Evolving networks as coupled differential equations: A phenotypic model of Dipterans during embryogenesis

Jeremy Rothschild

Master of Science

Department of Physics

McGill University

Montreal,Quebec

2017-04-15

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfilment of the requirements for the degree of Master of Science

©Jeremy B Rothschild, 2017

DEDICATION

This document is dedicated to my many close friends and family who have seen me through the good times and the bad in completing this work. It isn't always easy to dedicate your time to such a project, but having people that support you and make you smile can get you through anything.

ACKNOWLEDGEMENTS

I would like to thank Dr. Paul François for his patience working with me throughout my whole time in his research group. His constant guidance has equipped me with the many skills and the knowledge necessary to complete this work as well as prepare me for all the future work I mean to undertake. I can never thank you enough for the opportunity and the experience working with you has provided.

I would also like to thank the rest of the members of the Paul Françcois group for their (sometimes unknowing) input on many aspects of this work. Notably, the contributions of Peter Tsimiklis and Mathias Beaupeux, who offered innovative solutions for the problems at hand and who were always happy to discuss about anything and everything. You were always a source of exceptional insight and never shied from answering any inquiries I might have, no matter how trivial.

ABSTRACT

An elaborate network of genes defines the genetic profile along the anteroposterior axis of certain individuals in the Diptera family, establishing a robust pattern that differs slightly between species. Differences in the segmentation gene pattern of *Drosophila* and *Anopheles* suggest that the parameters defining their networks evolved differently from the last common ancestor onwards. The study of the evolution of the network, defined by a set of differential equations, using our genetic algorithm reveals a phenotypic pathway that explains these dissimilarities while conserving the necessary conditions for viable species. This computational search through a model and parameter "hyperspace" using the genetic algorithm predicts homologies between different stripe modules for the invariant target gene, *eve*. Additionally, a network model is further developed to explain the polarity of the pair-rule gene pattern expressed in the embryo, which suggests that ancestral Dipteran exploited a dynamic network to establish a proper periodic pattern of the other segmentation genes with respect to *eve*.

ABRÉGÉ

Un réseau complexe de gènes définit le profil génétique selon l'axe antéro-postérieur de certains individus de la famille des diptères et établit un modèle robuste qui diffère légèrement entre les espèces. Les différences dans la configuration spatiale des gènes de segmentation de *Drosophila* et *Anopheles* suggèrent que les paramètres définissant leurs réseaux ont évolué différemment à partir de leur dernier ancêtre commun. L'étude du réseau, défini par un ensemble d'équations différentielles, en utilisant notre algorithme génétique révèle une trajectoire phénotypique qui explique ces différences tout en conservant des conditions nécessaires pour garder des espèces viables pendant l'évolution. Cette recherche à travers un "hyperespace" de paramètres en utilisant l'algorithme génétique prédit des homologies entre les différents modules contrôlant les bandes de gène cible invariant *eve*. De plus, un modèle de réseau est développé pour expliquer la polarité du motif de gènes de segmentation exprimés dans l'embryon, ce qui suggère que cette éspèce ancestrale exploitait un réseau dynamique pour établir un motif périodique des autres gènes de segmentation par rapport à *eve*.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

ABSTRACT iv

ABRÉGÉ v

LIST OF TABLES viii

LIST OF FIGURES ix

1 Introduction 1

1.1 Developmental biology 1
1.2 Evolution and the genetic algorithm 2
1.3 Motivation 4
1.4 Thesis overview 7

2 Theoretical Framework 9

2.1 The French Flag Model 9
2.2 Dipteran embryogenesis 13
2.3 Gene regulatory networks 16
2.4 The *Drosophila* gene network 20

3 The genetic algorithm 28

3.1 Structure 28
3.2 Initial Population Configuration 29
3.3 Fitness Selection 31
3.4 Genetic Operations 34
3.5 Termination of the Algorithm 34

4 Results 36

4.1 *Drosophila* to *Anopheles* 37
4.2 *LCA* to *Drosophila* 41
4.3 *Clogmia* 43
4.4 *Drosophila* to *Anopheles* with *ftz* 45
4.5 Pair-rule gene polarity 48

5 Discussion 52

5.1 Pair-rule genes and polarity 52
5.2 The computational evolution 53
5.3 Conclusion 54

Appendix A: *Drosophila* Network 57

Appendix B: *Anopheles* network 58

Appendix C: The Fitness Function *(Cont'd)* 59

References 63

LIST OF TABLES

Table page

5–1 *Drosophila* Network Parameters 57

5–2 *Anopheles* Network Parameters 58

LIST OF FIGURES

Figure		page
1–1	Phylogenetic tree of the order of Diptera	3
1–2	Experimental images of *eve* in *Anopheles* and placement of gap genes in *Anopheles* and *Drosophila* with respect to *eve* stripes	6
2–1	Wolpert's French Flag Model	11
2–2	Positional Information of profiles	12
2–3	Cascade of the gene network during embryogenesis along anteroposterior axis	24
2–4	Gene Expression Profiles and Networks	25
2–5	Examples of Hill functions	26
2–6	*eve* modules and the whole *eve* pattern	27
4–1	Predictive evolutionary tree and *eve* stripes across Dipteran species	38
4–2	Simulated evolutionary pathway (label F1 on 4)	39
4–3	Simulated evolutionary pathway (label F2 on 4)	40
4–4	Simulated evolutionary pathway (label LC on 4)	43
4–5	Clogmia Data	44
4–6	Clogmia	45
4–7	Simulated evolutionary pathway (label Ftz on 4)	47
4–8	Evolutionary Tree and *eve* stripes across Dipteran species	50
5–1	Example of genetic algorithm fitness for an *eve* simulation	61
5–2	Example of genetic algorithm fitness for an *eve* and ftz simulation	62

CHAPTER 1
Introduction

1.1 Developmental biology

It is an impressive feat of life itself that such unparalleled complexities, constrained by physical laws of nature, are attained in biological systems. The fact that laws governing the dynamics of the inanimate establish an environment conductive to the initiation and development of living organisms on different length scales is spectacular. Given the prodigious task of studying and understanding such composite organisms and the limitations on data, it is no surprise that traditionally the biological fields have offered qualitative descriptions of systems rather than focus on the intricate connections weaving the physical laws into animate descriptions of all things living. However, the success of the physical sciences has thus far been due to their methodical study of first principles and ability to offer quantitative descriptions to exploit the different laws and concepts explored. In the last century, much work has been done in developing mathematical frameworks to model biological phenomena. Pioneers in the field have applied concepts and models from other disciplines to develop frameworks in which to study biology. For example, Hans Meinhardt noted how the formation of patterns from homogeneous initial conditions was not exclusive to organic matter: we find similar patterning in sand dunes and forms of erosion (to name a few). This influenced him in applying similar models from physics to study pattern formation in biology[1]. Explaining the provenance of phenomena with a

physics mindset is still not a practice that is common across all fields in biology and yet it has yielded numerous successful results and is useful for developing functional models.

The field of developmental biology focuses on studying the formation of structure in biological systems and establishing an understanding of how complex multicellular systems emerge from single cells[2]. This is a broad topic as systems undergo many steps to form fully developed organisms and that, given the diversity of life, there exists a multitude of different pathways for which species develop into their adult structures. Nevertheless, many different organisms display similar tendencies and apparent gene homologies throughout their growth suggest some universality[3][4].

Genetic expressions obtained experimentally give insight into the similarities between certain species however constructing models for genetic regulatory networks and phylogenies remains a difficult problem for most organisms. This leads us to the question: *is it possible to infer ancestral phenotypes and dynamics through data acquired from their descendants, species emerging later in the phylogenetic tree*?

1.2 Evolution and the genetic algorithm

At the core of any universal model for development lies the genes, blueprints of the biological world. The genes are in constant contact with a dynamic environment, whose interaction with said genes causes the formation of certain products. These gene products hold many different functions within the cell[6] depending on what information the genes have been provided by their environment. Hence, a description of how these genes and their products interact as a whole is crucial to understanding the steps of biological processes on a molecular level. From a global point of view,

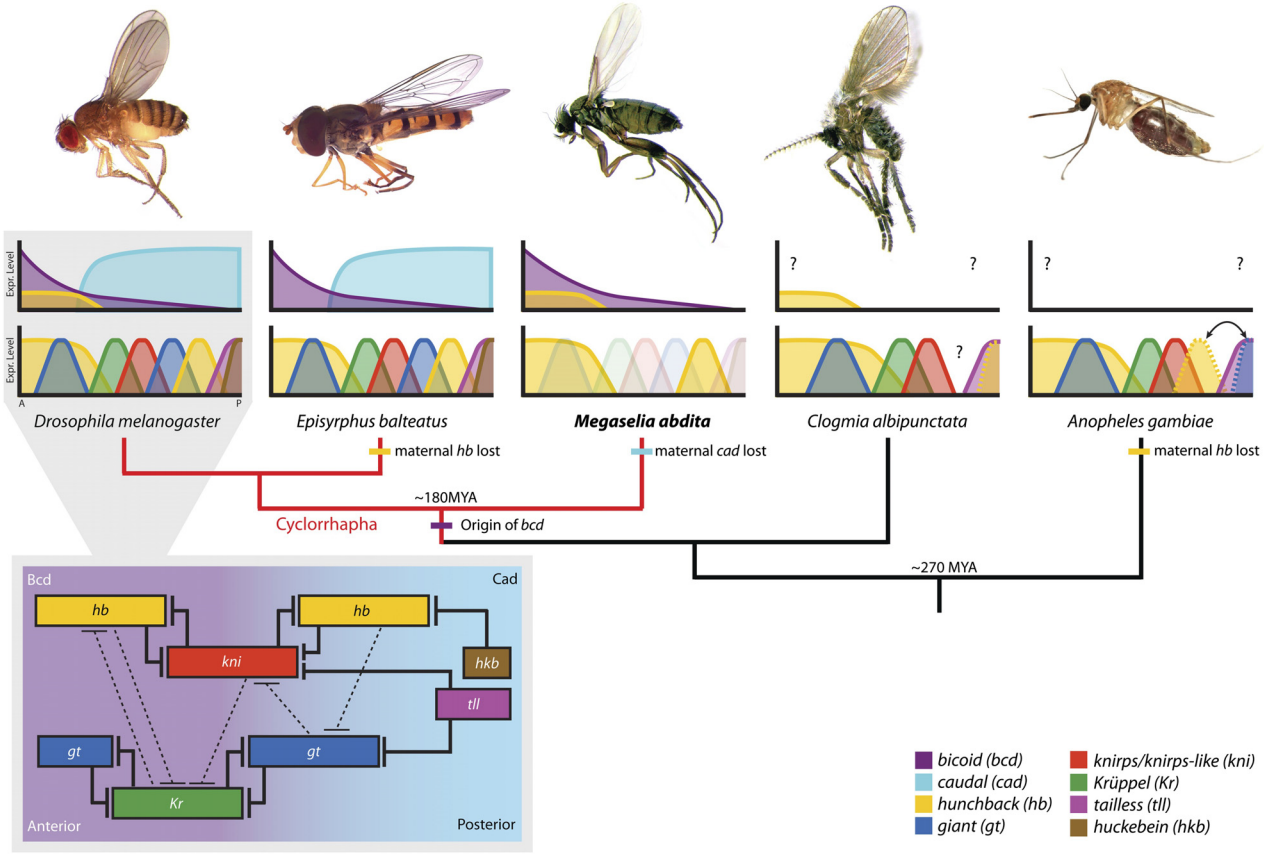

Figure 1–1: Above is a family tree of certain well studied Dipterans, placed according to their proximity in their evolution. This shows the difference that the species exhibit in terms of their gene profile along the anteroposterior axis in the embryo, while maintaining a similar body structure. Below each picture of different species is a graph of the maternal input genes and, below that, the graph of gap genes. Note that . The question marks indicate locations along the axis where the profile of either gap or maternal genes is not very well known. The pale outlines indicate weaker gene expressions. Figure reproduced as in *Quantitative system drift compensates for altered maternal inputs to the gap gene network of the scuttle fly Megaselia abddita* by Wotton [5].

much can be learnt by finding the proper regulation scheme of the gene regulatory

network (GRN). The problem with constructing adequate descriptions of these GRNs is the scarcity of the data necessary to comprehend the interactions of these genes.

With an increase in computing power of late, development of computational models for complex genetic networks has become a more feasible feat even if most systems exhibit a multitude of possible states. The genetic algorithm[7] will serve as our main instrument for exploring the genetic landscape of different species. In a sense, the algorithm mimics the dynamics of Darwin's evolution, evolving a population through many generations using mutations and selection. However searching through this complicated state space to find the proper network is not a trivial task and, given the sparsity of the data, it is crucial to not exaggerate our model by overfitting[8]. Finding a balance between variance and bias is a key concept of machine learning and will justify our choice of network selection.

1.3 Motivation

Traditionally, experiments have been conducted primarily on model organisms such as fruit flies (*Drosophila melanogaster*), e.g. mapping the genome sequence[9], and zebrafish, (*Danio* (*Brachydanio*) *rerio*), e.g. studying the stages of embryonic development in vertebrae[10], to infer the developmental processes. For this reason, the study we conducted focuses on Dipteran embryogenesis and, more specifically, on the case of *Drosophila* and some of its cousin species, namely *Anopheles* and *Clogmia*, which have only recently become the topic of further investigation within the experimental biology community[3][11].

Although data concerning the gene regulatory networks is sparse outside of *Drosophila*, imaging of the gene expression for different species has provided us

with sufficient information to construct phenotypic models of the cascade of genes involved through embryogenesis. As such, we have spatial and temporal descriptions of gene domains that pattern the embryo along different axis. Interesting questions arise from these gene expression profiles: namely concerning the differences between species pregastrulation. Much work has been put into modelling the networks during these early stages of embryogenesis[12]. Although the anterior gap gene motif along the anteroposterior axis remains invariant in all these species, the relative position of two genes, *gt* and *hb*, is inverted in *Anopheles* and absent in *Clogmia* within the posterior domain of the embryo. Consequently, pair-rule genes such as *eve* that are presumably controlled by these gap genes exhibit dissimilarities in their pattern: prior to gastrulation *Clogmia* expresses only six *eve* stripes, *Drosophila* has 7 and *Anopheles* can have up to 8. Given these observations, it is not a completely trivial task to construct networks with the proper regulation schemes to fit the data. For example, it seems contradictory that *eve* 5, the module responsible for the fifth stripe of *eve* in the *Drosophila* embryo[13], is repressed by *gt* and that we find a regular striped *eve* at roughly that same position in both the posteriors of *Clogmia* and *Anopheles*. We would expect that, given that the domain of *gt* is absent in *Clogmia* and more posterior in *Anopheles*, the lack of a repressor for this *eve* 5 module would result in an elongated stripe in both species. However this is not what is observed experimentally as can be seen in Figure 1-2.

