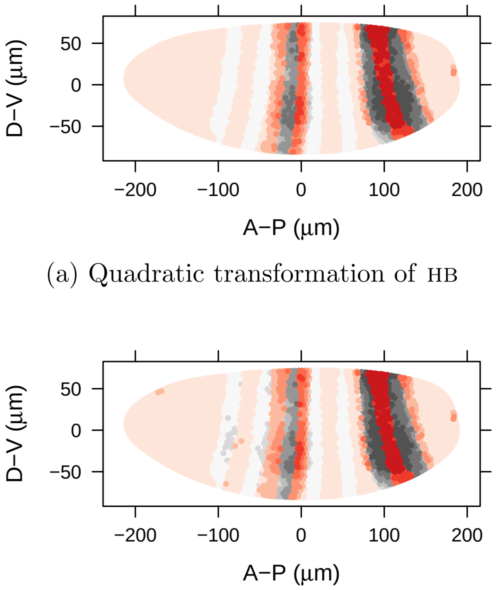


Figure 4.11: Model training and prediction for all pairs of stripes using HB, *kni* and *tll* and a quadratic HB term.

4.3.2 *Dual regulation by Hunchback*

A possible explanation for the importance of the quadratic term is that HB's contribution needs to be magnified at higher values. This could, for example, be necessary if measurement is attenuated at higher protein levels. One way to test this is to retain the quadratic term for HB in the minimal model of figure 4.9, but exclude the corresponding linear term. This prevents HB from having dual-regulatory capabilities, but still allows the quadratic transformation to be tested. Figure 4.12 compares the prediction of this model with the purely linear model. As can be seen, the quadratic term enhances repression in the anterior and helps define stripe 3 a little more strongly. However, in order to achieve the more pronounced improvement in precision seen in figure 4.9 it is necessary to add the linear term; this indicates that dual regulation might be an important aspect. (From here on, all quadratic models will include the relevant lower order linear terms as well.)



(a) Quadratic transformation of HB

(b) No transformation

Figure 4.12: Comparison of the predictions of two logistic regression models, with HB, *kni* and *tll* as regulators.

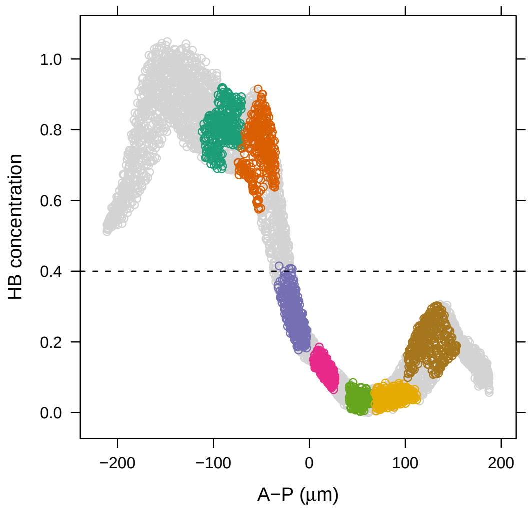


Figure 4.13: Change in HB in all nuclei according to anteroposterior position. The coloured nuclei form the *eve* stripes (stripes 3 and 7 are purple and brown respectively), the grey nuclei are those in between. The nuclei below the dashed line are used to form the training data for figure 4.15.

To explore dual regulation further, it is worth considering the contribution of HB to the expression of stripes 3 and 7 in the minimal model that includes the quadratic and linear terms for HB. Figure 4.13 shows the level of HB protein in all nuclei of the embryo against their position on the A-P axis. By comparing this to figure 4.14 it is apparent that HB is functioning as an activator in the model of *eve* stripes 3 and 7. It is only repressing anterior to stripe 3, indicating the need for additional repression to the posterior of stripe 7 (section 6.3 explores this in the context of model selection). Interestingly, since the concentration of HB is low in many nuclei between the stripes, this lends assistance to repression by KNI.

It is revealing to restrict the training set to those nuclei with HB expression less than 0.4. In this case, the model does not have to repress expression of *eve* in the anterior, for which repression by HB would be useful. And, in fact, the relevant linear model does treat HB as an activator[¶]. This suggests that HB activation is valuable in the region

[¶] Coefficient of 110.

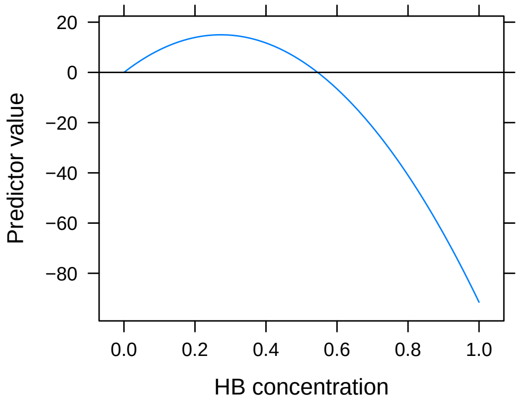
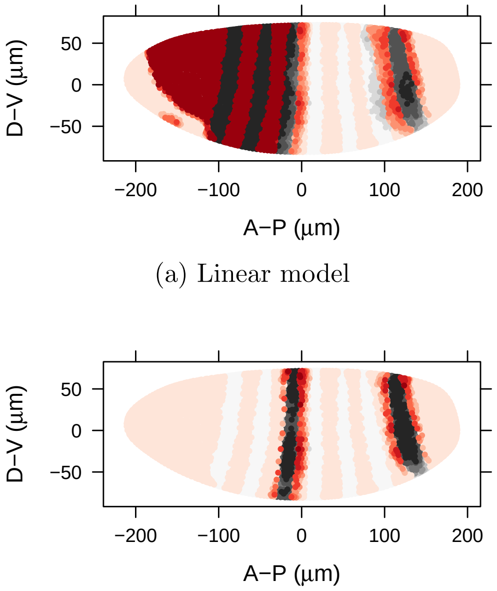


Figure 4.14: HB behaviour when only it has a quadratic term in the logistic regression model of *eve* 3+7.

of stripes 3 and 7. However, the linear model can no longer generalise to the whole embryo (see figure 4.15a) since it lacks a repressor for the anterior region. Remarkably, the model including a quadratic term for HB does generalise across the whole embryo (figure 4.15b). This demonstrates that the dual regulatory ability of HB is important over its activating range (as found in the region of *eve* stripes 3 and 7) and that this is compatible with the requirement for repression in the anterior. This is not to say that the exact shape of the quadratic is correct as far as the actual enhancer function is concerned. For example, repression at high concentrations might not be so strong. Nevertheless, it captures the dual regulatory capabilities of HB for *eve* 3+7 within the range of values in the data set. Further data sets mis-expressing HB at various concentrations would be useful in determining the regulatory function at greater accuracy, but for the purposes of this thesis, it suffices to justify a quadratic term in the logistic regression model. This model can now be used to search for plausible regulators of the *eve* enhancers.



(a) Linear model

(b) Quadratic HB model

Figure 4.15: Models predictions based on training data where HB is less than 0.4.

5

MODEL SELECTION: REGULATOR DISCOVERY

The previous chapters have explained how the response of *eve* can be modelled using logistic regression. This chapter takes advantage of the fact that the fitting process is quick, thus making it possible to test many different models. The process of selecting a small set of the best models is called *model selection*. Since each model differs by which regulators are included, model selection is, in effect, regulator selection. In order to select models, it is first necessary to score each model according to how well it classifies the data[*].

5.1 MODEL SCORING

There are two main ways of scoring a logistic regression model that is used for classification: classification- and likelihood-based scoring.

5.1.1 *Classification-based scoring*

The logistic regression models of the previous chapters can be used to assign nuclei to one of two classes: ON or OFF. This can be done by picking a threshold, which is then applied to the fitted probability of each nucleus. If the fitted probability is above the threshold, the nucleus is ON, otherwise it is OFF. The model then becomes a binary classifier, which can be evaluated using a range of methods. One, for example, considers the ratio of true positives to all positive predictions, but this only evaluates the model at a single threshold. Receiver Oper-

[*] The training data.

ating Characteristics (ROC) curves (see Fawcett, 2006) instead assess performance across the full range of possible thresholds (from 0 to 1). Overall performance can then be summarised by the Area Under the Curve (AUC). This was found to perform well and was particularly useful for comparing the results of Support Vector Machines (SVMs) with those of logistic regression. It confirmed a result of the previous chapters: logistic regression is able to perform as well as SVMs for these data. However, the AUC measure produced similar results to likelihood-based scoring, and yet is computationally more expensive. Thus this method is not used further.

5.1.2 *Likelihood-based scoring*

Logistic regression is a statistical model and can therefore be evaluated as such. The likelihood of the model is the probability of observing the data given the model and its parameter values (the βs of section 3.3). The fitting procedure for logistic regression is based on finding the values of the β that maximise the log likelihood; this final log likelihood is called the maximised log likelihood. For logistic regression with a binary response (like ON or OFF for *eve*), the maximised log likelihood can be written as:

$$\log \hat{\mathcal{L}} = \sum_i y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i) \tag{5.1}$$

where $\hat{p}_i$ is the fitted value (or prediction) for the ith nucleus, and y_i is 0 or 1 depending on whether the ith nucleus is OFF or ON respectively. The $\hat{p}_i$ are determined using the βs (equation 3.2 on page 55).

Deviance

One measure based on maximised log likelihood is *deviance*, which is

$$-2(\log \hat{\mathcal{L}}_c - \log \hat{\mathcal{L}}_f)$$

where $\hat{\mathcal{L}}_c$ is the maximised likelihood of the model being considered, and $\hat{\mathcal{L}}_f$ is the maximised likelihood of the model that fits the data perfectly. Deviance is the natural score for measuring the goodness of fit of a logistic regression model, but it is not directly applicable in the case of binary response data[†].

It is, though, applicable in the case of nested models. Nested models are where one model includes all the regulators of another. The deviance for each model can be calculated from the maximised log likelihood and the difference between the deviances of the two models can be tested using a χ^2 test. Since an intercept-only model (only with β_o) is nested in all other models, it is possible to subtract the deviance of the model of interest from the deviance of the intercept-only model, and then perform a χ^2 test. However, almost every single-regulator model was found to produce a statistically significant change in deviance compared to the intercept-only model—30 out of 38 single-regulator models resulted in a significant change in deviance at 1%. Thus, different methods of assessing model performance were considered.

Two methods will be described here: stepwise selection and a custom model selection method, both of which depend on maximum likelihood scoring. Stepwise selection depends on a penalised form and this is described next.

[†] And it does not follow an approximate χ^2 distribution (Collett, 2002).

Penalised maximum log likelihood

According to Dobson and Barnett (2008), Akaike's information criterion[‡] (AIC) is a measure that can be used to compare generalised linear models, whether they are nested or not. When regulators are added to a model, maximum likelihood never decreases and will usually increase. This means that comparing deviance (or log likelihood) alone will lead to models with more regulators being preferred. AIC addresses this by adding a penalty of 2 for each parameter in the model to the value of $-2 \log \hat{\mathcal{L}}$ (the maximum log likelihood, $\log \hat{\mathcal{L}}$, is a negative number, so the AIC is positive). A stricter penalty can be used per regulator, such as for the Bayesian Information Criterion (Schwarz, 1978), which is $\ln(N)$ where N is the number of observations. This can vary depending on the size of the training data, but when all nuclei are included from a single cohort of the *Virtual Embryo* the penalty is 8.7.

5.2 STEPWISE SELECTION

AIC can be used for stepwise model selection as shown in Venables and Ripley (2003). In stepwise selection, using AIC, a model is compared to all models with one regulator added (forward selection) or removed (backward selection); if this results in a better AIC the best one is selected and the process repeated. When no further improvement can be found, the final model is returned.

A convenient function `stepAIC` is provided in the R package *MASS* (Venables and Ripley, 2003). Using this, stepwise selection was performed for three relatively well-characterised enhancers: *eve* 2, *eve* 3+7 and *eve* 4+6. *eve* 2 has been introduced in chapter 3; its primary regulators are BCD and HB, two activators, and KR and GT, two repressors (Small et al., 1992). *eve* 3+7 was introduced in chapter 4; its regulators

[‡] Proposed by Akaike (1974) and developed under the name An Information Criterion.

are thought to include two repressors (HB and KNI), which are shared by *eve* 4+6 (Clyde et al., 2003).

The training data were selected as described in section 3.4.4. The full set of candidate regulators listed in table 1.2 on page 22 were considered during stepwise selection, but since dual regulation by HB might be important, model selection was also run where a quadratic HB was added as one of the available regulators. Using the standard AIC penalty produced large models (with about 15 regulators). This is too lenient, since the goal of this chapter is to find the most important primary regulators. A stricter penalty based on the Bayesian Information Criterion still produced large models (around 11 regulators). Thus, in order to evaluate the resulting models it was necessary to use a stricter, although arbitrary, penalty of 40 per regulator added. The final models resulting from stepwise selection are shown in tables 5.1–5.5. The tables show the regulators that were selected, their estimated coefficients and their standard error.

	Estimate	Std. Error
Intercept	12.12	2.76
kni	-32.88	9.03
HB	35.36	3.93
BCD	-44.79	8.8
GT	-43.78	4.01
KR	-47.37	4.79
zen	-11.17	1.44
tll	-45.71	9.64

Table 5.1: Final model after stepwise selection for *eve* stripe 2.

5.2.1 *Discussion of results*

By comparing the regulators resulting from stepwise selection with the candidate regulators from table 1.2, it is apparent that gap and maternal genes feature heavily, whereas there are no regulators from the

	Estimate	Std. Error
Intercept	17.4	1.08
sob	-7.74	0.8
kni	-64.61	6.43
tll	-38.3	2.68
GT	-22.9	1.52
HB	-13.56	0.99
twi	5.76	0.48
slp2	19.43	1.82
KR	-5.04	0.5
knrl	-31.09	3.69

Table 5.2: Final model after stepwise selection for *eve* stripes 3 and 7.

	Estimate	Std. Error
Intercept	3.14	1.78
HB sq	-302.98	21.24
HB	162.45	11.6
tll	-61.57	4.13
KR	-13.07	1.24
GT	-23.58	2.32
cad	-12.9	1.58
kni	-47.26	7.54
twi	3.9	0.5

Table 5.3: Final model after stepwise selection for *eve* stripes 3 and 7 with a quadratic term for HB.

'Other' category. This is an interesting result since the gap and maternal genes have largely been discovered through segmentation phenotype assays of mutants (e.g., Nüsslein-Volhard and Wieschaus, 1980). Given that the *eve* stripes are crucial for the segmentation of the embryo (see section 1.2), these results demonstrate that the expression patterns of these genes can also identify them as important for segmentation in *Drosophila*.

Further, as can be seen from table 5.1, all the main regulators of stripe 2 are recovered. However, a negative coefficient for BCD identifies it as a repressor rather than an activator as expected. This is discussed further in section 6.2.

	Estimate	Std. Error
Intercept	-3.39	0.43
HB	-59.45	4.36
KR	13.62	0.87
GT	11.67	0.7
sob	-6.63	0.61
slp2	9.74	1.3

Table 5.4: Final model after stepwise selection for *eve* stripes 4 and 6.

	Estimate	Std. Error
Intercept	-0.04	0.61
HB sq	-488.22	29.79
KR	17.75	1.03
kni	-8.18	0.89
sob	-6.79	0.65
GT	8.54	0.95

Table 5.5: Final model after stepwise selection for *eve* stripes 4 and 6 with a quadratic term for HB.

For *eve* 3+7 and *eve* 4+6 (tables 5.2 and 5.4), one of the key regulators, HB, is found, but another, *kni*, is only found for *eve* 3+7. Interestingly, *kni* is selected for both of these enhancers after adding a quadratic term for HB (tables 5.3 and 5.5). This suggests that dual regulation might be important for *eve* 4+6 as well.

Thus, overall, stepwise selection performed quite well, but there are a few problems. The main difficulty is that the penalty used is arbitrary. Even with the relatively strict penalty, there are regulators in each model in addition to the known regulators and it is not clear whether these are relevant or not. A further weakness of the current approach is that when adding a quadratic term (such as for HB) during stepwise selection, it might be best if the linear term was included if it was not in the model already.

In spite of these weaknesses, the results do suggest that model selection is a valid approach for recovering regulatory relationships. However, rather than enhancing the algorithm for automatic stepwise selection to address its shortcomings, it is simpler and more transparent to take a more manual and careful approach to stepwise selection. This is done in chapter 6. Another option is to search more comprehensively for good models. This is the basis for a custom approach which is presented later in this chapter. Both of these depend on an algorithm that can constrain the parameters of the models, and this is now introduced.

5.3 ALGORITHM FOR CONSTRAINED PARAMETER OPTIMISATION

As shown in section 4.2.2, a quadratic model with too many regulators can lead to a model that does not generalise well to mutant predictions. One approach is to keep the number of regulators in the model small, examining the behaviour of each closely. However, if one wishes to search across many different regulators without scrutinising each one, it is necessary to introduce some restrictions to the behaviour of the quadratic term. This section explains the restrictions that were introduced and an algorithm that implements them.

5.3.1 *Maximum likelihood and parameter constraints*

The basis for the algorithm of this section is that logistic regression models can be fitted using maximum likelihood. The standard algorithm used thus far (via `glm` in R) uses iteratively reweighted least squares (IRLS) to maximise the likelihood, but a more direct optimisation approach can be used as suggested in Venables and Ripley (2003). In particular, the BFGS algorithm[§] was used, as supplied by the function `optim` in the package *stats* in R. The BFGS algorithm is a quasi-Newton method which requires the function and its gradient to

[§] Published simultaneously and separately in 1970 by Broyden, Fletcher, Goldfarb and Shanna. See Nocedal (2006).

be given. It then numerically optimises the function by finding where its gradient is zero.

The BFGS algorithm, however, does not support constraints on the parameters. One option is to use the L-BFGS-B (Byrd et al., 1995) algorithm, which is a modification of BFGS to support *box constraints*; that is upper and lower bounds can be placed on each parameter. A more flexible alternative (which can also handle box constraints) is the function `constrOptim` (Lange, 1999), again in the *stats* package of R. This provides a wrapper to BFGS, and allows linear inequality constraints to be enforced. This means that linear combinations of the parameter values can be restricted to lie above or below a specified value. Which constraints, then, are biologically relevant?

5.3.2 *Restrictions on quadratic parameters*

Hunchback dual regulation is supposed to activate at low concentrations and repress at high, rather than the converse. This might be explained by a mechanism where the regulator forms dimers (on or off DNA) at higher concentrations thus masking its own activating sites. Fortunately, it is relatively straightforward to relate dual regulatory mechanisms to the quadratic terms. To make it more concrete, quadratic and linear terms for x, the concentration of a regulator, are

$$ax^2 + bx \tag{5.2}$$

where a and b are the parameters of the quadratic and linear terms respectively. This function will be called the *regulator function.* If the regulator function is concave (opening downwards), then the regulator activates (positive values for the log odds of being ON) at low concentrations and represses at high. This corresponds to a negative coefficient a of the quadratic term. Preventing the coefficient of the quadratic from being positive allows it to function, over the range of concentrations in the data, as an activator or repressor (if $a = 0$), or as a dual regulator that activates at low concentrations and represses at high. Thus the constraint being added is that $a \leq 0$.

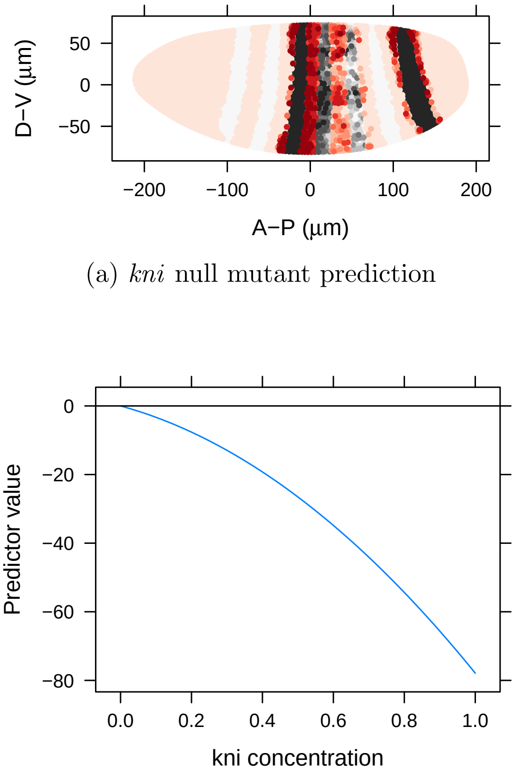
Constraints can address a further problem that was previously observed. One of the problematic behaviours shown in section 4.2.2 (figure 4.8) is that *kni* is an activator only for very low concentrations becoming an extreme repressor at intermediate values of concentration. This might fit the data, but it is biologically implausible. It is thus useful to introduce a moderating constraint. One option is to control where the function crosses the x-axis (i.e., reaches zero), which is at $x = 0$ and $x = \frac{-b}{a}$. Restricting the minimum value of the latter would correspond to allowing dual regulation as long as the regulator is an activator over a reasonable range of values in the data set. The difficulty with this is that the constraint on x should allow the optimisation procedure to set $a = 0$ (that is, the regulator behaves linearly), but this would produce division by zero in the constraint. An alternative constraint is to restrict the minimum that the function reaches over the range 0 to 1 (the concentrations in the *Virtual Embryo* are relative and are scaled between 0 and 1; see section 1.5). Since a is negative (or zero), the minimum that matters is at $x = 1$ (the value of the intercept term is not considered here, so the function is zero at $x = 0$). Thus, if c is the minimum value, the relevant linear inequality constraint is $a + b \geq c$.

Finally, in chapter 6, it is desirable to restrict the coefficients of certain linear regulators according to whether they are activators or repressors. These can be implemented simply by restricting the b coefficient to negative values for repressors and positive for activators.

Implementation details

These constraints were implemented in a wrapper function for the R function `constrOptim`. A different implementation was also tried using only *box constraints* (with `optim` and L-BFGS-B). This was found to produce similar results (although much more quickly), since in most cases the minimum constraint on the regulator function was not necessary. However, the minimum constraint is useful for exploring alternative models as in chapter 6. A further difference was that `constrOptim` can handle infinite and NaN[¶] values for the likelihood function. This is relevant in the case where the model produces a fitted value of 0 or 1 for any nucleus (at the level of numerical accuracy). When this occurs, the log likelihood function (equation 5.1) will be minus infinity $(0 \times \log 1 + 1 \times \log 0)$ when it disagrees with the training data for the nucleus, or NaN in the case that it agrees $(1 \times \log 1 + 0 \times \log 0)$. In both cases, `constrOptim` rejects this as a solution, and continues with optimising other parameters. L-BFGS-B, however, requires these cases to be dealt with specifically. Thus, since `constrOptim` can handle this situation and is more flexible, it was used to produce the results of this thesis. In order to improve convergence (usually in the case of models with poor fits), the `outer.eps` parameter was set to 1×10^{-3} instead of the default 1×10^{-5}. This controls the relative convergence of the outer algorithm. Specifically, this means that if the relative improvement in log likelihood is less than this, the algorithm is considered to have converged. This change does not practically affect the results since precise values of the parameters are not important, yet it does help with convergence in certain (poor-fitting) cases.

[¶] Such as $0/0$.



(a) *kni* null mutant prediction

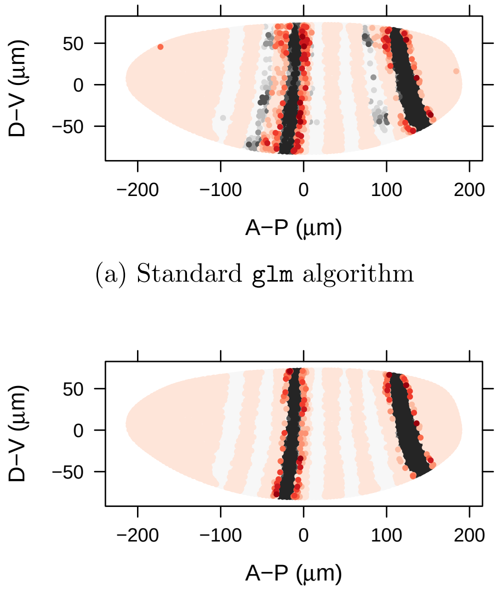
(b) *kni* behaviour in the model

Figure 5.1: Model results for *eve* 3+7 with 11 regulators and constrained parameters. For comparison with figures 4.7 and 4.8.

5.3.3 *Algorithm proof of concept*

It is now instructive to test the algorithm on the problem of section 4.3 on page 79 where training with a number of quadratic terms led to an unlikely *kni* null mutant prediction. Section 4.3.1 showed that one solution is to restrict the number of quadratic terms added to the model. This section demonstrates that constraining the parameters improves the plausibility of the model predictions even with many regulators. In order to test the algorithm, the minimum value of the regulator function was set at -300 (i.e., $ax^2 + bx \geq -300$).

The first test is with the set of 11 regulators used for the predictions of figures 4.6, 4.7 and 4.8 (from page 79). In the unconstrained predictions, *kni* functioned as an activator at low concentrations, but quickly



(a) Standard `glm` algorithm

(b) Custom fit with constrained parameters

Figure 5.2: Comparison of fits from `glm` and the custom algorithm for *eve* 3+7 with 38 regulators, all with quadratic terms.

became a strong repressor. Correspondingly, the null mutant prediction suggested that *kni* was an activator for stripe 3. By constraining the parameters, the results are biologically much more reasonable. The results are shown in figure 5.1 where *kni* functions clearly as a repressor between the stripes as expected (Clyde et al., 2003). In fact, it is relatively linear over its concentration range in the data.

The second test for the constrained parameter algorithm is that it performs well even with all 38 candidate regulators. Figure 5.2 contrasts the results of the standard `glm` algorithm to the fit of the custom algorithm. It is apparent that when the parameters are constrained the overall prediction is more reasonable.

Thus, it is clear that constraining the parameters is a reasonable strategy for automatic model selection. Interestingly, although the minimum value of the regulator function $(ax^2 + bx)$ was constrained here, the results of this section hold even when this is not the case. In other words, it is the restriction on the dual regulatory form that makes the substantial difference. The other form of dual regulation (i.e., repression at low concentration and activation at high) leads to less plausible predictions.

5.3.4 *Flexibility of the model in fitting pairs of stripes*

Before using the algorithm for model selection, as an aside, it is interesting to consider what type of flexibility is inherent in the model for fitting any pair of stripes. This is similar to what was done to produce figure 4.11 on page 83 using a restricted set of regulators. Here the results are for all 38 candidate regulators. Stripes 3+7 do not fit as well when the minimum of the regulator function is restricted (see section 6.3), and so in this case, the regulator function was relatively unrestricted‖. Figure 5.3 shows the results with all quadratic terms added, whereas figure 5.4 has quadratic terms only for HB and BCD (see section 6.2 for why BCD has been included).

There are some observations that suggest interesting areas of further research (see section 8.3), but for now, it is clear that pairs of stripes do not have unlimited flexibility in being combined together. Stripes 4+6 and 3+7 are known to be regulated together, so the main exceptions are that 2, 5 and 7 appear to have regulator concentrations in common, and that stripe 1 can be linked with many stripes (this is most clear in figure 5.4).

In preparation for the next section, the most relevant conclusion from these results is that including all quadratic terms does not lead to better results. On the one hand, this suggests that large scale model selection need not consider dual regulation from factors other than BCD and HB, but on the other, it suggests that this might not harm the model selection process. However, since it is easier to interpret models with only two dual regulating factors, it is this simpler approach that will be taken for the next section.

‖ Minimum value of -2000.

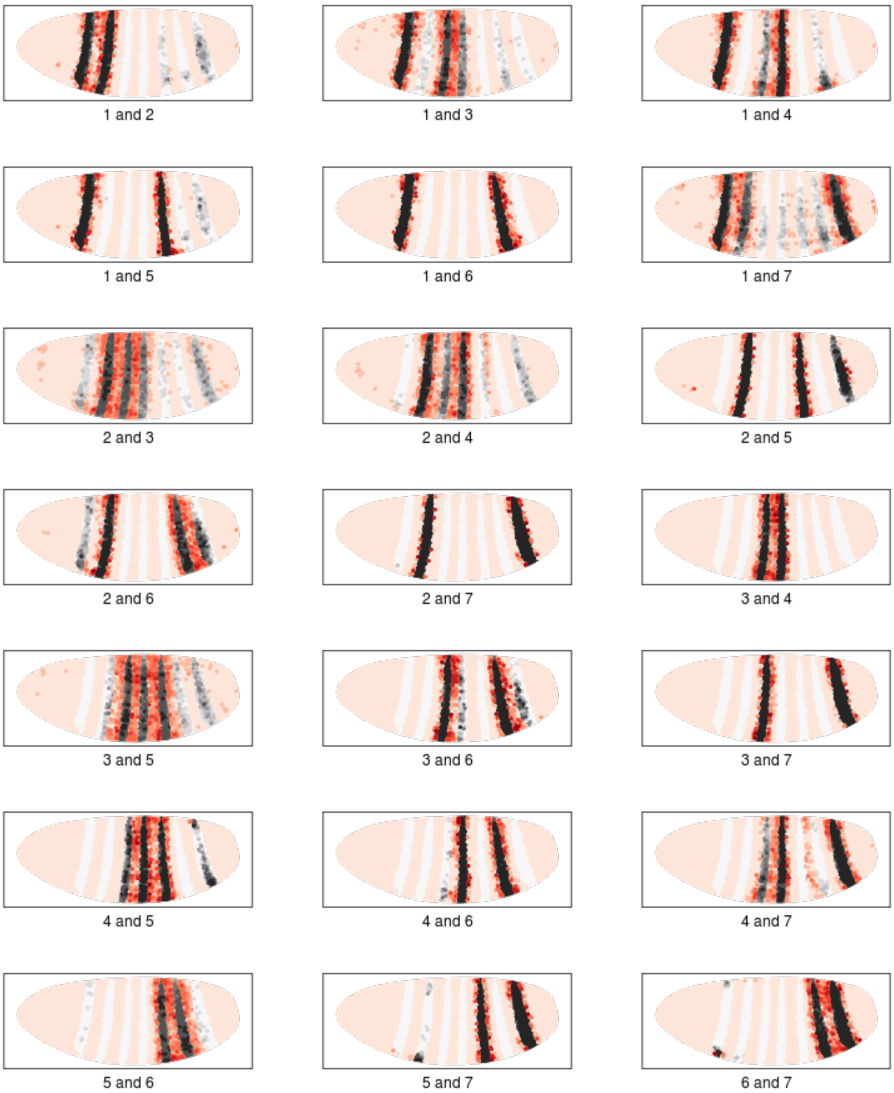


Figure 5.3: Model training and prediction for all pairs of stripes with the custom algorithm with 38 regulators and all quadratic terms.

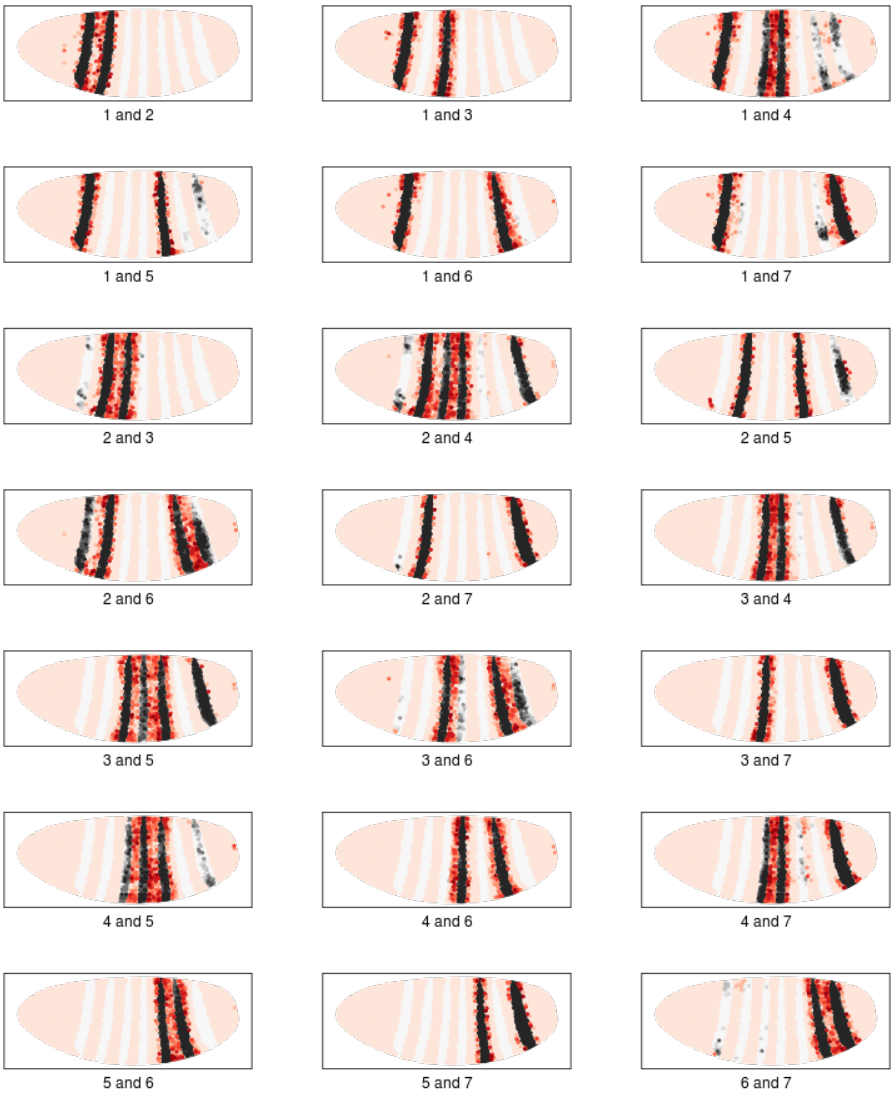


Figure 5.4: Model training and prediction for all pairs of stripes with the custom algorithm, 38 regulators and quadratic terms for HB and BCD.

5.4 EXHAUSTIVE SUBSET SELECTION

It is not feasible to consider all possible models. For 38 regulators, this amounts to 2.7×10^{11} models, which, at one second a model, would take over 8,000 years to evaluate. Stepwise selection dramatically reduces the number of models considered by only adding or removing one regulator at a time. This section takes a different approach which, as will be seen, is particularly effective for discovering the primary regulators of the *eve* stripes. It considers all possible models (*exhaustive*) of a particular size (the *subset*).

5.4.1 *Size of the models in the subset*

Fitting small numbers of regulators to the *eve* stripes is usually sufficient to produce a good fit. Only four regulators are required for a good fit for stripe 2, and three for *eve* 3+7 (see chapter 6). It is therefore reasonable to suppose that four is a good size for the models of the subset. For 38 regulators, this results in 73,815 possible models. Considering a subset model size of five results in far more: 501,942 models. Since each model can be fitted quickly (of the order of seconds) this makes it possible to consider all models of four regulators within a day or two. All models of five would take a week or two, and as will be demonstrated, this is not necessary for finding the primary regulators.

Thus, all four regulators were combined to consider all possible linear models, but as shown in chapter 4, dual regulation might be important. Therefore, a further subset of models were tested. Whenever HB or BCD were included in a model, their corresponding quadratic terms were included as well. (The inclusion of BCD is justified in section 6.2.) The algorithm of section 5.3 was used for fitting. Since only two quadratic terms were included with only four regulators considered, these models were not restricted in the minimum value the regulator function could contribute, but were only restricted to follow the standard dual regulatory form (activation at low concentration, and repression at high). Model selection was also done for the subset where all quadratic terms are added, and the results are shown in the appendix, section B.2.

5.4.2 *Scoring models*

The models were all scored and compared using their maximised log likelihood, which is equivalent to using AIC. AIC can be used for comparing different models even if they are not nested (Dobson and Barnett, 2008), and since the penalty is the same (they are all the same size), comparing AIC is equivalent to comparing maximised log likelihood.

5.4.3 *Filtering of the model scores*

The main difficulty is filtering all the model scores to make a meaningful conclusion. One way is to examine the top models (say, the top 100). This works well (results given in the appendix, section B.1) but the cutoff is arbitrary. A more informative approach is introduced here, which summarises all the model scores and identifies groups of regulators that work well together. This is done by taking the best model prediction for each pair of regulators across all possible models of four regulators that include the pair. The best score for each pair is then shown in a heat map. The basis for selecting a pair is a key assumption of how the *eve* stripes are defined: it is supposed that the borders are controlled by pairs of repressors—GT and KR in the case of *eve* 2, and HB and KNI in the case of *eve* 3+7 and *eve* 4+6 (Small et al., 1992; Clyde et al., 2003). Importantly, these regulators depend on broadly distributed activators, and so regulator context is important.

This provides further justification for considering models no bigger than four regulators. If one is interested in the maximum score of each pair in the context of other regulators, then it is possible to provide a relatively straightforward interpretation of the heat map in the case where there are no more than two other regulators for each pair. Each best score of the heat map is based on four regulators, which for the sake of explanation will be considered to consist of the *fixed pair* (which can be read off from the axes on the heat map) and the *context pair* (the hidden pair that provides the best score for the fixed pair). All scores on the heat map are relative so that intensity scales from the minimum actual score to the maximum actual score. Thus, if a fixed pair scores relatively poorly on the heat map, it means it provides little improvement on the best performing regulator pair (since every model represented in the heat map will be at least as good as the best performing model of only two regulators; every model in the heat map can make use of this pair). Conversely, if a fixed pair scores well then at least one of the fixed pair makes an important contribution to the context pair. If the identity of the fourth regulator is relatively unimportant, then one can expect to see darker bands on the heat map corresponding to the important third regulator. Better scores (darker patches) within these bands will identify where the fourth regulator is more informative. This is, in fact, what is seen in the results. It should also be noted that a useful by-product of this approach is that in more complicated situations when things are less clear-cut, the heat map itself will be less clear. It is thus a useful way of assessing the overall success of model selection.

5.4.4 *Results*

The heat map (figure 5.5, page 110) for the linear models of stripe 2 is remarkably clear-cut. Broad bands of darker colour run across the heat map for three regulators: BCD, HB and GT. These are three of the four primary regulators from Small et al. (1992). The fourth regulator, KR does appear as a darker patch within the bands of the other three, but it does not stand out from other possibilities. However, including a dual-regulating HB and BCD brings it to the fore. This, and the absence of *slp1*, a proposed regulator in Andrioli et al. (2002), are discussed further in section 6.2.

It is thought that KNI and HB act as repressors to define both *eve* stripes 3+7 and stripes 4+6 (Small et al., 1996; Clyde et al., 2003). The results for stripes 4+6 (figure 5.7, page 112) are consistent with this, although they indicate additional roles for GT and KR, two other gap genes. The stripes 3+7 results (figure 5.6, page 111) are a little more complicated. In the linear case, GT and *kni* are important, as is *tll*. HB only stands out in the context of these regulators. In the quadratic case, *kni* remains important, although somewhat less so. KR replaces GT as valuable and *tll* gains further importance. These issues are discussed in depth in section 6.3.

eve stripe 1 and stripe 5 are not examined in detail in this thesis. The results for stripe 1 (figure 5.8, page 113) require further investigation, especially since some of the regulators, such as *trn*, might be downstream of *eve* rather than vice versa. It is worth noting, however, that a quadratic BCD could be important in its regulation, along with KR, *slp1*, *slp2* and *knrl*.

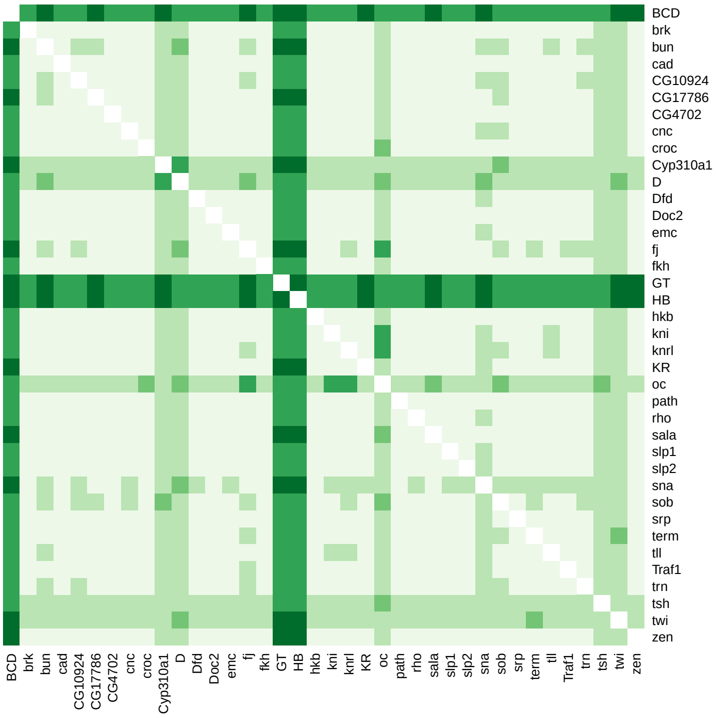
The results, though, for *eve* 5 (figure 5.9, page 114) are clear. Likely regulators include BCD, GT, KR, *kni* and *knrl*. Dual regulation by HB or BCD does not seem important.

Finally, it is worth briefly commenting on the model selection results for all quadratic terms (see the appendix, section B.2). Overall the results are broadly consistent, although they are not as easy to relate to the known regulators. It would seem that only a few regulators benefit from having quadratic terms. It is not clear whether dual regulation by these regulators is of significance or not, but as the next chapter will demonstrate, it is not necessary to include these for good fits of *eve* stripes 2 and 3+7.

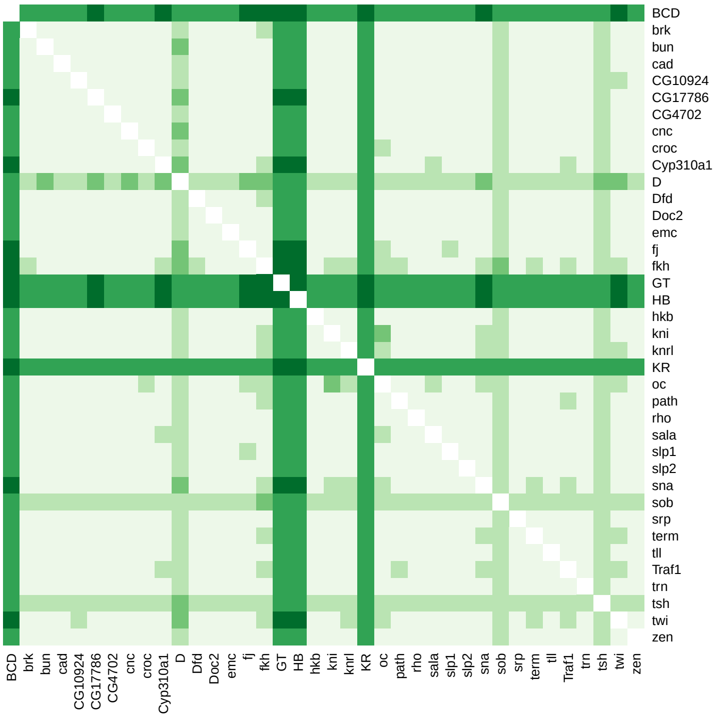
5.4.5 *Comparison with BDTNP regulator discovery*

Before examining the *eve* 2 and *eve* 3+7 enhancers in more detail in the next chapter, it is worth noting possible reasons for the success of the model selection approach presented here compared with the method in the paper announcing the *Virtual Embryo* data (Fowlkes et al., 2008). Their results are shown in figure 1.8 on page 20.

Their method was similar in that they also used protein measurements where available, but they considered a smaller set of regulators (17). These regulators did, however, include all the regulators that the models of this thesis have found to be important. Their method was further similar in the form of their function, but the first major difference is that their training data used a continuous response for *eve*. As introduced in chapter 2, the discretisation approach used here may be more robust to inaccuracies in the levels of *eve* mRNA (primarily from spatial and temporal registration). The issue of temporal correspondence between nuclei presents a further difference. This method is cohort-specific in order to mitigate the uncertainty; their method appears to average data across all cohorts, which will blur regulatory relationships if there are problems linking nuclei and measurements across the cohorts. Finally, one of the hypotheses of this thesis is that dual regulation is an important regulatory mechanism in the case of HB and BCD. Fowlkes et al. (2008) did try model fitting with quadratic terms and found the fits better, but did not present those results as they were harder to interpret.

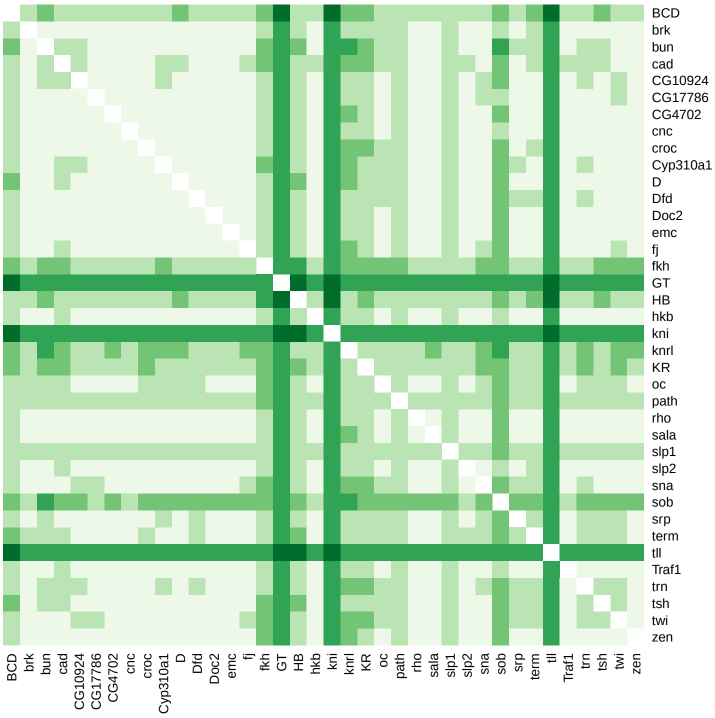


(a) Linear

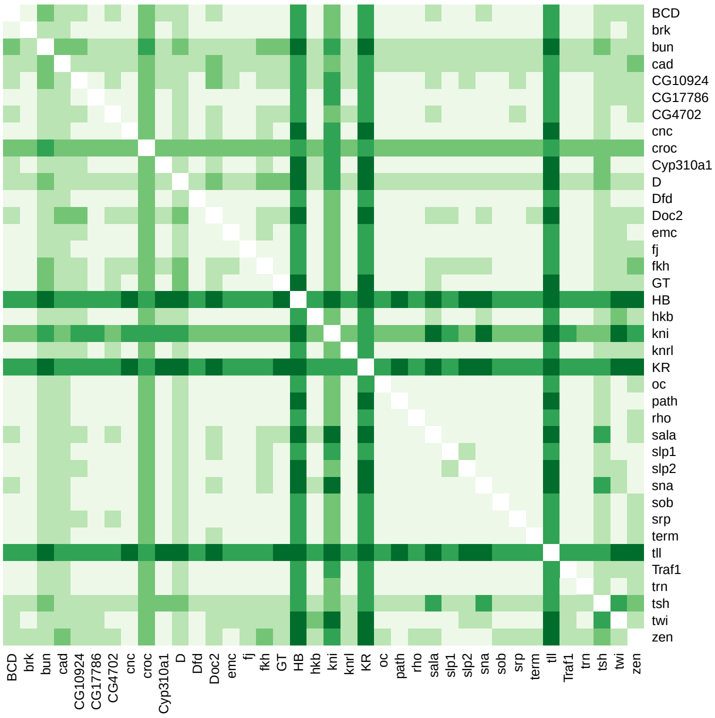


(b) With quadratic HB and BCD

Figure 5.5: Best performing regulators for *eve* stripe 2.

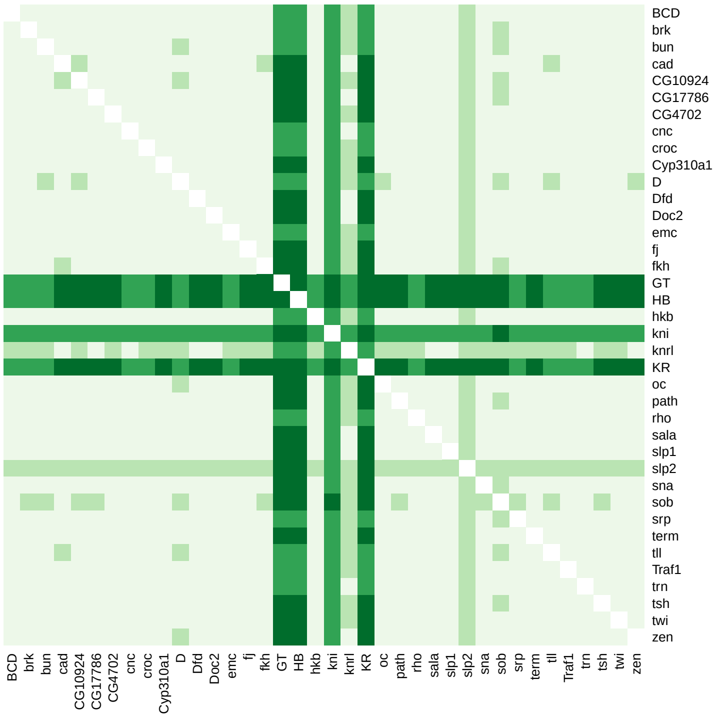


(a) Linear

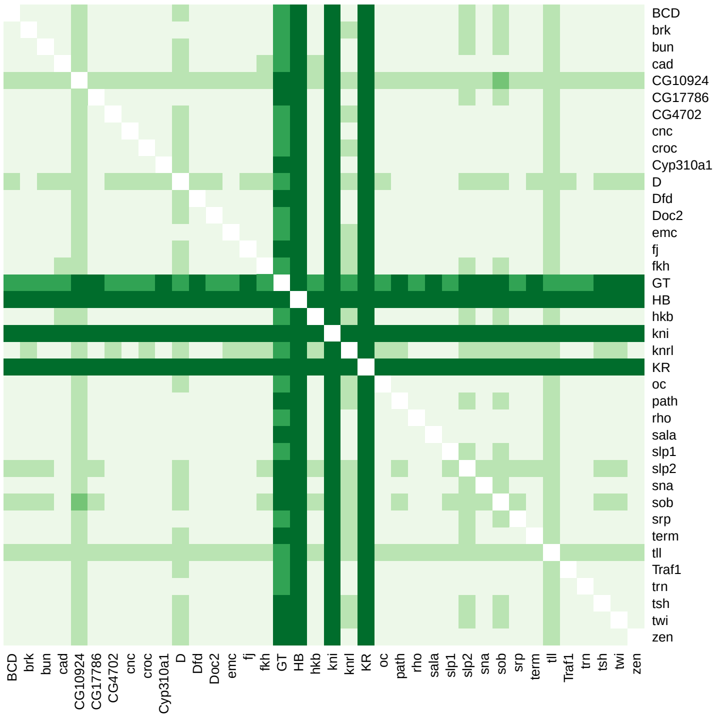


(b) With quadratic HB and BCD

Figure 5.6: Best performing regulators for *eve* stripes 3+7.

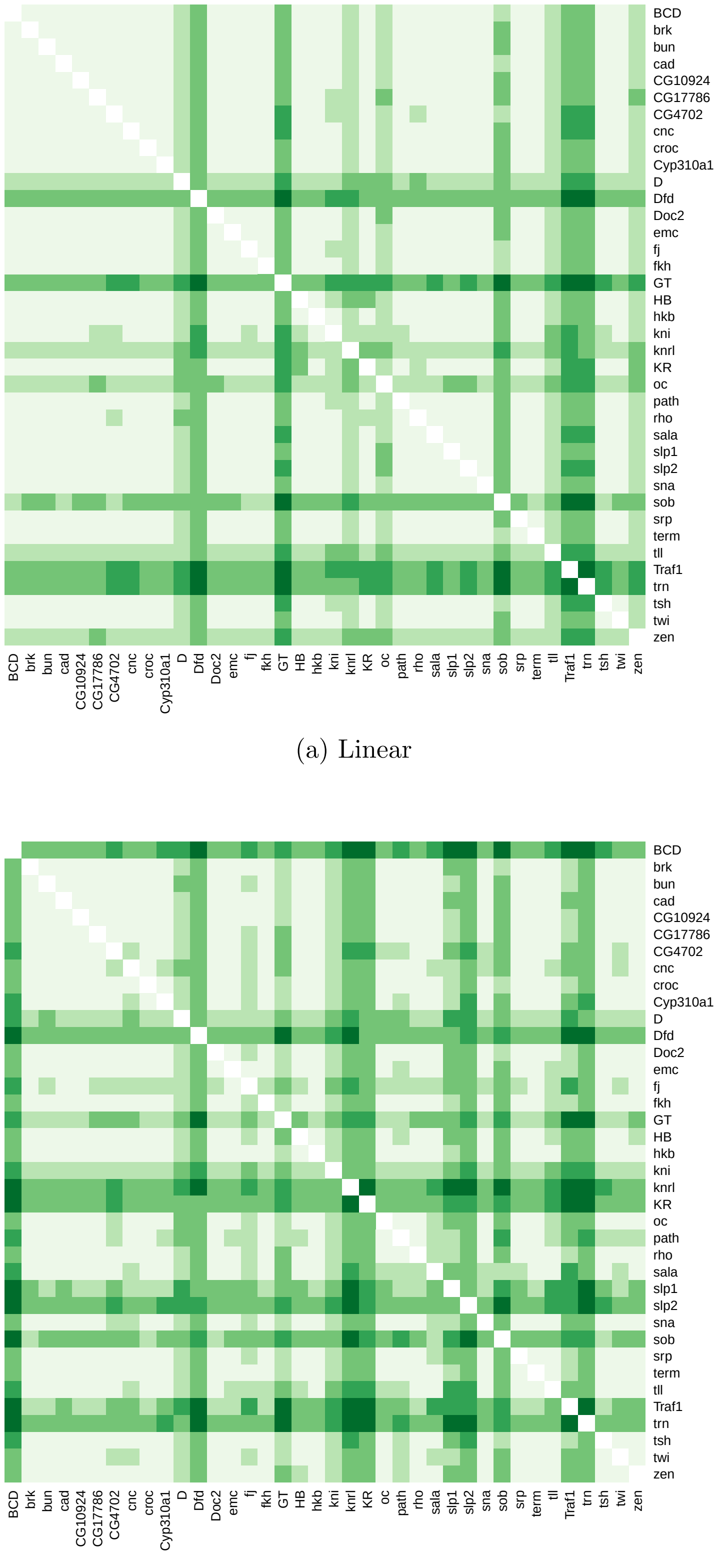


(a) Linear



(b) With quadratic HB and BCD

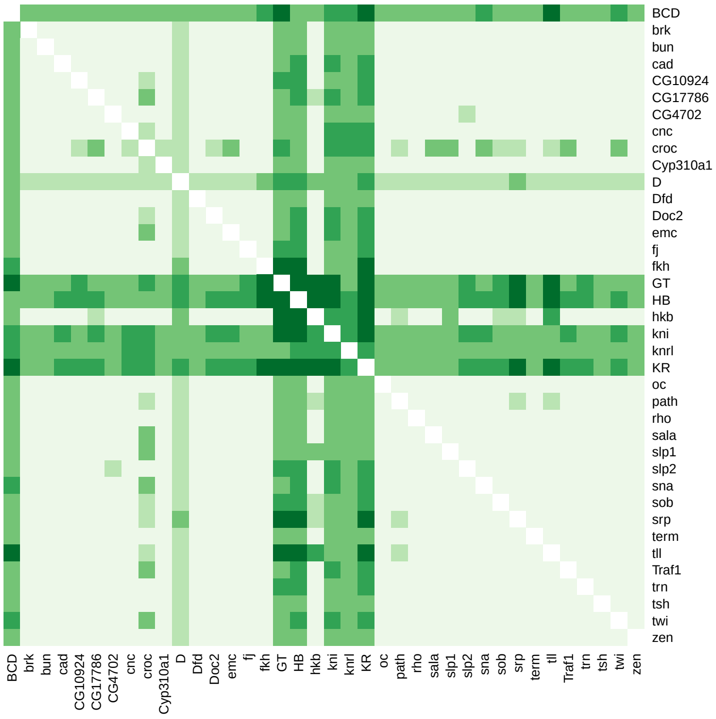
Figure 5.7: Best performing regulators for *eve* stripes 4+6.



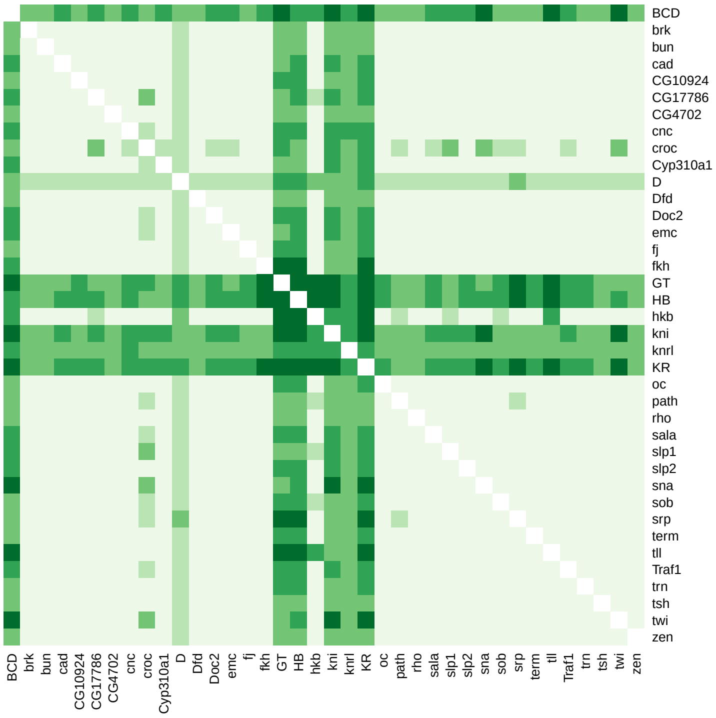
(a) Linear

(b) With quadratic HB and BCD

Figure 5.8: Best performing regulators for *eve* stripe 1.



(a) Linear



(b) With quadratic HB and BCD

Figure 5.9: Best performing regulators for *eve* stripe 5.

6

MODEL SELECTION: BIOLOGICAL EXPLANATION

The emphasis of the previous chapter was regulator discovery. In this context, automatic stepwise selection was found to be limited, and so an alternative approach was introduced that was far more effective. The focus of this chapter, however, is biological explanation: the model should be able to explain experimental observations for both wild-type and mutant embryos. As will be shown here, stepwise selection can be remarkably successful in building up a successful explanatory model when a more careful, biologically-informed approach is taken.

The process of model selection also provides the opportunity to consider some important questions regarding enhancer function and the specification of the *eve* stripes. In particular, the following are considered:

- Can a single function describe enhancer function across the whole embryo? Andrioli et al. (2002) introduce a qualitative model that suggests this is not feasible for *eve* 2. An alternative model will be proposed here that shows it is indeed possible if BCD is a dual regulator of *eve* 2.

- A prevalent model is that the borders of the *eve* stripes are defined by opposing gradients of repressors. This has been proposed for *eve* 2 (Stanojevic et al., 1991; Small et al., 1992; Andrioli et al., 2002), *eve* 3+7 (Small et al., 1996; Clyde et al., 2003) and *eve* 4+6 (Fujioka et al., 1999; Clyde et al., 2003). This chapter will demonstrate the suitability of this model for *eve* 2, but that it is less satisfactory for *eve* 3+7. Instead, it is proposed that dual regulation by HB is important for precision.

- Section 5.4 found *tll* as a potential regulator of *eve* 3+7 in the dual-regulating model. This chapter evaluates this more fully, showing that repression by TLL is consistent with experimental observations.

6.1 INTRODUCTION TO THE PROCEDURE

Similar to the method of section 5.2, the procedure followed in this chapter begins with an intercept-only model* and adds regulators progressively. However, rather than relying on automatic selection and stopping according to a penalised likelihood-based score (e.g., AIC), the results of each selection step are considered carefully. In this chapter, the log likelihood score of the top models of each step will be shown in a table (e.g., table 6.1 on page 120). If the coefficient of a putative regulator is positive (an activator), its regulator name is shown in orange. Blue is used for negative coefficients (repressors) and black for dual-acting regulators. The top model is then selected if it is obviously better than the alternatives. Model selection is considered complete when no further regulators stand out and the visual fit of the model is appropriately good.

Models are evaluated using two types of experiments: null mutant expression results and misexpression studies. Both perturb the inputs to the enhancer function, and hence, if the enhancer function is modelling mechanism it is reasonable to expect that it should predict direct relationships correctly. Indirect relationships, however, will require more care in interpretation. In order to understand the values that are used for producing model predictions, it is helpful to be reminded that the measurements in the *Virtual Embryo* are relative. They are scaled from

* No regulators in the model. β_0 is the only parameter.

0 to 1 for each probe where 1 is the maximum measurement for that probe across all time points. Thus, for example, a value of 0.2 means 20% of the maximum value of all nuclei at all time points for that probe. The procedure will now be applied to *eve* 2 and *eve* 3+7 in turn.

6.2 *eve* 2

The main weaknesses of the linear classifier model of chapter 3 is that Bicoid (BCD) functions as a repressor (see figure 3.14 on page 72), although it is known as an activator (Small et al., 1992). Yet, in spite of this, it is still found as a primary regulator during automatic selection of linear models (see section 5.4.4) suggesting that repression by BCD is important.