Furthermore, little is known about the regulation of pair-rule genes in most other Dipterans[14], whether their positioning is directed by the maternal input and gap genes[15] or dynamically controlled by a set of primary pair-rule genes (notably

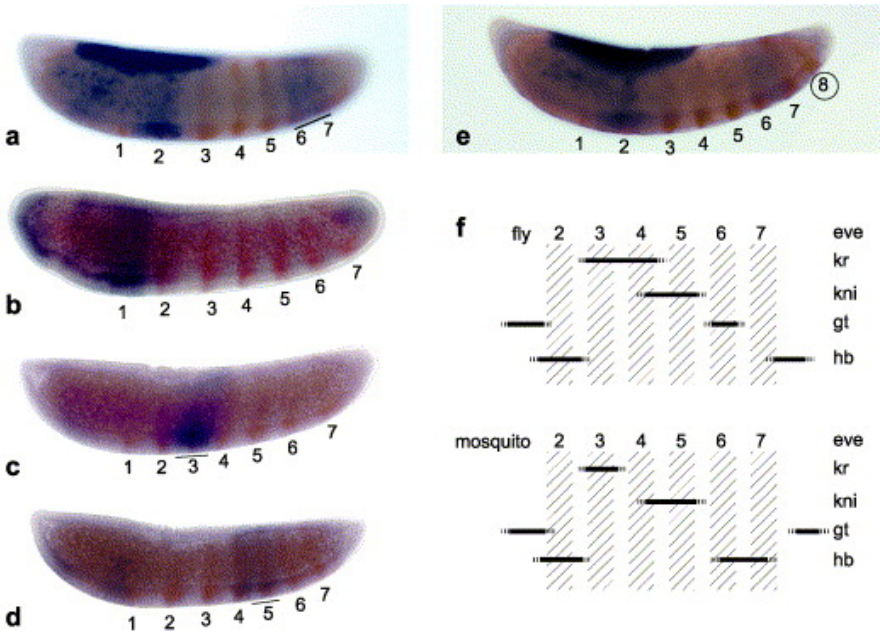

Figure 1–2: The formation of *eve* stripes in mosquito along the anteroposterior axis, from left to right dorsal up. The scheme in (f) indicates the different expression of the *eve* stripe with respect to the location of the gap genes. Note how the gap gene varies considerably between the two, especially anterior *gt*, suggesting a change in regulation scheme. Panels (a)-(d) show the expression of *eve* (stained in red) in mosquito with the respective gap gene stained in blue: (a) *hunchback*, (b) *giant*, (c) *Kruppel* and (d) *knirps*. We can see in (e) an eighth stripe forming in the anterior. Figure reproduced as in *Different combinations of gap repressors for common stripes in Anopheles and Drosophila embryos* by Goltsev [3].

eve) as in *Drosophila*[16]. Indeed, the complexity in the experiments required to probe these gene regulatory networks in most organisms make them difficult to study, however analogies can be drawn from known cases such as the *Drosophila* network and searching for ancestral genotypes can give insight on how this regulation changed through their split evolution.

Given the experimentally available data described above, we focus on a particular instance of the question posed previously. In the specific case of embryonic development of Dipterans, we search to answer the following: *can some evolutionary simulations generate phenotypic pathways from different species such as Drosophila and Anopheles to exhibit the network of their last common ancestor and what can we infer about the dynamics of the segmentation genes in this ancestral Dipteran*?

1.4 Thesis overview

Following this brief introduction in Chapter 1 of the field of biophysics and motivation for the study, the thesis will subsequently present in Chapter 2 an overview of the concepts necessary to tackle the questions regarding the segmentation genes, presenting both the biological and mathematical framework with which we will be working. The first concept described is Wolpert's French Flag model in Section 2.1, establishing a basis for obtaining positional information. Section 2.2 offers a biological description of Dipteran embryogenesis, focussing on the details pertinent to the study at hand. This is followed by a discussion on Gene Regulatory Networks in Section 2.3, both a conceptual and mathematical description as well as the particular modelling chosen for our GRNs. Section 2.4 describes what is known about the regulation scheme of the *Drosophila* network, giving the minimalistic network with which we will be working.

Chapter 3 delves into the the genetic algorithm, starting with an overall view of the structure of the algorithm. Sections 3.2 onwards describe each step in the algorithm in detail and the choices that were made in each step regarding our own study.

The results of the conducted research are presented in Chapter 4 with each section focussing on one particular simulation. Section 4.1 deals with the simulated evolution from *Drosophila* to *Anopheles* and Section 4.2 with the results of the reverse simulations, that is from the last common ancestor to *Drosophila*. Section 4.4 presents the results that revisit the simulation from *Drosophila* to *Anopheles* with the addition of a second pair-rule gene, ftz in the system to investigate incorporating other segmentation genes to the network. Subsequently, the study of the polarity and regulation of two more second pair-rule genes by the network is the subject of Section 4.5.

Finally, Chapter 5 offers a discussion on the results presented in Chapter 4. First dealing with the results from the evolution and the study of the polarity in the embryo, we later discuss the genetic algorithm briefly in Section 5.2. To conclude this thesis, Section 5.3 expands on potential future directions that may be explored.

CHAPTER 2
Theoretical Framework

Historically, due to the lack of quantitative information, biological fields have offered qualitative descriptions of systems studied, however in recent years the influx of data and collaborative work with researchers from quantitative sciences have expanded our understanding of these biological systems. For the questions addressed in this thesis, a mathematical model of the gene network is necessary to probe the underlying dynamics of the phenotypic evolution.

2.1 The French Flag Model

It is essential for cells within a developing organism to obtain some "positional information" to allow for the functionality of the cell to be determined in terms of its location in the system. By "positional information", we mean the ability of cells within an organism to detect where they are situated in the system. From an exterior point of view it is clear to an external observer what is the front from back of an embryo or to differentiate specific positions in the egg, but for cells within the embryo there is no obvious way a priori to get a global view of their environment and pinpoint their location. Thus there must be a mechanism for cells to differentiate between locations of the body simply by probing their immediate surroundings as this is their only source of information.

Wolpert's French Flag Model aims to explain this cell localization through the use of morphogen gradients[17]. As the cell detects different molecules in its environment it gets a sense of the concentration of these different components, termed morphogens, in its surrounding. The fate of the cell will depend on the concentration sensed by the cell of particular morphogens. For example, imagine a morphogen gradient along a certain axis in an embryo that is highly concentrated on one end and gradually diffuses to low concentrations on the other end, such as in Figure 2-1. The cells have further been programmed to detect this concentration gradient and know that at high concentrations (above a certain threshold) they are to be "blue" cells, whereas at low concentrations (below a certain threshold) they are to be "red" cells. In between these two thresholds, the cells are fated to be "white".

Thus from the morphogen gradient already in place, we see that the cells gain some form of positional information and their cell fates establish a French Flag pattern along the specified axis of the gradient. This is the principle of the model put forward by Wolpert and it has since been used to successfully model many phenomena, from limb regeneration[19] as well as pattern formation in *Drosophila*[20].

The positional information can be increased by the number of thresholds determining cell fate[21]. This is generally speaking more difficult to do for one particular morphogen as the system might have a limited accuracy in differentiating certain concentrations. However, given multiple morphogens, the positional information can potentially increase by the cells ability to read the concentrations of each separate morphogen and the subsequent thresholds those morphogens might have determined for a particular positional marker. In fact, the question of whether or not multiple

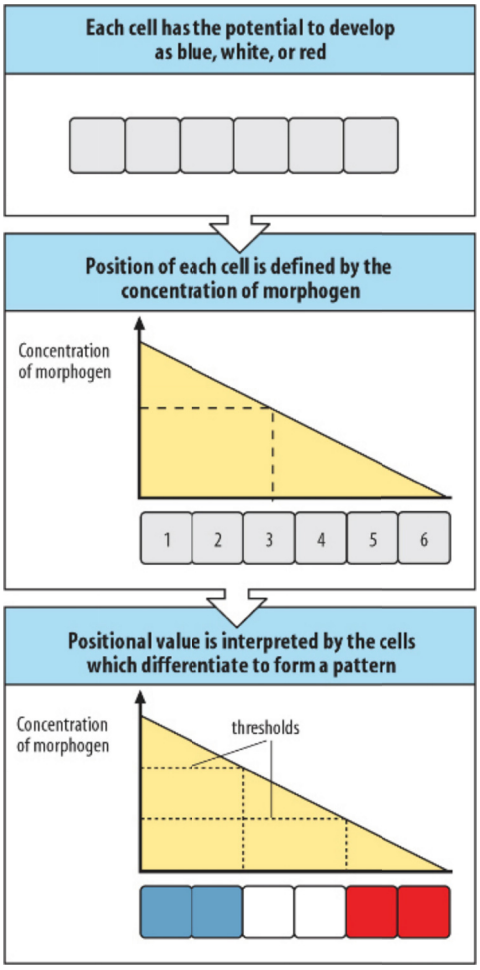

Figure 2–1: The French Flag model seeks to explain the acquisition of positional information through the use of morphogens. The depiction in the figure shows how a gradient of a morphogen can determine some boundaries within the system through the use of thresholds in its concentration and establish a pattern along the axis of this morphogen. Figure reproduced as in *Principles of Development* by Wolpert [18].

genes can increase the amount of positional information is a delicate one. It will depend on the ability of the system to distinguish between concentrations as well as the concentration profile of the input[22].

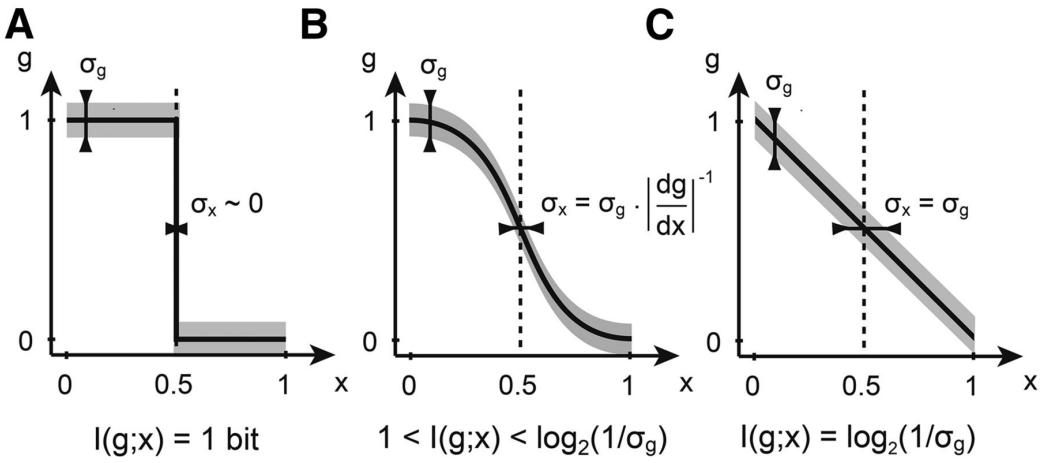

Figure 2–2: The information that a profile concentration carries (I(g;x)) can vary greatly from profile to profile, where g is the concentration of the morphogen. Panel (A) depicts a on-off and 1 bit of information as there are simple 2 regions. The sigmoidal function in (B) increases the information as there are a multitude of different positions along the axis given certain concentrations. Clearly the positional information is greater than 1 bit since there exists a cutoff at 0.5 which distinguishes the two regions. Similarly for the linear gradient in (C). The Figure reproduced as in *Positional Information*, *Positional Error*, *and readout Precision in Morphogenesis* : *A Mathematical Framework* by Tkacik[22].

In Figure 2-2, the mean gene expression profile (plotted as a solid black line) determines a certain position in the cell and the grey zone represents the variability (σ_g) of the profile, in other words how well the system can distinguish a certain concentration. As the variance increases and the profile of the gene concentration changes, the information the concentration carries varies greatly. Questions regarding the transfer of information are of great interest in developmental biology. It is not the focus of our study here and for the purpose of this thesis it is enough to assume that information about the product concentrations is transmitted to subsequent genes in the cascade of gene expressions[23].

2.2 Dipteran embryogenesis

The developmental process for many Dipterans, including *Drosophila*, starts off well before the fertilization of the egg. The unfertilized egg is subject to an initial polarization of maternal input along its different axis[24] depending on its position within the mother's ovary. In *Drosophila*, maternal-effect genes, such as *bicoid* (*bcd*), *nanos* (*nos*)[25]and *torso* (*tor*)[26], produce morphogen gradients of their subsequent RNA and protein products throughout the egg along the anteroposterior and dorsoventral axis. Starting at the moment of fertilization, the nucleus undergoes multiple mitotic divisions[18] and the number of nuclei in the embryo grow exponentially. Unlike other zygotic embryos, the individual cells will not form for each nuclei until later in development, thus allowing for protein products encoded by the genes of these nuclei to diffuse freely within the embryo. After 8 nuclear divisions, 256 nuclei are formed within the egg and migrate to the cellular membrane of the embryo[27], where mitosis continues to occur. These nuclei spread in a relatively uniform distribution along the periphery of the cell and begin transcription.

Transcription is the process through which genes synthesize RNA. The location along the DNA that initiates this transcription for the gene is referred to as the promoter for the gene. Transcription factors are proteins possessing specific activator and/or repressor sequences of nucleotides that will only bind to their associated promoter sites, initiating/respressing the transcription of that gene. Once attached to the DNA, the RNA polymerase copies the sequence to form a messenger RNA (mRNA) which is then released into the cytoplasm to be synthesized into a protein. This whole mechanism, the "central dogma" of molecular biology, is the basis for all

other operations that occur within a cell, as it produces the necessary molecules that drive the machinery of the organism.

Thus, this production of a protein by any particular nucleus is determined by the concentration of regulating transcription factors at the position of that specific nucleus. It is this protein production that is modelled in this study, neglecting the mechanisms of translation and other dependencies in the cell. Along the antero-posterior axis, the original maternal-effect gene gradients serve as input for, mostly, activation[28] and, in some cases, repression[29] of a first set of genes in the genetic transcription cascade, called gap genes. These gap genes encode for the production of other protein products that will themselves serve to regulate expression of a set of segmentation genes[30], as is depicted in Figure 2-3.

This is the structure of our gene regulatory network. The particular gene expression of a nucleus can be viewed as the concentration of the products of transcription by this gene at a given position. Along the anteroposterior axis a profile of the concentration of gene expressions appears at this stage of embryogenesis. After the thirteenth nuclear division, cellular membranes begin to form around each nucleus at the edge of the embryo, enclosing the nuclei with their surrounding cytoplasm. Cells can be viewed as machines that work with a function depending on their nature and the components enclosed within the membrane of the cell at this stage determine the cells functionality in the organism, defining their development and fate thereafter.

The role of the segmentation genes, according to their concentration levels at a position, is to establish segments within the embryo that will define the different structural sections of the adult body[31]. In *Drosophila*, the segmentation gene

profile splits the embryo into 14 segments generated by sets of pair-rule genes with seven evenly spaced stripes having similar width. Certain of these segmentation pair-rule genes appear before others, regulating the secondary pair-rule genes in the same manner the gap genes regulated their own expression. In this hierarchy of segmentation genes, *even-skipped* (*eve*) is one of the first to appear[32] and is the main focus of this thesis as its regulation is well understood in *Drosophila*[33].

Although all the pair-rule stripes appearing in the anteroposterior view are the expression of the same gene, they are in fact regulated by different transciption factors. That is to say that along a DNA strand there is one loci that encode for *eve*, however this loci is controlled by multiple modules, i.e. promoters sequences[30]. Thus the gene regulatory network that controls the gene expression of one particular stripe might not be the same as another stripe. For example, in *Drosophila* there are five *eve* modules: *eve* 1, *eve* 2, *eve* 3+7, *eve* 4+6 and *eve* 5.

As seen previously, different members of the Dipteran family have very different gene expression profiles in their embryos and this is the motivation for this thesis. For the most part they maintain similarities in terms of the nature of the genes that can be found throughout the egg during embryogenesis, however it is the gene regulatory network that demonstrates dissimilarities in terms of the strength of the interactions between gene expressions[35]. This leads to varying gene expression patterns across the anteroposterior axis. Notably, the position of gap genes is considerably different between the species *Drosophila* and *Anopheles* (mosquito) while the target pair-rule gene *eve* is invariant. This lead us to hypothesise that there are, between the two species, a displacement and slight redistribution of the *eve* modules.

By redistribution, we mean that the modules expressing certain *eve* stripes in the *Drosophila* profile seem to now be homologous to other stripes in the *Anopheles* profile. We propose that *Drosophila*s' *eve* 5 does not exist in *Anopheles*[3] and predict an additional stripe forming in the back of *Anopheles* to maintain a 7 stripe profile.

2.3 Gene regulatory networks

With the already high complexity of the environment only increasing as cells undergo development it is hard to study and characterize individual reactions, however it is apparent that some underlying structure maintains some order that can be exploited in the system. Most activity in the cell occurs through the complicated regulation scheme of the reactants and different components that interact with each other to form a network and define the systems state. The products and connections that relate the interactions between these different substances is known as a gene regulatory network[36]. It is a graphical model where the network is completely defined by these transitional probabilities and state probabilities of the state variables.