This section will consider model selection under two different hypotheses that offer a solution to this discrepancy. The first hypothesis is the qualitative model of Andrioli et al. (2002). In this model (see figure 6.1), the embryo anterior to stripe 2 is divided into three regions. The region most adjacent to the stripe requires GT for repression, which defines the border of stripe 2. The next anterior region requires SLP1 for repression, together with an unknown regulator. Finally, the anterior-most region requires a factor, perhaps Torso, that interferes with BCD-dependent activation.

This is contrasted with a second hypothesis which is introduced here: BCD is a dual regulator of *eve* 2. As will be shown, this hypothesis is consistent with the experimental data considered and can explain the functioning of the *eve* 2 enhancer without requiring additional position-specific mechanisms.

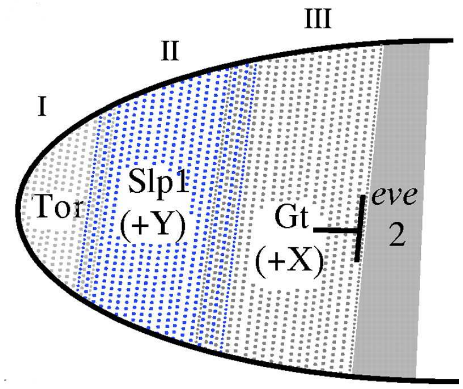


Figure 6.1: The *eve* 2 model of Andrioli et al. (2002) proposes three regions of repression. The activity of GT and at least one other factor (X) is required for repression very near the anterior border. Anterior to this domain, SLP1 and at least one other factor (Y) mediate *eve* 2 repression. At the anterior pole, TOR activity may downregulate the activity of BCD, the primary activator of *eve* 2. Figure from Andrioli et al. (2002), reprinted by permission from *Development*.

6.2.1 *Search for a repressor pair*

Even though the model of Andrioli et al. (2002) divides the embryo into three regions, the specification of the borders of the stripe follows the earlier models of Stanojevic et al. (1991) and Small et al. (1992): the borders of the stripe are precisely defined by a pair of repressors (GT and KR). It is thus instructive to begin model selection with a search for all pairs of repressors that are able to define the stripe border. To do this, the training data were restricted to the stripe and closely neighbouring nuclei. Since the *eve* stripes are not orthogonal to the A-P axis, but are slanted, it is not appropriate to define this using A-P axis alone (particularly if the goal is to model expression around the whole stripe). Thus, a graph (or network) of nuclei was constructed from the nuclei adjacency information in the *Virtual Embryo*. The training data was then set up to include *eve* stripe 2 and all nuclei within three neighbours† of the stripes. (Additionally, the expression

† Graph edges.

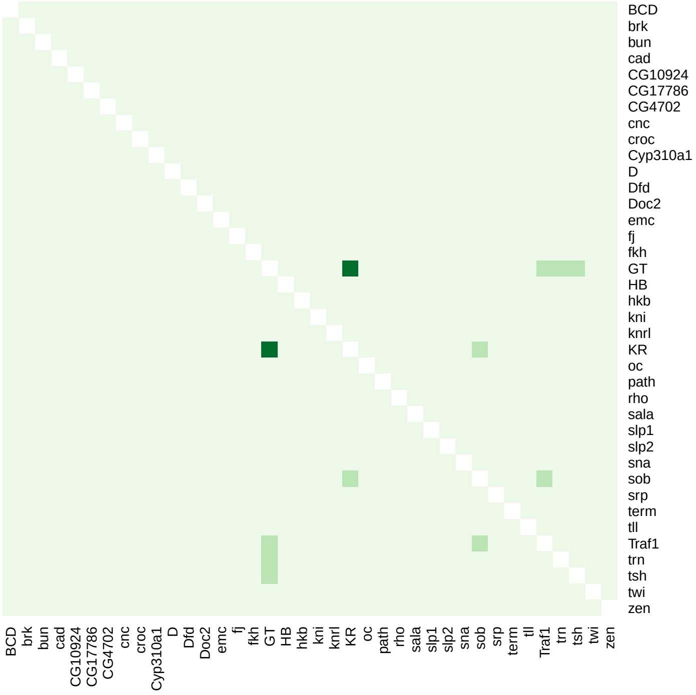


Figure 6.2: Best performing *eve* 2 models with only two repressors.

of all nuclei outside of stripe 2 were set to OFF since a few nuclei from stripes one and three were included.) Finally, the algorithm described in section 5.3 allows constraints to be placed on the regulators. For the initial search, the regulators were constrained to be repressors, in other words, to have negative coefficients. As can be seen from the heat map of the resulting scores of the relevant two regulator models (figure 6.2), it is quite clear that GT and KR are the only repressor pair able to define the borders of *eve* 2. This strongly supports the model that the borders of the stripe are defined by a repressor pair.

6.2.2 *Search for broadly-distributed activators*

However, according to Small et al. (1992), the repressor pair is not the only contribution to positional information; other regulators act over a wider region, including HB and BCD as activators. Thus, the next step was to look for activators (or perhaps other repressors) that are

Probe	HB	BCD	*path*	*tsh*	*slp2*	*croc*
Log likelihood	-164	-595	-838	-950	-965	-1063

Table 6.1: Best regulators for *eve* 2 in addition to KR and GT for a restricted region.

Probe	HB	*path*	*tsh*	BCD	*fj*	*croc*
Log likelihood	-407	-917	-965	-1072	-1082	-1084

Table 6.2: Best regulators *eve* 2 in addition to KR and GT for a larger region.

important when the whole embryo is considered. To achieve this, the training set was widened across the embryo (including other stripes, which were set to zero), but in accordance with Andrioli et al. (2002), the anterior-most region was excluded‡. The models considered all included GT and KR (still constrained to function as repressors). The remaining regulators were unconstrained. Table 6.1 shows the results. (In this chapter, orange regulator names are for activators, blue for repressors and black for dual-acting regulators). HB and BCD stand out as the best regulators, with importantly, BCD as an activator. However, as expected from (Andrioli et al., 2002), the result is sensitive to the region selected for training. If the region extends further to the anterior§, then BCD is no longer as useful. The result of this is shown in table 6.2.

If HB and BCD are added, but constrained to be activators, keeping GT and KR as repressors and using the same region as for table 6.2, then SLP1 is a potential repressor, although not convincingly so, as shown in table 6.3. Continuing with stepwise selection does not lead to the inclusion of SLP1, and nor does using different regions for training¶.

‡ Excluding those nuclei with an x coordinate of less than -110.

§ Including $x > -140$.

¶ Each region considered included *eve* stripe 2 so that there were positive examples in the training data.

Probe	*knrl*	*oc*	*kni*	*srp*	*CG10924*	*slp1*
Log likelihood	-282	-312	-317	-326	-330	-337

Table 6.3: Best regulators for *eve* 2 in addition to KR, GT, BCD and HB for the larger region.

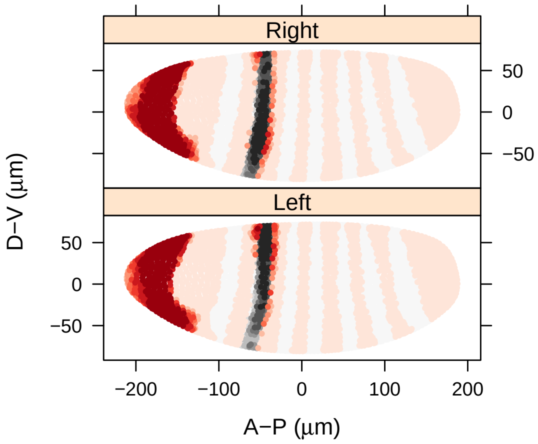


Figure 6.3: Prediction of *eve* 2 using a linear model with KR, GT, HB, SLP1 and BCD. BCD and HB are constrained to be activators, and KR and GT repressors.

These results, then, are broadly consistent with Andrioli et al. (2002). However, although regulator discovery has been based on a linear classifier, successful prediction across the whole embryo cannot be made using the same approach (see figure 6.3). In essence, the embryo needs to be divided into regions, each with a separate model with different parameters for the regulators so that BCD can function as a repressor in the anterior-most region and as an activator around stripe 2. Finally, each of these are stitched together to produce an overall prediction. This is dissatisfying since the goal is to model the behaviour of an enhancer with a single function that applies to the whole embryo (i.e., for any nucleus of the embryo). One possible solution is to introduce interactions, perhaps with unknown factors. However, including a single dual regulator can produce a simpler explanation, and this is what shall now be done.

Probe	BCD	HB	*tsh*	*fj*	*sala*	*kni*
Log likelihood	-290	-945	-1053	-1125	-1143	-1164

Table 6.4: Best regulators for *eve* 2 in addition to KR and GT with dual-regulating BCD.

Probe	BCD	HB	*tsh*	*D*	*kni*	*slp2*
Log likelihood	-290	-537	-809	-912	-931	-1023

Table 6.5: Best regulators for *eve* 2 in addition to KR and GT with all others dual regulating.

6.2.3 *Dual-regulating Bicoid*

An alternative model is to suppose that BCD has dual regulatory capabilities similar to that introduced for HB in chapter 4. Model selection can then begin from the model that includes GT and KR as repressors, as with the searches leading to the results of tables 6.1 and 6.2. However, the whole embryo is now used for training data. For the purposes of this section, the minimum of the dual regulator function of BCD is unrestricted (see section 5.3.2) since this does not have unwanted effects on other regulators, which are all linear in the models considered. Also, restricting BCD repression was not found to affect the findings significantly. The results of the initial search are shown in table 6.4. Strikingly, BCD is now the best regulator. This is not purely because BCD has the most flexibility to fit the data. If all regulators (excluding KR and GT) are allowed to have dual regulatory capacity, then BCD is still the best regulator in addition to KR and GT (see table 6.5). One noticeable feature of table 6.4 (in contrast to table 6.5) is that HB is not clearly the next best regulator; for example, *tsh* is not that far behind, suggesting that HB benefits from having dual regulating capability more than the other regulators. Either way, BCD is a good choice, and is thus added at this step.

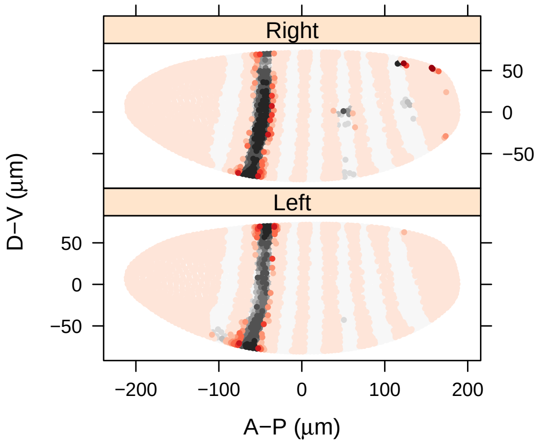


Figure 6.4: Prediction of *eve* 2 with KR, GT and a dual-regulating BCD.

Probe	HB	*path*	*slp2*	*croc*	*CG10924*	*D*
Log likelihood	-145	-257	-262	-263	-270	-270

Table 6.6: Best regulators for *eve* 2 in addition to KR, GT and a dual-regulating BCD.

Figure 6.4 shows that the resulting fit is still lacking precision. Table 6.6 shows that HB is clearly the next best choice, and this leads to a good visual fit (figure 6.5). This is confirmed by the results of table 6.7—no further regulator stands out.

Thus, all the primary regulators of *eve* 2 have been recovered (BCD, HB, KR and GT). As shown previously (figure 3.14, page 72), a linear classifier is unable to explain the results of a *bcd* mutant as BCD functions as a repressor in the model. Further, if BCD is constrained to function as an activator, it is unable to predict expression across the whole embryo (see figure 6.3). In contrast, the addition of dual-regulating BCD recovers the well-characterised regulators of *eve* stripe 2, and is

Probe	*path*	*zen*	*Traf1*	*slp2*	*sala*	*bun*
Log likelihood	-116	-123	-129	-134	-134	-135

Table 6.7: Best regulators for *eve* 2 in addition to KR, GT, HB and a dual-regulating BCD.

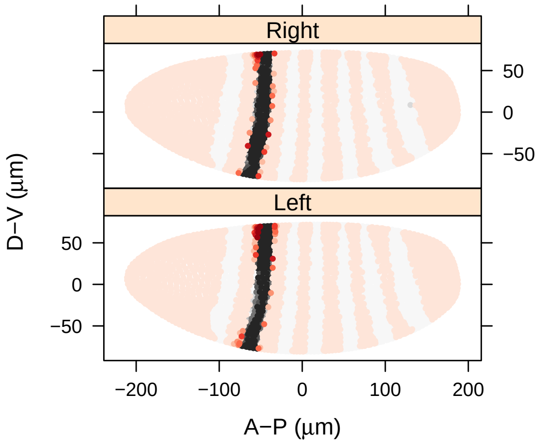


Figure 6.5: Prediction of *eve* 2 with KR, GT, HB and a dual-regulating BCD.

able to describe the function of the enhancer in all nuclei. This is arguably a simper result than dividing the embryo up into regulatory regions where different enhancer functions are applicable. Moreover, not only does it provide a simpler explanation, it is also able to explain mutant behaviour.

6.2.4 *Mutant predictions*

This section will consider a number of experimental observations relevant to the functioning of the *eve* 2 enhancer. Throughout, the qualitative model of Andrioli et al. (2002) will feature for comparison and evaluation.

bcd null mutant

Even though BCD is dual regulating, it still functions as an activator of stripe 2: in the null mutant prediction, stripe 2 is removed (see figure 6.6). It should be noted, though, that increasing BCD concentration does not reproduce shifts in stripe patterns as described in Driever and Nüsslein-Volhard (1988a). This is because BCD affects the gap genes that are crucial in defining the borders of the stripes in vivo; this

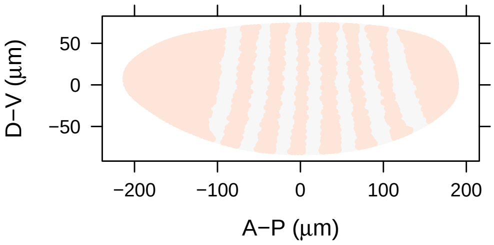


Figure 6.6: BCD null mutant prediction for *eve* 2 with dual-regulating BCD.

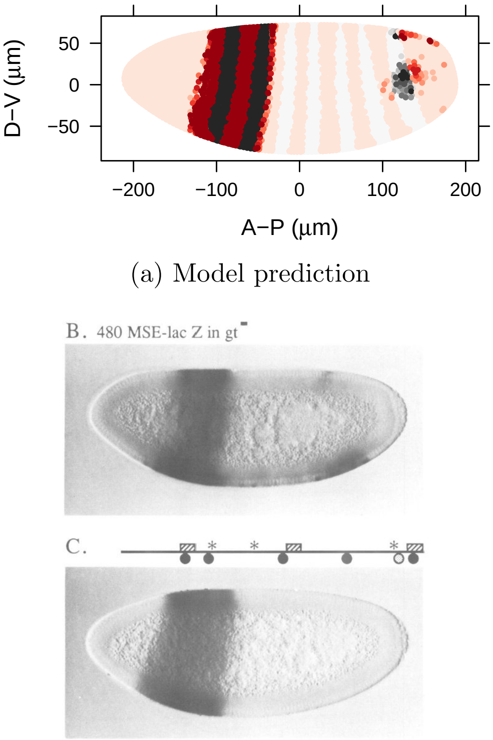
highlights the limitation of the current modelling approach for handling indirect effects.

gt null mutant

The next mutant to consider is *gt*. One of the reasons proposed in Andrioli et al. (2002) for multiple domains of repression is that removal of GT causes modest expansion in the region adjacent to the stripe, rather than all the way to the posterior tip. The dual-regulating BCD model reproduces this as shown in figure 6.7, which shows that to achieve anterior repression it is sufficient to have dual-regulating BCD. As an aside, it is interesting to note that introducing a dual-regulating KR moderates the effect of the null *gt* mutation, reducing the extent of anterior expansion.

Binding site mutation

However, anterior repression is only one aspect of the case for different domains of repression given in Andrioli et al. (2002). There is also evidence for SLP1 repression of *eve* 2. For example, a reporter gene containing a deleted site, presumably a binding site for SLP1, results in expression of *eve* in a similar region to where *slp1* is expressed. Although the explanation of SLP1 as the cause cannot be ruled out, it is possible to present an alternative explanation in the context of the dual-regulating BCD model. If one speculates that the binding site change affects the sensitivity of the enhancer to BCD, reducing the



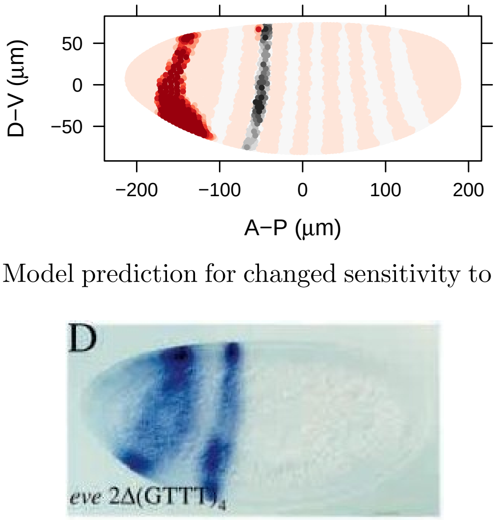
(a) Model prediction

(b) In situ hybridisation. Image B shows expression of an *eve* 2 reporter gene in a *gt* null mutant (*gt*$^{\text{YA82}}$ allele). Image C is the same reporter gene but with mutated GT binding sites. Figure from Small et al. (1992), reprinted by permission from Macmillan Publishers Ltd: The EMBO Journal © 1992.

Figure 6.7: *gt* null mutants for *eve* 2.

effective BCD concentration, say by 50%, then the model prediction qualitatively agrees with the experimental data (see figure 6.8). This prediction is somewhat sensitive to which part of the training data is used. In particular, the predictions can vary quantitatively depending on which part of the dorsoventral axis and which side of the embryo is included. However, the majority of these training subsets still produce a qualitatively similar result. The prediction shown was trained on the full embryo.

The main qualitative aspect not captured correctly is the direction of the 'curve' of *eve* expression in the anterior (and the weaker expression of stripe 2), indicating that the effect of mutating the binding site is not quite captured by changing the sensitivity of the model to a dual-regulating BCD. However, it is relevant that each lateral side of



(a) Model prediction for changed sensitivity to BCD

(b) Expression of *eve* 2 where a putative SLP1 binding site has been deleted. Figure from Andrioli et al. (2002), reprinted by permission from *Development*.

Figure 6.8: Comparison of expression in a *eve* 2 transgenic reporter with the model prediction.

the *Virtual Embryo* differs in how the contours of BCD measurements are oriented with respect to the *eve* stripes (figure 6.9). It is possible that measurements of gradients that are aligned differently to the *eve* stripes might be less accurate. Also, the different width of the embryo near the anterior pole might affect aspects such as signal attenuation during image acquisition. Nevertheless, in spite of these uncertainties, the model has demonstrated that it is conceivable for a change in the interpretation of BCD by *eve* 2 to lead to the result observed in figure 6.8b.

slp1 misexpression

The other primary evidence given in Andrioli et al. (2002) for the importance of SLP1 was that misexpression of *slp1* from the ventral side caused repression of stripes 1, 2 and 3. However, this effect was only observed with the endogenous *eve* gene (figure 6.10), and not with the *eve* 2 transgenic enhancer. This indicates that repression of *eve* 2 by SLP1 might be mediated through higher-level interactions within or

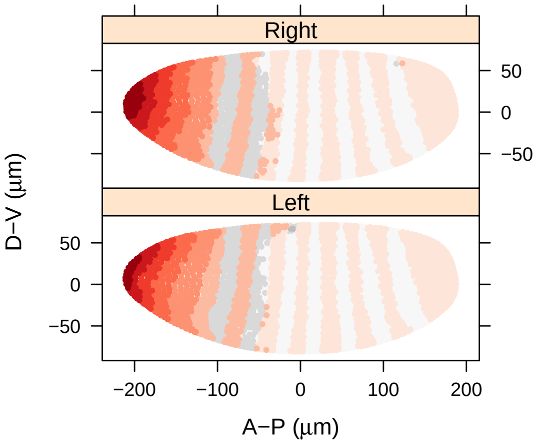


Figure 6.9: BCD expression for cohort 3, with darker colours for higher expression.

with the *eve* locus. For example, it is interesting that with ectopic *slp1*, *eve* stripe 1 expression is disrupted and shifted towards the region of stripe 2. This implies that the *eve* 1 enhancer might be active near the region where *eve* 2 normally activates and perhaps this causes interference in the endogenous locus. An alternative explanation is that SLP1 might regulate other gap genes, which in turn might affect *eve* 2, although there is little evidence for this (see Andrioli et al., 2004).

Conclusion

In conclusion, more evidence is required to demonstrate direct SLP1 repression of the *eve* 2 enhancer. An alternative model has been proposed involving dual regulation by BCD that is able to explain experimental observations reasonably well, thus demonstrating its plausibility. It is,

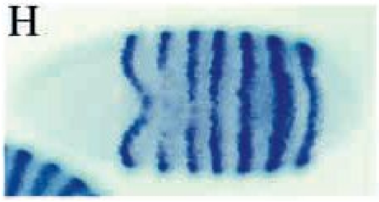


Figure 6.10: Ventral view of *eve* RNA expression pattern with ectopic *slp1*. Figure from Andrioli et al. (2002), reprinted by permission from *Development*.

in a sense, a simpler explanation than the introduction of SLP1 as a key regulator. Of course, further quantitative data would be useful, such as the distribution of the protein product of *slp1*. In the meantime, the model proposed here functions as an example of how quantitative data can be used to test hypotheses more rigorously than simply assessing micrographs by eye.

6.3 *eve* 3+7

This section will use model selection to evaluate the qualitative model that the border of *eve* stripes 3 and 7 are defined in a similar way to those of *eve* stripe 2, namely by a repressor pair (Small et al., 1996; Clyde et al., 2003). In this model, the inner borders are defined by Knirps (KNI), and the outer borders by Hunchback (HB), but the other regulators of stripes 3 and 7 are less clear. This model will be constrasted with a second proposal, that of dual regulation by HB (Papatsenko and Levine, 2008). This will also provide an opportunity to explore whether restricting the minimum value of the regulator function (section 5.3.2) is helpful or not.

6.3.1 *Search for a repressor pair*

Searching for repressors that define the borders precisely worked well for *eve* stripe 2 in the previous section. The same approach, using only the region near the borders as part of the training data, does not work straightforwardly for 3 and 7. HB and *kni*[‖] are not found as the obvious repressor pair. Instead, as shown in figure 6.11, *D* and *tll* seem important, although not as clear-cut as the results from the same search done for *eve* stripe 2 (see figure 6.2). This could indicate that spatial

[‖] Different font styles are used to illustrate what measurements are available in the *Virtual Embryo*; see section 1.5.8.

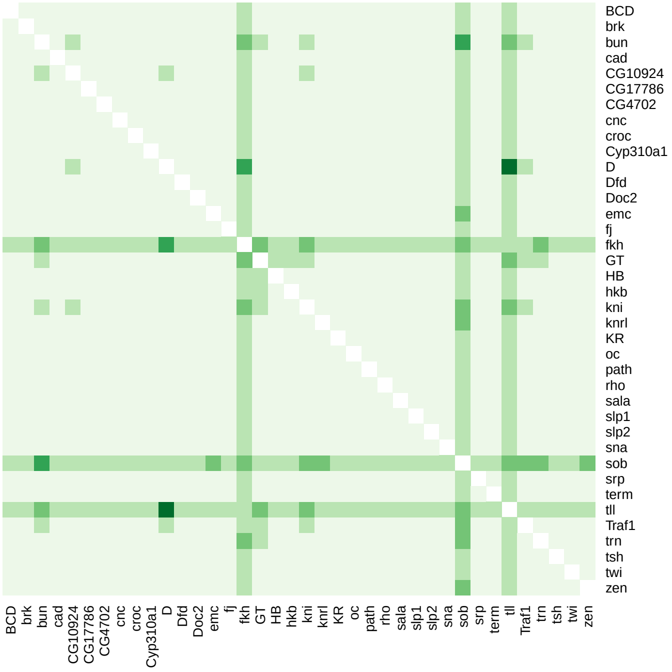


Figure 6.11: Best performing *eve* 3+7 models with only two repressors.

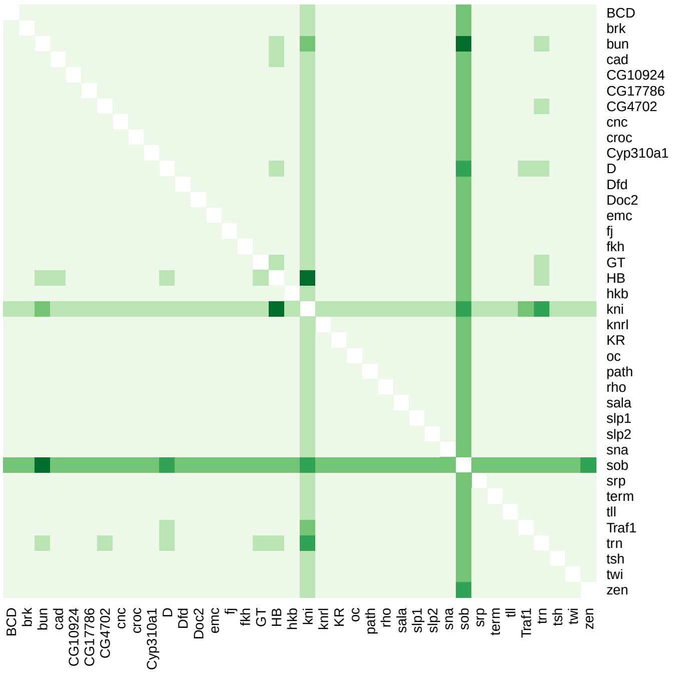
registration was not good enough, particularly since fine spatial registration is done for the mRNA probes, but not for protein (see section 1.5.5). However, the protein stains for GT and KR *were* able to define the borders of stripe 2 precisely. Another problem could be that the stripes are moving. As discussed in section 2.5, *eve*'s response might have shifted in relation to its regulators. Both these problems would be compounded by the use of a pair of stripes, making it necessary for repressor concentration measurements to be in register at multiple borders. It is therefore relevant to consider whether a search for repressors is more successful on the stripes individually.

Figures 6.12a and 6.12b show the result of a search at each stripe individually. It is apparent that, in this data set at least, *kni* and HB acting as repressors can provide more information for the borders of stripe 3 (along with other factors like *sob* and *bun*) than those of stripe 7, where *fkh*, GT and *tll* stand out.

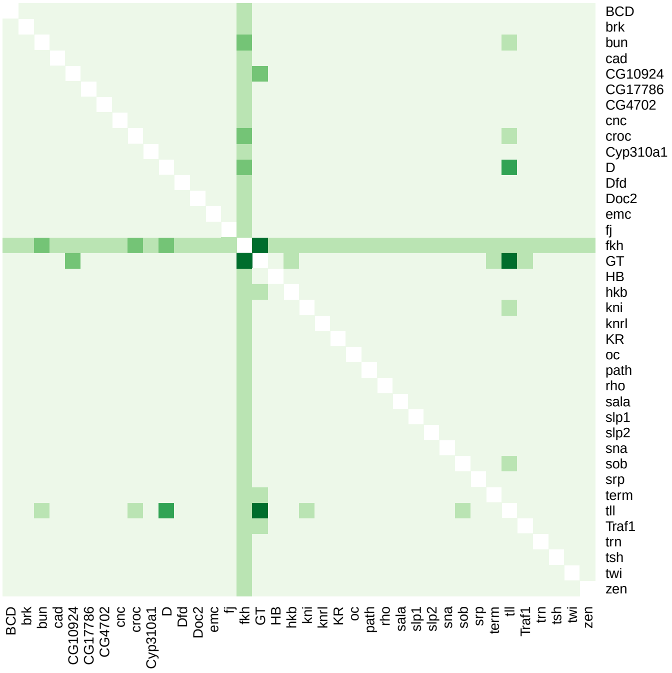
In chapter 4 it was demonstrated that including dual regulating capabilities for HB significantly improved the visual fit, and adding this capability certainly improves the performance of HB in defining the borders of stripe 7 as shown in figure 6.13. *kni*, though, is still not important for defining the borders. However, if the region between the stripes is included in the training data, it becomes important as a repressor (see figure 6.14).

There is one aspect which is not apparent from figures 6.13 and 6.14. If the dual regulatory function** of HB is less restricted (or given 'stronger' dual regulatory capabilities), the importance of *kni* for the region between the stripes decreases. This is because a dual-regulating HB is able, on its own, to ensure less expression between the stripes (see section 4.3.2 on page 84). Similarly, when HB is not in the model, *kni* is important (shown by the band for *kni* in figure 6.14). Thus, *kni* is important in models without HB and in models were HB is closer to linear (a 'milder' or 'weaker' form of dual regulation). Since HB is important even with mild, dual regulating capability (see figure 6.13) and given the trade-off between the strength of HB's dual regulating capabilities and the importance of *kni*, it is not as yet clear which model is to be preferred. Nevertheless, when HB has no dual regulatory capabilities, it does not appear as an important regulator and the visual fit is rather imprecise even with many other regulators added (see figure 4.3 on page 74). Thus, the remainder of this section will consider the case for a dual-regulating HB, without yet preferring the strong or weak form.

** $ax^2 + bx$ where x is the concentration of the regulator; see section 5.3.2.



(a) Stripe 3



(b) Stripe 7

Figure 6.12: Best performing models with only two repressors for *eve* stripes 3 or 7.

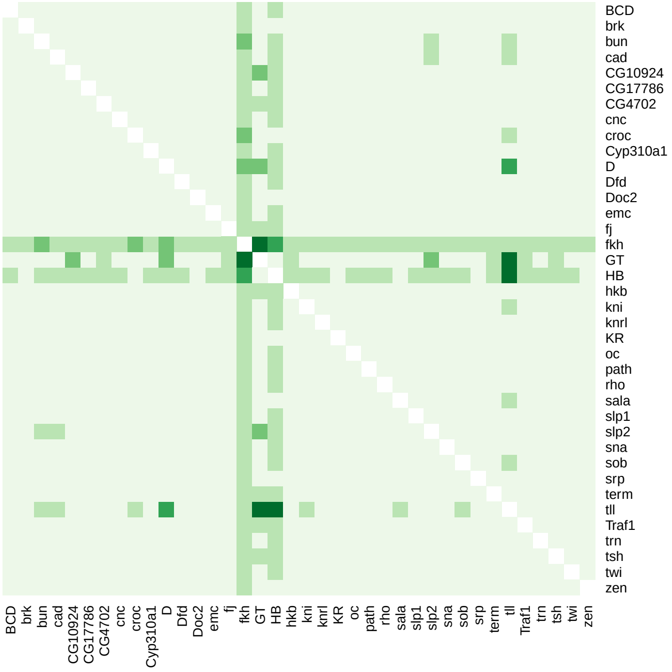


Figure 6.13: Best performing repressor pairs for stripe 7 with a mild, dual-regulating HB.