By studying the gene expressions and correlations between the different products in the network it is possible to model the appropriate regulation scheme defining the network[37], abstracting the complicated chemical dynamics into a simpler illustration of cause and effect. The gene regulatory network is often described as a collection of nodes (representing the genes) with edges (representing reactions) that connect the nodes. The input and output of each node control the state of the nodes much like a neural network[38]. The edges, essentially interactions between genes,

can be separated into two categories: inhibitory and inductive. Inductive interactions allow for a sigmoidal increase of the output whereas inhibitive interactions lead to inversely sigmoidal changes in output given an input. The nodes can be regarded as functions that take the input through some combination of basic functions to produce the output and characterize the dependence of the environment on each genes state. The gene expression is defined by the output of each corresponding node. A variety of edge schemes connecting the nodes allow for the modelling of different phenomena in this graphical view. For example, an edge connecting a node to itself is an instance of a feedback loop: a process where the output of a gene would, through some mechanism, come back as input to the gene as in [39]. By adding more edges we change the complexity of the systems structure and vary the connections between nodes. Defining these nodes and how they process their input determines the output of each gene and it is the study of these models that uncovers the complex's dynamics which can be tested experimentally to study the correlations between genes.

A common mathematical approach to modelling these gene regulatory networks is through the use of ordinary differential equations[40]. The dynamics of each node is expressed as an ordinary differential equation that governs the concentration of each gene expression. Thus a set of differential equations is necessary to explain the network and steady state solutions represent the systems stable states. Having a functional form for each state as a function of time allows for the study of the dynamics of these states out of equilibrium and their response to perturbations in the input. The production term is defined by the node and is a function of all the

input to that node. Additionally, to be biologically viable, a degradation term is present to account for the natural turnover of the proteins.

The production terms in our differential equations model the gene expressions response to the input. In a Boolean network where the state of each gene is either ON or OFF, these can be defined as products of Heaviside step functions where the input is either there or it isn't. It is possible to model a more continuous state space where the state of the gene is defined by the concentration of the gene expression. A more smooth (that is to say less discontinuous than a Heaviside step function) way to evaluate the information provided by the input is to use sigmoidal or Hill functions to regulate the production of the gene expression.

Hill functions describe the response of a certain concentration of input to an established threshold of activation. In biochemistry, they are mostly used to explain the binding of certain molecules to receptors[41]. We consider the reaction

$$A_{free} + nX \underset{k_i}{\overset{k_+}{\rightleftharpoons}} A_{bound} \tag{2.1}$$

where n is the number of molecules of type X that it takes to bind to the binding site A to produce the bound state $A*$ with rate k_+ and the opposite reaction at a rate k_-. The expression that relates the appropriate concentration for each of the reactants and products is

$$[A_{free}][X]^n = k_+[A_{bound}]. \tag{2.2}$$

Combining this with the condition that the concentration of receptors A(either bound or unbound) do not change

$$[A_{free}] + [A_{bound}] = [A*] \tag{2.3}$$

we can derive the fraction h of bound and free particles to the total amount of receptors:

$$h(bound) = \frac{[A_{bound}]}{[A*]} = \frac{X^n}{C^n + X^n} \qquad h(free) = \frac{[X]}{[A*]} = \frac{X^n}{C^n + X^n} \tag{2.4}$$

where $C^n = k_+$. These are Hill equations, monotonically increasing and decreasing functions of the concentration of the reactant $[X]$ which give the fraction of free or bound states. These can be written in a convenient way to summarize the dynamics of repression or activation by a particular gene X through the following equation

$$hill(X, C, n) = \frac{1}{1 + (X/C)^n}. \tag{2.5}$$

where we have generalized n to range the positive and negative real numbers. In the context of our regulatory gene network, C, the concentration at half maximum, is a threshold over which our gene X has an effect on the studied gene in it's production term. The strength of this effect is governed by the magnitude of the Hill coefficient n and the effect itself is established by the sign of n. A negative Hill coefficient denotes an activation whereas a positive Hill coefficient designates a repression by the gene.

In our *Drosophila* network, different activators and repressors manage the concentration of the gene expression production. Although repressors work in unison

to inhibit the activity of a particular gene, the activators compete to activate their particular DNA sequence, thus only the highest rate of production for activators is processed. This is as modelled in previous work by François et al. [37].

$$\frac{dX_i}{dt} = \max_{n_{ij}<0}\left\{\frac{1}{1+(X_j/C_{ij})^{n_{ij}}}\right\}\prod_{n_{ik}\geq 0}\frac{1}{1+(X_k/C_{ik})^{n_{ik}}} - \delta_i X_i \tag{2.6}$$

In some cases the dispersal of the genes across the environment is non-negligible, thus a diffusion term can be added to this differential equation to more accurately define the dynamics of the products, with diffusion coefficient D. We obtain a PDE which can be solved using numerical methods in a similar fashion to the previously presented ODE. We will see that for most of the genes in our gene network it will not be necessary to model diffusion.

$$\frac{dX_i}{dt} = \max_{n_{ij}<0}\left\{\frac{1}{1+(X_j/C_{ij})^{n_{ij}}}\right\}\prod_{n_{ik}\geq 0}\frac{1}{1+(X_k/C_{ik})^{n_{ik}}} - \delta_i X_i + D\frac{\partial^2 X_i}{\partial x^2} \tag{2.7}$$

Using this formalism, we can describe the effect of morphogens on the genes as is suggested in the French Flag model, with the concentration at half maximum defining the threshold and the Hill coefficient determining how steep this pattern is at a certain position. Thus we can acquire positional information from these different gradients.

2.4 The *Drosophila* gene network

The gene network of Drosophila during embryogenesis, although extensive, is well understood in a qualitative manner: The regulation of the essential genes for development are well mapped and studied. Our idealized *Drosophila* network is one that contains the strictly necessary genes to establish a minimal profile of the studied

pair-rule genes *eve* and, eventually, *fushi tarazu ftz*. Although other genes, such as shadow enhancers[42], exist in the network and adjust our gene expressions, we argue that they are not necessary for the qualitative pattern to arise, thus are not included in our network.

The maternal gradients are supplied by *bcd* in the anterior and *caudal* (*cad*) in the posterior. In the evolutionary simulations, we assume that some other maternal gradient such as *orthodenticle* (*otd*) replaces *bcd*[43]. *cad* is itself regulated through a slight repression by *bcd*, keeping it maintained in the posterior. Additionally, two posterior gradients *huckebein* (*hkb*) and *tailless* (*tll*) allow for additional repression and activation of subsequent posterior genes such as *hb*[44][45].

These initial inputs regulate the downstream gap genes giant (*gt*), hunchback (*hb*), Krüppel (*Kr*) and knirps (*kni*) which are the sufficient components to define our segmentation gene pattern. All of these gap genes are activated by different thresholds of *bcd*. Additionally, *gt* is activated by *cad* in the posterior[46] while being repressed by *Kr* situated in the middle of the embryo and an anterior repression suppied by *tll*. Other than the *bcd* activation, the rest of the regulation scheme of *hb* is not entirely understood and different models attempt to explain its gene expression. Thus, we simply model its posterior dynamic through an activation by *tll* and a repression by *hkb*, which is sufficient to obtain the desired profile. *kni* is repressed by *Kr*, *tll* and *hb*[46][47] whereas *Kr* is repressed by *gt*[48][49] and *hb*[50]. In many Dipterans the anterior segmentation gene expression in the embryo is relatively well conserved whereas the posterior is subject to more differences. This will be a key concept in our later discussion and is a consequence of the fact that the anterior and

posterior domain of genes such as *gt* and *hb* are regulated by different modules of repressors and activators.

As stated previously, the gap genes work in unison to control the production of the segmentation genes *eve* and ftz. Each module has its own regulation scheme that is summarized in Figure 2-4 along with the regulation of the gap genes. There are five different modules to model: *eve* 1, *eve* 2, *eve* 3+7, *eve* 4+6 and *eve* 5[47][51][52]. Note that as it is primarily the gap gene expressions in the posterior that vary between the species: the *eve* 1 module is invariant and thus not present in most of our simulations. For the other modules, the posterior repressions of these modules are assumed to come from some posterior gradient, thus we've chosen to use *tll*[13]. When no activator is indicated, it is assumed some constant activator is present in the system, such as DSTAT[53] or Zelda[54]. Figure 2-6 shows how each of the *eve* modules are combined to form the complete *eve* profile, which is the same way the ftz modules do the same.

Note that the ftz 4 stripe represents a special case where the module does not seem to have a stripe pattern element through its gap gene regulators, experiments suggest that it is controlled instead as a secondary pair-rule gene[14] (downstream of the *zebra* pattern in the genetic cascade) while still maintaining some repression by the gap genes *tll* and *hb*. We've added a slight repression by *eve* to illustrate the second pair-rule gene nature of ftz 4.

Thus we have our initial network that is defined as the minimal gene regulatory network of *Drosophila* for our target genes *eve* and ftz. This is the network whose parameters we will evolve through generations of mutations to hopefully obtain a

network consistent with the gene profile of *Anopheles*. Although some of the posterior genes are regulated by different genes, for the sake of this study it is only necessary that we get the posterior regulation in both species. It could be that some new regulation scheme involving different genes developed in a last common ancestor that explains this discrepancy between the species[43].

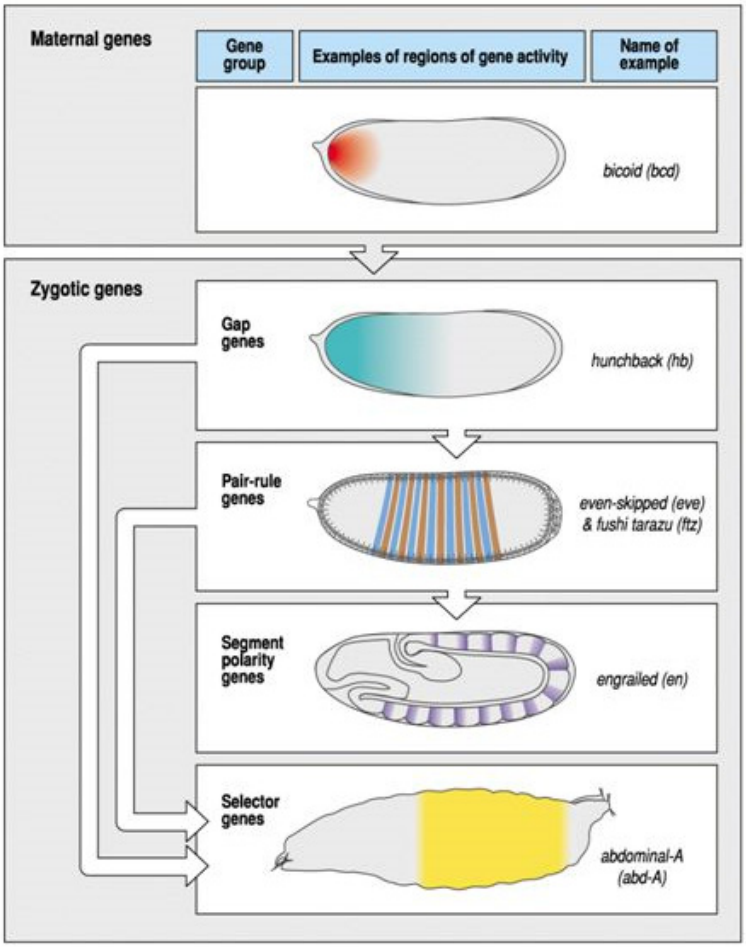

Figure 2–3: All the information for the segments in the body plan of *Drosophila* postgastrulation is initially specified by the mother and cascades from the maternal input. The top panel has an example of one particular maternal gene, *bicoid*, which the mother provides to the embryo. This input passes on the positional information to gap genes such as *hunchback* (subsequent panel) who themselves regulate the segmentation gene pattern (third panel from the top). This hierarchy of genes along the anteroposterior axis establishes the parasegments and morphology of the egg, depicted in the last panels. The arrows connecting the panels indicate the gene cascade, that is to say the progression in the regulation scheme. Figure reproduced as in *Principles of Development* by Wolpert [18].

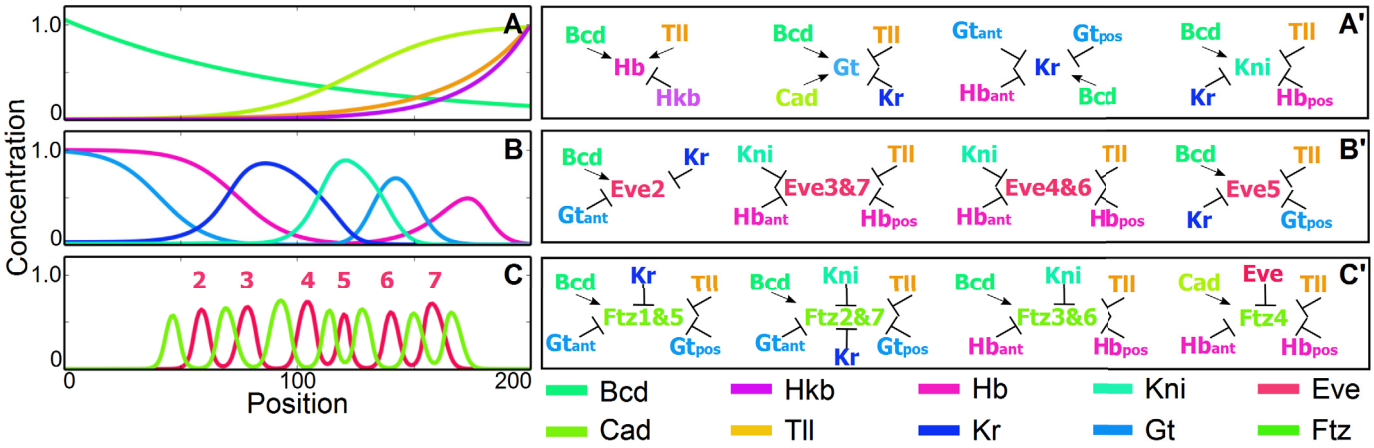

Figure 2–4: (A)-(B) Simulated gene expression profile of the maternal input and the gap genes across the embryo. (C) The profiles of the segmentation genes *eve* and *ftz*. Position 0 indicates the head of the embryo and position 200 the tail. (A')-(C') The corresponding network structure used in our simulations and described more in detail in the text. The strength of the interactions are tabulated in Appendix A. Whenever no activator is present, a generic spatially uniform activator is assumed. Figure from our published article *Predicting Ancestral Segmentation Phenotypes from Drosophila to Anopheles Using In Silico Evolution*[34]

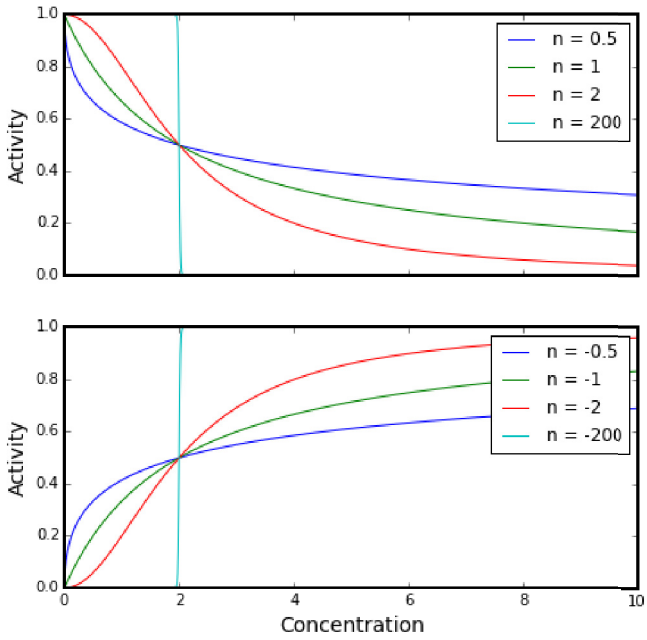

Figure 2–5: Examples of activator and repressor hill functions accoring to different values of the hill coefficient, n in the hill equation $f(x) = \frac{1}{1+(x/C)^n}$. $C = 0.5$ in these examples as can be noted that it is at 0.5 along the abscissa that the function attains it's half-maximum value. Note that as the $|n|$ increases, the function becomes more steep, in fact $\lim_{n\to\infty} f(x)$ is a Heaviside function.