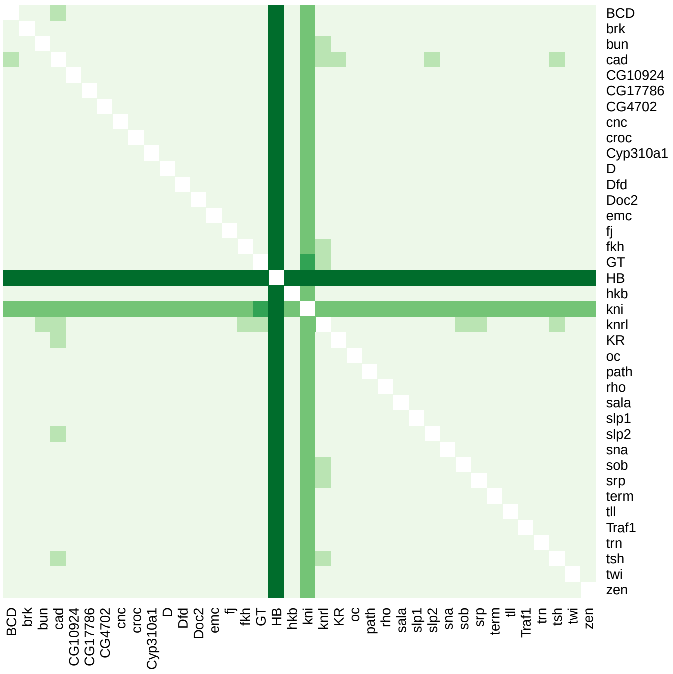


Figure 6.14: Best performing repressor pairs for stripes 3+7 and the region in between, with a strong, dual-regulating HB.

6.3.2 *Requirement for posterior repression*

Given uncertainty concerning the role of *kni*, it is thus instructive to look for single repressors that can complement HB in defining the borders of the stripes using the same training data as above for both stripes. With both mild and strong HB dual regulation, two useful repressors for this region are both gap genes: *tll* and *fkh* (see tables 6.8 and 6.9). And as described previously, *kni* does better in the case of mild dual regulation. The reason that *fkh* and *tll* are included is that they have a region of expression to the posterior of stripe 7, which is being used to define this border (the training set only includes a few nuclei posterior of stripe 7).

Probe	*fkh*	*tll*	*slp2*	*Traf1*	*sob*	*bun*	*kni*
Log likelihood	-927	-938	-978	-980	-981	-989	-995

Table 6.8: Best regulators for *eve* 3+7 after including a mild dual-regulating HB, for the borders of stripes 3+7.

Probe	*tll*	*fkh*	*slp2*	*CG10924*	*slp1*	*D*	*cnc*
Log likelihood	-692	-712	-815	-840	-841	-845	-853

Table 6.9: Best regulators for *eve* 3+7 after including a strong dual-regulating HB, for the borders of stripes 3+7.

6.3.3 *Knirps is still useful*

Probe	*D*	*bun*	*kni*	*croc*	*CG4702*	*cad*
Log likelihood	-822	-847	-848	-873	-884	-889

Table 6.10: Best regulators for *eve* 3+7 after including *fkh*, *tll* and a mild dual-regulating HB, for the *borders* of stripes 3+7.

Probe	KR	*kni*	*croc*	*bun*	*D*	*zen*
Log likelihood	-532	-564	-597	-601	-605	-609

Table 6.11: Best regulators for *eve* 3+7 after including *fkh*, *tll* and a strong dual-regulating HB, for the *borders* of stripes 3+7.

Probe	*kni*	*cad*	*bun*	*croc*	BCD	*knrl*
Log likelihood	-907	-1038	-1038	-1038	-1041	-1052

Table 6.12: Best regulators for *eve* 3+7 after including *fkh*, *tll* and a mild dual-regulating HB, for the *whole embryo*.

Both *tll* and *fkh* fulfil similar roles, so for the purposes of model selection judgement can be suspended for the time being on whether one or the other is preferable: they can both be included. From tables 6.10 and 6.11, it can be seen that *kni* is still not the top choice for defining the borders precisely, although it is relatively more important, especially for the stronger form of dual regulation. If the training data are expanded to the whole embryo (tables 6.12 and 6.13), *kni* becomes more important for the mild dual-regulating model. This suggests that the model with a stronger dual-regulating HB is more consistent near the borders and between the stripes. In contrast, in the weaker model, *kni* repression plays a more prominent role between the stripes than at the borders. In the strong form of dual regulation, *eve* expression is significantly reduced between the stripes simply because there is too little HB (see section 4.3.2 on page 84). Thus, it is noteworthy that the stronger form of dual regulation provides better evidence that *kni* helps define the borders of the stripes than the weaker model does.

Probe	KR	*kni*	*croc*	*bun*	*zen*	*CG10924*
Log likelihood	-544	-566	-634	-634	-639	-656

Table 6.13: Best regulators for *eve* 3+7 after including *fkh*, *tll* and a strong dual-regulating HB, for the *whole embryo*.

Probe	*kni*	*Kr*	*croc*	*bun*	*zen*	*CG10924*
Log likelihood	-566	-576	-634	-634	-639	-656

Table 6.14: Results for the same search reported in table 6.13, but with *Kr* mRNA instead of protein.

Finally, it is useful to see that these results are somewhat sensitive to the accuracy of spatial registration. If the search for regulators in the strong model is done with *Kr* mRNA instead of protein measurements, then KR is reduced a little in relative importance (table 6.14). Ideally, this search would use a finely-registered *hb* as well, but this is not straightforward since the translation of *hb* is inhibited in the posterior (Irish et al., 1989).

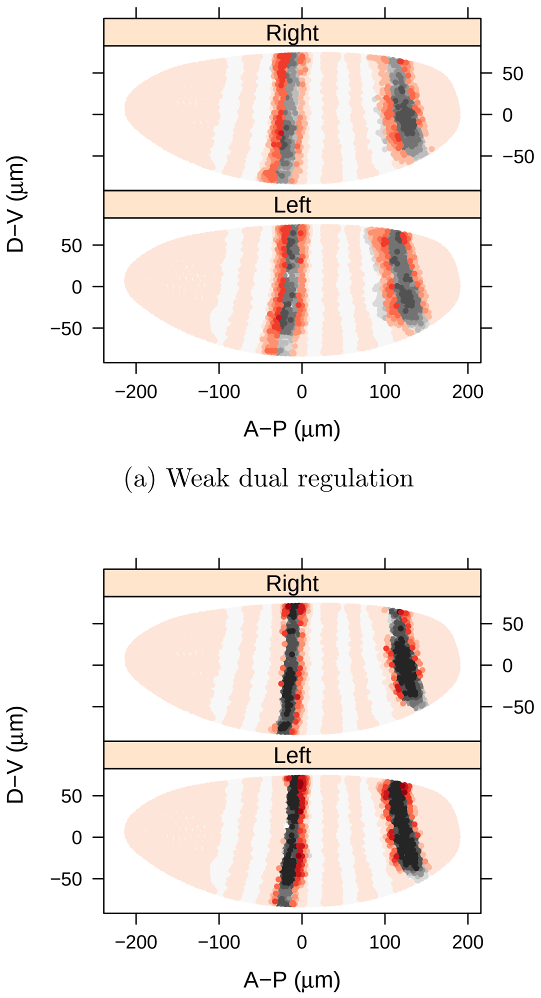
6.3.4 *Visual fits*

Thus, the following has been established for the models evaluated so far:

- HB dual regulation is plausible and the strong form is most likely.
- Posterior repression is required; *tll* and *fkh* are good candidates.
- *kni* is an important repressor (and KR might be important if the strong dual regulation model is reasonable.)

It is now appropriate to evaluate these with visual fits. Figure 6.15 shows that strong dual regulation is able to define the borders much more precisely than the weak form. As discussed in chapter 4, alternative explanations are possible, but this demonstrates that strong dual regulation is, in principle, able to provide significantly improved precision.

The difference between strong and weak dual regulation is clear to see, but the difference between *fkh* and *tll* is not as obvious (see figure 6.16).

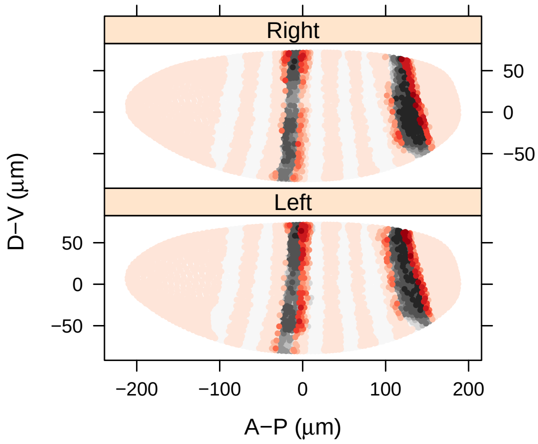


(a) Weak dual regulation

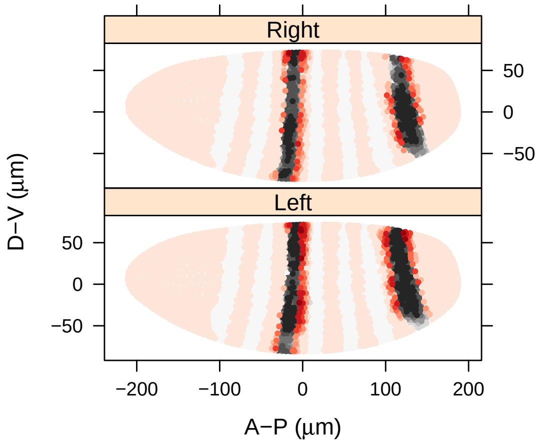
(b) Strong dual regulation

Figure 6.15: Prediction for *eve* 3+7 with *kni*, *fkh*, *tll* and HB dual regulation.

tll defines the border of stripe 7 more clearly, and leads to a darker (higher probability) in stripe 3, but the difference is small. From a log likelihood perspective, the difference is more pronounced (see table 6.15), but given the uncertainties surrounding temporal and spatial registration, further evidence is required to separate the two. For this reason, a 3+7 transgenic reporter was designed with *tll* binding sites perturbed (see appendix C). However, besides this, including *tll* has the additional benefit of explaining previous experimental observations of mutant embryos, which are now described.



(a) With *fkh*



(b) With *tll*

Figure 6.16: Prediction for *eve* 3+7 with *kni*, HB with dual regulation, and either *fkh* or *tll*.

Probe	*tll*	*fkh*	*CG10924*	*slp2*	*D*	*hkb*
Log likelihood	-566	-695	-864	-893	-894	-955

Table 6.15: Best regulators after including *kni* and a strong dual-regulating HB, for the whole embryo.

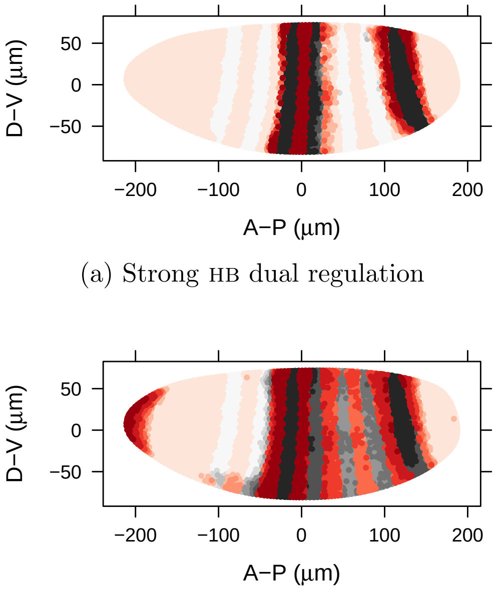
6.3.5 *Mutant predictions*

This section considers the suitability of the model in explaining previous mutant experimental results; in particular, the model with a dual-regulating HB, *kni* and *tll.* Some mutant predictions were for null mutants, in which case, the relevant regulator concentration was set to 0. In other cases, the concentration of the regulator was set to a specific value, say 0.2. Since the measurements are relative, this means that the value was set to 20% of the maximum concentration of that regulator for all nuclei over the whole time period of the *Virtual Embryo.*

kni

Strong dual regulation can fit the data more precisely than the more restricted form. However, a strongly dual-regulating HB can itself ensure low expression of *eve* between the two stripes without relying on *kni*, specifically in the region between stripes 5 and 6 (see figure 6.17a). This does not agree with in situ hybridisation data from a *kni* mutant (figure 6.18) where expression extends all the way between stripes 3 and 7. Certainly, in the model prediction there is some expression towards the posterior from stripe 3, but it is not as extensive as in the mutant embryo. The weak and strong forms of dual regulation can be compared in figure 6.17.

There is, though, a rather important fact that needs to be taken into account. KNI is a repressor of *hb*: in a *kni* mutant, the posterior domain of *hb* expression extends somewhat towards the anterior (Clyde et al., 2003). It is not clear how far this is quantitatively, but if it is supposed that HB expression is mild (0.2) between stripes 3 and 7, then the resulting phenotype is correctly predicted (see figure 6.19). Nevertheless, this prediction relies on speculation on how far and at what strength *hb* extends towards the anterior, and so it would be



(a) Strong HB dual regulation

(b) Weak HB dual regulation

Figure 6.17: Model predictions for *eve* 3+7 for *kni* mutants.

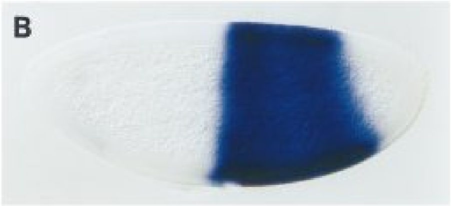


Figure 6.18: In situ hybridisation for *eve* in a *kni* mutant (*kni*^{10} allele). Reprinted from Small et al. (1996) by permission Elsevier: *Developmental Biology* © 1996.

useful to have quantitative data on the behaviour of *hb* in the mutant *kni* embryo.

As an aside, it is worth mentioning that if HB is a pure repressor, as in the Clyde et al. (2003) model, then one might expect some compensating repression by HB in the *kni* mutant embryo in the region including at least stripe 7, but this is not in fact observed.

The effect of KNI on *hb* should presumably not be relevant for an *eve* 3+7 transgenic reporter where KNI binding sites have been perturbed. And, in this case, the resulting expression does agree with the strong dual regulation model: *eve* expression does not extend all the way (see figure 6.20). Clyde et al. (2003) suggest that this is because not all KNI

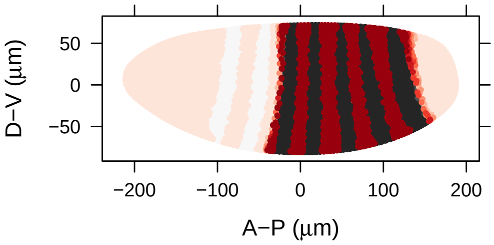


Figure 6.19: Model prediction for *eve* 3+7 for a *kni* mutant with strong dual-regulating HB, and low level of HB between stripes 3 and 7.

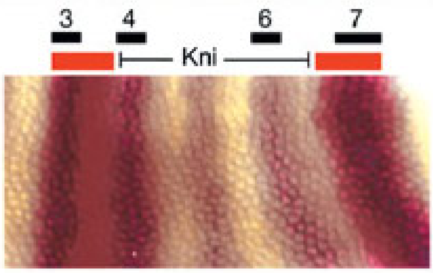


Figure 6.20: Expression of endogenous *eve* mRNA (black) compared with *lacZ* reporter mRNA (red) driven by a mutant 3+7 enhancer. Reprinted from Clyde et al. (2003) by permission from Macmillan Publishers Ltd: *Nature* © 2003.

binding sites were mutated. It would be interesting to know whether mutating more KNI sites does in fact extend expression across the whole region between the stripes.

Finally, the role of *kni* rests on another experimental observation. When *kni* is misexpressed from the ventral side, it is found to act as a strong repressor of stripes 3 and 7 confirming the hypothesis that KNI is an important repressor. And in the model, *kni* is indeed a strong repressor as demonstrated in figure 6.21. This is from a relatively mild increase of *kni* (0.2).

hb

Expression of *eve* is difficult to predict in mutants for *hb*, since HB affects many regulators of *eve*. Further, the changing distribution of *hb* in the early embryo is complex and difficult to characterise (see, for example, Margolis et al., 1995). Nevertheless, the model does provide

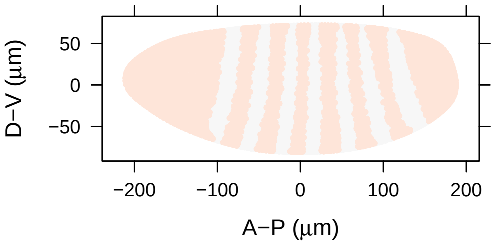


Figure 6.21: Model prediction for *eve* 3+7 for slightly increased *kni* levels.

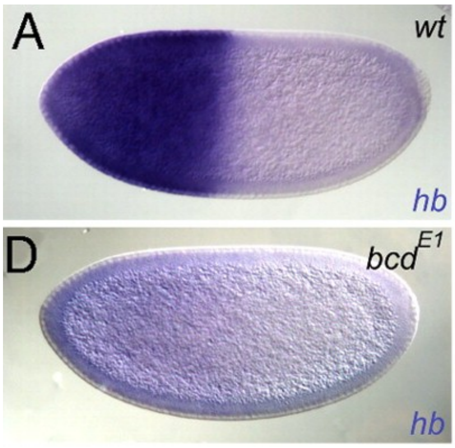
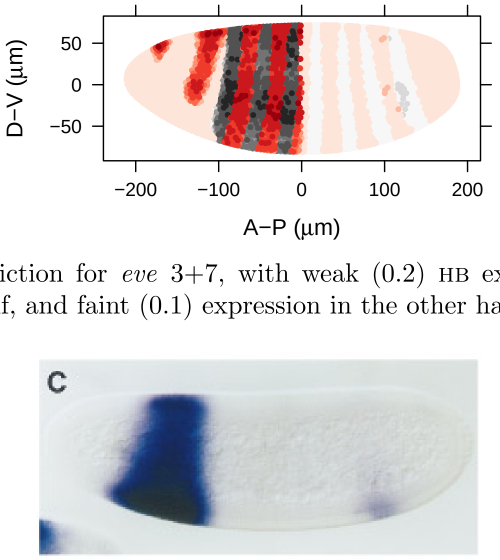


Figure 6.22: *hb* expression in an embryo produced by a *bcd* mutant female. Image A is an embryo from wild-type; image D is an embryo from a *bcd* mutant female. Figure from Ochoa-Espinosa et al. (2009). © 2009 by The National Academy of Sciences of the USA.

a basis for explaining relevant mutant phenotypes. For example, Small et al. (1996) describe the expression of the *eve* 3+7 enhancer in an embryo produced by a *bcd* mutant female. It has a weaker stripe 7 and an expanded stripe 3 pattern that does not expand into the anterior-most regions of the embryo (figure 6.23b). As can be seen from figure 6.22, the distribution of *hb* is affected in embryos produced by these mutants. If one assumes a weak (say, 0.2) expression of *hb* in the anterior half of the embryo, and faint (0.1) expression in the other half, then the model prediction is as in figure 6.23, which accords with the observation in Small et al. (1996).

Small et al. (1996) also describe a mutant lacking zygotic *hb* activity where the expression of stripe 3 is similar, but stripe 7 is broader than



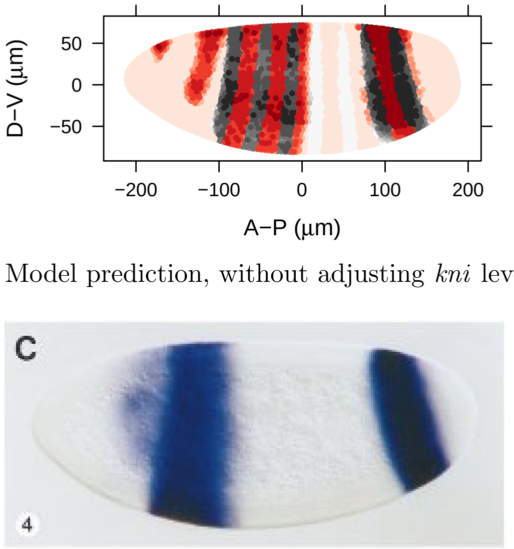
(a) Model prediction for *eve* 3+7, with weak (0.2) HB expression in the anterior half, and faint (0.1) expression in the other half.

(b) In situ hybridisation of the *eve* 3+7 reporter. Reprinted from Small et al. (1996) by permission Elsevier: *Developmental Biology* © 1996.

Figure 6.23: *eve* 3+7 expression in an embryo produced by a *bcd* mutant.

wild type, expanding towards the posterior. This led them to infer a ubiquitous activator. Without quantitative data it is difficult to supply the correct regulator concentrations for a model prediction. However, according to Margolis et al. (1995), in embryos lacking zygotic HB activity, *hb* expression around stripe 3 is reduced compared to normal, but the posterior stripe of *hb* is expanded and intensified. In accordance with this, if one considers an embryo with a uniform and weak (0.2) expression of HB across the embryo, the resulting prediction is broadly consistent with the mutant data (figure 6.24), including a broader stripe 7, but without a shift to the posterior. More complicated distributions of HB attempting to match the putative HB distribution produce the same prediction. However, in the case of faint HB expression (say 0.1), almost no *eve* expression is predicted, which, if representative, would support the hypothesis that a further activator is required.

It is also instructive to compare model predictions with those from Clyde et al. (2003) where *hb* was misexpressed on the ventral surface of the embryo using the *snail* (*sna*) promoter (see figure 6.25). Stripe 3 moves to the posterior with a weak increase in HB, and it is removed



(a) Model prediction, without adjusting *kni* levels

(b) In situ hybridisation. Reprinted from Small et al. (1996) by permission Elsevier: *Developmental Biology* © 1996.

Figure 6.24: *eve* 3+7 expression in a mutant lacking zygotic *hb* activity.

at a higher concentration. The model predictions (figure 6.26) agree with this result. The model also predicts that with a small increase of HB, stripe 7 extends towards the anterior, and with a slightly higher increase that it moves towards the anterior, but the experimental results (figure 6.25) only show these effects at the higher concentration of *hb* mRNA. However, since *hb* translation is repressed in the posterior (Irish et al., 1989), the same increase in *hb* mRNA will not lead to the same increase in HB. This is a plausible explanation for why the model requires a relatively lower level of HB to produce the same effect for stripe 7 as seen in figure 6.25. Finally, it should be pointed out that the model of this section is able to predict shifts in the patterns and the differential sensitivity of *eve* 3 and 7 in the misexpression study of the endogenous *eve* locus without recourse to the effects of perturbation on the expression of stripes 4 and 6.

tll

It has been proposed (Frasch and Levine, 1987; Small et al., 1996) that TLL is an activator required for stripe 7 expression, which contradicts

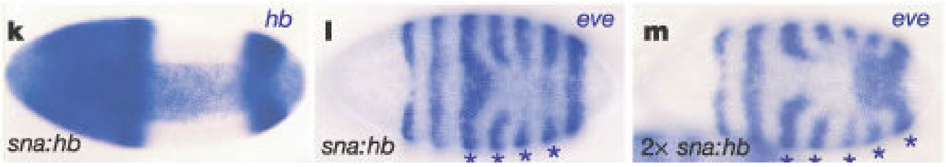
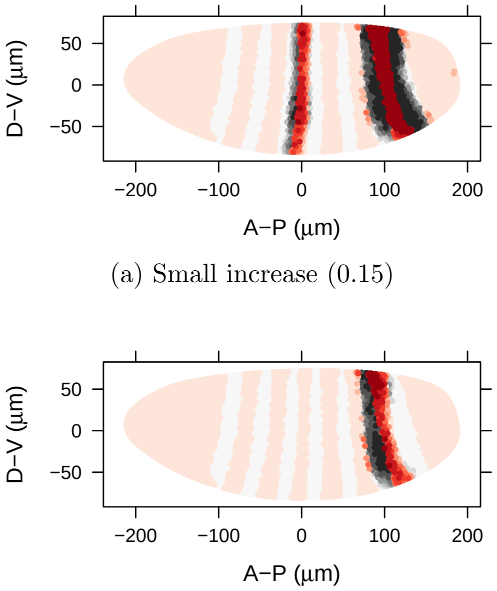


Figure 6.25: Ventral views showing stripe-specific repression with one and two copies of the *sna:hb* construct. Reprinted from Clyde et al. (2003) by permission from Macmillan Publishers Ltd: *Nature* © 2003.



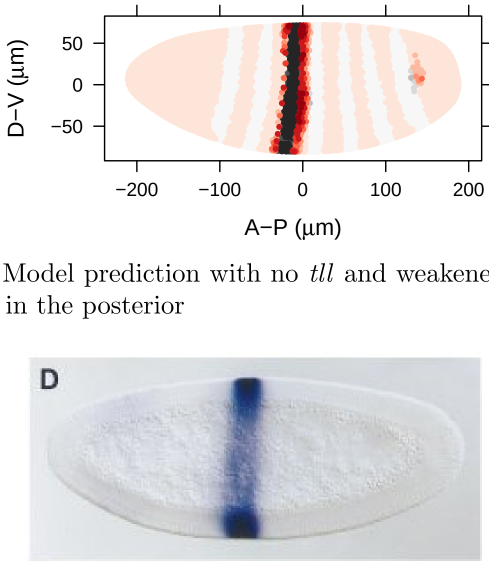
(a) Small increase (0.15)

(b) Larger increase (0.3)

Figure 6.26: Prediction of *eve* 3+7 expression with increased HB.

the model proposal here that TLL is a repressor. In the *Virtual Embryo*, *tll* expression (and thus, likely, the protein) are strongly expressed adjacent to the posterior border of stripe 7, but not within the stripe.

Since this is wild-type expression, the lack of *tll* expression within the stripe presents a difficulty for any model requiring direct TLL activation. Indirect effects such as diffusion of the TLL protein can be invoked, but effects such as these can similarly be used to explain a model with TLL as a repressor. Thus, for example, TLL might affect the expression of *hb*, perhaps as an activator (Margolis et al., 1995) or potentially indirectly. As figure 6.27 shows, a decrease of HB by 0.2 in the posterior half



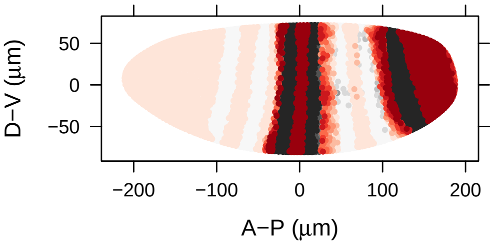
(a) Model prediction with no *tll* and weakened HB in the posterior

(b) In situ hybridisation. Reprinted from Small et al. (1996) by permission Elsevier: *Developmental Biology* © 1996.

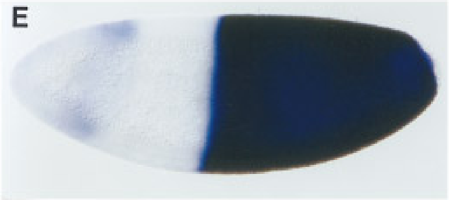
Figure 6.27: *eve* 3+7 expression in a *tll* null mutant.

is sufficient to abolish stripe 7 (it is only the level within the stripe that matters). Since the model includes a dual regulatory capability for HB, it should be noted that an increase of 0.4 produces the same effect. Finally, perhaps the most direct explanation (since it does not require diffusion) is that the removal of *tll* results in a posterior shift of *kni*—this too results in the same prediction using the same model.

Further support of the model including TLL repression is that in Small et al. (1996), the result of a *tor* and *kni* double mutant is described, where it is stated that a *tor* mutant contains no tll^+ function. Surprise was expressed that in this embryo, *eve* expression extends all the way to the posterior pole. However, this is what the model predicts. The result is shown in figure 6.28, but for ease of interpretation, it has not been adjusted for the effect of KNI on HB, which changes the expression of *eve* between 3 and 7 (see figures 6.17 and 6.19). After this adjustment, the prediction is in agreement with the experimental result. Again, it should be noted that the actual expression level of HB in this mutant is unknown, and thus it is difficult to ascertain what cross-regulation effects might be relevant.



(a) Model prediction with no *tll* or *kni*. HB is unadjusted.



(b) In situ hybridisation for an *eve* 3+7 reporter. Reprinted from Small et al. (1996) by permission Elsevier: *Developmental Biology* © 1996.

Figure 6.28: *eve* 3+7 expression in a *tor* and *kni* double mutant.

Finally, a study (Moran and Jimenez, 2006) has demonstrated that converting TLL to an obligate repressor yields similar results to normal TLL function, which, if valid, would further support the role of TLL as a repressor.

6.3.6 *Conclusion*

In conclusion, Clyde et al. (2003) propose that KNI defines the posterior of stripe 3 and the anterior of stripe 7, and HB the anterior of 3 and the posterior of 7. However, as demonstrated here, KNI does not seem able to define these borders precisely without the help of a dual-regulating HB. Also, the region posterior to stripe 7 requires an additional repressor. The alternative model proposed here, a dual-regulating HB with KNI and TLL as repressors, is able to explain experimental observations to the extent that quantitative data are available.

7

FRENCH FLAG PROBLEM RECONSIDERED

The previous chapters have shown that the position of the *eve* stripes can be specified in terms of the concentrations of their regulators, many of which occur in gradients across the anteroposterior axis of the embryo. The French flag problem (section 1.1) has long been used as a basis for reasoning about gradients of factors that provide spatial information (often called morphogen gradients), and it is therefore fitting to reconsider the French flag problem here. This chapter introduces the *tipping point model* for understanding positional information, interpretation, and patterning in the *Drosophila* embryo. This will be contrasted with the most common alternative for specifying different regions in the developing embryo: one or more thresholds per morphogen gradient, which here will be called the *thresholded morphogen gradient model.*

The specification of positional information is a perennial and popular topic in developmental biology*, and at times quite technical. This chapter's focus is therefore a preliminary perspective of how the French flag problem can be addressed using the model of this thesis. In particular, it will touch on aspects of precision and robustness and the relative independence of positional information and its interpretation. This will be discussed in the context of stripe patterns as well as the regulation of *hunchback* (*hb*), a common experimental system for studying this problem in *Drosophila.*

* More than 700 papers on morphogen gradients were published between 1989 and 2004 (mentioned in Ephrussi and Johnston, 2004).

7.1 THE TIPPING POINT MODEL INTRODUCED

Much discussion regarding positional information, either implicitly or explicitly, assumes that individual regulators are thresholded. The model of enhancer[†] function proposed here, the *tipping point model*, places the threshold (or sharp transition) at the level of the enhancer rather than at the level of the regulators. The ability of each regulator to nudge the enhancer over the threshold then depends on whether the enhancer is near the *tipping point* or not, and this is determined by the context—the concentrations of all relevant factors. As a result, the requirement for other factors emerges naturally, without needing physical interactions between factors or carefully positioned DNA binding sites within the enhancer.

In the work here, a simple sigmoidal transition suffices to describe the transition of the enhancer. This can be expressed as

$$\frac{1}{1+e^{-\eta}} \tag{7.1}$$

where η is a linear combination of regulator concentrations,

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$

x_j is the concentration of the jth transcription factor and β_j is the corresponding strength of that factor's contribution. Thus, the enhancer has a natural tipping point which corresponds to a value of η.

Similar (perhaps more accurate) forms can be derived from thermodynamic considerations, such as in Bintu et al. (2005). However, of primary relevance here is how the regulators contribute near transition. The essential aspect of the formula is that the same relative change in different regulators contributes to the *transition* at different strengths,

[†] Or regulatory DNA more generally.

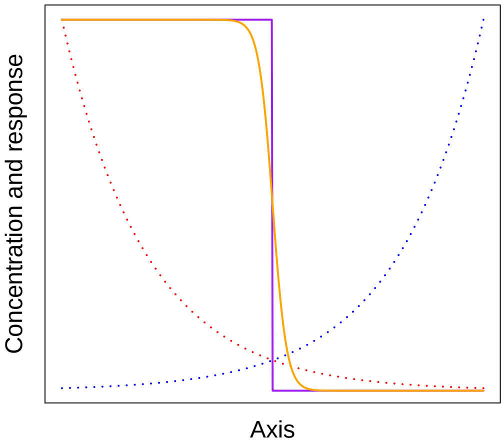
the β parameters. These parameters can be interpreted as representing the affinity of each regulator for the available binding sites as well as the strength of its interaction with the necessary cofactors and other molecules required for activating transcription. The sign of each β indicates whether the factor contributes as an activator or repressor. Interestingly, and perhaps surprisingly, the results of this and the previous chapters indicate that repression can be modelled successfully as a simple additive offset against the contribution of the activators: increased repression can be offset by increased activation and vice versa.

Direct interactions between regulators are not required in the models of this thesis, but they can be included if required. A more important extension is that, as shown in this thesis, some regulators (e.g., BCD and HB) might act as dual regulators activating at low relative concentrations and repressing at high. For these, it has been demonstrated that a simple quadratic term suffices to capture their behaviour over the range of concentrations in the data set.

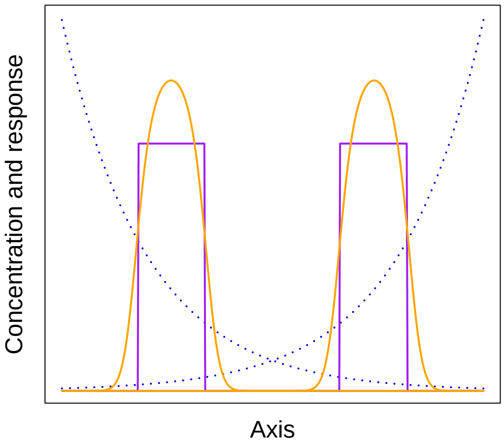
Although this formulation includes the thresholded morphogen gradient model as a special case (i.e., one regulator with one overall transition), the crucial difference arises when there is more than one regulator. Unlike the thresholded morphogen gradient model, transitions do not occur for each regulator (separate thresholds), but only at the level of the enhancer.

7.1.1 *Gap gene and stripe patterns*

In the thresholded morphogen gradient model the border between each region within the overall pattern requires a threshold. So, in the classic case of the French flag, two thresholds are required: one for the border between blue and white, and one between red and white. In *Drosophila*, the anterior *hunchback* expression domain has usually been studied as



(a) One-border patterns



(b) Stripe patterns

Figure 7.1: A comparison of the thresholded morphogen gradient model (purple) and the tipping point model (orange). The morphogen gradients are shown as dashed lines, which can be activators (red) or repressors (blue).

an example of a one-border pattern, with the relevant threshold applied to the Bicoid (BCD) gradient, an activator.

When Wolpert considered stripes, here defined as two-border patterns, it seemed that too many thresholds were required—twice as many as the number of stripes (Wolpert, 1989). Further, he thought it unlikely that this could arise progressively during evolution. As a result, he favoured a prepattern mechanism, such as one based on a reaction-diffusion model in the vein of Turing's model (Turing, 1952). A single threshold is then applied to the prepattern to achieve the requisite ON/OFF stripe pattern. Since then, it has been shown that different stripes, sometimes pairs of stripes, are controlled by discrete enhancers, and this has made it plausible that multiple thresholds are in operation. Specifically, a different threshold is applied to each of two opposing repressor gradients, and this defines the borders of a stripe or pair of stripes. This has been proposed for *eve* 2 (Stanojevic et al., 1991; Andrioli et al., 2002), eve 3+7 (Small et al., 1996; Clyde et al., 2003) and eve 4+6 (Fujioka et al., 1999; Clyde et al., 2003).

It is thus instructive to compare one-border and stripe patterns under the thresholded morphogen gradient model and the tipping point model (figure 7.1). Both can produce one-border and two-border patterns, and in both the location of the borders can be varied. In early discussion of the French flag model (see section 1.1), it was pointed out that robustness to different sizes of the embryo (scaling) requires positional information to make use of both ends of the relevant axis. Gradients of maternal factors in *Drosophila* do run from both ends (e.g., BCD from the anterior pole, and CAD from the posterior pole, see figure 1.2 on page 6), and this might have a role to play in the robustness of the embryo to scaling (but most likely via the gap genes). Further, as mentioned, opposing gradients are important in defining the stripes. Thus, for the purposes of comparing the tipping point model with the thresholded morphogen gradient model, it is useful to have two opposing gradients (the dashed lines in figure 7.1). However, it should be noted that for the thresholded morphogen gradient model, a one-border pattern depends on one morphogen gradient; it cannot make use of two gradients.

Thus, the one-border pattern of figure 7.1 can be viewed as representing the response of *hunchback* to Bicoid (the red, activator gradient). This will be examined more concretely in section 7.4. The two-border patterns are representative of the *eve* stripes, which respond to opposing gradients of repressors. The morphogen concentrations in figure 7.1 should not be interpreted as absolute levels; each morphogen concentration is plotted relative to its maximum. So, for example, the stripes are positioned for convenience rather than to illustrate how actual morphogen concentrations would appear relative to one another near the borders of the stripe.

7.1.2 *Implications of the tipping point model*

An important implication of the tipping point model is that experimentally increasing or reducing the concentration of any one factor might reveal whether it functions as a repressor or activator, but quantitative modelling requires a knowledge of the concentration of all relevant factors. Examining each gradient and its supposed thresholds separately will not reveal the overall functional behaviour of the enhancer. Rather it will remain obscure and difficult to explain.

This can be expressed more theoretically in terms of positional field, which is defined here as the contours of equal positional value in the developing embryo. The role of the enhancer is to interpret positional information provided by the available transcription factors to produce a positional value. In the tipping point model, interpretation plays a part in defining the positional field. Traditionally, a positional field could be defined independently of interpretation and reflected the contours of the morphogen concentration. However, in the *tipping point model* the field will depend on the weights involved (the βs), which depend partly on the affinity of the enhancer's binding sites for the available factors, and partly on the strength of those factors in initiating transcription. Thus, a contour of the resulting positional field might not follow the contours of any given factor.

One consequence of this is mentioned above: measuring the concentrations of a single factor will not reveal the positional field. But another consequence applies if the change in the concentration of one regulator is offset by the change in another, perhaps through cross-regulation. This could occur along the dorsoventral axis (in the case of a field specifying anteroposterior position) or it could occur across different embryos. This is quite a crucial point. The traditional French flag model operates in one dimension, but either does not explain how this pattern can be regulated in two dimension, or it relies on something like perfectly uniform diffusion from a line of cells. Further, it is unable to explain how each embryo can have the same absolute concentration of the morphogen. However, if positional field is specified by multiple regulators, and these regulators cross-regulate each other suitably (like the gap genes), then consistency can be provided, even from embryo to embryo. One aspect of this, precision and reproducibility, will be discussed further in the next section.

7.2 PRECISION AND REPRODUCIBILITY

The requirement for precise read-out at the borders of a pattern under the thresholded morphogen gradient model has been recognised as an important problem from early on (Wolpert, 1969): is it possible to implement such a precise read-out at the cellular level, and is the level of precision appropriate given natural variations in the level of the morphogen or the underlying cellular mechanisms? A recent example is Gregor et al. (2007). The consideration of how the organism could be buffered against natural fluctuations of the morphogen, both within and between individual embryo can be considered an issue of reproducibility: *robustness* or *tolerance* to different conditions, but the terminology can be confusing. For example, Houchmandzadeh et al. (2002) refer to the robust specification of the *hb* border as a problem of *precision* (since precision is inversely related to variability). It is also not immediately obvious whether discussions elsewhere regarding increasing *precision* of the borders of the gap genes over time refer to robustness or to a sharpening of their borders.

To make things clear, *precision* here does not refer directly to molecular precision. Precision in this chapter is a feature of interpretation by the enhancer. So, given some variability in the inputs, how variable is the response of the enhancer? A more precise response is less variable. However, precision here does not refer to system-level effects that increase robustness to variation in the initial conditions, such as the 'attractors' and 'canalisation' of Manu et al. (2009b). Therefore, the extent to which *interpretation* is robust to variation also reflects the extent to which less precision is required at the molecular level.

7.2.1 *Robustness to initial conditions*

One of the circumstances in which this is important is in accommodating different sizes of the embryo. This was, in fact, one of the considerations of the French flag model where robustness was seen as an aspect of the global properties of positional information, or the morphogen gradient. One solution, therefore, involves a source and a sink leading to a linear morphogen gradient, which is robust to scaling (Crick, 1970). However, although robustness to initial conditions is not a theoretical problem for the French flag model, exponential gradients like BCD do present a practical problem. In particular, the movement of the border of *hunchback* (*hb*)[‡], a target of BCD, in response to natural variations in the gradient of BCD is much less than predicted by the thresholded morphogen gradient model. Work like Manu et al. (2009b) has demonstrated that multiple transcription factor inputs to a target gene result in more precision than a single input alone. They conclude that this is from compensatory cross-regulation. The increase of an activator (like BCD) results in an increase of repressors of *hb*, which compensate for the increased activation, leading to a smaller shift in the location of *hb*'s border. This might be important in the case of stripes (as described in the next section), but the *tipping point model* demonstrates that for the gap genes much of the increased tolerance can be explained as a feature of *interpretation* without recourse to higher-level effects. These considerations are explored more concretely in section 7.4, specifically for *hunchback* regulation in the *Drosophila* embryo, but figure 7.2 demonstrates the principle. Figure 7.2a shows that doubling the activator gradient leads to a much smaller shift in the border for the tipping point model (orange) than predicted by a single threshold model (purple). The stripe patterns are affected much more (figure 7.2b). They have disappeared in the tipping point model (orange), and are greatly reduced in the thresholded morphogen gradient model (purple).

As a matter of interest, figure 7.3 demonstrates that small fluctuations in the level of the gradients does not remove the patterns (stripes or one-border patterns). It might be that temporal fluctuations (e.g., in the read-out) would be smoothed out over time, and that spatial fluctuations (if they exist) might be smoothed by diffusion. Further work is required to consider this more fully, but it can briefly be pointed out that the gene circuit model of Manu et al. (2009b) provides a solution for robustness to initial conditions, but not for small stochastic

[‡] The posterior border of the anterior expression domain.

fluctuations. Specifically, Jaeger and Martinez-Arias (2009) point out that:

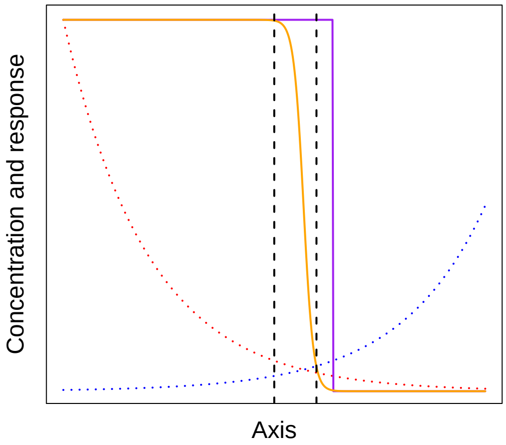
> For instance, while the mechanism presented[§] can reduce embryo-to-embryo variability in monotonically decreasing concentrations of Bcd, it seems unable to reduce non-monotonic stochastic fluctuations between nuclei in individual embryos.

7.2.2 *The necessity and function of the gap genes*

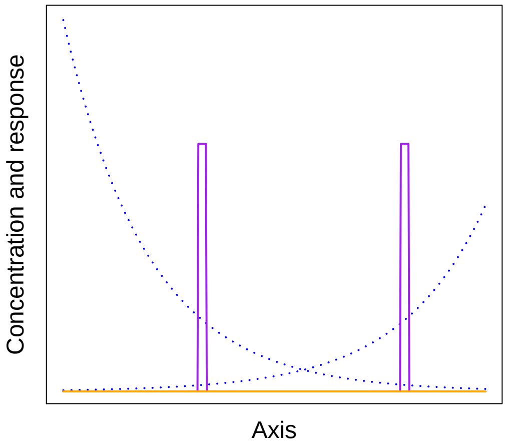
The stripe patterns of both models are not robust to variations in the concentrations of their regulators. Since the absolute levels of maternal factors are likely to vary, this suggests the necessity of an intermediate level of control: the gap genes. An important requirement is that the gap genes regulate each other in order to buffer this variability (e.g., Kraut and Levine, 1991). However, the two models differ in what type of buffering is required as well as their evolutionary implications.

The thresholded morphogen gradient model requires that the absolute concentrations of each regulator remain stable. Thus, the position of the stripe depends on the availability of repressors with the correct absolute strength of repression. Further, the border of each stripe is presumably specified independently with separate thresholds, but these thresholds have limited value except when tied to repressors with the precisely correct characteristics. This appears rather inflexible with regards to evolution. Further, as has been pointed out in Papatsenko (2009), the prerequisites of the two repressors are exacting and consequently the information gain in the step of *eve* stripe formation is limited. This has echoes of the prepattern hypothesis described above. Thus, it would seem necessary to assume that prepatterns arise commonly during evolution, and that the embryo takes advantage of this

[§] In Manu et al. (2009a,b).

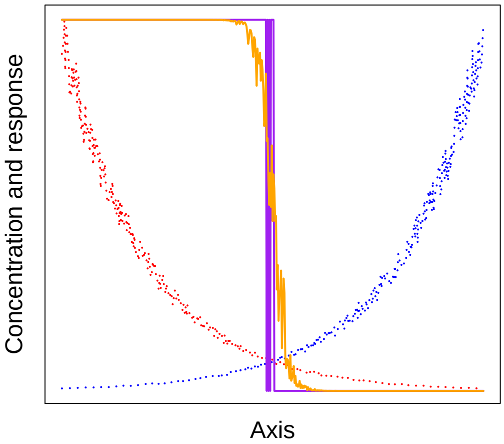


(a) One-border patterns

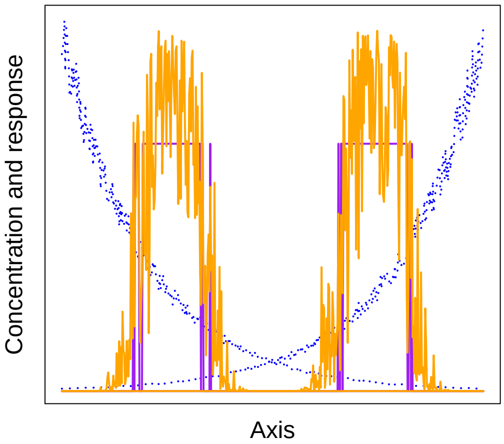


(b) Stripe patterns

Figure 7.2: The response of stripes and one border patterns to a doubling of one of the gradients. Colours are as in figure 7.1. The dashed vertical lines in (a) show 50% and 60% along the axis.



(a) One-border patterns



(b) Stripe patterns

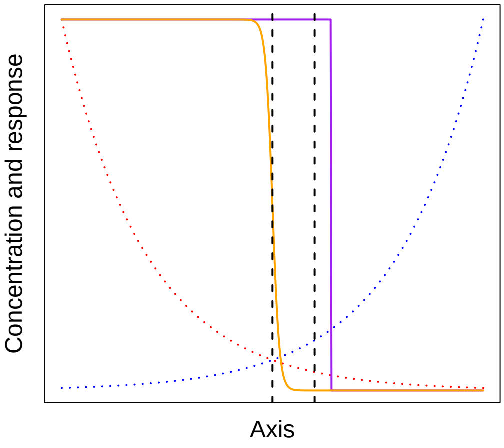
Figure 7.3: The response of stripes and one border patterns to random variations in regulator gradients (up to ±10%). Colours are as in figure 7.1.

to specify the stripes. But a simpler explanation is possible under the tipping point model.

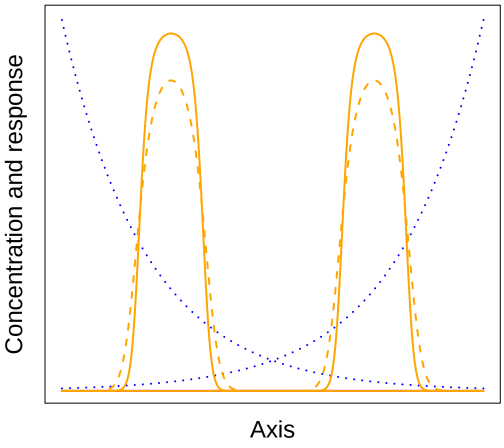
Figure 7.4 shows the response to a doubling of all regulators (including an implicit activator[¶] for the stripes). As expected, the patterns of the thresholded morphogen gradient model have disappeared. In contrast, for the tipping point model not only are the borders of the stripes and the one-border pattern stable, but the stripe borders have also sharpened. At once this provides an explanation for the increasing sharpness of the stripe patterns that is observed in the *Drosophila* embryo: it is an inevitable consequence of increasing concentrations of its regulators. If the regulators of the stripe are balanced (that is, increasing one leads to a proportionate increase in the others), then the location of the stripe will be stable, otherwise it will move. A related point is that positional information for the stripes is available early, even before the regulator concentrations have reached a supposed threshold. It would be of interest to explore to what extent the absolute gap gene concentrations are controlled across individual embryos. If they are not controlled, this would make the double threshold model unlikely, in which case it would be useful to see whether the *eve* stripes are in any way sharper in these embryos. Of course, further mechanisms of control might buffer the organism against this.

Finally, the tipping point model provides a mechanism for tuning stripe patterns over evolutionary time. Not only can the stripe be sharpened by increasing the concentrations of the regulators, but it can also be sharpened by increasing the affinity of binding sites to available transcription factors. If any broadly-distributed activator is available and affinity for this factor is increased, it is then possible for the overall affinity for the repressors to increase, thus leading to a sharpening of the stripe. Further points are made in section 8.2.

[¶] The intercept.



(a) One-border patterns



(b) Stripe patterns

Figure 7.4: The response of stripes and one border patterns to a doubling of both gradients. Colours are as in figure 7.1. The dashed vertical lines in (a) show 50% and 60% along the axis. The dashed stripes in (b) show the stripes before doubling of the regulators.

7.3 POSITIONAL INFORMATION CAN BE STUDIED

One of Wolpert's goals was to suggest that positional information is general and that the same information might be used in different contexts with different thresholds or interpretations (Wolpert, 1968, 1969). This might, for example, suggest flexibility in developing new patterns over evolutionary time. As a result, the French flag problem was stated statically. His emphasis was the separation of interpretation from the specification of positional information, recognising that time (of the order of hours) was necessary to set up positional information in the developing embryo. In other words, the question is: how is positional information interpreted to produce a pattern?

Given the problems of the thresholded morphogen gradient model, it is no surprise that this has been considered too simplistic. In the context of the *Drosophila* embryo, the *Drunken French Flag* model is a recent example. Jaeger and Reinitz (2006), elaborated by Jaeger (2009) and Jaeger and Martinez-Arias (2009), propose that positional information is tied to the dynamic behaviour of the gap genes, in particular their cross-regulation. Positional information is continuously shifting and becoming more precise. As a consequence it becomes inappropriate to specify positional information independently of its interpretation—and only a dynamic model can explain position in the developing embryo. This is illustrated by a quote from Jaeger and Reinitz (2006):

> In summary, positional information in the *Drosophila* blastoderm can be said to consist of dynamically changing combinations of maternal and zygotic protein concentrations, depending not only on maternal morphogens but also on shifting positions of segmentation domain boundaries due to zygotic downstream gene regulatory interactions. This implies an active, rather than a passive, mode of gradient interpretation and blurs the distinction between establishment and interpretation of positional information. . . . our version of the French Flag emphasizes the complexity and the unique character of each develop-

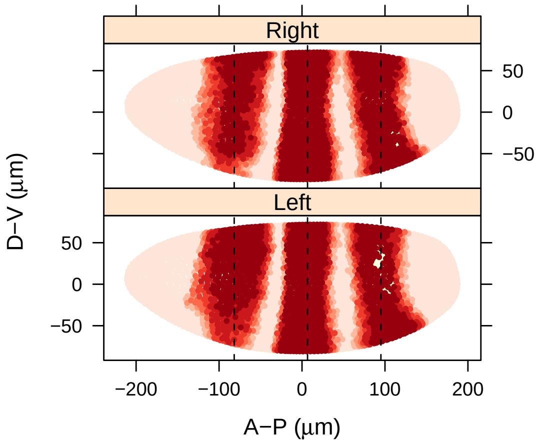
> mental field, rather than suggesting a universal regulatory mechanism of pattern formation.

However, it is argued here that this perspective is too pessimistic. A degree of separation between positional information and interpretation is useful. It makes it easier to discuss questions such as pleiotropy and how evolutionary selection of enhancers might occur independently of the positional information available to them. Further, it provides us with a conceptual toolkit for exploring problems in developmental biology such as scaling and regeneration —precisely the original goal of the French flag problem. Two examples of how positional information can be studied will now be given.

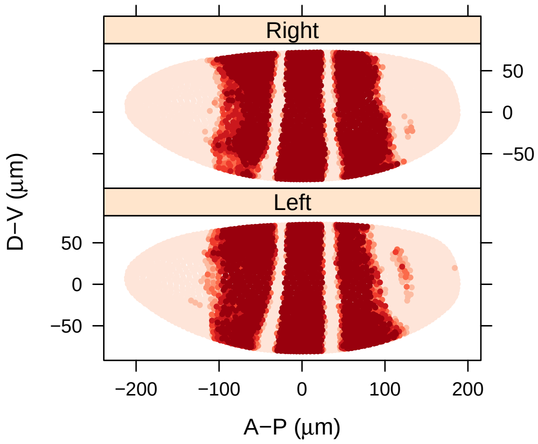
7.3.1 *Positional information is available to the embryo*

The essential feature of positional information is that it can be used in many ways, simply by changing the way it is interpreted. In the thresholded morphogen gradient model, this is done by changing the threshold. In the tipping point model it is done by changing the relative contributions of the regulators. When this is effected by changing the actual protein concentrations (*trans*), it is a change in positional information, and this has downstream consequences via all enhancers that interpret it. But if the relative weights of the regulators are changed only in the enhancer (*cis*), then interpretation has changed. So, does the embryo have much scope to use the available position information along the anteroposterior axis in various ways? As it turns out, yes, within limits.

Figure 7.5 shows a three stripe embryo constructed in two ways using positional information supplied by some maternal and gap genes. In both cases, a separate model was trained for each stripe, with the final prediction being a simple sum of the prediction of each artificial enhancer. The stripes of the training data for figure 7.5a were constructed by including nuclei within a rectangular region, orthogonal to the A-P axis. For figure 7.5b the borders of the training stripes follow the natural flow of the *eve* stripes. This was done by merging different stripes and the region between them.



(a) Stripes orthogonal to the A-P axis. The dashed line shows the centre of each stripe.



(b) Stripes constructed from merging *eve* stripes 1+2, 3+4, and 5+6.

Figure 7.5: A three stripe embryo, trained with BCD, HB, *kni*, GT, KR, *tll* and *cad*, with no dual regulation. Darker shades signify higher predicted expression.

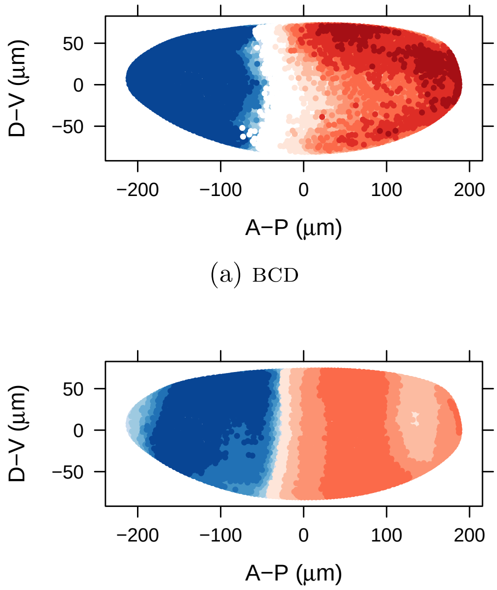
From the fuzziness of the borders in figure 7.5a and their tendency to be slanted, it is quite apparent that positional value is not strictly orthogonal to the A-P axis, but follows the actual borders of the stripes. This suggests that potential positional fields are not purely determined by interpretation but are constrained by the actual regulator concentrations. Secondly, from figure 7.5b it is apparent that for the regulators supplied, positional information is much clearer near the middle of the embryo than towards the poles. Interestingly, this can be remedied by making BCD dual regulating. The posterior border of the third stripe is difficult to improve, but it is much sharper if it is made to extend further to the border of *eve* stripe 7.

Thus new single stripe patterns can be constructed using existing positional information. Stripes of different sizes can be made, as long as their borders are oriented with respect to the relevant positional field. The next section will use a wider stripe pattern to illustrate that progressive refinement over time is possible.

7.3.2 *Extra factors add sharpness incrementally*

A stripe pattern has sharp borders, which are important for clearly defining location within the embryo. Creating these requires other regulators, such as the gap and maternal genes, but the borders of the gap genes are not as pronounced as those of the *eve* stripes. This can be explored by considering two one-border patterns at each end of the embryo, each controlled by a different enhancer (figure 7.6). It is interesting to observe that a single early regulator (like BCD or HB) can provide enough information for a one-border pattern with a diffuse border (the blue region). Adding a single additional regulator (KR) improves the pattern remarkably, as shown in figure 7.7. Finally, the border can be specified even more sharply by including additional gap genes as regulators (figure 7.8).

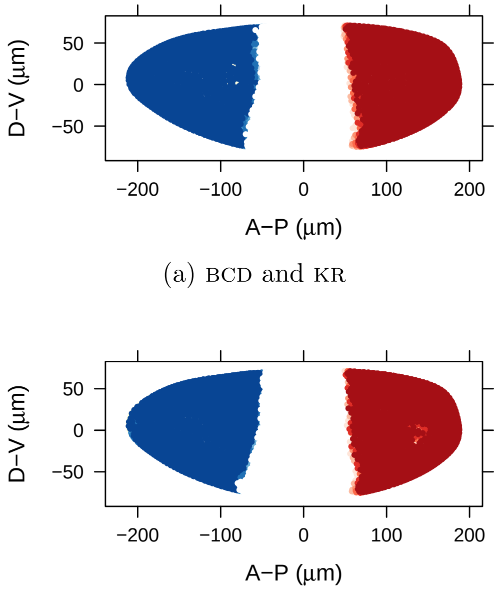
With respect to the biological implementation of the *tipping point model*, this result indicates that an enhancer with binding sites for a single factor can provide information for broad regions of expression and that binding sites for further factors can progressively refine the



(a) BCD

(b) HB

Figure 7.6: French flag with two enhancers (blue and red) using a single regulator, either BCD or HB. Darker shades signify higher predicted expression.



(a) BCD and KR

(b) HB and KR

Figure 7.7: French flag with two enhancers (blue and red) and two regulators, KR and either BCD or HB. Darker shades signify higher predicted expression.

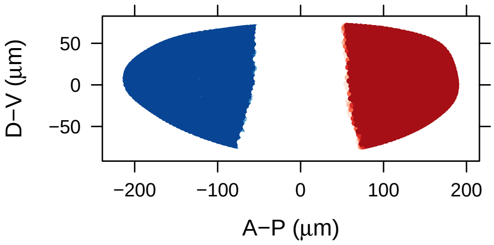


Figure 7.8: French flag with two enhancers (blue and red) using BCD, HB, KR, *kni* and GT. Darker shades signify higher predicted expression.

border. For comparison, in the thresholded morphogen gradient model, the threshold is sharp from the start. It can be shifted, but its boundary is immediately well-defined. An additional regulator can be added to specify the other border, again sharply from the start. Further regulators are then unnecessary, except perhaps to provide cross-regulation of each other.

7.4 A CONCRETE EXAMPLE: BICOID AND HUNCHBACK

Experimental work addressing the problem of precision has often examined the relationship between Bicoid (BCD) and its downstream target, *hunchback* (*hb*); for example: Frohnhöfer and Nüsslein-Volhard (1986), Houchmandzadeh et al. (2002) and Gregor et al. (2007). However, if precision is provided by the action of multiple factors on the target gene, then focusing on only BCD and HB would be misleading. Dynamic cross-regulation amongst the gap genes has been proposed as being essential to the resulting robustness and precision (Manu et al., 2009b). For example, increasing BCD leads to an increase in another repressor of *hb*, thus compensating for the original increase. But, as will be demonstrated shortly, the static aspect of cross-regulation—multiple regulators acting directly on *hb*—is sufficient to explain the observed precision.

7.4.1 *Continuous model*

First, it is necessary to introduce a continuous model relating *hb* to its regulators. Although the transition of *hb* from high expression to low expression is quite pronounced (changing over about 10% of the embryo's length) it is broader than the *eve* stripes (changing over a few nuclei, or about 2% or 3% percent of the embryo's length). This means that uncertainty over spatial and temporal registration (see section 2.5) is less of a concern. So, although the border can be captured by logistic regression, a continuous form is appropriate if it can be supported by the data[‖]. Therefore, equation 7.1 will be used directly and the training data will retain continuous values for the response variable, rather than being discretised into ON or OFF. This form still captures the essence of a threshold: there is a sharp transition from OFF to ON, but it has the advantage of being able to fit intermediate values better, which is helpful in the case of gap genes like *hb*. The parameters are then fitted to the data using non-linear least squares (with the R function `nls` in the package *stats*).