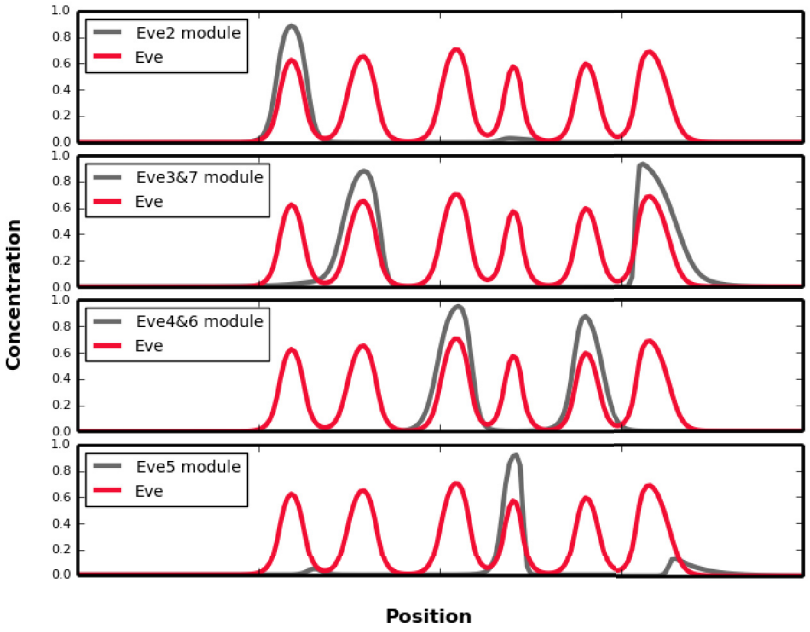

Figure 2–6: Individual modules for *eve* (in grey) activate the final common *eve* pattern (in red). The same mechanism is true for ftz, this allows for a more uniform pattern with regular sized stripes.

CHAPTER 3
The genetic algorithm

With an established model for the gene regulatory network, we now turn to the problem of evolution of this network. The genetic algorithm is our main tool in answering the questions this thesis addresses as it establishes a framework for tackling evolutionary inquiries and, in a way, simulates the evolution that we are trying to study. They are a subset of evolutionary algorithms, used in a variety of problems to optimize parameters[55]. This particular genetic algorithm is the same used by François et al. in previous works, which can be found in [56][57]

3.1 Structure

The genetic algorithm is a heuristic search, through some complicated state space, that simulates natural selection. In other words it is a computational method for evolving a population of genotypes, selecting at each generation the individual that optimizes for some characteristic that is defined by the problem at hand and then mutating these solutions to find additional candidates in further generations[58].

The first step in this algorithm is to initialize a population composed of individuals that have randomly distributed solutions of their genotype in some representation. This step is shortly followed by a selection process in which each individual solution is evaluated and given a score based on a constructed fitness function. This fitness is designed to optimize the target genotype appropriate to the study conducted. This is analogous to searching for a solution that minimizes the energy, defined by the

problem, studied through some complicated state space from physics. Subsequently, the selected individuals are used to create a second generation population through a combination of genetic operations such as mutation, selection and other operators. In a sense, this is the simulated "breeding" step in which the most fit individuals pass on the properties that deem them as fit in our evolutionary framework. The process of creating new generations continues until some termination condition is reached by the algorithm. This is in accordance with many genetical theories of nature selection[59]. Each step is described in more detail below.

3.2 Initial Population Configuration

As noted previously, the first step in the algorithm is designing an initial population. This involves finding a proper representation of the solution domain, keeping in mind that this representation will have to adequately define the genotypes for the problem at hand.

In our simulations, the initial population is completely comprised of individuals all possessing the *Drosophila* genotype that we are studying. Since we are exploring the evolutionary dynamics from *Drosophila* to *Anopheles* it is a logical choice that the starting point would not to be a random set of individuals but rather all individuals from the same species. This will allow for a more extensive search of the parameter space in later iterations of the algorithm. The genotype is defined by the set of parameters that relay the information concerning the network, i.e. the Hill coefficient and concentration at half maximum for each interaction. Thus, the genetic representation of our solution domain is an array of Hill parameters, θ, and

it is the numerical value of these parameters that will be explored and changed from generation to generation, defining new solutions.

As described in section 2.4, the interactions that define the network qualitatively are known for *Drosophila* and fixed in our simulations. However, exact quantitative results in terms of the strength of the interactions between the different genes is not well documented and in fact quite difficult to describe. For our model, what is important is that we can explain the difference of the segmentation gene patterns. A set of "kernel" functions are used to model the interactions between genes in the network: these are functions that define a certain relation between two genes. In our framework, these are the Hill functions that relate a repressor or activator to its target gene. Thus, the parameters of these kernels were manually tuned to provide a gene profile in the anteroposterior axis that readily resembles the profile in *Drosophila*. This is sufficient in the framework of this project. The parameters that comprise our initial gene regulatory network are tabulated in Appendix A.

As stated previously, the gene regulatory network that describes the expression of the gap and pair-rule genes is only slightly different in the posterior for *Anopheles*, however for the sake of simplicity our model has them defined as the same. Which is to say that the algorithm doesn't create new kernel functions or add new input to these kernels. Thus we would expect that to go from one species to the next we need only evolve the parameters to match the parameters of *Anopheles*, while maintaining what we will later define, through the fitness function, as live species throughout the evolution. Additionally, we ran the genetic algorithm to simulate the evolution from the LCA, Last Commom Ancestor, to *Drosophila* and in this case the starting point

for the initial population was a last common ancestor phenotype, exhibiting a very posterior gt. The parameters that establish the initial configuration were, as in the case for *Drosophila*, manually adjusted to qualitatively explain the gene expression profile in LCA and are tabulated in Appendix B.

In our simulations, the size of the population was maintained throughout the evolution at 50 individuals, relatively small for genetic algorithms, to cut on computation time. In fact after trying for a variety of sizes it was deemed the dimension of the population had very little impact in our situation on the range explored in the phenotype state space and the ensuing results, although different sizes of populations can be tried[60]. Running the algorithm on multiple machines and many times allowed for a broad exploration in a similar manner, as it is possible to pick up from the last solution found by a previous simulation and continue the algorithm from that point forward.

3.3 Fitness Selection

As stated previously, what differentiates the genetic algorithm from a random walk through the state space of the solution domain is the selection of the most adequate solutions by a fitness function. The fitness function defines a topology, a $|\theta|$-dimensional energy landscape for the representation of the phenotypes. The idea is to construct this topology such that it establishes a minimum at the desired optimal solution[61][62]. As such, the fitness corresponds to a measure of each solution on this topological space. Through iterative steps the genetic algorithm scores each individual in each generation and tries to find this minimal solution.

Our fitness reads the values of the Hill parameters for each individual, solves the differential equations defined in Section 2.3 and outputs the gene expression concentrations across the embryo. These concentration profiles are used to differentiate between the networks in a population and select the most fit individual. As stated previously, the target profile of *Anopheles* exhibits an expression of *gt* that is so late/posterior it seems irrelevant to the formation of the segmentation genes. Since it is basically non-existent in the target network, the integral of the gene expression of *gt* in the posterior is a good indicator in the fitness as minimizing the integral will provide the desired *gt* profile, i.e. eliminating it.

Furthermore, since the anterior of the *eve* profile is conserved, the difference between the anterior initial *eve* profile and anterior *eve* profile in the individual scored is also calculated for use in the fitness function. However to prevent this condition from becoming too restrictive in the evolution of parameters, only the maximum of the *eve* difference and the difference of *hb* in the individual scored to a predetermined *hb* target profile contributes to the score in the fitness function. This allows the profile of *eve* in the anterior to vary slightly only if it means that the *hb* profile moves towards the profile it would have in *Anopheles*.

It is important to note that the high dimensional energy landscape is rather complicated and allows for many different paths. However, given that we search for a viable pathway between the two species *Drosophila* and *Anopheles*, we need to assure that every generation in between contains a species that is biologically realistic. That is to say we want each Dipteran generated to be a live species. Given we cannot experimentally study the said phenotypes for ancestral Dipterans (the

very reason for this study), it falls to us to select what we mean by a live species in this context. This is where the number of stripes becomes relevant. Experiments have shown that although certain species in the Dipteran family, such as *Anopheles*, can have more than seven stripes of segmentation genes, having less of the stripes prior to gastrulation is in fact a problem. Indeed, *Drosophila* mutants and other Dipteran mutants exhibiting 6 or less stripes did not survive to adulthood. Thus our fitness assigns an incredibly large score to phenotypes that exhibit less than 7 *eve* stripes (similarly for ftz when it is present in the simulations).

Additionally, the simulations containing ftz included another condition to insure that a proper pattern is displayed, an alternation between the two segmentation genes. Similar to the condition on the number of stripes, if the pattern produced by the differential equations does not exhibit an alteration between *eve* and ftz, the network is assigned a high score so that it will be discarded by the genetic algorithm as a poor solution.

Finally, for the simulations concerning the inverse evolution from the least common ancestor to *Drosophila* the target solution is now the *Drosophila* segmentation pattern and so the choice of fitness function must reflect this new aim. Using our previously calculated concentration of gap genes for the *Drosophila* network, we define the fitness as the sum of the difference between the individuals gap gene profile and the *Drosophila* gap gene profile, as well as the condition for a minimum of 7 segmentation stripes. Further information, examples and detailed expressions for the fitness functions can be found in Appendix C.

3.4 Genetic Operations

Once each individual is scored and ranked within the population, the algorithm selects a subset of the current population to be carried over and fashioned, through some set of genetic operators[63], into the next generations population. This is the step that creates genetic diversity and is the exploration of the solution domain.

There are various ways to pass on the genetic code of a population onto the next generation. The first step is the selection of the individuals who carry over their genetic makeup (parameters θ_i). The fitness function indicates which individuals are more fit in context and so selecting the individuals according to their score allows for a subsequent population closer to the desired result. For our simulations the top half of the population, ranked according to score, were used to generate the next generation. Although crossovers, the process of recombining the genetic material (parameters) from different parents, are often common in genetic algorithms we have chosen to omit this recombination and stick to only introducing mutations into the new population. Only mutations, changing the parameters at a certain rate (which is transcribed in Appendix C), are employed by the algorithm to produce new individuals. Thus our new population is composed of the individuals carried over after selection and, for each of these individuals, a mutant that carries different parameters from its parent according to the rate of random mutations on the genes.

3.5 Termination of the Algorithm

The process of searching through the solution domain and selecting the best individuals to breed new generations continues until some termination condition is

reached by the algorithm. To increase efficiency, the convergence of genetic algorithms towards solutions can be studied to optimize the fitness function[64] There are many candidates for potential termination points, each suitable for the type of problem at hand. Ideally, the end step of the algorithm produces the target genotype.

In our situation, the fitness defined above serves as an optimization tool for which zero is a quantitative lower bound. However, this lower bound is difficult to achieve, considering the restrictions on the profiles of genes which define the fitness. Thus terminating the code only when this lower bound is achieved is an unreasonable demand. Finding a condition for convergence is equally complicated as the mutations are random and so simulations plateau for many generations at a certain fitness score before finding another combination of parameters that reduces the fitness score. The stochasticity and variability of the process means that finding a regular pattern in the evolution is difficult and defining some trend or tendency that the fitness exhibits is non-trivial. For these reasons, using a preprogrammed maximum number of generations for termination is a sensible choice. For more on the convergence of our genetic algorithm, Appendix C contains some examples of the fitness across some simulations which shows how different simulations can take considerably different routes to reduce their score.

The termination of the algorithm concludes the simulation and it is then possible to study the pathway of the most fit individuals throughout the evolution. It is mostly through the defining different fitness functions and genetic operations different pathways can be explored, however for the purpose of this study the ones defined previously are sufficient.

CHAPTER 4
Results

The results and figures described below were published in the paper *Predicting Ancestral Segmentation Phenotypes from Drosophila to Anopheles Using In Silico Evolution* Rothschild JB, Tsimiklis P, Siggia ED, Franois P (2016) Predicting Ancestral Segmentation Phenotypes from Drosophila to Anopheles Using In Silico Evolution. PLOS Genetics 12(5): e1006052. doi: 10.1371/journal.pgen.1006052[34]. As described previously, the difference in the gap gene expression pattern between *Drosophila* and *Anopheles* lies in the relative positioning in the posterior domains of *hb* and *gt*. In *Anopheles* there is a forward shift in the peak of posterior *hb* relative to it's positioning in *Drosophila* and as for posterior *gt*, the peak becomes expressed much further in the posterior making it an unlikely candidate for regulating any set of stripe boundaries. As for the rest of the gap gene network, very little changes and the rest of their profiles between the two species are generally conserved.

The difference in gap gene profiles across different Dipterans causes a distinction in the pair-rule gene description for each species. The formation of the *eve* 5 stripe (which has a posterior boundary defined by the *gt* posterior domain in *Drosophila*) is a problem in the evolution. As this region of *gt* is not located anteriorly enough in *Anopheles* and is completely absent *Clogmia*, it is implausible that such an interaction regulates the expression of the stripe in those species. The question thus becomes one of finding a pathway that "dissolves" the *gt* expression while allowing

for other stripes of *eve* to take the place of *eve* 5 in *Drosphila*. In accordance to this, the relation between other *eve* stripes in the posterior is rather different which could imply that the regulation scheme for the pair-rule gene has changed between these insects. For example, the sixth and seventh stripes are situated symmetrically on either side in the *gt* posterior domain in *Drosophila* (anterior to the *hb* domain) whereas in *Anopheles* the sixth and seventh stripes are found symmetrically about the *hb* domain. Additionally, a weak eighth *eve* stripe is present in the posterior of *Anopheles*.

In the description that follows, we simulated the phenotypic evolutionary pathway between *Drosophila* and *Anopheles* using computational tools and the genetic algorithm described above, inferring a LCA along the way. Solutions were found using appropriate fitnesses described in previous sections and were found multiple times across many simulations. An overview of the simulation scheme overlaying the evolution tree that we found as well as the *eve* modules in the different insects are displayed in Figure 4-1.

Note that in all these simulations, *eve* stripe 1 is not shown as it is primarily the posterior genes that vary and are the focus of the study. The maximum of each *eve* module is normalized to 1 for visualization purposes.

4.1 *Drosophila* to *Anopheles*

Our first set of simulations start from the *Drosophila* network defined in Figure 2-4. As described previously, the target designated by the fitness function for these simulations is the *Anopheles* gap gene profile as an end point. With no restrictions on size or position from the fitness function, the end profile appears to be uneven

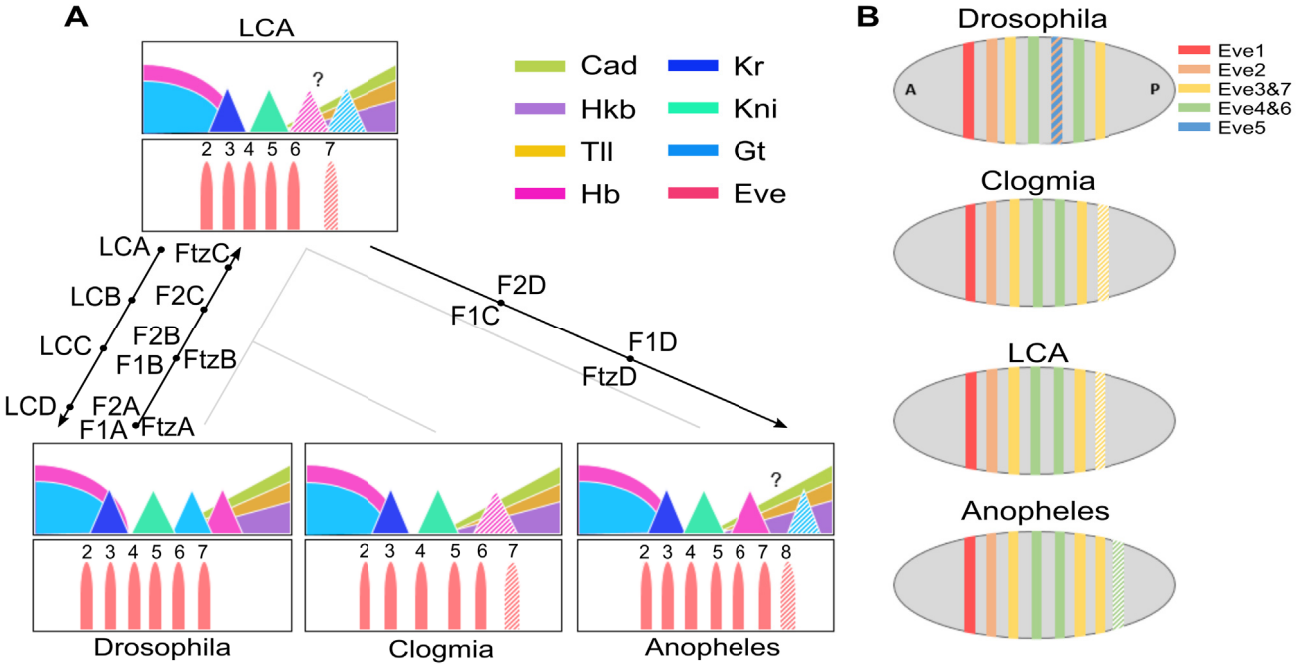

Figure 4–1: (A) Predictive evolutionary tree connecting *Drosophila*, *Clogmia* and *Anopheles* through some last common ancestor. The labelled points along the branches of this tree correspond to points in our simulation, discussed in the respectively labelled Figures found in subsequent sections. (B) Caricatures of the *eve* stripes in the different insects, color-coded according to their corresponding module in *Drosophila*

however it can be imposed starting from the last step shown in panels (**D**), obtained by regulating the strength of interactions at that time point.

An example of one of the simulations from *Drosophila* to *Anopheles* is found in Figure 4-2.

In this particular simulation, the posterior *hb* domain moves anteriorly to split the *eve* 7 stripe into an additional stripe as seen in 4.1C. This allows for stripe 5 to disappear while maintaining the condition that seven stripes are expressed in the

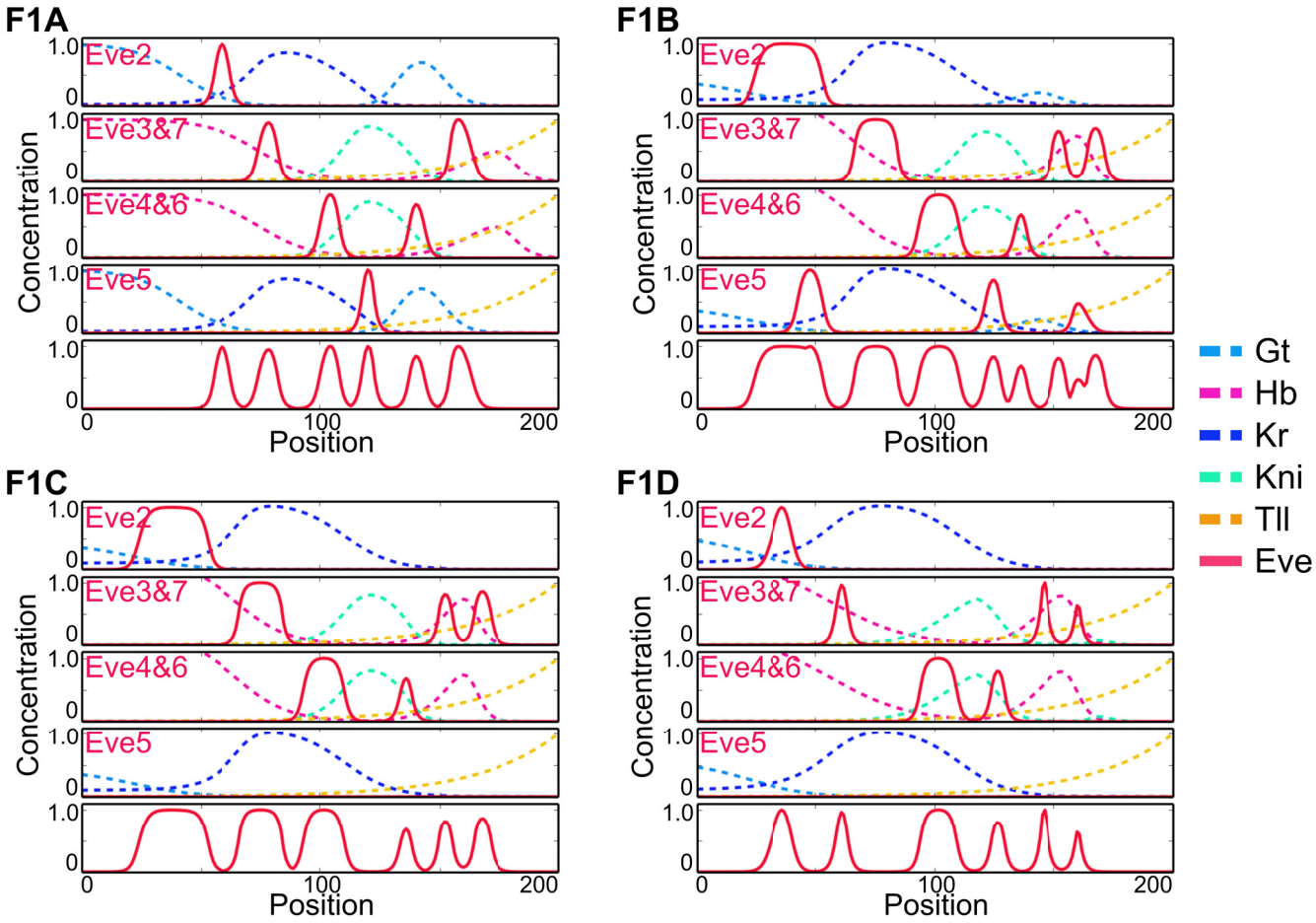

Figure 4–2: Simulated evolutionary pathway (label F1 on 4)

profile. Once this stripe vanishes, the posterior *gt* domain can become obsolete in the *eve* network and eventually dies out. We find ourselves with the module that previously defined stripes 3 and 7 now controlling the stripes 3, 6 and 7 whereas the *eve* module responsible for the control of the 4+6 stripes now controls the stripes 4 and 5.

Similarly, Figure 4-3 presents a different pathway taken by our simulation through phenotype space. However in this simulation, once a new *eve* stripe appears in the posterior of the embryo it is first the posterior domain of *gt* that disappears allowing for the stripe 5, repressed by *gt*, to extend towards the posterior and merge with what is at that time stripe 6. The module for *eve*5 can then completely disappear in further generations.

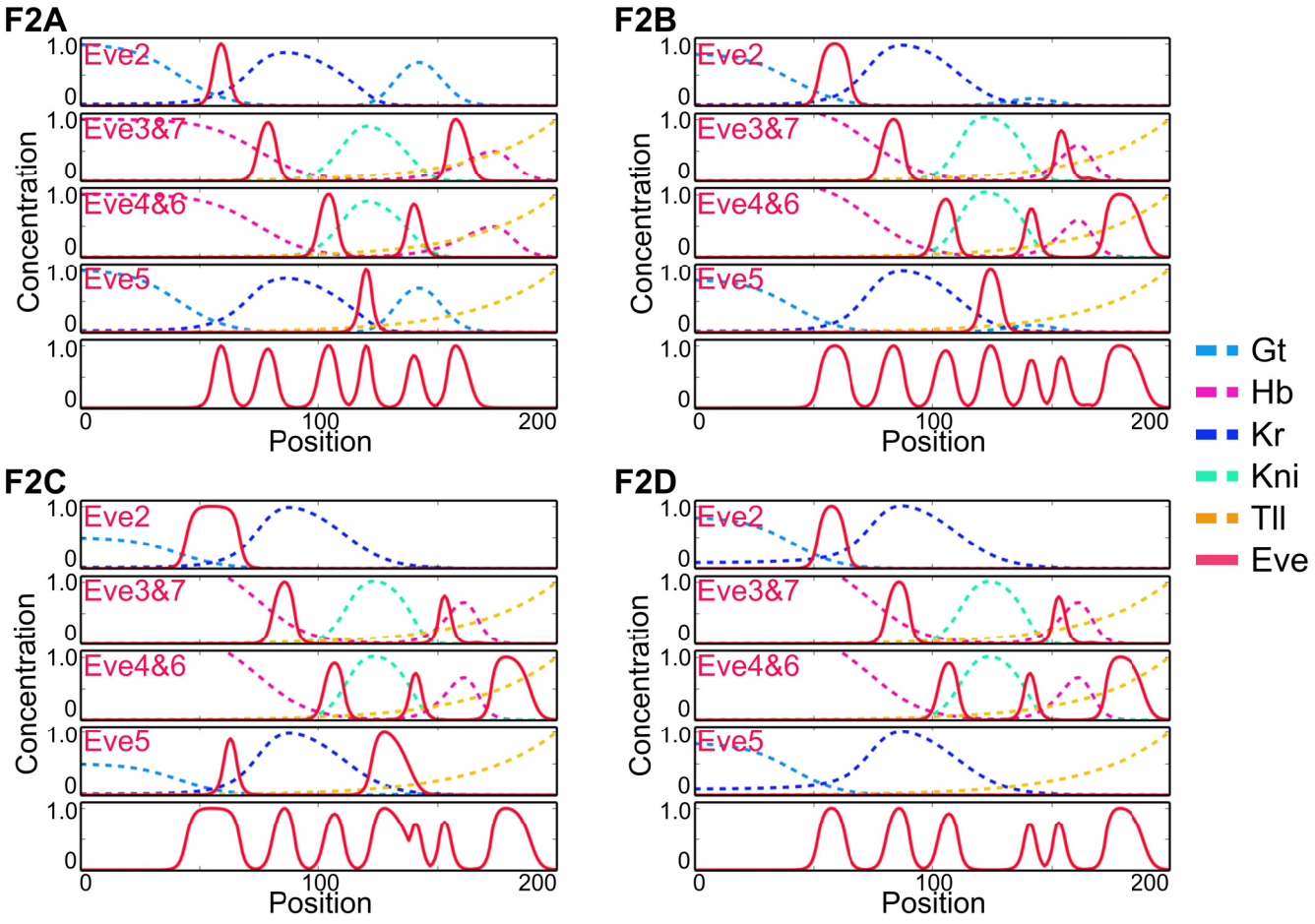

Figure 4–3: Simulated evolutionary pathway (label F2 on 4)

Both the evolutionary scenarios presented describe a phenomena that is a highly reproducible result from our simulations. To go from *Drosophila* to *Anopheles* it is necessary to create one if not more new posterior *eve* stripes in the existing modules to subsequently remove the *eve* 5 stripe module completely from the network while maintaining the constraint of a minimum of 7 stripes. In both cases, the evolutionary pressure of the posterior *hb* to shift forward consequently leads to the formation of a third stripe in either the module 3+7 or 4+6 symmetrically about this *hb* domain. This eventually gives rise to these modules expressing the strips 3,6 and 7 or 4, 5 and 7 respectively. The observation that the pair-rule pattern of *eve* in *Anopheles* is composed of eight stripes can most likely be explained by having both modules express a third stripe at the same time.

A prediction of our simulations is that intermediate Dipterans retain the modules, regulation scheme and logic of the *eve* modules present in *Drosophila*. For example, $eve4+5$ and $eve3+6$ are homologous to the $eve4+6$ and $eve3+7$ respectively in *Drosophila* which exhibits an additional module of *eve*5. As these stripes ($eve4+5$ and $eve3+6$) are repressed by *kni* and *hb*, we expect to find them symmetrically about the *kni* domain which is not the case in *Drosophila* where the fifth stripe, its own module not having any *kni* repression, is situated in the center of the *kni* peak.

4.2 *LCA* to *Drosophila*

In the previous simulated evolution, the *eve*5 module always lost all gene expression by the time the target phenotype was reached. Thus one might ask the

important question of how *eve*5 appears in the evolution from a least common ancestor to *Drosophila*. Is it possible that it is a complete new fabrication of the network or simply an alteration of the existing network? A hint to a possible solution is suggested in Figure 4-3 where the *eve*5 module had a weak stripe emerge at the exact position where the corresponding *eve*2 stripe is at that generation. This is consistent with the observation that both the *eve*5 and *eve*2 modules have the same *Kr* and *gt* repression scheme.

We started by defining a network with the appropriate last common ancestor characteristics for the gap genes, such as a very posterior *gt* domain, and *eve* modules corresponding to the ancestral gene homologous to their *Drosophila* counterparts, excluding *eve*5. This network produces the gene expression profiles present in Figure 4-4**A**. The rest of the panels of Figure 4-4 illustrate different key generations in the simulated evolution.

As the posterior *gt* domain shifts forward to a more anterior position than the posterior *hb* domain, the ancestral *eve* 2 (*aeve* 2) module forms a second stripe situated along the axis in a position that would make it the fifth stripe in the total ancestral eve pattern. This is the generation depicted in panel **C**. It is then possible for a distinct stripe 5 to later evolve through some genetic drift. Simultaneously, the appearance of this stripe and the back shift of posterior *hb* allows for the destruction of the second $aeve4+5$ stripe. Later, the $aeve4+5$ stripe reappears posterior to the new fifth stripe at the same time as the $aeve3,6+7$ module loses a stripe due to the *hb* domain moving more to the posterior in panel **D**. Thus the *Drosophila* pattern

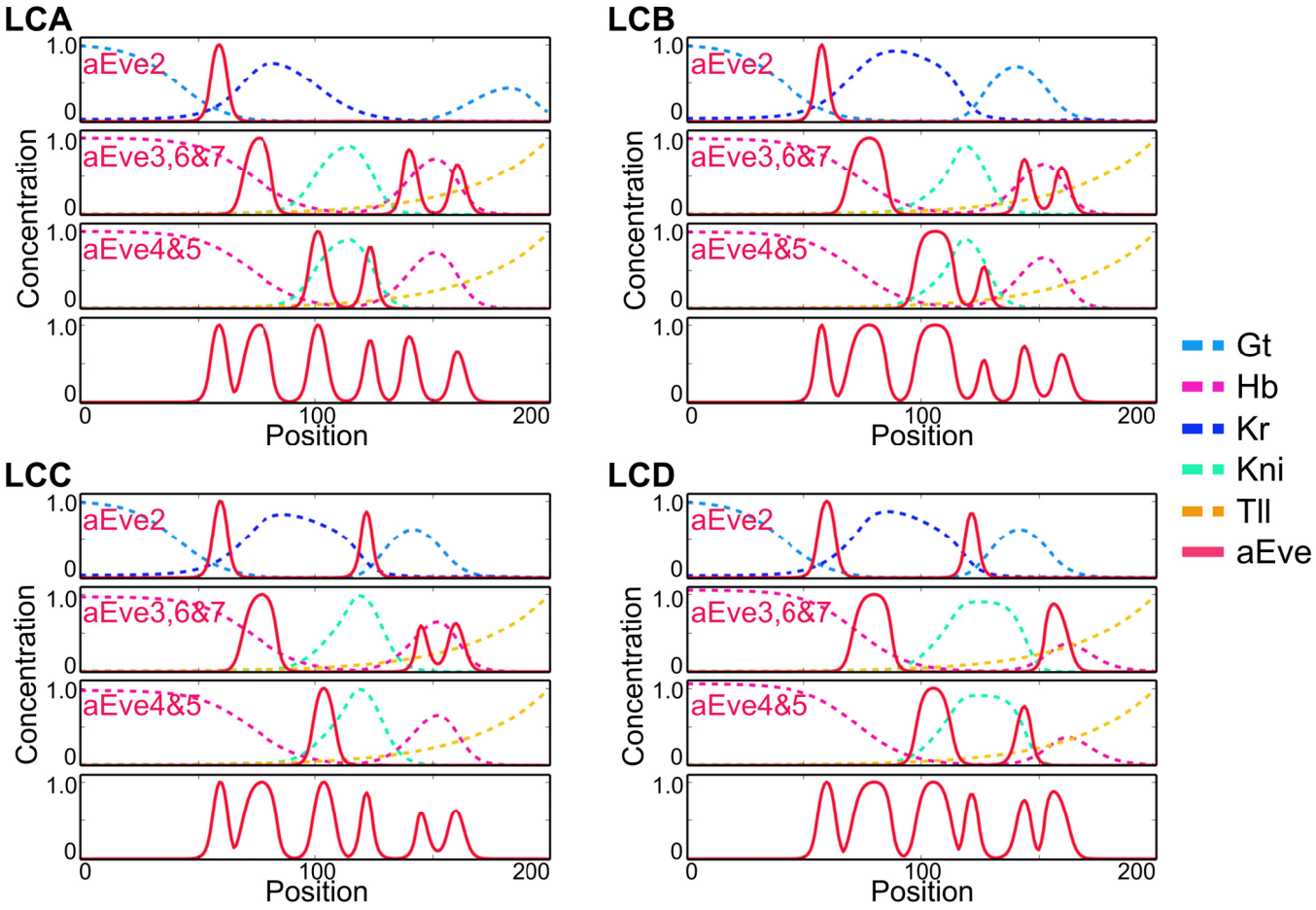

Figure 4–4: Simulated evolutionary pathway (label LC on 4)

of gap gene and segmentation gene is recovered.