Ideally, one would relate *hb* mRNA to the relevant regulators. However, over the time period considered, the *hb* mRNA patterns are complex, reducing in some regions and increasing in others, which could be the result of aspects not under consideration here, such as autoregulation. On the other hand, the pattern of HB is fairly stable and is similar to the expression discussed in Houchmandzadeh et al. (2002) and Gregor et al. (2007), making it a reasonable choice for modelling *hb* response. But since translation of *hb* into HB protein is subject to translational inhibition in the posterior (Irish et al., 1989), the training set will exclude the posterior domain of HB expression[**].

[‖] In this case, the considerations of appendix A become relevant.

[**] x < 60 for the training data.

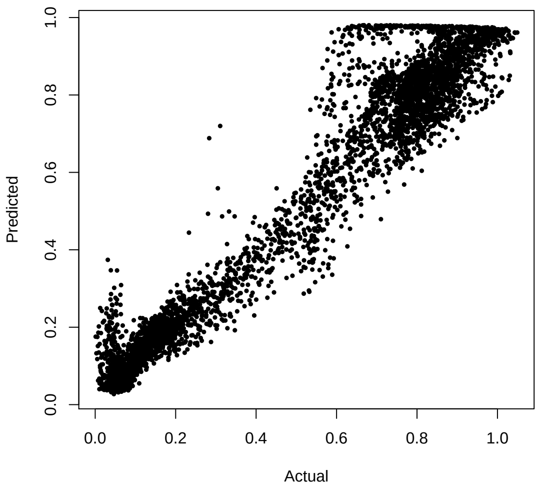


Figure 7.9: Actual concentration of HB in each nuclei of the training data plotted against the prediction of the non-linear model described in the text with BCD, KR, GT and *kni* as regulators.

Using the continuous form with only a few gap genes as regulators (BCD, KR, GT and *kni*), with a dual-regulating BCD, gives a good fit to the model. Figure 7.9 shows that, with the exception of extreme values, the prediction holds up well within the training data. Further, figure 7.10 shows that the prediction is good across the part of the embryo corresponding to the training data (the anterior half). Interestingly, the model predicts expression in the posterior (where there is negligible BCD), even though this was not part of the training data. It is intriguing that the prediction is higher than the actual protein levels, since as mentioned previously, there is translational repression of *hb* in the posterior. However, knowledge of the contribution of maternal *hb* to HB protein levels would be important in a more detailed model. Currently, the model makes use of constitutive expression of *hb*[††], with BCD compensating for increased repression by gap genes in the anterior domain. In other words, in the model at least, activation by BCD is not the only source of positional information. Also, it is worth noting that

[††] An alternative explanation is that it is activated by a uniform regulator, potentially translated maternal *hb*.

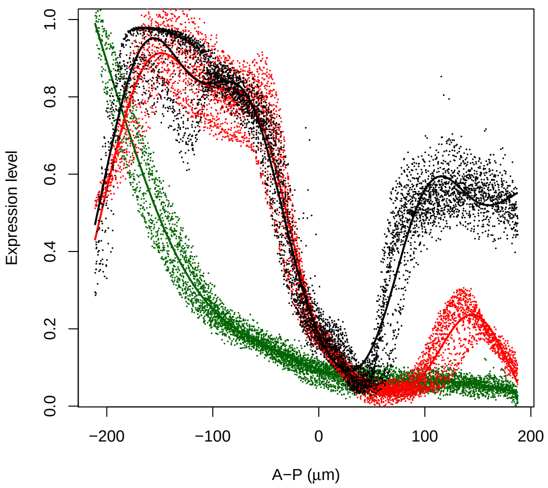


Figure 7.10: BCD, HB and predicted *hb* expression for all nuclei by position on the A-P axis. BCD is green, HB is red and predicted *hb* is black. The lines were fitted to the coordinates of these points using penalised regression splines.

a dual-regulating BCD is not essential to the observations of this section, but it does give a better fit near the anterior pole of the embryo[‡‡].

Before discussing the results of BCD perturbation, it is necessary to clarify the smooth curves that are fitted to the expression points in figures 7.10 and 7.11. These curves are fitted using penalised regression splines (with the R function `gam` in the package *mgcv*). The purpose of introducing them is to summarise the expression predictions of the nuclei (as in figure 7.11). They fit the expression points well (which are not shown in figure 7.11), except at the extreme edges. This is why the x-axes of the plots in figure 7.11 do not start from zero. To reiterate, the predictions of expression values are reasonable at the extremities, but the curve fitting summary is not. Nevertheless, the summary curves fit well in the region of interest—the border of *hunchback* expression.

[‡‡] One one can speculate that BCD positional information might be more important near the anterior pole. Perhaps this was its ancestral function.

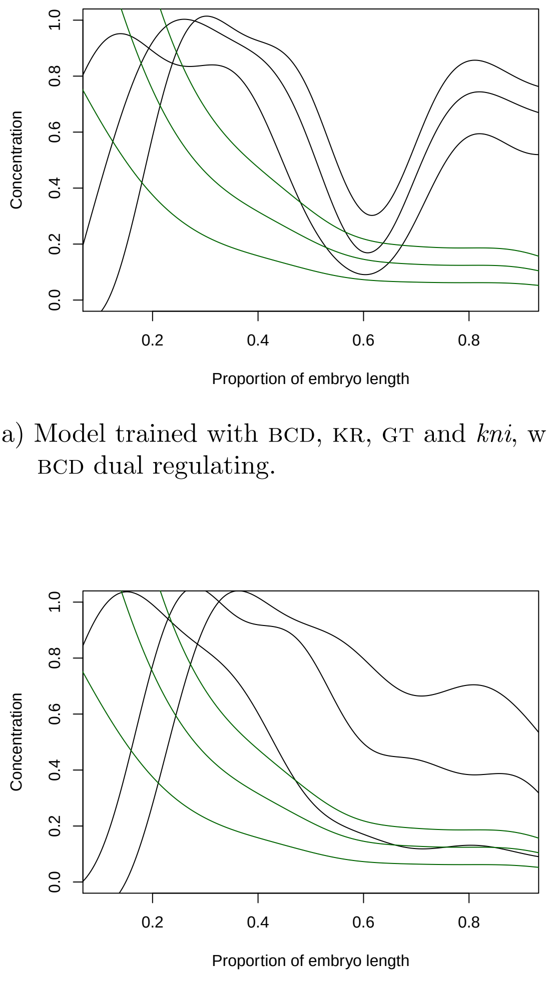
7.4.2 *Bicoid perturbation*

Now that a reasonable model has been fitted to the data, it is possible to compare the read-out of *hb* in the case of increasing BCD concentrations. BCD is often modelled as an exponentially decreasing gradient (e.g., Gregor et al., 2007; Alon, 2006), with the peak concentration determined by the maternally deposited mRNA. Thus, 4 and 6 copies of the *bcd* gene can be modelled by multiplying the normal BCD concentration by 2 and 3 respectively.

The results are shown in figure 7.11. It is striking how more reliable the read-out is in the case of multiple regulators. Each doubling of the dose of BCD does shift the anterior expression of *hb* to the posterior, but by far less than expected in the classic threshold model: it shifts by less than 10% egg length, which is as seen experimentally by Houchmandzadeh et al. (2002) and earlier (Struhl et al., 1989). This demonstrates that dynamic, compensatory, cross-regulation is not essential to describe this aspect of robustness. Multiple regulators alone are sufficient.

7.4.3 *Double and single mutants*

There is much more that could be done using this approach than is covered in this thesis. This will be illustrated with one more observation. In Houchmandzadeh et al. (2002), many single mutants were tested, and none of the gap genes were found to alter the border or the precision of *hunchback* expression significantly. Yet Manu et al. (2009b) show experimentally that double mutants of *Kr* and *kni* do have increased variability. Their conclusion is that the double mutant is best explained by dynamic compensatory regulation. In other words, increasing BCD results in more repressors of *hb*, which balance the increased activation.



(a) Model trained with BCD, KR, GT and *kni*, with BCD dual regulating.

(b) Model trained with a dual-regulating BCD only.

Figure 7.11: Model prediction of *hb* to increasing BCD profiles. The green curves show increasing BCD concentration. The black curves show the predicted *hb* response. See text for details.

This conclusion, however, is not in fact necessary as shown by the mutant predictions of the static model of this section (figures 7.12, 7.13 and 7.14). In order to relate these predictions to experimental data, it is necessary to consider that the minimum concentration of *hunchback* over the whole embryo might in fact vary between different embryos (as per the predictions). Since fluorescence measurements are often affected by the minimum and maximum intensities, it is quite possible that the effective threshold used for analysing experimental results might in fact not be absolutely constant across different embryos. This

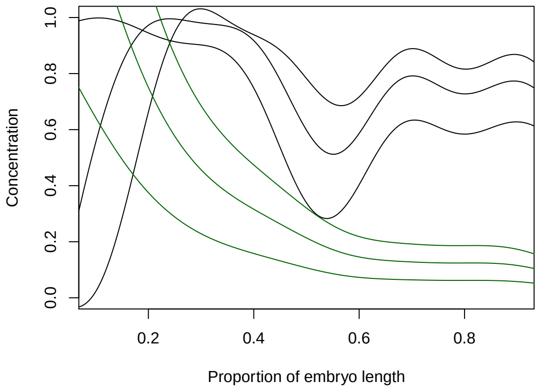


Figure 7.12: *kni* mutant prediction for the full *hb* model.

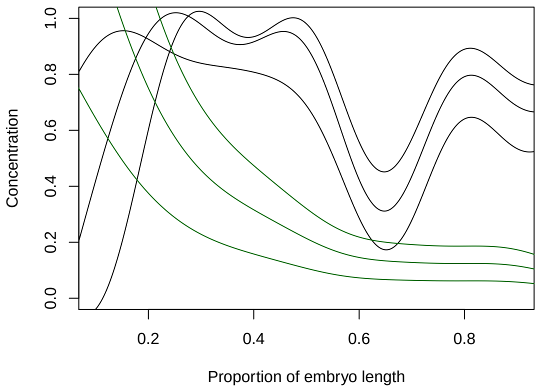


Figure 7.13: *Kr* mutant prediction for the full *hb* model.

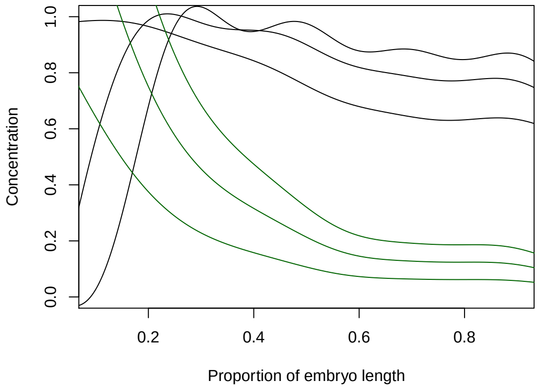


Figure 7.14: Double mutant of *Kr* and *kni* for the full *hb* model.

is relevant to determining the location of the *hb* border from normalised expression measurements, as done by Houchmandzadeh et al. (2002) and Manu et al. (2009b). In this case, normalisation of the intensity of *hb* will obscure the change in minimum across embryos that is seen in the predictions. In particular, in the double mutant the predicted *hb* profile is fairly flat. In this case, slight variations in the resulting *hb* expression profile will lead to large changes in the position of the thresholded border after normalisation. This would lead to a loss of precision in agreement with Manu et al. (2009b).

The model presented here is not complete. For example, the model predicts a posterior shift in the border of *hb* in the *Kr* null mutant, yet Houchmandzadeh et al. (2002) do not see this. Of course, this could be a result of indirect regulation not captured in the model, the most plausible being that *gt* expression moves anteriorly in the relevant *Kr* null mutant (Struhl et al., 1992). Since *gt* is a repressor in the model of this section, this would compensate by moving *hb* back towards the anterior. However, although the model cannot directly explain the stability of the border position in the case of the *Kr* null mutant, figures 7.12, 7.13 and 7.14 show that the relevant single mutants do not lead to a loss in precision after normalisation, but that the double mutant does.

In conclusion, the model demonstrates that increased precision can indeed be explained by the *interpretation* of positional information, and does not require dynamic compensatory mechanisms.

8 CONCLUSION

This thesis has demonstrated that the tipping point model is able to explain the behaviour of the *eve* enhancers. With the inclusion of dual regulatory capabilities for BCD and HB it was also shown to be consistent with previous biological observations. In the process of elaborating the model, a number of additional predictions and observations have been made. This chapter begins by summarising these and concludes with a discussion of the implications and future directions of research with respect to evolution, enhancer structure and dynamic modelling.

8.1 SUMMARY OF REGULATORY HYPOTHESES

This section summaries the various hypotheses that have been proposed in connection with the *eve* enhancers and gene regulation in general.

8.1.1 *eve 2 enhancer*

It was found that the main known regulators, BCD, KR, GT and HB, are indeed the best regulators for defining *eve* stripe 2. Counterintuitively, BCD is important as a repressor in the anterior. This has been identified previously in connection with a model proposing SLP1 as an additional repressor (Andrioli et al., 2002). However, as shown in section 6.2, if BCD has dual regulatory capabilities, then this hypothesis is not necessary. The four main regulators alone are sufficient to explain the experimental observations considered here.

8.1.2 *eve 3+7 enhancer*

As confirmed in chapter 5, HB and KNI are important regulators of *eve* 3+7. However, as explained in chapter 4 and in section 6.3, introducing dual regulation by HB significantly improves the sharpness of the read-out. HB becomes an important activator of stripe 3, but represses in the anterior. KNI remains important as a repressor between the two stripes. Further, the model demonstrated that HB and KNI are not sufficient for explaining repression to the posterior of stripe 7. Within the *Virtual Embryo* data set, *tll* was the most likely candidate. Although TLL has usually been assumed to be an activator of stripe 7, section 6.3 demonstrated that repression by TLL is consistent with the experimental data presented.

8.1.3 *Other enhancers*

Although the regulatory discovery results for *eve* 4+6 and *eve* 5 in section 5.4 were relatively clear-cut, these models were not explored in further detail. Here it is worth mentioning some results not shown. The visual fit for stripe 5 was good, but the borders for stripes 4 and 6 were somewhat fuzzy. Stripes 4 and 6 are quite close together, thus exacerbating any inaccuracies in registration. It is thus relevant that the fit was much more precise when measurements for *Kr* and *gt* mRNA were used instead of protein measurements, since, as explained in section 1.5.5, the mRNA measurements are finely registered with respect to *eve*, whereas the protein measurements are not. Nevertheless, it would be beneficial to explore this enhancer in more detail.

Stripe 1, on the other hand, did not show clear visual fits. The most probable cause is that stripe 1 expression appears early in the data set (cohort 1) and is broader than a mature *eve* stripe. By cohort 3 it is narrowing, thus reducing in expression in places. Therefore, it is likely that the regulatory relationships for this stripe might best be revealed by considering data from the earlier cohorts. The results from this are intriguing, but are only mentioned briefly here. If the training data consist of the broader stripe 1 from early cohort 1, then the prediction extends towards stripe 2 producing a larger stripe reminiscent of the larger primary stripe 1 in *Tribolium* (Patel et al., 1994). The minimal regulators required are a dual-regulating BCD and two repressors, SLP1 and KR. Perhaps repression by another pair rule gene is important in refining this particular stripe. Also, the role of SLP1 in repressing the anterior border of this larger stripe might go some way to explaining the results of SLP1 repression in the vicinity of stripe 2 for the endogenous locus (see section 6.2.4).

8.1.4 *Dual regulation*

Dual regulation has been shown to be important in the case of the *eve* 2 and *eve* 3+7 enhancers. However, it was not necessary to postulate widespread dual regulation. In fact, it has thus far been useful to introduce dual regulation for only two transcription factors: *bcd* and *hb*. The other stripes were not examined in detail, so this is not necessarily a complete list. Nevertheless, it is interesting that both of these factors are thought to make use of cooperative binding (e.g., Lebrecht et al., 2005; Papatsenko and Levine, 2008), and both are important in setting up early positional information along the A-P axis.

8.1.5 *Interactions are not necessary*

The models of the previous chapters have shown that the gap and maternal protein concentrations contain precise spatial information for the early *Drosophila* embryo. Interestingly, the relationship is simple to describe and it is not necessary to introduce interactions between factors (although as mentioned above, dual regulation by a few factors is important). This is relevant in discussions regarding synergy, such as between HB and BCD (Simpson-Brose et al., 1994). The models here demonstrate that simple additive effects can capture the requirement for both regulators: both are required to push the enhancer over a threshold, but increasing either would suffice. This is confirmed by work such as that showing HB is able to substitute for BCD in forming thoracic segments in *Drosophila* (Wimmer et al., 2000).

8.1.6 *Functional description at the level of the enhancer is tractable*

The modelling work presented has demonstrated that the functional behaviour of enhancers can be explained with relatively simple relationships, whereas this was not successful at the level of the gene. This has some implications for modelling gene regulatory networks in general. If a gene's behaviour can be described simply, it might be that only a single enhancer is in operation for that data set. More complicated relationships would benefit from knowledge of the gene's enhancer structure. However, it has been crucial that the data have been at the level of the nucleus; the *eve* stripes are known to be defined by different enhancers so it was possible to assign nuclei to their appropriate enhancer. In essence, this made it possible to look only at the inputs and outputs that were relevant to the enhancer being modelled. However, the simplicity of the relationships, and the apparent constraint on which stripes could be trained together, does suggest that a more general method could be applied when the exact enhancer structure is not known. It might be possible, for a data set that is at the resolution of individual cells, to evaluate enhancer structures based on the relative simplicity with which they explain the data.

8.1.7 *Anteroposterior and dorsoventral axes*

The relationship between the anteroposterior (A-P) and dorsoventral (D-V) axes was not studied specifically in this thesis. However, this work has succeeded in modelling across the entire *Drosophila* embryo, including both axes, so it is worth making a few comments. It is apparent that the maternal and gap genes do indeed contain sufficient information to describe stripe position around the whole embryo. Notably, the angle of the *eve* stripes was shown to follow positional information provided by these genes. Specific information from genes that pattern the D-V axis were not required. Of course, it is of interest to understand why positional information in the A-P patterning genes does follow the stripes, but this is unclear from the work presented thus far. For example, *hb* expression around the whole embryo could also apparently be explained purely by maternal and other gap genes. It would be of interest to see whether the same is true for the other gap genes and to understand how this pattern arises.

8.1.8 *A combinatorial model is not necessary*

The word *combinatorial* in connection with gene regulation is sometimes used to mean that the gene has *multiple regulators* that are all important. However, its standard usage in mathematics and computer science is to refer to discrete and countable entities. So, a natural interpretation of combinatorial regulation is that each regulator has a threshold leading to multiple discrete inputs, which are combined combinatorially to produce the target gene's response. Then, for a small number of genes, many alternative states can be specified. In the case of binary thresholds, 5 regulators could specify $2^5 = 32$ states. For 20 genes, there are over a million possible states.

It would seem this is what Wolpert (1989) had in mind:

> It is also far from clear whether the specification of cell state or cell differentiation is combinatorial or not. In a combinatorial system the number of signals required, or genes activated, to specify a cell would be small in relation to the total number of specified states. For *Drosophila* at least, while more than one gene is used to specify cell states in early development, the number of genes seems to be similar to the number of states.

The combinatorial explosion of combinatorial regulation makes reconstructing regulatory networks computationally challenging with many different approaches presented for mitigating this. One method, for example, is to model the enhancer as a logical function that combines its inputs with logical AND or OR. This can significantly reduce the problem. However, it is difficult to see how this maps plausibly to mechanism. In this case, the enhancer is required to implement independent thresholds for each transcription factor and then on top of that to impose higher-level logical functions (e.g., with adapter proteins or DNA looping) to produce the desired outcome. Fortunately, the model presented here has demonstrated that a combinatorial model is not necessary in the case of the *Drosophila* segmentation network.

A logical model might be valuable in approximating gene regulatory behaviour in order to model large-scale theoretical network dynamics (see section 1.4.2), but this abstract modelling approach should be separated from the hypothesis that the actual mechanism is indeed combinatorial. A simpler, non-combinatorial mechanism can underlie the model presented here. Each regulator interacts with the enhancer according to its strength of contribution to the enhancer's transition, and hence the mechanism of transition need only exist at the level of the enhancer. Thus, it would seem that further evidence should be required before a combinatorial mechanism is taken for granted, at least in connection with the *Drosophila* segmentation network.

8.2 EVOLUTIONARY IMPLICATIONS

From an evolutionary point of view, an enhancer can be viewed as a function that reads the transcription factor concentrations and produces an output. Over evolutionary time, *cis* or *trans* mutations modify the function. It is therefore of interest to consider what constraints the function is under.

8.2.1 *Optimisation*

It is remarkable that the key regulators of the various *eve* enhancers can be discovered from gene expression measurements. The search for key regulators looked for the simplest models, and these seem to correspond to actual enhancer functions. This suggests that, in some way, the concentrations of transcription factors and their interpretation have been optimised. It is further striking that the *eve* stripes can be explained by an apparently small repertoire of transcription factors, which are used multiple times in different combinations* to produce different results.

It would be interesting to consider to what extent this is a feature of the concentrations of the available transcription factors (*trans*) or a reflection of the constraints on enhancer binding sites (*cis*). For example, could an enhancer accumulate binding sites for many factors that each provide a little information, such as for many broadly distributed activators? In this case, perturbing any one of these factors would produce a limited change in expression, which would make them harder to discover experimentally. However, it also means that there is less constraint on the removal of the corresponding binding sites. This suggests that evolutionary optimisation might make it inevitable that enhancer functions can be discovered.

∗ In a non-combinatorial manner.

8.2.2 *Origin of patterns*

This section will outline a few speculative ideas with the intention of showing that studying positional information can be useful when considering the origin of patterns.

As suggested by chapter 7, any reasonable gradient can suffice to create broad regions of expression, but opposing (exponential) gradients are rather flexible. Not only can they make broad expression patterns like the gap genes, but they can be used to create stripes of variable size. If broad expression domains, like the gap genes, cross-regulate each other, there is ample scope for setting up patterns that are robust to scaling and individual variability. It is possible that individual and environmental variability is the driver for the origin of gap gene cross-regulation, rather than it being a requirement for positional information *per se.* The importance of this is that the organism can make use of gradients of positional information even before complex cross-regulatory relationships have arisen. A related point is that the addition of a regulator to the enhancer function does not require it to be essential immediately. It can begin by making a weak contribution and, if that contribution is helpful, it can be strengthened over time.

Another more specific example is *bcd.* BCD is not essential to the spatial positioning of eve stripe 2 (in fact, if the model of section 6.2 is correct, regulation nearer the anterior pole is more important). *Hunchback* positioning is also not as tightly controlled by BCD as first supposed (with the thresholding model). In fact, if a dual-regulating BCD model is accepted, BCD positional information (not purely repressive) might be more important near the anterior pole of the embryo than towards the centre of the embryo where the focus has usually been regarding its positional information. *bcd* might have acquired a broader activating role over time. This could have some relevance to the origin of patterning in long germ insects like *Drosophila.*

Although the examples provided here are necessarily incomplete and unsatisfactory, they do provide an idea of how further work might proceed.

8.2.3 *Binding site changes*

The *eve* enhancers have clusters of binding sites for their known regulators (Berman et al., 2002). Also, studies of spatially-averaged transcription factor binding (ChIP-chip) have shown that gap and maternal factors bind extensively in the enhancer regions (Li et al., 2008; MacArthur et al., 2009). This had led to the conclusion that multiple binding sites are important for regulatory behaviour, but the meaning behind the arrangement of binding sites—sometimes referred to as the grammar of transcription factor binding—has been elusive (Lusk and Eisen, 2010).

The positional model of chapter 7 requires that the coefficients for the same regulators differ across enhancers. Since the coefficients correspond to varying strengths of interaction of each regulator with the enhancer and the transcriptional machinery, it is plausible that the strength of in vivo 'coefficients' can be tuned if there are multiple binding sites. If this is the case, it would also buffer the organism against random mutations, making many changes relatively neutral, and yet not completely so.

A further suggestion of the model of this thesis is that the relative positions of binding sites for different transcription factors might not be crucial; interactions between factors were not found to be necessary. This, again, makes evolutionary change less brittle. The effect of changes within the binding sites of one factor will be relatively independent of those for other factors. Thus each factor can be tuned separately over evolutionary time. This still applies in the case of dual regulation, where closely situated binding sites might be important, as only one factor is being affected by the change.

Finally, the model suggests that since the transition is at the level of the enhancer, one factor can substitute for another, in proportion to the strength of its interaction with the enhancer and the transcriptional machinery. This provides much greater flexibility to the routes that evolution might take. It also goes some way to explaining the degree of optimisation seen in the data. If a factor provides good positional information and its binding sites start appearing in an enhancer, it can be useful immediately. There is no need for it to have the 'correct' threshold. Instead, it can act to buffer or support existing transcription factors, and if these, as a result, prove to be less useful, they can slowly reduce in importance—again, without having to put a vital threshold in danger.

Clearly, then, the model proposed in this thesis is compatible with the requirement for evolutionary flexibility in the interpretation of positional information.

8.3 ENHANCER STRUCTURE

One interesting aspect that has not received much focus here is how the overall function of the gene has been decomposed into individual enhancers, which will be referred to as *enhancer structure*. For example, is it merely coincidental that stripe 1 can pair with a number of other stripes (see figure 5.4 on page 104), or does this reflect mechanism or evolutionary history? And why are some stripes controlled in pairs? Although these questions are not explored here, it is worth making a few preliminary comments showing how enhancer structure can be investigated using positional information.

8.3.1 *eve 2 and eve 3+7*

It is particularly noteworthy that stripe 7 can be trained either with stripe 2 or with stripe 3, but all three cannot be trained together. This is rather interesting because, as described in Small et al. (1996), the sequences for stripe 7 are thought to be distributed across the *cis*-regulatory region that includes the *eve* 2 and *eve* 3+7 enhancers. Also, at times, a transgenic reporter for *eve* 2 will produce weak expression in stripe 7. Yet, the minimal enhancers for stripe 2 and stripes 3+7 cannot be placed right next to each other without a spacer sequence. The expression of the one furthest from the promoter is usually disrupted, which mirrors the model findings strikingly. It suggests that enhancer structure might be the result of evolutionary optimisation, reflecting what is possible given the available positional information. If this is the case, then studying what can and cannot be captured by the model could shed light on enhancer structure and evolution.

8.3.2 *Stripes 2, 5 and 7*

In this regard, the results for stripes 2, 5 and 7 are interesting and unexpected: stripe 5 can be trained with stripes 2 and 7. The fit is so good that when two out of the three are selected, parts of the third will sometimes be predicted (see figures 5.3 and 5.4 starting on page 103). This strongly indicates that they have highly compatible regulator concentrations, particularly since no other selection of three stripes can be trained together. However, stripe 5 is not regulated in vivo by *eve* 2 or *eve* 3+7. In fact, the *eve* 5 enhancer is located downstream of the *eve* gene, on the opposite side to the *eve* 2 and *eve* 3+7 enhancers (figure 1.5, page 9). This could be suggestive of higher-level structure in the *eve* locus, such as DNA looping; for example, the requirement for only one of *eve* 3+7 or *eve* 2 to be active at a time (because of their incompatibility), but with *eve* 5 able to be active with either separately. It could also reflect duplication of an ancestral DNA regulatory region and subsequent specialisation of each to fewer stripes. Synthetic enhancers and comparisons with other insects will be helpful in explaining these observations.

8.4 IMPLICATIONS FOR DYNAMIC MODELS

One of the limitations of the model of this thesis is that it does not describe the dynamic progression of gene expression in *Drosophila*. Surprisingly, though, the static nature of the model (and the use of measurements from the same cohort) has not been a significant limitation to discovering regulatory interactions, explaining mutant behaviour and analysing positional information in the embryo. This deserves consideration.

The simplest explanation is that the gap gene patterns are relatively stable over the time period considered. However, shifts of gap gene expression domains have been observed in another experimental context (Jaeger et al., 2004), and increasing levels of concentration and sharpening are also observed in the *Virtual Embryo*. It would be interesting to explore this in further detail, but two comments can be made suggesting how these observations are consistent with the models of this thesis:

- The modelling approach presented here may have met its original goal: it is robust to slight inaccuracies in the data and relatively modest movement of the gap gene patterns. And with the exception of the posterior GT domain, the movements of the gap gene patterns observed experimentally by Jaeger et al. (2004) do appear to be modest over a period of 10 or 20 minutes.

- An intriguing possibility is that *eve*'s regulators are largely in balance (as explained in section 7.2.2 on page 158). This would explain the apparent constant transcription rate of *eve* (see the conclusions of section 2.6), the sharpening of the *eve* stripes, and indeed the sharpening of the gap genes themselves.

If the latter point is valid, it suggests that some features of *Drosophila* anteroposterior patterning that have been attributed to dynamic properties of the regulatory network are instead the result of a static function: the interpretation of positional information by regulatory DNA. This might be true for the sharpening of patterns and, as shown in chapter 7, for aspects of the embryo's robustness to variation.

Dynamic properties are important for representing effects such as diffusion and moving expression patterns and nuclei. However, modelling these effects is computationally complex and including them has prevented models from explaining *Drosophila* anteroposterior patterning in all nuclei of the embryo. Thus, as this thesis has demonstrated, static models are an important alternative, even for a dynamic system like the developing embryo.

8.5 CONCLUDING REMARKS

In conclusion, this thesis has presented evidence that positional information in the *Drosophila* blastoderm is relatively stable. This validates the use of static models such as the one presented here, which is encouraging for a number of reasons. Firstly, these models have fewer parameters and are typically more accessible than dynamic models. This makes them more open to challenge and subsequent improvement, particularly since they suggest hypotheses and can guide future experimental work. Secondly, the success of the static positional model indicates that further theoretical and general principles might well be applicable in the field of developmental biology. And finally, the explanatory simplicity of the model is attractive for understanding gene regulation more generally. If a gene responsible for a spatially complex pattern can be described relatively simply, then surely there must be many more apparently complex biological systems that can ultimately be understood.

A

FINDING RATE FROM RATIO MEASUREMENTS

For simplicity, some analyses of transcriptional data use mRNA level as a proxy for rate. This was done, for example, in the initial analysis of regulatory relationships in the *Virtual Embryo* (Fowlkes et al., 2008) and in Segal et al. (2008). There are two conditions under which this can be done.

A.1 THE STEADY STATE ASSUMPTION

In many analyses (for example, comparison of two conditions in microarray data, or evaluation of fold change for site occupancy models; Bintu et al., 2005), it is usual to suppose that the rate of transcription is stable, although dependent on the experimental condition under study. In other words, $f(\mathbf{x}) = \alpha$ where α is constant over the time period. In this case, the mRNA level will reach a steady state when the amount of mRNA lost through decay is equal to the transcription rate i.e. $\alpha = \beta y$. This gives

$$y = \frac{\alpha}{\beta}$$

Thus, with a constant mRNA decay rate β one can interpret two different levels of mRNA abundance at steady state as the result of two different transcription rates, α_1 and α_2. In this, case the ratio of y_1 and y_2 will equal α_1/α_2, providing a direct measure of the ratio of transcription rates. This result holds even if the measure of abundance is relative as long as the scaling factor is the same between the two measurements and both measurements are zero when there is no mRNA. In other words, if the actual abundance (say, molecular count) of mRNA is Z but y is measured where $y = \gamma Z$, then the resulting ratio of mRNA measurements *at steady state* will still yield a ratio of transcription rates.