4.3 *Clogmia*

We initiated a gap gene network without the posterior *hb* domain and check that such a gap gene network (which resembles the gap gene expression profile of *Clogmia*) generates the appropriate eve profile.

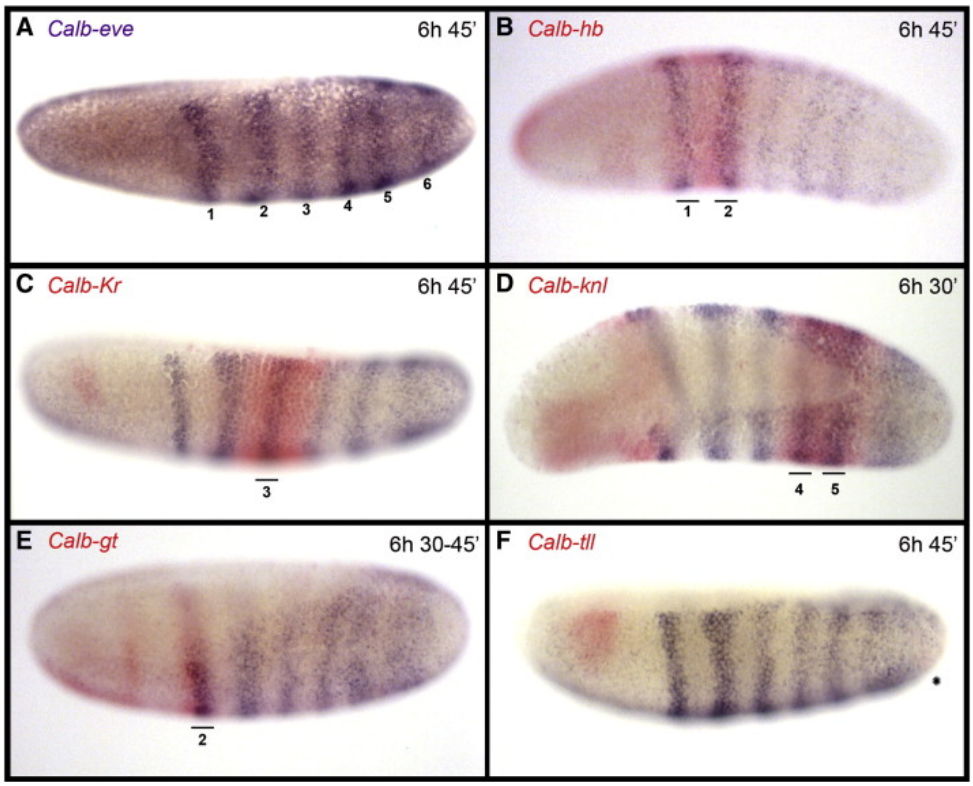

Figure 4–5: The *Clogmia eve* gene expression profile (in purple) positioned from anterior to posterior, right to left. Note there are 6 stripes prior to gastrulation. Panels B-D show the expression of these *eve* stripes with respect to the gap gene expression profiles (red). *eve*4&5 are positioned symmetrically about *kni* (D) and no posterior *gt* is present (E), whereas *eve* stripe 3 is centered about *kr* (C). The figure is reproduced as in *A systematic analysis of the gap gene system in the moth midge Clogmia albipunctata* by Garcia-Solache[65].

As seen in Figure 4-5, *Clogmia* exhibits only 6 stripes prior to gastrulation and our model is indeed consistent with this observation.

Much like in the final simulation steps targeting *Anopheles*, the *Clogmia*, 4-6(A), profile doesn't have any *eve*5 module from *Drosophila* and it is in fact not necessary to obtain the correct *eve* profile. What differentiates this profile with

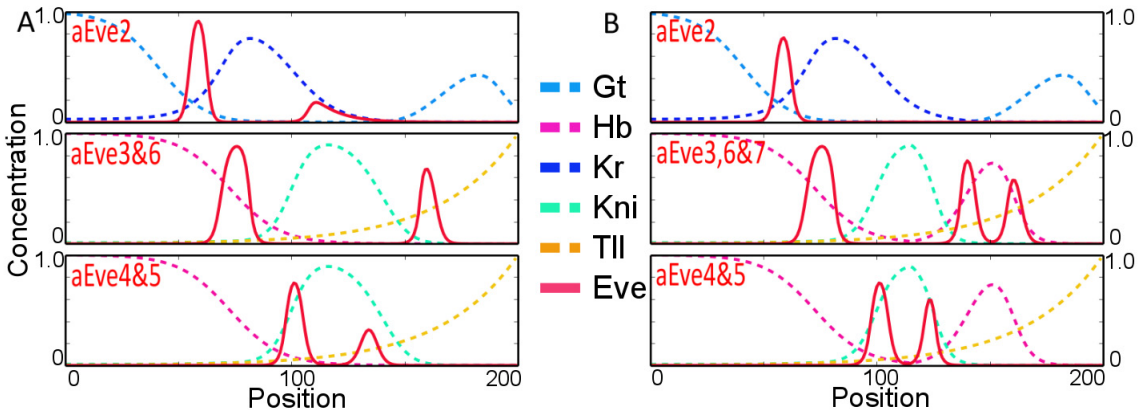

Figure 4–6: The *Clogmia* simulated gene expression profile in (A) compared to the LCA in (B). These are found in the evolution of our networks.

the previous LCA gene expression profiles is simply the lack of a posterior *hb* domain. Our model, as expressed in the figure, predicts that the ancestral *eve* modules *aeve*3 + 6 and *aeve*4 + 5 corresponding to *eve*3+ and *eve*4 + 6 in *Drosophila* are indeed laid out symmetrically about the *kni* domain that represses them.

4.4 *Drosophila* to *Anopheles* with *ftz*

As described previously, pair-rule genes define the location of anteroposterior parasegments in dipterans. Thus far, steps in the evolutionary paths have resulted in the destruction and creation of additional par-rule stripes. The problem with this picture is that biologically we could not find viable mutant flies with two consecutive segments missing. A relevant example to our simulations is the disappearance of *eve*5: the stripe is responsible for the A4 posterior and A5 anterior boundary. There must be some way for the embryo to retain information on the polarity of the proper

parasegments, in this case maintain the polarity with A4 anterior and A5 posterior. Our hypothesis is that another pair-rule gene that is out of phase with *eve* is present and provides input for the segment polarity even when certain stripes disappear. Thus, when a certain stripe disappears, the neighbouring stripes from the out of phase pair-rule gene will combine and only one parasegment is lost. The segmentation gene ftz was added to the model to play this role: the network for the ftz genes are defined in Figure 2-4. Note that for the first set of simulations, we postulated that a pure ftz stripe 4 be regulated by *hb* in it's anterior and *gt* in it's posterior, not as defined in Figure 2-4 however this did not work. For now only these two pair-rule genes are present in the model, additional ones will be considered below in a broader discussion of polarity in the embryo.

However, simulating ftz as a primary pair-rule gene in this way (regulated by *hb* in it's anterior and *gt* in it's posterior) led to problems: the evolutionary pathway which were obtained in simulations are not achievable in the constraints set by our new fitness, namely the alternating striped ftz and *eve* expression. First of all, both the presumed $ftz4$ and $eve5$ modules depend on the posterior domain of *gt* which is trying to disappear. Additionally the crucial constraint of having an alternation of two pair-rule genes poses a problem in terms of the creation of additional *eve* stripes in the posterior that previous simulations exhibit. The difficulty lies herein adding a stripe that is bounded by two other stripes of the interchanging gene, a non-trivial task in the evolution. Trials of simulations led us to conclude that simulations with ftz as a primary pair-rule gene are unsuccessful in preserving an interspersed pattern of eve/ftz.

An important fact in this context, and one not used in previous simulations, is that the ftz stripe 4 appears only with the 7 stripe *zebra* pattern in *Drosophila*. We can get this secondary pair-rule behaviour by allowing repression of the $ftz4$ stripe by *eve* and not *gt*. The $ftz4$ stripe will be regulated by the *eve* profile in the posterior and thus functions as a secondary pair-rule gene. An example of the simulated evolution of such a network is found in Figure 4-7.

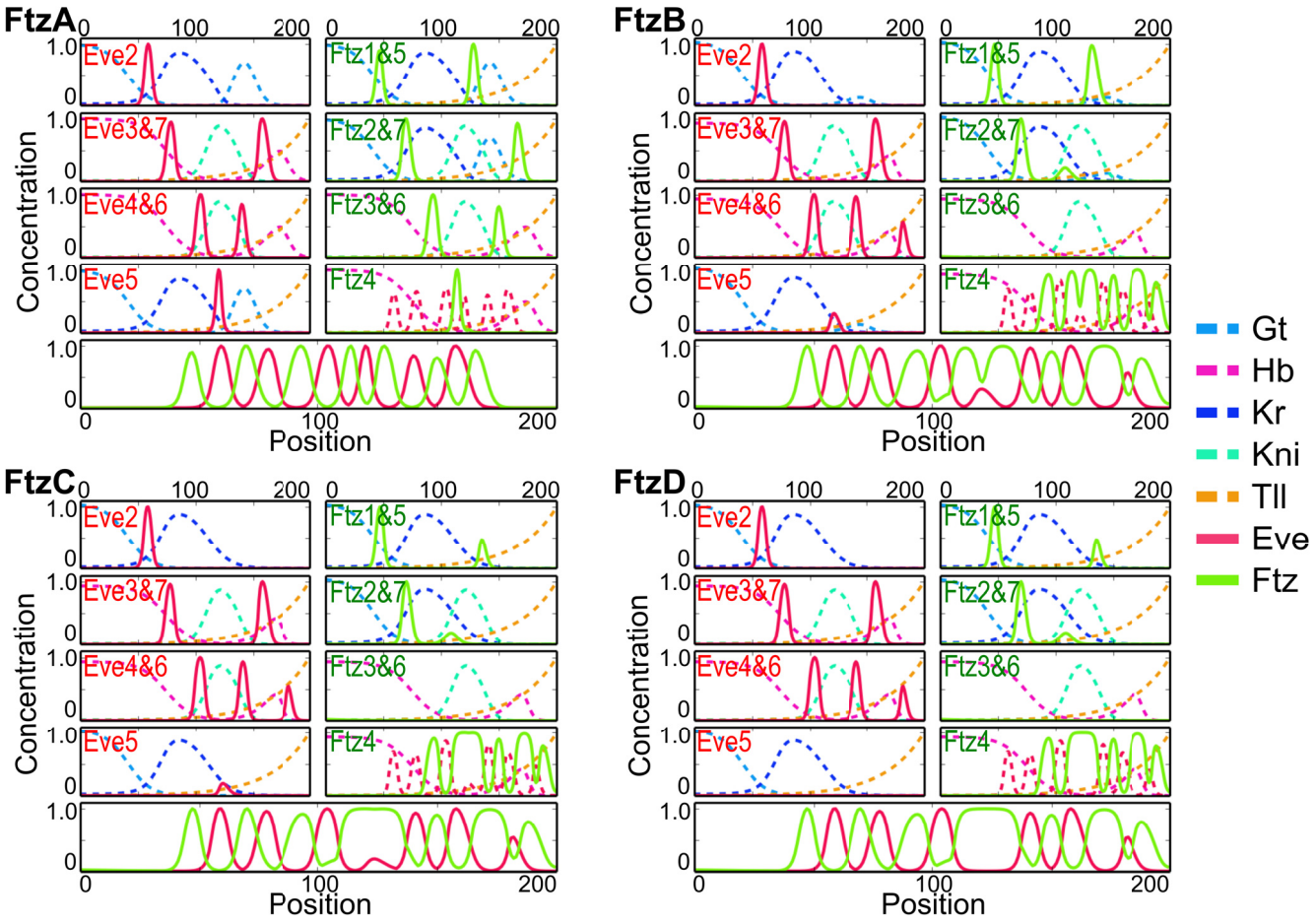

Figure 4–7: Simulated evolutionary pathway (label Ftz on 4)

In these simulations, the posterior ftz stripes, namely 5, 6 and 7 that are normally independent modules regulated by the gap genes gradually disappear and are replaced by the pair-rule $ftz4$ module. The ftz modules controlled by gap genes, now completely anterior, vary little in the evolution as the anterior pattern of *eve* doesn't either. This keeps the intertwining pattern of the segmentation genes sustained. However, as the posterior profile is now determined by the $ftz4$ module, any *eve* stripe dying out will have the flanking ftz stripes merge together into one stripe and thus maintain the alternating segmentation gene pattern that is required. This is exactly what happens to *eve*5 bounded by the $ftz4$ in Figure 4-7 **B** and **C** and hence the evolutionary pathway exhibits ftz and *eve* alternation at every generation from *Drosophila* to *Anopheles*, a reproducible result over many simulations.

4.5 Pair-rule gene polarity

In our study of establishing polarity in the embryo it was noted that evolutionary simulations fail as we add a primary pair-rule gene (ftz), succeeding only when the said pair-rule gene was made secondary. This suggests that doing the same would be true for additional pair-rule genes added to the network. It is nonetheless an interesting question to ask how the phase of multiple pair-rule genes could be conserved in such a network. This is also relevant to the question of polarity in the embryo as a 2 pair-rule network is not sufficient to distinguish the front from the back of our organism. Looking at a section defined by an *eve* and ftz alternation, it is impossible to say what is the anterior or posterior. However having additional striped genes will give some front-to-back ordering. In other words, adding the genes

hairy (h) and *runt* (run) to obtain the sequence *eve*, *run*, ftz and h distinguishes an anterior and posterior since the sequence is not identical when switched around (as opposed to having only *eve* and ftz). Following the previous work done with ftz, we propose that the pair-rule regulation scheme is much more dynamic in the posterior of the LCA compared to that of *Drosophila*. Our quantitative model of the dynamic regulation will combine h and run to the pre-existing network.

It is a shift of the *eve* stripes from posterior to anterior in *Drosophila* and *Clogmia* prior to gastrulation that suggests potential dynamic nature of this pair-rule regulation. If this is indeed the case, then there would be no correspondence, except for *eve*, between *Drosophila*'s gap gene regulated pair-rule genes and the dynamically regulated pair-rule genes in the LCA. The idea is that an appropriate combination of repressions between the segmentation genes *eve*, *run*, ftz and h will utilize the forward shift and establish the correct pattern as seen in *Drosophila* without any additional gap gene input. The additional pair-rule regulation scheme and the strength of repressions are characterized in Figure 4-8 **A**. These are in addition to the already existing gene regulatory network defined in our previous simulations. Coupled to temporal oscillations of the concentration of *eve* within a cell, this model exhibits an alternating phase of higher to lower concentrations of the different pair-rule genes within the cell, visualized in Figure 4-8 **B**. One cycle of this oscillation of *eve* is synonymous to the passing of one *eve* stripe through the cell, induced by some shift towards the anterior within the embryo.

It is through the control of the maternal gradients that the shift of gene concentrations to the anterior is introduced to our model. Further down the cascade

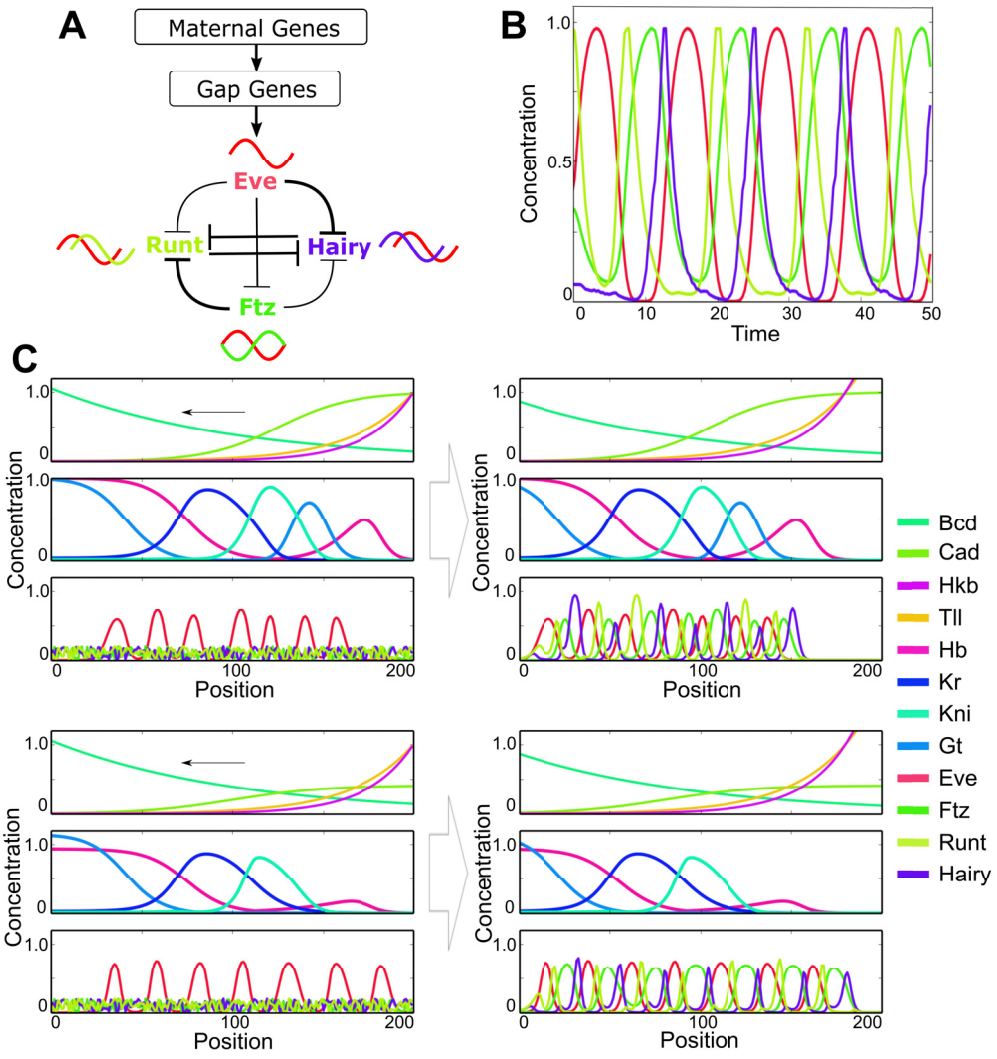

Figure 4–8: (A) Network. (B) Concentration in time. (C) Top *Drosophila*, Bot *LCA*.

of genes, the gap and, eventually, pair-rule genes also experience a forward shift as they are dragged by the displacement of the maternal genes. Thus by adjusting the location of maternal gradients dynamically it is possible to mimic the qualitative

behaviour of the genes through the early stages of embryogenesis. This shift was implemented in the *Drosophila* and *Anopheles* networks, as seen in the left panels of Figure 4-8 **C**, with random initial concentrations of pair-rule genes excluding *eve*. The interactions of the genes and sliding of the gap genes cause the noisy pair-rule genes to stabilize to their known relative phases in *Drosophila*, as the right panels of Figure 4-8 **C** show. Although this model for a dynamic ordering of the segmentation gene pattern was applied to the entire anteroposterior axis, we can expect that the anterior gap gene regulations persist while the posterior domain expresses this more dynamic behaviour, much like in Figure 4-7.

CHAPTER 5
Discussion

5.1 Pair-rule genes and polarity

The main result of this work is the prediction of the dynamics of the pair-rule genes throughout the evolutionary process that differ between species and yet shows homology. We have shown that certain *eve* modules (*eve* 3+7 and *eve* 4+6) of the *Drosophila* network can form, through some mutations, into analogous modules in *Anopheles* (respectively *eve* 3,6+7 and *eve* 4,5+8). Although it might seem unsurprising that this would be the case, the fact that a complex gene network is the backbone of this pattern means that the exact phenotypic pathway through which this change could have occurred is non-trivial. It is promising evidence in favour of our model that the correct pattern for *Clogmia* can also be attained through this evolutionary method, even with our minimal network.

In terms of the LCA, we have formulated that certain posterior pair-rule gene, *ftz*, *run* and *hairy* in this case, would have to be secondary as they would derive their phases from an anterior drift of the pre-existing gene network, notably *eve*. This quantitative system drift, i.e. physical shift of the gene profile to the anterior, would cause the The constructed feedback loop of these four segmentation genes is reminiscent of a segmentation clock which would regulate the phase through the use of a similarly constructed cyclic network. Indeed, a recent paper explains that in certain species, such as *Tribolium castaneum*, that although the anterior segmentation

gene profile is provided positional information, the posterior pattern growth employs a segmentation clock to develop it's cyclic pattern[66]. Additionally, *Nasonia vitripennis*, which is known to employ a similar structure as *Drosophila* in the anterior, displays a phasing of the pair-rule genes much like a segmentation clock as of the fifth segment[67]. This is precisely the location where our simulations become dependent on this form of second-pair rule gene for the segmentation genes other than *eve*. This all suggests that the mechanism we have modelled in *Drosophila*, where *eve* is a first pair-rule gene regulated by the gap genes that fixes the phase for subsequent pair-rule genes, is a strong potential candidate for the regulation in *Anopheles* and is something that could be tested in *Tribolium*[68].

5.2 The computational evolution

This study further illustrates the power of computational methods in inferring phenotypes and even ancestral genetic networks, offering predictions as to the dynamics that govern the development of such living organisms. In the context of Dipteran embryogenesis it is impossible to experimentally study the dynamics of the segmentation genes of the ancestral species. However, these computational methods derived from machine learning and certain principles in physics allow the development of predictive models from available data. Thus the study of these different evolutionary pathways is in itself a method to explore the dynamics of these GRNs as we can compare the pair-rule gene dynamics of the different models and see how they differ from their observable quantities.

It must be mentioned that by no means do we state that the result given here is unquestionably conclusive evidence that our model is the correct one. The genetic

algorithm has provided us with a simple possible pathway that would explain the differences in these species. There is of course the possibility that this is not the proper formalism, but given that is a simple mechanism it is a promising approach. Notably, further data from the different modules in *Anopheles* and *Clogmia* must be extracted so that predictions of our study can be compared to the biology in those systems. Furthermore, work on *Tribolium* has shown that there is in fact a conservation of certain gene enhancers homologous to the ones in *Drosophila*[69], suggesting that it might not be such an uncommon phenomenon within the family order of Diptera. Regardless, the ability to predict qualitative features using quantitative models highlights the importance of using more quantitative frameworks in approaching biological problems.

5.3 Conclusion

The gene regulatory network used to model the expression profiles of the different genes allowed us to find evolutionary pathways between two species of the family of the order Diptera. Through these simulations, ancestral phenotypes were obtained, potential networks that can explain. This is the power of computational evolution: finding these predictive models which otherwise cannot be observed. From these models, other ancestral dynamics can be tested. However, many assumptions were made to model the genetic network of the species studied, notably the structure of the network itself. As stated previously, there are many unknown factors contributing to the complete regulation scheme of the GRN of *Drosophila* during embryogenesis and although a lot of data exists concerning the system, deriving proper models is non-trivial. Extensive study of the system as a whole has revealed

many shadow enhancers, independent components that seem to have a pretty significant effect of the robustness of the anteroposterior patterning[70]. Experiments perturbing the initial maternal input have also shown that in fact the cascade of genes seems to exhibit a more responsive interpretation than previously thought, adjusting dynamically to these perturbations[15]. This would contradict the current threshold-dependent model, thus the underlying mechanism of this pattern formation is as of yet still uncertain.

Consequently, there is still a lot of work to be done in understanding the organization and structure of the gene regulatory network key to the segmentation of the *Drosophila* embryo along the anteroposterior axis. Starting with experimental data, it is possible to lean on available techniques, as Dubois et al. do with their more information theory based approach[71]. The problem here lies in constructing models that do not overcomplexify the solution, that is to say overfit the observable data. No doubt probabilistic models and interpretations, such as Bayesian Statistical Inference, can help tackle these issues and cleverly configured evolutionary algorithms have a lot to offer in terms of function optimization and model comparison: they need only be applied in the right representation for the problems conceived. We suggest that utilizing a version of the Bayesian Information Criterion[72] into the fitness function of a GA would help construct a sufficient model for the gene regulatory network of *Drosophila*, without overfitting said model. It is up to researchers from more quantitative sciences to take an interest in these yet unanswered questions in biology, from which some very fundamental concepts can be learnt, and to apply

their knowledge as well as techniques to bring some meaning to this plethora of data available.

Appendix A: *Drosophila* Network

	Regulated by								
Gene expression	*bcd*	*tll*	*hkb*	*cad*	*hb*	*gt*	*Kr*	*kni*	*eve*
cad	0.3/5	-	-	-	-	-	-	-	-
hb	0.5/-9	0.4/-3	0.45/6	-	-	-	-	-	-
gt	0.7/-10	0.23/7	-	0.6/-10	-	-	0.9/3	-	-
Kr	0.35/-10	-	-	-	0.8/4	0.1/1	-	-	-
kni	0.24/-10	0.2/5	-	-	0.1/2	-	0.6/4	-	-
*eve*2	0.45/-10	-	-	-	-	0.2/10	0.3/10	-	-
$eve3+7$	-	0.35/10	-	-	0.55/10	-	-	0.018/10	-
$eve4+6$	-	0.25/7	-	-	0.1/7	-	-	0.5/10	-
*eve*5	0.08/-2	0.25/5	-	-	-	0.07/10	0.3/10	-	-
$ftz1+5$	0.1/-10	-	-	-	-	0.46/10	0.1/10	-	-
$ftz2+7$	0.1/-10	0.43/10	-	-	-	0.08/7	0.7/7	0.03/7	-
$ftz3+6$	0.1/-10	0.29/7	-	-	0.25/7	-	0.13/5	-	-
$ftz4$	-	0.1/10	0.01/-10	-	0.04/10	-	-	-	0.2/10

Table 5–1: *Drosophila* Network Parameters

Gap gene network parameters for *Drosophila*. Each interaction is tabulated as the $(concentration\ threshold)/(hill\ coefficient)$. Positive hill coefficients denote a repression, whereas negative hill coefficients denote an activation. We've added *cad* too as it expresses a slight *bcd* repression.

Appendix B: *Anopheles* network

	Regulated by							
Gene expression	*bcd*	*tll*	*hkb*	*cad*	*hb*	*gt*	*Kr*	*kni*
cad	0.3/5	-	-	-	-	-	-	-
hb	0.51/-9	-	0.2/6	-	-	-	-	-
gt	0.7/-10	0.8/7	-	0.95/-15	-	-	0.9/3	-
Kr	0.39/-10	-	-	-	0.8/4	0.1/1	-	-
kni	0.25/-10	-	-	-	0.1/2	-	0.7/4	-
*aeve*2	0.4/-10	-	-	-	-	0.2/10	0.32/10	-
*aeve*3, 6&7	-	0.34/10	-	-	0.6/10	-	-	0.04/10
*aeve*4&5	-	0.15/7	-	-	0.1/7	-	-	0.74/10

Table 5–2: *Anopheles* Network Parameters

Gene network parameters for *Anopheles*. The *aeve* modules correspond to the 'ancestral' eve pattern. Each interaction is tabulated as the $(concentration\ threshold)/(hill\ coefficient)$. Positive hill coefficients denote a repression, whereas negative hill coefficients denote an activation. We've added *cad* too as it expresses a slight *bcd* repression.

Appendix C: The Fitness Function *(Cont'd)*

In terms of the fitness function for the evolution from *Drosophila* to *Anopheles* , certain conditions had to be demanded of the network as stated previously. First, the amount of eve stripes had to be maintained to at least 7 stripes, as it is consistent for both the initial and final profiles. Second, to be biologically relevant the model needs to express an alternation of the pattern of the segmentation genes even-skipped and fushi tarazu. Thus if either of those conditions are not met, the network is irrelevant and attributed a high score. To further direct the network through a relevant genetic pathway, a certain score is attributed to the quantity of giant in the posterior region of the embryo. Expressing the concentrations of giant in the cells in the posterior as values in a vector, the score is assigned as the dot product of this vector. Since there is no Giant in the posterior of the mosquito, seeing this dot product minimized will give the required result. Similarly, a condition on Hunchback was imposed whereas the difference between the current profile and profile in the posterior of anopheles was attributed a score. Finally, as the profile of even-skipped differs very little in the anterior of either species the difference of the current eve profile and the initial eve profile is another condition in the fitness. Note that as to not be too restricted in the evolution only the maximum of the difference of eve ($diff_{Eve}$) and difference

of hunchback ($diff_{Hb}$) is added to the score:

$$Score = C_1 \sum_{i \in posterior} (Gt_i)^2 + max(C_2 diff_{Hb}, C_3 diff_{Eve}) + \begin{cases} 0 \text{ if alternation and} \geq 7 \text{stripes} \\ 1000 \text{ otherwise} \end{cases} \tag{5.1}$$

This is what the fitness function tries to minimize in order to obtain the most promising network. The different constants C_1, C_2 and C_3 are there to scale the strength of each individual component of the score, depending on how strong the condition needs to be. Different values of the constants were used across many simulations, however for the final simulation the values used were of $C_1 = C_2 = 1$ and $C_3 = 0.1$ Note that when simulations were conducted without $fushi\ tarazu$, the condition of alternating the stripes was removed and the rest of the fitness was maintained.

As for the opposite evolution from the LCA to fly described earlier, the profiles have notably the positions of *hunchback* and *giant* inverted in the posterior. The function that we wish to minimize aught to reflect this, thus to keep things simple, the score was determined as the subtraction of the current concentration of gap genes (*knirps*, *Kruppel*, *giant* and *hunchback*) in each cell from the ones in the embryo of the fly. Practically what this means is that for each gap gene the concentrations were arrayed by cell number in the embryo and then subtracted from an array of the concentrations of that same gene in the fly, obtaining an array of the differences, $diff_{gap}$. the sum of the dot product of the arrays is our score, since as the profiles of the gap genes converge, the score converges to zero. Similar to the previous case, a high score was attributed if less than 7 stripes of Eve were observed in the embryos profile.

$$Score = \sum_{\text{gap genes}} (diff_{\text{gap genes}})^2 + \begin{cases} 0 \text{ if } \geq 7\text{stripes} \\ 1000 \text{ otherwise} \end{cases} \tag{5.2}$$

The following 2 figures show examples of different networks on different evolutionary simulations. One can see that the convergence towards a certain plateau of constant fitness if rather sporadic and hard to predict, given the nature of the genetic algorithm.

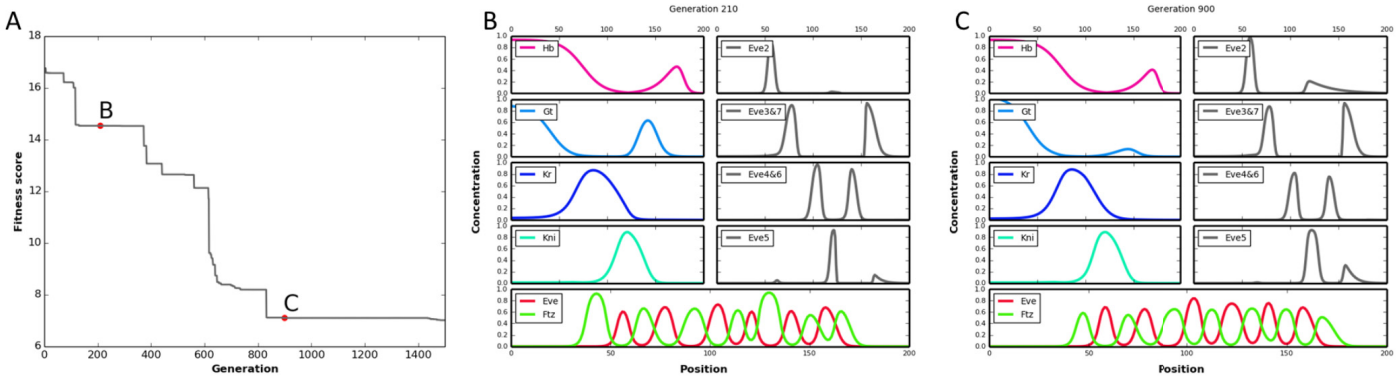

Figure 5–1: Example of the evolution from the fly to mosquito where (A) illustrates the minimization of the score that is attributed each network along the given evolutionary pathway as well as denoting where (B) and (C) lie on this path. The network illustrated in (B) is earlier and hence has a larger score, as attributed by Equation 6, because of a larger posterior concentration of Giant than in the network (C) where it is nearly non-existent. Note that both profiles have alternating patterns and at least 7 *eve* stripes (counting the *eve*1 stripe which is not depicted).