The *Virtual Embryo* data set contains measurements of relative abundance scaled in proportion to the maximum measurement over all time points (see section 1.5.3). It is further reasonable, as discussed in Fowlkes et al. (2008), that the measurements are proportional to actual abundance. This means that each measurement that is near *steady state* can be viewed as a direct measure of transcription rate in the form of a ratio to the maximum rate. Thus measurements of the same probe may be supposed to have the same (unknown) scaling factor. If the maximum measurement is $y_{max} = \gamma Z_{max}$, then another measurement is $y_i = \gamma Z_i$, but since y_i is expressed as a ratio of y_{max}

$$y_i = \frac{Z_i}{Z_{max}}$$

In terms of the above assumptions of steady state, it is measuring

$$y_i = \frac{\alpha_i}{\alpha_{max}}$$

and thus

$$\alpha_i = \alpha_{max}\, y_i \tag{A.1}$$

where α_{max} is unknown

The ideal steady state is never actually reached, but rather approached ever more closely. For this reason, it is generally an approximation, which becomes better the longer the dynamics proceeds. The speed

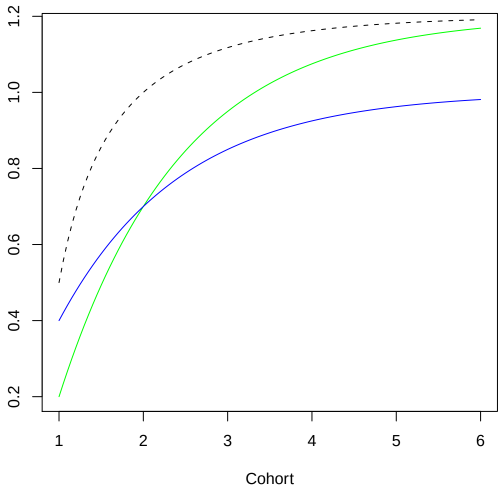


Figure A.1: Estimating rate with the steady state approximation. The solid lines represent the time course of two nuclei with different transcription rates (and a half-life of about 10 minutes) leading to different steady states. The dashed line shows the steady state approximation of their relative transcription rates, which approaches 1.2

that the mRNA approaches steady state is determined by the decay rate (and not the transcription rate). When the differential equation:

$$\frac{dy}{dt} = \alpha - \beta y$$

is solved, with $y = y_0$ at $t = 0$, then

$$y = \frac{\alpha}{\beta} - e^{-\beta t}(\frac{\alpha}{\beta} - y_0) \tag{A.2}$$

Thus, the difference between steady state α/β and the current level y decays exponentially, where the rate of decay is determined by β. The accuracy of estimation using the steady state assumption is shown in figure A.1 for the time periods of the *Virtual Embryo* data set.For a constant mRNA half-life (10 minutes), the time course for two different transcription rates with different starting values is plotted: $\alpha/\beta = 1.2$ (green) and $\alpha/\beta = 1$ (blue) leading to two different steady states. The approximation to the relative rate of transcription (which is equal to 1.2) is also plotted (dashed line).

In this particular situation it can be seen that the approximation is only good after the mRNA levels are fairly high—and this is with quite a fast mRNA decay rate.

A.2 SIMILAR START TIMES

Since the concentrations of HB in the nuclei of the *Virtual Embryo* are apparently changing over time, it is surprising that the continuous model of section 7.4 is able to fit the data so well. One condition under which the probe measurement (which is relative) can be identified with a rate is described here.

The estimate from the previous section is more accurate if both nuclei start transcription at similar times, and each maintain their own constant transcription rate. To show this, let $t = 0$ be the start of transcription of the *second* nucleus. At this point, the first nucleus can be assigned an mRNA level of y_0. Then the ratio of their levels at any time t can be given by:

$$\frac{y_1}{y_2} = \frac{\frac{\alpha_1}{\beta} - e^{-\beta t}(\frac{\alpha_1}{\beta} - y_0)}{\frac{\alpha_2}{\beta} - e^{-\beta t}(\frac{\alpha_2}{\beta})}$$

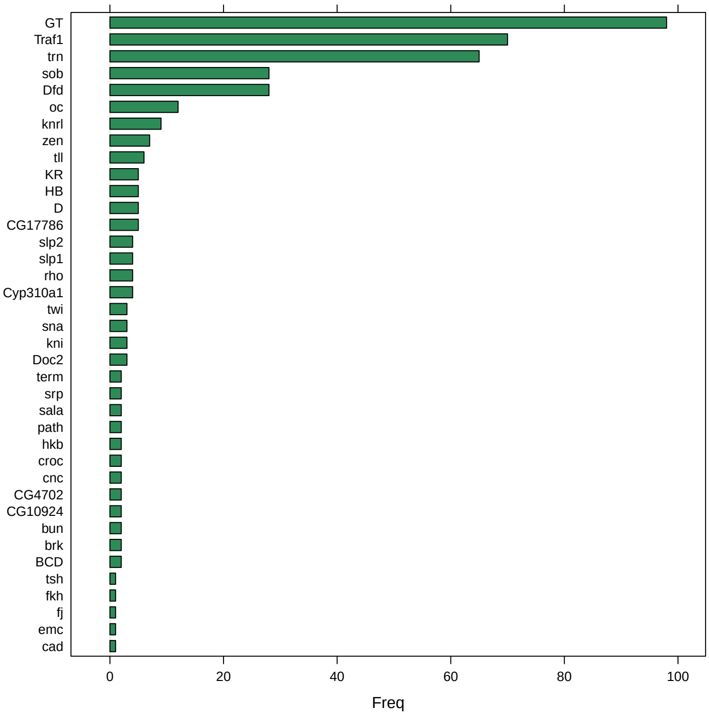
From this it can be seen that, as before, as $t \to \infty$ (steady-state), then $y_1/y_2 \to \alpha_1/\alpha_2$. But now it can also be seen that in the case where $y_0 = 0$ (which is true when transcription of the two nuclei has begun at the same time), then $y_1/y_2 = \alpha_1/\alpha_2$ for all t. This means that if nuclei are switched on at a similar time it is also possible to use relative mRNA levels to discover relative rates, even if they are not yet near steady state.

B

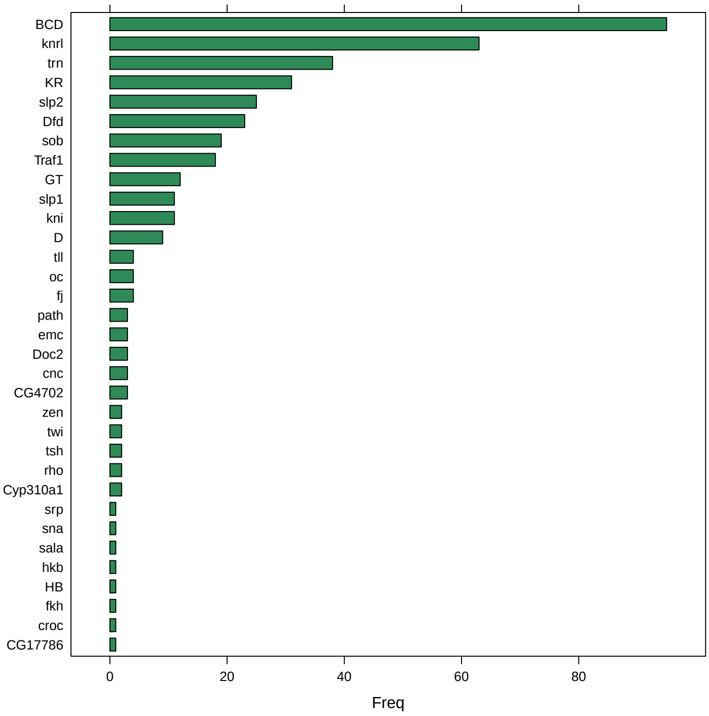
EXHAUSTIVE SUBSET SELECTION

B.1 REGULATORS OF THE TOP SCORING MODELS

Section 5.4.4 on page 107 used heat maps to highlight the best performing regulators considering all models of a certain size. This section summarises the same results by showing how many times each regulator occurs in the top 100 scoring models.

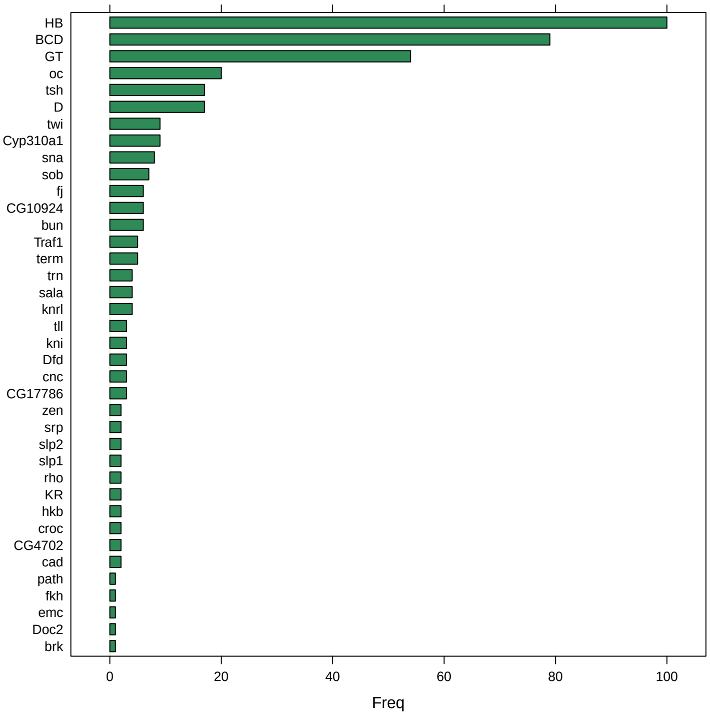


(a) Linear

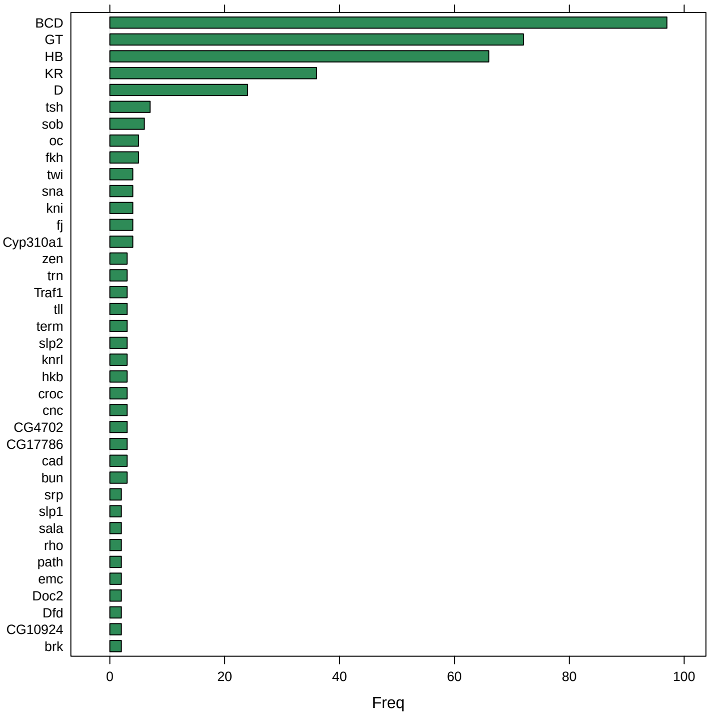


(b) With quadratic HB and BCD

eve stripe 1.

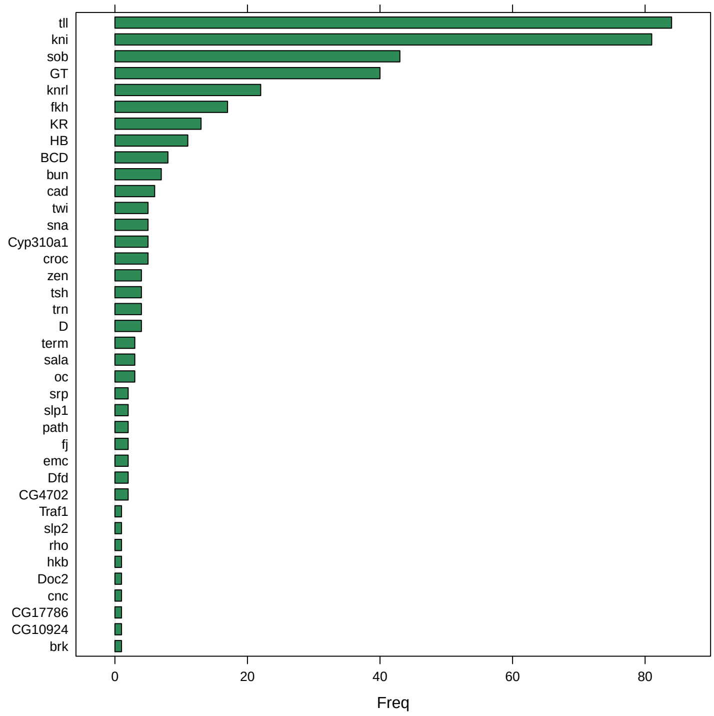


(a) Linear

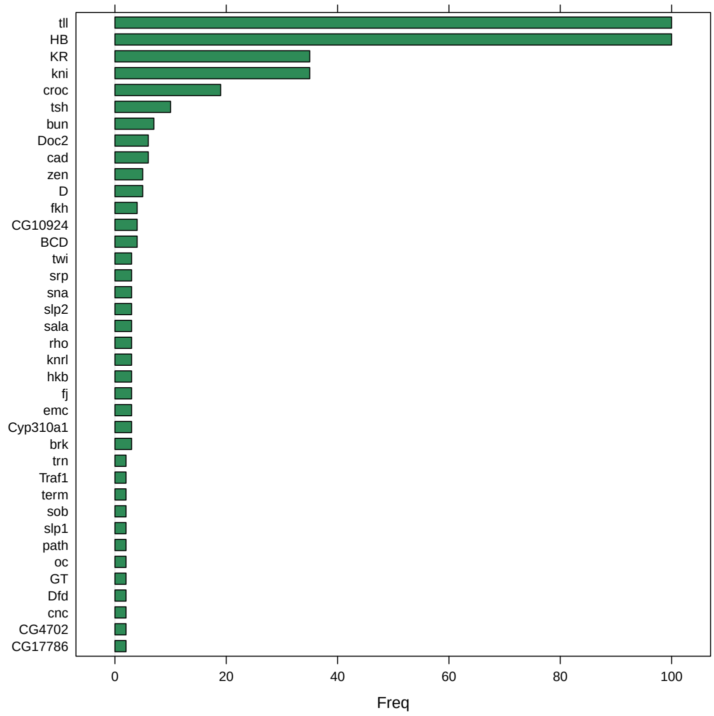


(b) With quadratic HB and BCD

eve stripe 2.

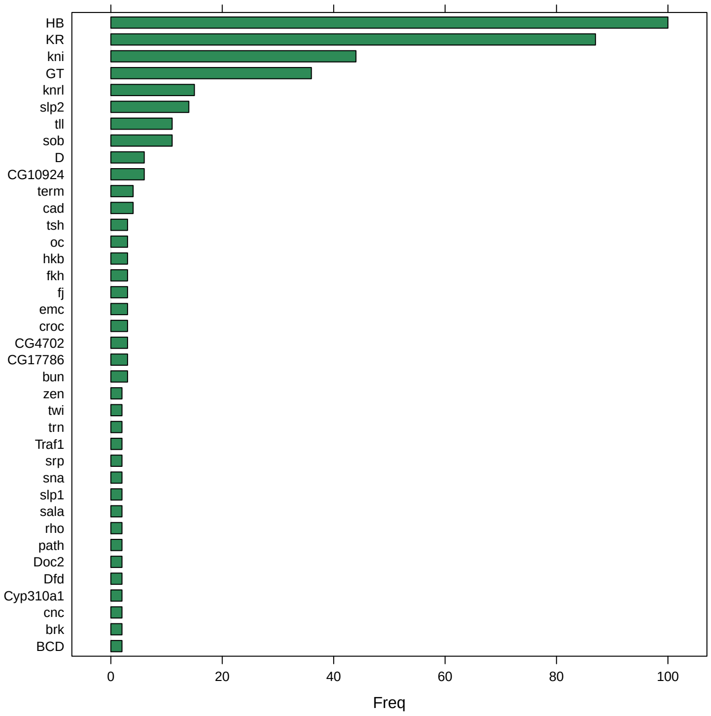


(a) Linear

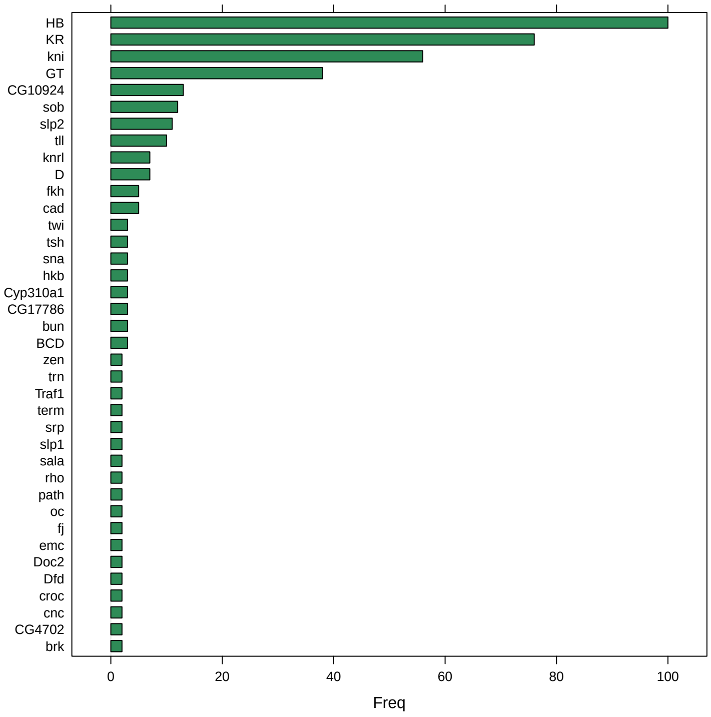


(b) With quadratic HB and BCD

eve stripes 3+7.

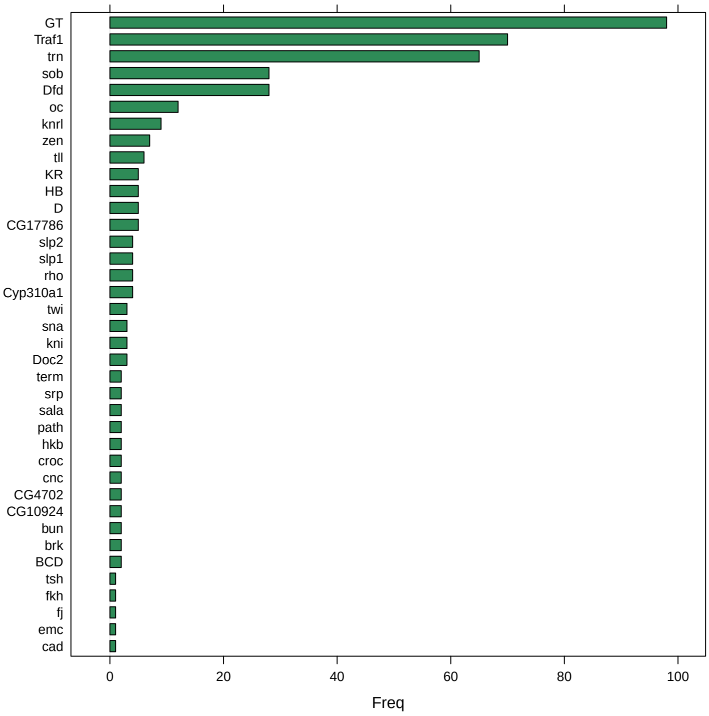


(a) Linear

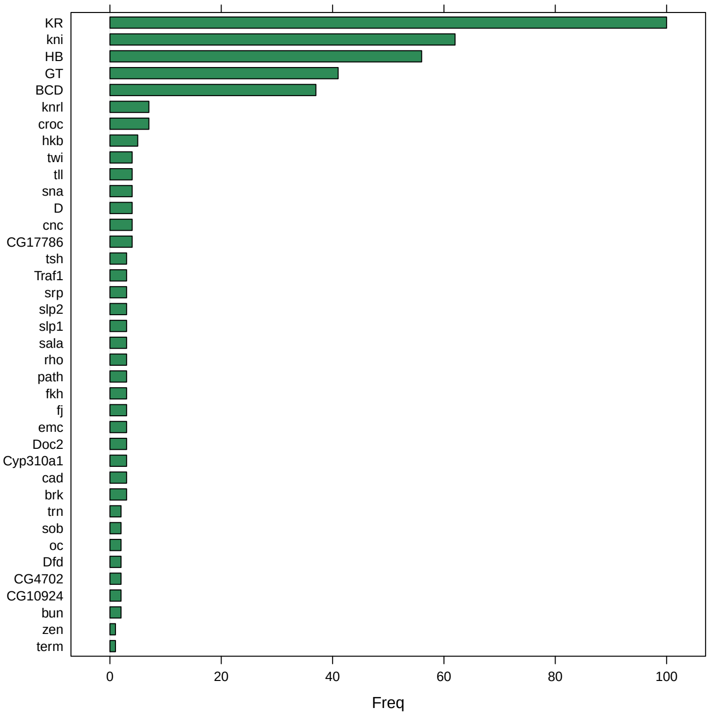


(b) With quadratic HB and BCD

eve stripes 4+6.



(a) Linear

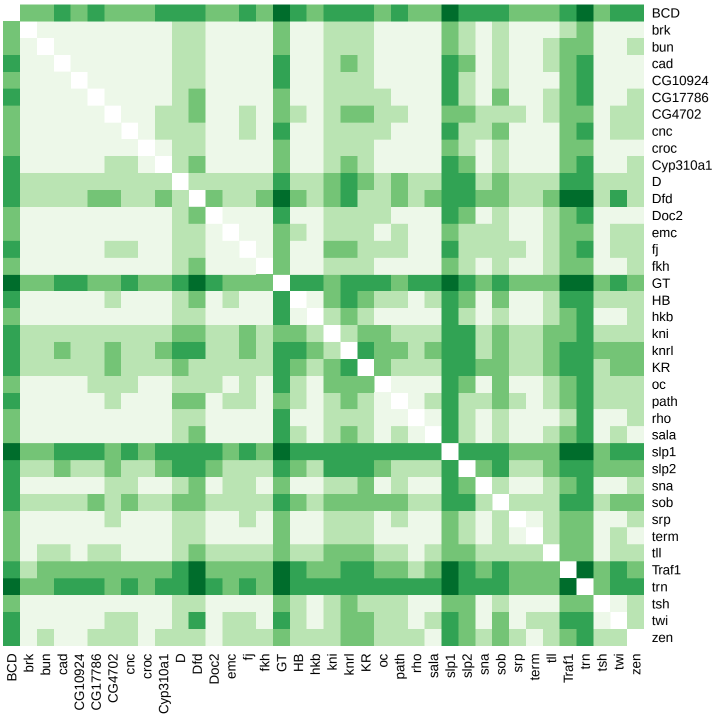


(b) With quadratic HB and BCD

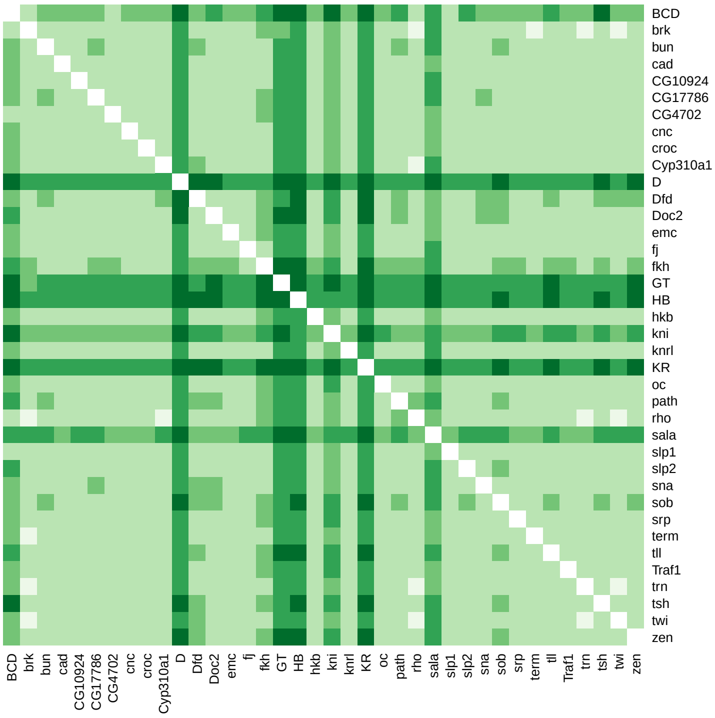
eve stripe 5.

B.2 FULLY QUADRATIC MODELS

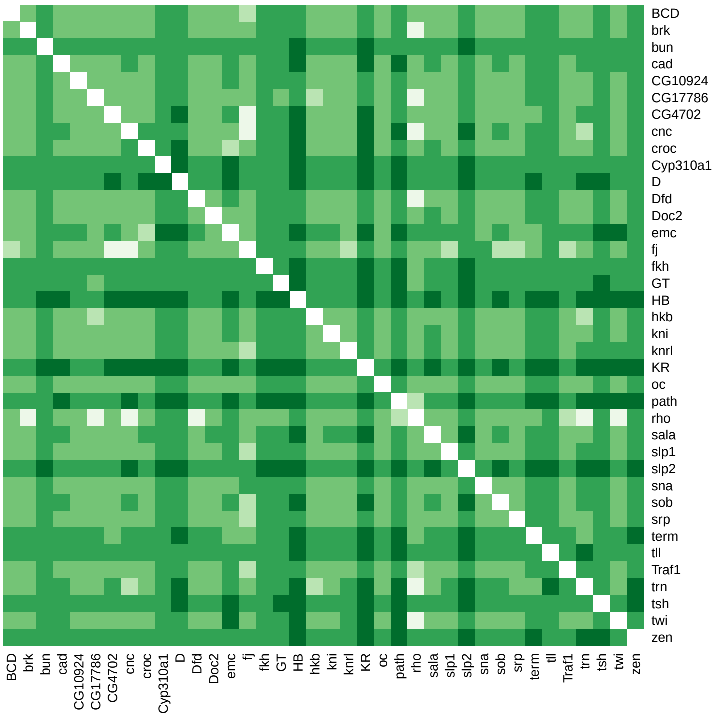
The results of section 5.4.4 on page 107 only included quadratic terms for HB and BCD. In this section, quadratic terms are added for every regulator considered. For these results, the minimum value of the regulator function was restricted to -300.



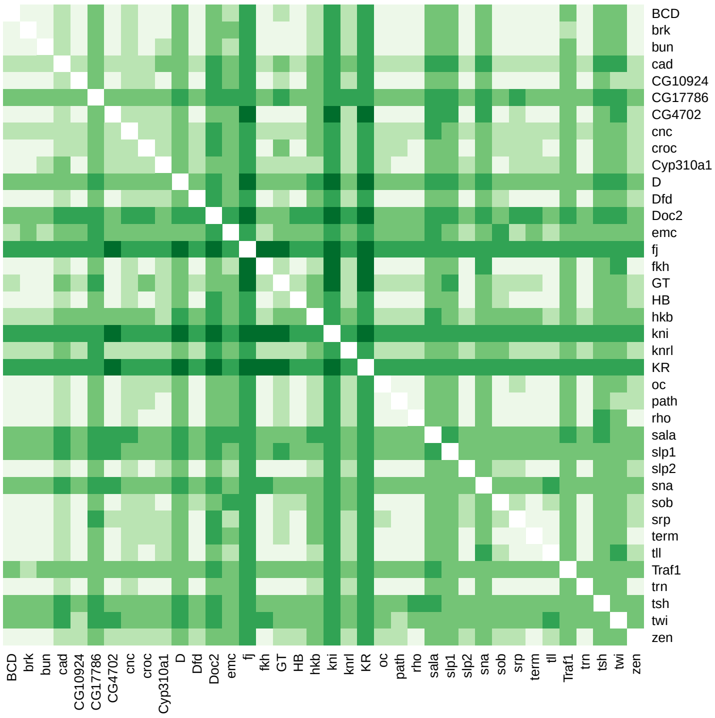
Best performing regulators for *eve* stripe 1.



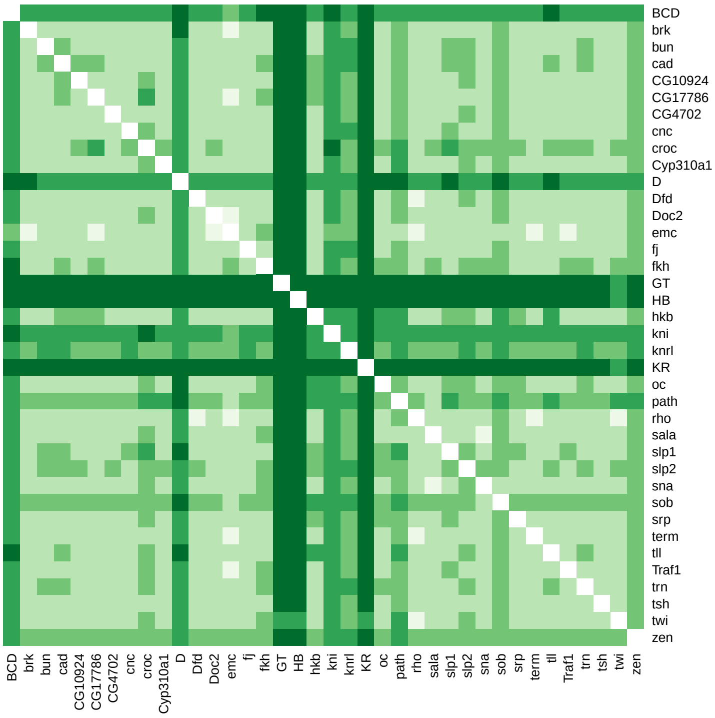
Best performing regulators for *eve* stripe 2.