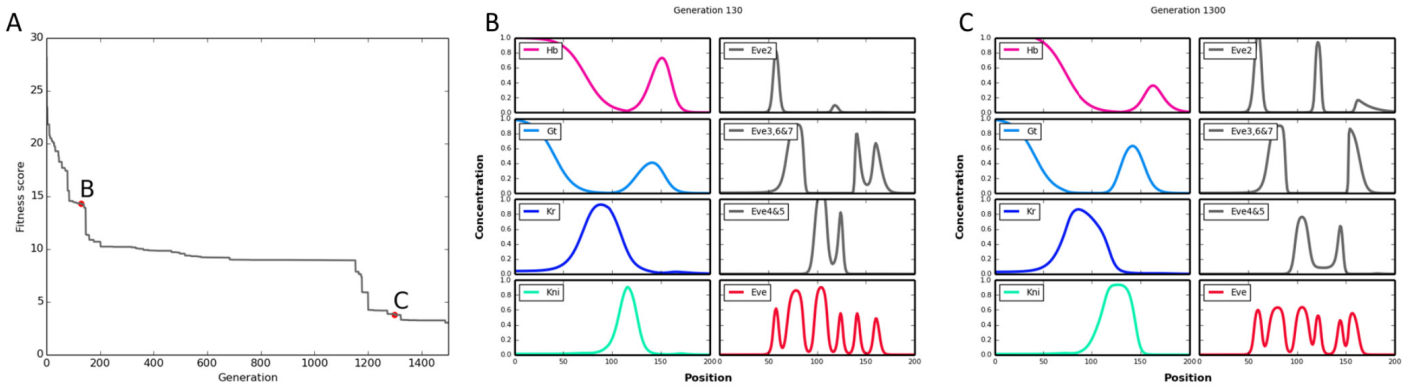

Figure 5–2: Example of the evolution from the presumptive last common ancestor to fly. As in Figure 3, the score of the evolution through the landscape of phenotypes is portrayed in (A) as well as the location in this evolution where the networks of (B) and (C) can be found. Following the score defined in Equation 7, the gap gene profiles in (B) are not as close to the profile we expect in the fly than are the profiles of (C).

References

[1] Hans Meinhardt. *Models of biological pattern formation*. Citeseer, 1982.

[2] J.E. Sulston, E. Schierenberg, J.G. White, and J.N. Thomson. The embryonic cell lineage of the nematode caenorhabditis elegans. *Developmental Biology*, 100(1):64 – 119, 1983.

[3] Yury Goltsev, William Hsiong, Gregory Lanzaro, and Mike Levine. Different combinations of gap repressors for common stripes in anopheles and drosophila embryos. *Developmental Biology*, 275(2):435 – 446, 2004.

[4] Mónica García-Solache, Johannes Jaeger, and Michael Akam. A systematic analysis of the gap gene system in the moth midge clogmia albipunctata. *Developmental Biology*, 344(1):306 – 318, 2010.

[5] Karl R Wotton, Eva Jiménez-Guri, Anton Crombach, Hilde Janssens, Anna Alcaine-Colet, Steffen Lemke, Urs Schmidt-Ott, and Johannes Jaeger. Quantitative system drift compensates for altered maternal inputs to the gap gene network of the scuttle fly *Megaselia abdita*. *eLife*, 4:e04785, jan 2015.

[6] Ronald J. Konopka and Seymour Benzer. Clock mutants of drosophila melanogaster. *Proceedings of the National Academy of Sciences*, 68(9):2112–2116, 1971.

[7] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1989.

[8] Douglas M. Hawkins. The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12, 2004. PMID: 14741005.

[9] Mark D. Adams and Celniker. The genome sequence of drosophila melanogaster. *Science*, 287(5461):2185–2195, 2000.

[10] Charles B. Kimmel, William W. Ballard, Seth R. Kimmel, Bonnie Ullmann, and Thomas F. Schilling. Stages of embryonic development of the zebrafish. *Developmental Dynamics*, 203(3):253–310, 1995.

[11] Hilde Janssens, Ken Siggens, Damjan Cicin-Sain, Eva Jiménez-Guri, Marco Musy, Michael Akam, and Johannes Jaeger. A quantitative atlas of even-skipped and hunchback expression in clogmia albipunctata (diptera: Psychodidae) blastoderm embryos. *EvoDevo*, 5(1):1–13, 2014.

[12] Paul François and Eric D Siggia. Phenotypic models of evolution and development: geometry as destiny. *Current Opinion in Genetics & Development*, 22(6):627 – 633, 2012. Genetics of system biology.

[13] M. Fujioka, Y. Emi-Sarker, G.L. Yusibova, T. Goto, and J.B. Jaynes. 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*, 126(11):2527–2538, 1999.

[14] Mark D. Schroeder, Christina Greer, and Ulrike Gaul. How to make stripes: deciphering the transition from non-periodic to periodic patterns in drosophila segmentation. *Development*, 138(14):3067–3078, 2011.

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

[16] M Frasch and M Levine. 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, 1987.

[17] L. Wolpert. *Positional information and the spatial pattern of cellular differentiation*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1971.

[18] Lewis Wolpert, Cheryll Tickle, and Alfonso Martinez Arias. *Principles of development*. Oxford University Press, USA, 2015.

[19] Matthew Towers and Cheryll Tickle. Growing models of vertebrate limb development. *Development*, 136(2):179–190, 2008.

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

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

[22] Mikhail Tikhonov, Shawn C. Little, and Thomas Gregor. Only accessible information is useful: insights from gradient-mediated patterning. *Royal Society Open Science*, 2(11), 2015.

[23] Dmitry Krotov, Julien O. Dubuis, Thomas Gregor, and William Bialek. Morphogenesis at criticality. *Proceedings of the National Academy of Sciences*, 111(10):3683–3688, 2014.

[24] John Reinitz, Eric Mjolsness, and David H. Sharp. Model for cooperative control of positional information in drosophila by bicoid and maternal hunchback. *Journal of Experimental Zoology*, 271(1):47–56, 1995.

[25] RP Wharton and G Struhl. Rna regulatory elements mediate control of drosophila body pattern by the posterior morphogen nanos. *Cell*, 67(5):955967, November 1991.

[26] Jordi Casanova and Gary Struhl. Localized surface activity of torso, a receptor tyrosine kinase, specifies terminal body pattern in drosophila. *Genes & development*, 3(12b):2025–2038, 1989.

[27] José A Campos-Ortega and Volker Hartenstein. *The embryonic development of Drosophila melanogaster*. Springer Science & Business Media, 2013.

[28] David S Burz, Rolando Rivera-Pomar, Herbert Jäckle, and Steven D Hanes. Cooperative dna-binding by bicoid provides a mechanism for threshold-dependent gene activation in the drosophila embryo. *The EMBO journal*, 17(20):5998–6009, 1998.

[29] Einat Cinnamon, Devorah Gur-Wahnon, Aharon Helman, Daniel St Johnston, Gerardo Jiménez, and Ze'ev Paroush. Capicua integrates input from two maternal systems in drosophila terminal patterning. *The EMBO journal*, 23(23):4571–4582, 2004.

[30] Mark D Schroeder, Michael Pearce, John Fak, HongQing Fan, Ulrich Unnerstall, Eldon Emberly, Nikolaus Rajewsky, Eric D Siggia, and Ulrike Gaul. Transcriptional control in the segmentation gene network of drosophila. *PLoS Biol*, 2(9), 08 2004.

[31] Rolando Rivera-Pomar and Herbert Jãckle. From gradients to stripes in drosophila embryogenesis: filling in the gaps. *Trends in Genetics*, 12(11):478–483, 1996.

[32] Katherine Harding, Christine Rushlow, Helen J Doyle, Timothy Hoey, and Michael Levine. Cross-regulatory interactions among pair-rule genes in drosophila. *Science*, 233(4767):953–959, 1986.

[33] Hilde Janssens, Shuling Hou, Johannes Jaeger, Ah-Ram Kim, Ekaterina Myasnikova, David Sharp, and John Reinitz. Quantitative and predictive model of transcriptional control of the drosophila melanogaster even skipped gene. *Nature genetics*, 38(10):1159–1165, 2006.

[34] Jeremy B. Rothschild, Panagiotis Tsimiklis, Eric D. Siggia, and Paul François. Predicting ancestral segmentation phenotypes from drosophila to anopheles using in silico evolution. *PLOS Genetics*, 12(5):1–19, 05 2016.

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

[36] Guy Karlebach and Ron Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, 2008.

[37] Paul François, Vincent Hakim, and Eric D Siggia. Deriving structure from evolution: metazoan segmentation. *Molecular Systems Biology*, 3(1), 2007.

[38] Jiri Vohradsky. Neural model of the genetic network. *Journal of Biological Chemistry*, 276(39):36168–36173, 2001.

[39] Tony Yu-Chen Tsai, Yoon Sup Choi, Wenzhe Ma, Joseph R Pomerening, Chao Tang, and James E Ferrell. Robust, tunable biological oscillations from interlinked positive and negative feedback loops. *Science*, 321(5885):126–129, 2008.

[40] Athanasios Polynikis, SJ Hogan, and Mario di Bernardo. Comparing different ode modelling approaches for gene regulatory networks. *Journal of theoretical biology*, 261(4):511–530, 2009.

[41] Moises Santillán. On the use of the hill functions in mathematical models of gene regulatory networks. *Mathematical Modelling of Natural Phenomena*, 3(2):85–97, 2008.

[42] Michael W. Perry, Alistair N. Boettiger, Jacques P. Bothma, and Michael Levine. Shadow enhancers foster robustness of drosophila gastrulation. *Current Biology*, 20(17):1562 – 1567, 2010.

[43] Jeremy Lynch and Claude Desplan. Evolution of development: beyond bicoid. *Current Biology*, 13(14):R557–R559, 2003.

[44] Jonathan S Margolis, Mark L Borowsky, Eiríkur Steingrímsson, Chung Wha Shim, Judith A Lengyel, and James W Posakony. Posterior stripe expression of hunchback is driven from two promoters by a common enhancer element. *Development*, 121(9):3067–3077, 1995.

[45] Jordi Casanova. Pattern formation under the control of the terminal system in the drosophila embryo. *Development*, 110(2):621–628, 1990.

[46] Gary Struhl, Paul Johnston, and Peter A Lawrence. Control of drosophila body pattern by the hunchback morphogen gradient. *Cell*, 69(2):237–249, 1992.

[47] Dorothy E Clyde, Maria SG Corado, Xuelin Wu, Adam Paré, Dmitri Papatsenko, and Stephen Small. A self-organizing system of repressor gradients establishes segmental complexity in drosophila. *Nature*, 426(6968):849–853, 2003.

[48] Xuelin Wu, Rajesh Vakani, and Stephen Small. Two distinct mechanisms for differential positioning of gene expression borders involving the drosophila gap protein giant. *Development*, 125(19):3765–3774, 1998.

[49] M Capovilla, Elizabeth D Eldon, and Vincenzo Pirrotta. The giant gene of drosophila encodes a b-zip dna-binding protein that regulates the expression of other segmentation gap genes. *Development*, 114(1):99–112, 1992.

[50] Dmitri Papatsenko and Michael Levine. The drosophila gap gene network is composed of two parallel toggle switches. *PLoS One*, 6(7):e21145, 2011.

[51] Dusan Stanojevic, Stephen Small, and Michael Levine. Regulation of a segmentation stripe by overlapping activators and repressors in the drosophila embryo. *Science*, 254(5036):1385–1387, 1991.

[52] Paolo Struffi, Maria Corado, Leah Kaplan, Danyang Yu, Christine Rushlow, and Stephen Small. Combinatorial activation and concentration-dependent repression of the drosophila even skipped stripe 3+ 7 enhancer. *Development*, 138(19):4291–4299, 2011.

[53] Melissa M Harrison, Xiao-Yong Li, Tommy Kaplan, Michael R Botchan, and Michael B Eisen. Zelda binding in the early drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS Genet*, 7(10):e1002266, 2011.

[54] Chung-Yi Nien, Hsiao-Lan Liang, Stephen Butcher, Yujia Sun, Shengbo Fu, Tenzin Gocha, Nikolai Kirov, J Robert Manak, and Christine Rushlow. Temporal coordination of gene networks by zelda in the early drosophila embryo. *PLoS Genet*, 7(10):e1002339, 2011.

[55] James A Foster. Evolutionary computation. *Nature Reviews Genetics*, 2(6):428–436, 2001.

[56] Paul François and Vincent Hakim. Design of genetic networks with specified functions by evolution in silico. *Proceedings of the National Academy of Sciences of the United States of America*, 101(2):580–585, 2004.

[57] Paul François. Evolving phenotypic networks in silico. In *Seminars in cell & developmental biology*, volume 35, pages 90–97. Elsevier, 2014.

[58] Darrell Whitley. A genetic algorithm tutorial. *Statistics and computing*, 4(2):65–85, 1994.

[59] Ronald Aylmer Fisher. *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press, 1930.

[60] R. L. Haupt. Optimum population size and mutation rate for a simple real genetic algorithm that optimizes array factors. In *Antennas and Propagation Society International Symposium, 2000. IEEE*, volume 2, pages 1034–1037 vol.2, July 2000.

[61] Heinz Mühlenbein, M Schomisch, and Joachim Born. The parallel genetic algorithm as function optimizer. *Parallel computing*, 17(6-7):619–632, 1991.

[62] Richard A. Neher and Boris I. Shraiman. Statistical genetics and evolution of quantitative traits. *Rev. Mod. Phys.*, 83:1283–1300, Nov 2011.

[63] William M Spears et al. Crossover or mutation. *Foundations of genetic algorithms*, 2:221–237, 1992.

[64] Günter Rudolph. Convergence analysis of canonical genetic algorithms. *IEEE transactions on neural networks*, 5(1):96–101, 1994.

[65] Mónica García-Solache, Johannes Jaeger, and Michael Akam. A systematic analysis of the gap gene system in the moth midge clogmia albipunctata. *Developmental Biology*, 344(1):306 – 318, 2010.

[66] Ezzat El-Sherif, Michalis Averof, and Susan J. Brown. A segmentation clock operating in blastoderm and germband stages of tribolium development. *Development*, 139(23):4341–4346, 2012.

[67] Jeremy A Lynch, Ezzat El-Sherif, and Susan J Brown. Comparisons of the embryonic development of drosophila, nasonia, and tribolium. *Wiley Interdisciplinary Reviews: Developmental Biology*, 1(1):16–39, 2012.

[68] Chong Pyo Choe, Sherry C. Miller, and Susan J. Brown. A pair-rule gene circuit defines segments sequentially in the short-germ insect tribolium castaneum. *Proceedings of the National Academy of Sciences*, 103(17):6560–6564, 2006.

[69] Christian Wolff, R Schroder, Cordula Schulz, Diethard Tautz, and Martin Klingler. Regulation of the tribolium homologues of caudal and hunchback in drosophila: evidence for maternal gradient systems in a short germ embryo. *Development*, 125(18):3645–3654, 1998.

[70] Michael W. Perry, Alistair N. Boettiger, and Michael Levine. Multiple enhancers ensure precision of gap gene-expression patterns in the drosophila embryo. *Proceedings of the National Academy of Sciences*, 108(33):13570–13575, 2011.

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

[72] David Posada and Thomas R Buckley. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, 53(5):793–808, 2004.