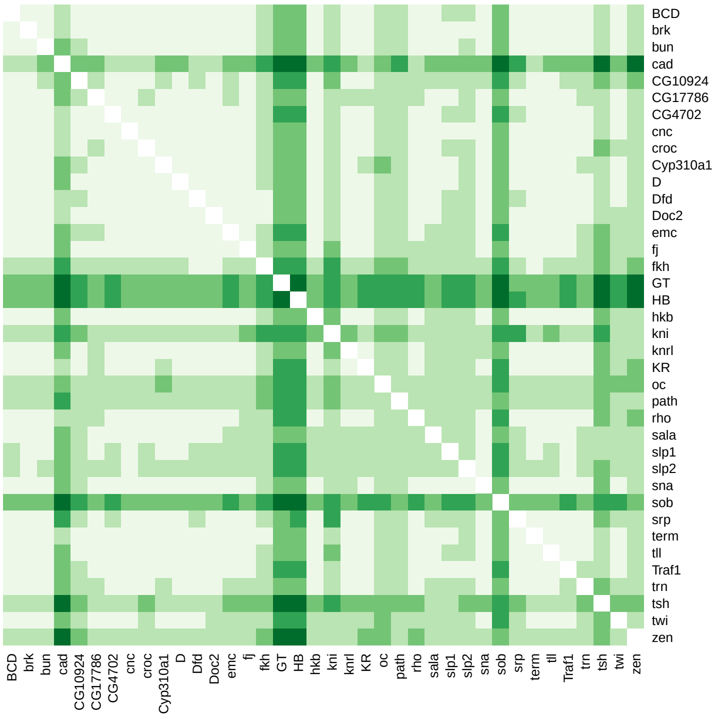
Best performing regulators for *eve* stripe 3.



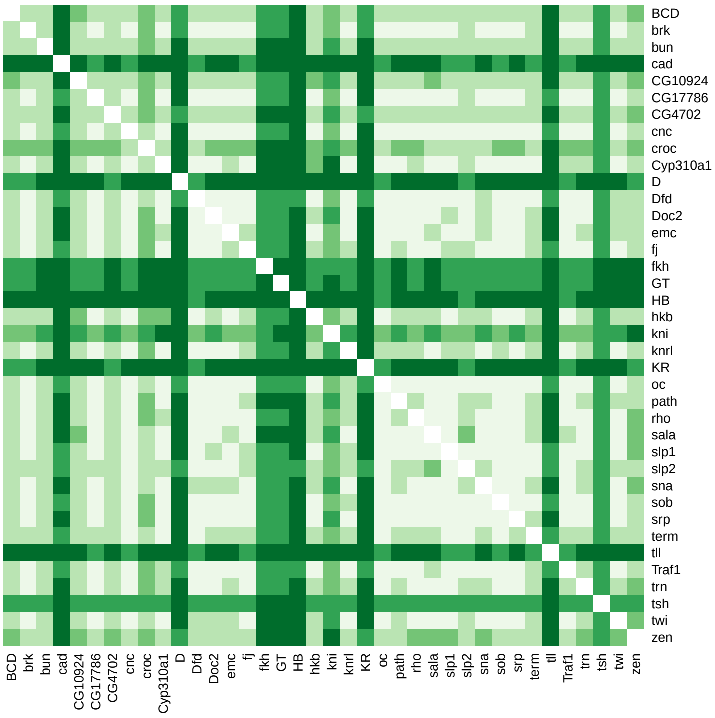
Best performing regulators for *eve* stripe 4.



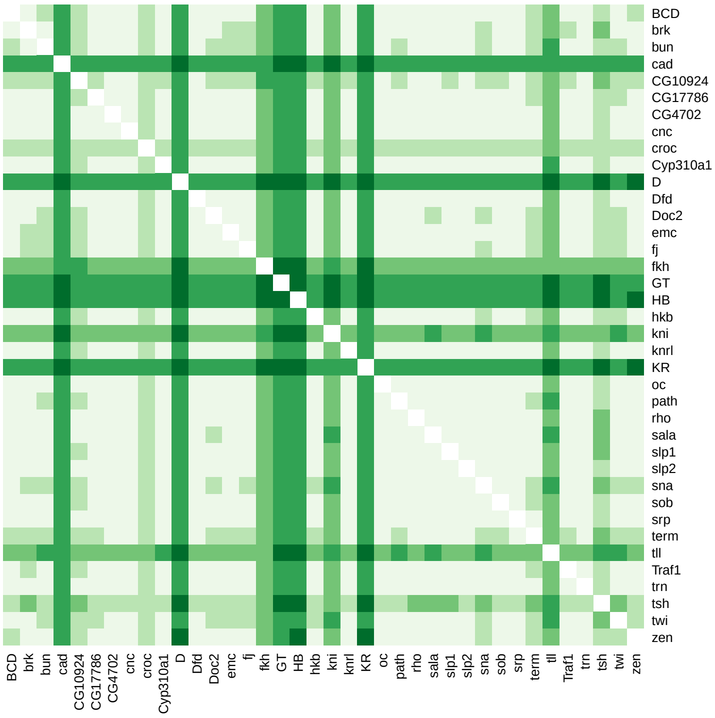
Best performing regulators for *eve* stripe 5.



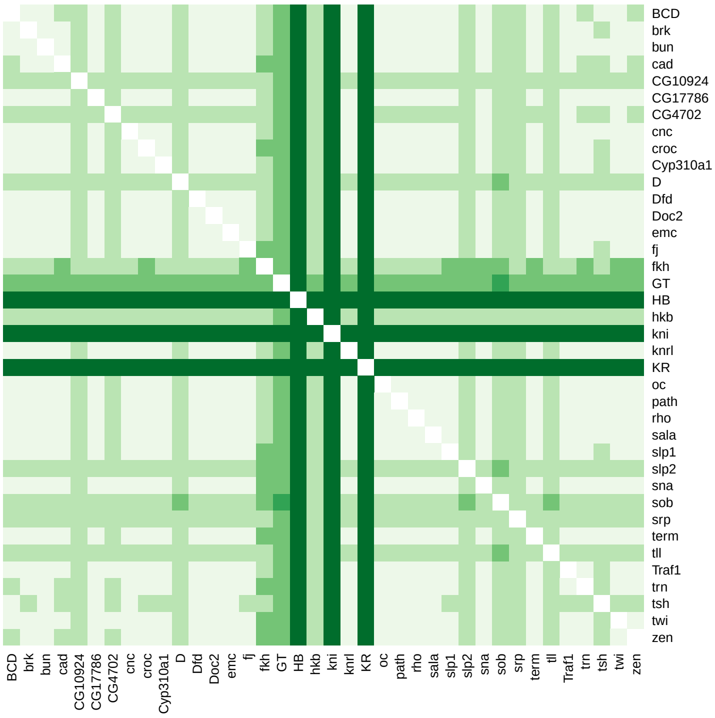
Best performing regulators for *eve* stripe 6.



Best performing regulators for *eve* stripe 7.



Best performing regulators for *eve* stripe 3+7.



Best performing regulators for *eve* stripe 4+6.

C

DESIGN OF A TRANSGENIC REPORTER

C.1 BINDING SITE SELECTION

One way to test the hypothesis that TLL is important for posterior repression of *eve* stripe 7 is to use a transgenic reporter based on the *eve* 3+7 enhancer, but with potential TLL binding sites mutated. The design of such an enhancer is described here.

The first step was to identify TLL, HB and KNI binding sites in the region of the stripe 3+7 enhancer. This was done using the *Drosophila melanogaster*, April 2006 (BDGP Release 5/dm3) assembly provided by the Berkeley Drosophila Genome Project (Celniker et al., 2002). The motifs used for searching were from the Berkeley Drosophila Transcription Network Project web site (BDTNP, 2007). Searching was done with `patser-v3e` (Hertz and Stormo, 1999), in both directions, for the genome sequence 20,000 base pairs long starting from position 5,852,335. Base pair prior frequencies were set at 0.29 for A and T, and 0.21 for C and G. The matches for scores above 4 in the vicinity of the *eve* stripe 3+7 enhancer are shown in figure C.1. The location of the hits within part of the sequence are shown in figure C.3.

In order to remove TLL binding sites, but preserve binding sites for KNI and HB, the following base pair changes were proposed:

```
5863075 - 5863077 to CAA
5863348 - 5863350 to GCG
5863403 - 5863405 to AGG
```

This was verified by performing the same `patser` search on the original and the mutated DNA (figures C.1 and C.2).

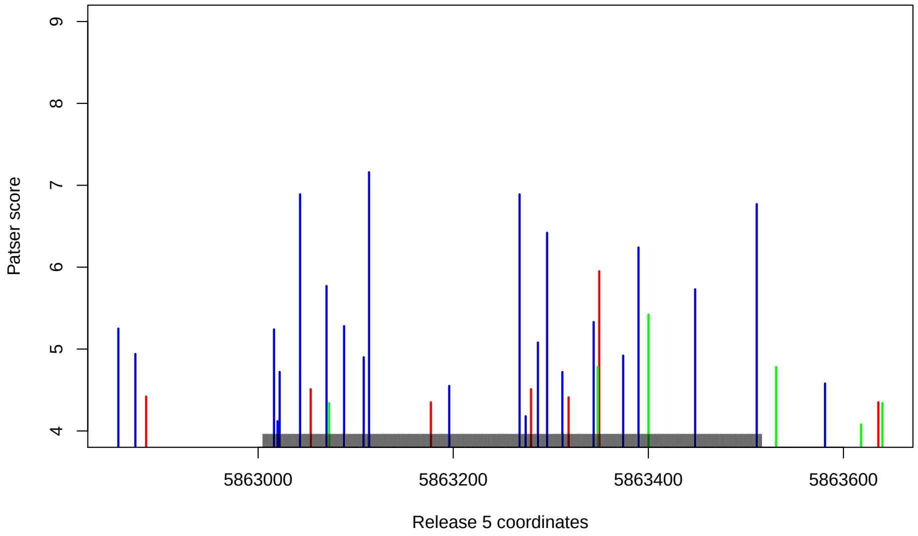


Figure C.1: Patser results using the TLL (green), HB (red) and KNI (blue) position weight matrices. The dark band on the axis shows the location of the *eve* 3+7 enhancer, annotation from ORegAnno (Griffith et al., 2008).

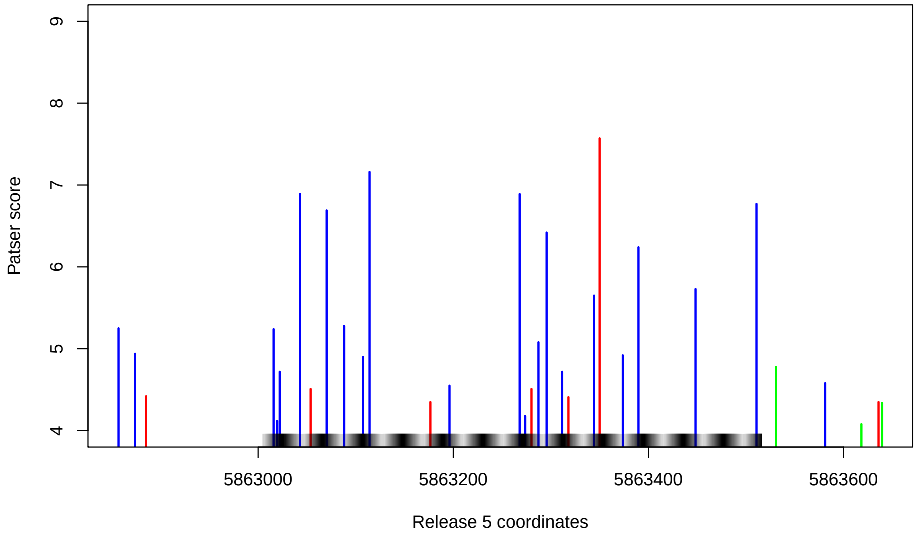


Figure C.2: Analysis for figure C.1 repeated on mutated DNA.

```
>chr2R:5862800-5863999
TTAGTCGTTGTCCGGGACAGGAGAGTATGCGGAAGGACATGCGTGAGTTTATTGCCCGCT
CGAATTTCCACTAAAAATTGGGCCGAAAAAAAAACAACTAGGTAGGACTAGGAACTGCAA
ACTAGCAAAGCGGACGCGCCTTTTTATTGGTGCACCTTCGGCGGAACCGCAGGATAACAG
CAGTAAAAGCGACGACGAGGACACAAGGATCCTCGAAATCGAGAGCGACCTCGCTGCATT
AGAAAACTAGATCAGTTTTTTGTTTTGGCCGACCGATTTTTGTGCCCGGTGCTCTCTTTA
CGGTTTATGGCCGCGTTCCCATTTCCCAGCTTCTTTGTTCCGGGCTCAGAAATCTGTATG
GAATTATGGTATATGCAGATTTTTATGGGTCCCGGCGATCCGGTTCGCGGAACGGGAGTG
TCCTGCCGCGAGAGGTCCTCGCCGGCGATCCTTGTCGCCCGTATTAGGAAAGTAGATCAC
GTTTTTTGTTCCCATTGTGCGCTTTTTTCGCTGCGCTAGTTTTTTTCCCCGAACCCAGCG
AACTGCTCTAATTTTTTAATTCTTCACGGCTTTTCATTGGGCTCCTGGAAAAACGCGGAC
AAGGTTATAACGCTCTACTTACCTGCAATTGTGGCCATAACTCGCACTGCTCTCGTTTTT
AAGATCCGTTTGTTTGTGTTTGTTTGTCCGCGATGGCATTCACGTTTTTACGAGCTCGTT
CCTTCGGGTCCAAAATTATGCCAGTTTGTTTTGTCTCTGGCAATTATTGGAAATTTCATT
GGGTCGATTTCGCTGCCTTCCTTGCTCTTCCCTTGAGAAAAGTGAATAGGTTGTGCCATA
AAAATCGCTGCTCCTGAAGACCAAATGAAATGGATTTGTGTAAGCATTAAAAACGCGAGG
CAAGCCCCAAGATTCCTCCACTGCTTTTTTTATATTGCCCACTGCTAAATGCAGCTAATT
CGTCGATTGTTTAAAAATTAAATTACTTATGTTGCCATTCATACATCCCCTCACATTTTA
TGGCCATTTGAGTGCGGGGTGCACAGTTCTGTCTTAAGTGGCGGATGGAAACCACCACAT
TTACTCGAGGGATGATGTGCTCTAATATCTCCTCATCAAATGGGATGGTTTCTATGGAAA
GGCAAAATCGTTGTAAAGTGAGGCGGAGTTAAAAAATACCTTGTTATAGCCTTTTTAAAA
```

Figure C.3: *eve* 3+7 regulatory genomic sequence from the UCSC Genome Browser (Rhead et al., 2010). The underlined sequence is the minimal 3+7 enhancer as annotated by ORegAnno (Griffith et al., 2008). Patser hits are coloured red for HB, green for TLL and blue for KNI. Overlapping sites are indicated by the secondary colours (cyan for TLL and KNI, yellow for TLL and HB, and magenta for HB and KNI. An overlap of all three is coloured grey (the sequence ATTT).

BIBLIOGRAPHY

Ackers, G. K., Johnson, A. D. and Shea, M. A. (1982). Quantitative model for gene regulation by lambda phage repressor, *Proceedings of the National Academy of Sciences of the United States of America* **79**(4): 1129–1133.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE transactions on automatic control* **19**(6): 716–723.

Akam, M. (1987). The molecular basis for metameric pattern in the *Drosophila* embryo, *Development* **101**(1): 1–22.

Albert, R. and Othmer, H. G. (2003). The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*, *Journal of Theoretical Biology* **223**(1): 1–18.

Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*, 1st edn, Chapman & Hall/CRC.

Andrioli, L. P. M., Vasisht, V., Theodosopoulou, E., Oberstein, A. and Small, S. (2002). Anterior repression of a *Drosophila* stripe enhancer requires three position-specific mechanisms, *Development* **129**(21): 4931–4940.

Andrioli, L. P., Oberstein, A. L., Corado, M. S., Yu, D. and Small, S. (2004). Groucho-dependent repression by Sloppy-paired 1 differentially positions anterior pair-rule stripes in the *Drosophila* embryo, *Developmental Biology* **276**(2): 541–551.

Arnosti, D. N. (2003). Analysis and function of transcriptional regulatory elements: insights from *Drosophila*, *Annual Review of Entomology* **48**(1): 579–602.

BDTNP (2007). Berkeley *Drosophila* Transcription Network Project, `http://www.webcitation.org/query.php?url=http://bdtnp.lbl.gov/&refdoi=10.1186/gb-2006-7-12-r123`.

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M. and Eisen, M. B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome, *Proceedings of the National Academy of Sciences of the United States of America* **99**(2): 757–762.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J. and Phillips, R. (2005). Transcriptional regulation by the numbers: models, *Current Opinion in Genetics & Development* **15**(2): 116–124.

Bishop, C. M. (2007). *Pattern Recognition and Machine Learning*, Springer-Verlag New York Inc.

Bolouri, H. (2008). *Computational Modeling Of Gene Regulatory Networks – A Primer*, 1st edn, Imperial College Press.

Bolouri, H. and Davidson, E. H. (2003). Transcriptional regulatory cascades in development: Initial rates, not steady state, determine network kinetics, *Proceedings of the National Academy of Sciences* **100**(16): 9371–9376.

Brody, T. B. (2010). The Interactive Fly, `http://www.webcitation.org/query?url=http://www.sdbonline.org/fly/aimain/1aahome.htm&date=2010-05-29`.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization, *SIAM Journal on Scientific Computing* **16**(5): 1190–1208.

Campos-Ortega, J. A. and Hartenstein, V. (1997). *The Embryonic Development of Drosophila melanogaster*, 2nd edn, Springer.

Celniker, S. E., Wheeler, D. A., Kronmiller, B., Carlson, J. W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S. P., Frise, E., Hodgson, A., George, R. A., Hoskins, R. A., Laverty, T., Muzny, D. M., Nelson, C. R., Pacleb, J. M., Park, S., Pfeiffer, B. D., Richards, S., Sodergren, E. J., Svirskas, R., Tabor, P. E., Wan, K., Stapleton, M., Sutton, G. G., Venter, C., Weinstock, G., Scherer, S. E., Myers, E. W., Gibbs, R. A. and Rubin, G. M. (2002). Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence, *Genome Biology* **3**(12).

Chang, C. and Lin, C. (2001). LIBSVM : a library for support vector machines.

Clyde, D. E., Corado, M. S. G., Wu, X., Pare, A., Papatsenko, D. and Small, S. (2003). A self-organizing system of repressor gradients establishes segmental complexity in *Drosophila*, *Nature* **426**(6968): 849–853.

Collett, D. (2002). *Modelling Binary Data*, 2nd edn, Chapman and Hall/CRC.

Crick, F. (1970). Diffusion in embryogenesis, *Nature* **225**(5231): 420–422.

Dobson, A. J. and Barnett, A. (2008). *An Introduction to Generalized Linear Models*, 3rd edn, Chapman and Hall/CRC.

Driever, W. and Nüsslein-Volhard, C. (1988a). The bicoid protein determines position in the *Drosophila* embryo in a concentration-dependent manner, *Cell* **54**(1): 95–104.

Driever, W. and Nüsslein-Volhard, C. (1988b). A gradient of bicoid protein in *Drosophila* embryos, *Cell* **54**(1): 83–93.

Ephrussi, A. and Johnston, D. S. (2004). Seeing is believing:: The bicoid morphogen gradient matures, *Cell* **116**(2): 143–152.

Fawcett, T. (2006). An introduction to ROC analysis, *Pattern Recognition Letters* **27**(8): 861–874.

Fowlkes, C. C., Luengo Hendriks, C. L., Keränen, S. V., Weber, G. H., Rübel, O., Huang, M., 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. and Malik, J. (2008). A quantitative spatiotemporal atlas of gene expression in the *Drosophila* blastoderm, *Cell* **133**: 364–374.

Fowlkes, C. and Malik, J. (2006). Inferring nuclear movements from fixed material, *Technical Report UCB/EECS-2006-142*, University of California, Berkeley.

Frasch, M. and Levine, M. (1987). Complementary patterns of even-skipped and fushi tarazu expression involve their differential regulation by a common set of segmentation genes in *Drosophila.*, *Genes & Development* **1**(9): 981–995.

Frohnhöfer, H. G. and Nüsslein-Volhard, C. (1986). Organization of anterior pattern in the *Drosophila* embryo by the maternal gene bicoid, *Nature* **324**: 120–125.

Fujioka, M., Emi-Sarker, Y., Yusibova, G. L., Goto, T. and Jaynes, J. B. (1999). Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients, *Development (Cambridge, England)* **126**(11): 2527–38.

Gregor, T., Tank, D. W., Wieschaus, E. F. and Bialek, W. (2007). Probing the limits to positional information, *Cell* **130**(1): 153–164.

Griffith, O. L., Montgomery, S. B., Bernier, B., Chu, B., Kasaian, K., Aerts, S., Mahony, S., Sleumer, M. C., Bilenky, M., Haeussler, M., Griffith, M., Gallo, S. M., Giardine, B., Hooghe, B., Loo, P. V., Blanco, E., Ticoll, A., Lithwick, S., Portales-Casamar, E., Donaldson, I. J., Robertson, G., Wadelius, C., Bleser, P. D., Vlieghe, D., Halfon, M. S., Wasserman, W., Hardison, R., Bergman, C. M., Jones, S. J. and Consortium, T. O. R. A. (2008). ORegAnno: an open-access community-driven resource for regulatory annotation, *Nucl. Acids Res.* **36**(suppl_1): D107–113.

Hertz, G. Z. and Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences, *Bioinformatics (Oxford, England)* **15**(7-8): 563–577.

Hill, T. (1985). *Cooperativity Theory in Biochemistry: Steady-State and Equilibrium Systems*, 1st edn, Springer.

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

Hsu, C. W., Chang, C. C. and Lin, C. J. (2010). A practical guide to support vector classification, `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`.

Ingham, P. W. (1988). The molecular genetics of embryonic pattern formation in *Drosophila, Nature* **335**(6185): 25–34.

Irish, V., Lehmann, R. and Akam, M. (1989). The *Drosophila* posterior-group gene nanos functions by repressing hunchback activity, *Nature* **338**(6217): 646–648.

Jaeger, J. (2009). Modelling the *Drosophila* embryo, *Molecular BioSystems* **5**(12): 1549–1568.

Jaeger, J. and Martinez-Arias, A. (2009). Getting the measure of positional information, *PLoS Biol* **7**(3): e1000081.

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

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. and Reinitz, J. (2004). Dynamic control of positional information in the early *Drosophila* embryo, *Nature* **430**(6997): 368–371.

Janssens, H., Hou, S., Jaeger, J., Kim, A., Myasnikova, E., Sharp, D. and Reinitz, J. (2006). Quantitative and predictive model of transcriptional control of the *Drosophila melanogaster* even skipped gene, *Nat Genet* **38**(10): 1159–1165.

Keränen, S., Fowlkes, C., Luengo Hendriks, C., Sudar, D., Knowles, D., Malik, J. and Biggin, M. (2006). Three-dimensional morphology and gene expression in the *Drosophila* blastoderm at cellular resolution II: dynamics, *Genome Biology* **7**(12): R124.

Kraut, R. and Levine, M. (1991). Mutually repressive interactions between the gap genes giant and krüppel define middle body regions of the *Drosophila* embryo, *Development (Cambridge, England)* **111**(2): 611–621.

Lange, K. (1999). *Numerical Analysis for Statisticians*, 1st edn, Springer.

Lebrecht, D., Foehr, M., Smith, E., Lopes, F. J. P., Vanario-Alonso, C. E., Reinitz, J., Burz, D. S. and Hanes, S. D. (2005). Bicoid cooperative DNA binding is critical for embryonic patterning in *Drosophila*, *Proceedings of the National Academy of Sciences of the United States of America* **102**(37): 13176–13181.

Lemon, B. and Tjian, R. (2000). Orchestrated response: a symphony of transcription factors for gene control, *Genes & Development* **14**(20): 2551–2569.

Levine, M. (2008). A systems view of *Drosophila* segmentation, *Genome Biology* **9**(2): 207.

Li, X., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C. L., Chu, H. C., Ogawa, N., Inwood, W., Sementchenko, V., Beaton, A., Weiszmann, R., Celniker, S. E., Knowles, D. W., Gingeras, T., Speed, T. P., Eisen, M. B. and Biggin, M. D. (2008). Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm, *PLoS Biology* **6**(2): e27.

Luengo Hendriks, C., Keranen, S., Fowlkes, C., Simirenko, L., Weber, G., DePace, A., Henriquez, C., Kaszuba, D., Hamann, B., Eisen, M., Malik, J., Sudar, D., Biggin, M. and Knowles, D. (2006). Three-dimensional morphology and gene expression in the *Drosophila* blastoderm at cellular resolution I: data acquisition pipeline, *Genome Biology* **7**(12): R123.

Lusk, R. W. and Eisen, M. B. (2010). Evolutionary mirages: Selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers, *PLoS Genet* **6**(1): e1000829.

MacArthur, S., Li, X., Li, J., Brown, J. B., Chu, H. C., Zeng, L., Grondona, B. P., Hechmer, A., Simirenko, L., Keränen, S. V., Knowles, D. W., Stapleton, M., Bickel, P., Biggin, M. D. and Eisen, M. B. (2009). Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions, *Genome Biology* **10**(7): R80.

Manu, Surkova, S., Spirov, A. V., Gursky, V. V., Janssens, H., Kim, A., Radulescu, O., Vanario-Alonso, C. E., Sharp, D. H., Samsonova, M. and Reinitz, J. (2009a). Canalization of gene expression and domain shifts in the *Drosophila* blastoderm by dynamical attractors, *PLoS Comput Biol* **5**(3): e1000303.

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

Margolis, J. S., Borowsky, M. L., Steingrímsson, E., Shim, C. W., Lengyel, J. A. and Posakony, J. W. (1995). Posterior stripe expression of hunchback is driven from two promoters by a common enhancer element, *Development (Cambridge, England)* **121**(9): 3067–3077.

Mjolsness, E., Sharp, D. H. and Reinitz, J. (1991). A connectionist model of development, *Journal of Theoretical Biology* **152**(4): 429–453.

Moran, E. and Jimenez, G. (2006). The tailless nuclear receptor acts as a dedicated repressor in the early *Drosophila* embryo, *Mol. Cell. Biol.* **26**(9): 3446–3454.

Nasiadka, A., Dietrich, B. H., Krause, H. M. and DePamphilis, M. L. (2002). Anterior-posterior patterning in the *Drosophila* embryo, *Gene Expression at the Beginning of Animal Development*, Vol. Volume 12 of *Advances in Developmental Biology*, Elsevier, pp. 155–204.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society. Series A (General)* **135**(3): 370–384.

Nocedal, J. (2006). *Numerical optimization*, 2nd edn, Springer, New York.

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

Ochoa-Espinosa, A., Yu, D., Tsirigos, A., Struffi, P. and Small, S. (2009). Anterior-posterior positional information in the absence of a strong bicoid gradient, *Proceedings of the National Academy of Sciences* **106**(10): 3823–3828.

Papatsenko, D. (2009). Stripe formation in the early fly embryo: principles, models, and networks, *BioEssays* **31**(11): 1172–1180.

Papatsenko, D. and Levine, M. S. (2008). Dual regulation by the hunchback gradient in the *Drosophila* embryo, *Proceedings of the National Academy of Sciences* **105**(8): 2901–2906.

Patel, N. H., Condron, B. G. and Zinn, K. (1994). Pair-rule expression patterns of even-skipped are found in both short- and long-germ beetles, *Nature* **367**(6462): 429–434.

Perkins, T. J., Jaeger, J., Reinitz, J. and Glass, L. (2006). Reverse engineering the gap gene network of *Drosophila melanogaster*, *PLoS Comput Biol* **2**(5): e51.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*, MIT Press.

Reinitz, J., Hou, S. and Sharp, D. (2003). Transcriptional control in *Drosophila*, *Complexus* **1**(2): 54–64.

Rhead, B., Karolchik, D., Kuhn, R. M., Hinrichs, A. S., Zweig, A. S., Fujita, P. A., Diekhans, M., Smith, K. E., Rosenbloom, K. R., Raney, B. J., Pohl, A., Pheasant, M., Meyer, L. R., Learned, K., Hsu, F., Hillman-Jackson, J., Harte, R. A., Giardine, B., Dreszer, T. R., Clawson, H., Barber, G. P., Haussler, D. and Kent, W. J. (2010). The UCSC genome browser database: update 2010, *Nucleic Acids Research* **38**: D613–619.

Ripley, B. D. (2008). *Pattern Recognition and Neural Networks*, 1 edn, Cambridge University Press.

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

Sánchez, L. and Thieffry, D. (2003). Segmenting the fly embryo:: a logical analysis of the pair-rule cross-regulatory module, *Journal of Theoretical Biology* **224**(4): 517–537.

Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics* **6**(2): 461–464.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. and Gaul, U. (2008). Predicting expression patterns from regulatory sequence in *Drosophila* segmentation, *Nature* **451**(7178): 535–540.

Simpson-Brose, M., Treisman, J. and Desplan, C. (1994). Synergy between the hunchback and bicoid morphogens is required for anterior patterning in *Drosophila*, *Cell* **78**(5): 855–865.

Small, S., Blair, A. and Levine, M. (1992). Regulation of even-skipped stripe 2 in the *Drosophila* embryo, *The EMBO Journal* **11**(11): 4047–4057.

Small, S., Blair, A. and Levine, M. (1996). Regulation of two pair-rule stripes by a single enhancer in the *Drosophila* embryo, *Developmental Biology* **175**(2): 314–324.

Stanojevic, D., Small, S. and Levine, M. (1991). Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo, *Science* **254**(5036): 1385–1387.

Struhl, G., Johnston, P. and Lawrence, P. A. (1992). Control of *Drosophila* body pattern by the hunchback morphogen gradient, *Cell* **69**(2): 237–249.

Struhl, G., Struhl, K. and Macdonald, P. M. (1989). The gradient morphogen bicoid is a concentration-dependent transcriptional activator, *Cell* **57**(7): 1259–1273.

Thomas, R. (1973). Boolean formalization of genetic control circuits, *Journal of Theoretical Biology* **42**(3): 563–585.

Turing, A. M. (1952). The chemical basis of morphogenesis, *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences (1934-1990)* **237**(641): 37–72.

Venables, W. and Ripley, B. (2003). *Modern Applied Statistics with S*, 4th, corr. 2nd printing edn, Springer.

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

Wimmer, E. A., Carleton, A., Harjes, P., Turner, T. and Desplan, C. (2000). Bicoid-independent formation of thoracic segments in *Drosophila*, *Science* **287**(5462): 2476–2479.

Wolpert, L. (1968). The French flag problem: A contribution to the discussion on pattern development and regulation, *Towards a theoretical biology* **1**: 125–133.

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

Wolpert, L. (1989). Positional information revisited, *Development (Cambridge, England)* **107 Suppl**: 3–12.

Wolpert, L. (1996). One hundred years of positional information, *Trends in Genetics* **12**(9): 359–364.

Zinzen, R. P., Senger, K., Levine, M. and Papatsenko, D. (2006). Computational models for neurogenic gene expression in the *Drosophila* embryo, *Current Biology* **16**(13): 1358–1365.