Stony Brook University

Breaking the genomic *cis*-regulatory code by an experimental and theoretical analysis of *eve* enhancer fusions

A Dissertation Presented
by

Ah-Ram Kim

to
The Graduate School
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

in

Biochemistry and Structural Biology

Stony Brook University
August 2012

Stony Brook University
The Graduate School

Ah-Ram Kim

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Professor John Reinitz, Advisor
Department of Applied Mathematics & Statistics

Professor Peter Gergen, Committee Chairperson
Department of Biochemistry & Cell Biology

Distinguished Professor Rolf Sternglanz, Committee Member
Department of Biochemistry & Cell Biology

Associate Professor David Green, Committee Member
Department of Applied Mathematics & Statistics

Professor Stephen Small, Outside Committee Member
New York University

This dissertation is accepted by the Graduate School.

Charles Taber
Interim Dean of the Graduate School

Abstract of the Dissertation

Breaking the genomic *cis*-regulatory code by an experimental and theoretical analysis of *eve* enhancer fusions

by

Ah-Ram Kim

Doctor of Philosophy

in

Biochemistry and Structural Biology

Stony Brook University

2012

Encoded within DNA sequence is the *cis*-regulatory logic responsible for controlling gene expression in metazoans. The precise and predictive decryption of this code is on going endeavor at the heart of modern genomics. Even though state of the art technologies in genomics have been generated tremendous amount of data, how the interplay of multiple transcriptional mechanisms give rise to the complex expression changes has remain elusive. This dissertation presents a theoretical model that reconstitutes *even-skipped* transcriptional control *in silico* by implementing molecular regulatory mechanisms that are essential for the *even-skipped* gene expression, then applies the model to *even-skipped* enhancer fusions in order to elucidate the underlying rules governing the transcriptional control of the *Drosophila* genome. Rearrangements of about 2.5 kb of regulatory DNA located 5' of the transcription start site of the *Drosophila even-skipped* locus generate large scale changes in the expression of *even-skipped* stripes 2, 3 and 7. The most radical effects are generated by juxtaposing the minimal stripe enhancers MSE2 and MSE3 for stripes 2 and 3 with and without small "spacer" segments less than 360 bp in length. The model reproduced gene expression of the arrangements with high fidelity and was able to predict expression patterns driven by a variety of segments of the genomic DNA totaling 50 kb for gap and pair-rule genes, *even-skipped* enhancers not included in the training set, stripe 2, 3 and 7 enhancers from various Drosophilidae and Sepsidae species. These results suggest that the molecular mechanisms implemented in the model are essential not only for *Drosophila melanogaster even-skipped* but also for many genes of early *Drosophila* and Sepsid embryo development. In addition, the model predicted gene expression of long segments of *even-skipped* regulatory DNA which contain

multiple enhancers. This result opens the door to quantitative and predictive models of entire loci, the physiological units of the genome. The model demonstrated that two mechanisms, short-range quenching and coactivation, are key mechanisms conferring the independent action of enhancers in the large *even-skipped* regulatory DNA. I establish that elevated expression driven by a fusion of MSE2 and MSE3 is a consequence of the recruitment of a portion of MSE3 to become a functional component of MSE2, demonstrating that *cis*-regulatory "elements" are not elementary objects. Finally, I demonstrate that the conservation of stripe 2 expression driven by six *Drosophila* and Sepsid stripe 2 enhancers requires novel molecular interactions, not seen in the *Drosophila* melanogaster S2E, presenting a clear example of compensatory adaptation with a precise mathematical description of the essential molecular mechanisms.

Contents

List of Figures

List of Tables

I praise you because
I am fearfully and wonderfully made.
Your works are wonderful,
I know that full well.

Psalm 139:14

Chapter 1

Introduction

Transcription is a fundamental process in all living things. Within organisms as minimal as symbiotic bacteria or as massive as a blue whale in the Pacific Ocean, transcription is the critical initial step in transforming the genetic information encoded in their DNA into biological functions. Transcription from the genome is essential for their development and for the ongoing homeostasis. One of the most prominent characteristics of eukaryotic transcription systems is the precise spatiotemporal control of gene expression. Regulatory DNAs in eukaryotes, located upstream or downstream of transcripts, control gene expression by integrating information from a variety of regulatory protein-DNA and protein-protein interactions. Almost all biological processes, including development, differentiation, proliferation, apoptosis and aging, require proper transcriptional control. Deregulation of gene expression can lead to developmental defects or serious diseases [1, 2].

The central players of the control of transcription are sequence specific transcription factors (TFs). A large number of TF genes have been cloned in the past 30 years and their protein functions have been extensively investigated.

For many developmentally essential TFs, their binding sites, the binding site arrangements in their target regulatory sequences and the transcriptional consequences of their activity have been characterized in great detail. Generally, TFs bind to regulatory elements within DNA sequences, guide the unpacking of the chromatin, recruit adaptor factors, such as Mediator [3, 4], which in turn facilitate transcription initiation by interacting with transcription machinery including RNA polymerase II. This basic paradigm stands as a nearly universal mode of eukaryotic gene regulation. Genomic technologies have further expanded our understanding of transcription by generating a wealth of information about the transcriptional regulatory system dynamics extending to chromatin states, enhancer-promoter occupancy by TFs, and quantitative mRNA expression. Despite the increasing volume of knowledge, we still do not understand how the activities of the regulatory proteins are integrated on the regulatory DNA to control gene expression precisely in space and time.

To understand such a fundamental process, we must understand how the diverse components of transcription control system such as TFs, coacivators, transcription machinery and regulatory sequences operate in concert and regulate gene expression. The aim of this dissertation is to arrive at such an integration for the better understanding of the *cis*-regulatory control of metazoan gene expression using enhancer fusions of *Drosophila melanogaster even-skipped* and an *in silico* transcription system.

1.1 *cis*-regulatory control of metazoan genes

Transcriptional control of eukaryotic genes is a complex process, especially in metazoan, that requires the precise orchestration of the interactions of numerous proteins such as chromatin proteins, TFs, transcription machinery including RNA polymerase II (RNAPII) and adaptor proteins, through which TFs regulate the action of the transcription machinery [5]. It is known that regulatory DNA which controls the transcription of genes in higher eukaryotes can frequently be divided into functionally distinct contiguous regions defined by their ability to direct expression independently when placed in reporter constructs. When assayed in this manner, each fragment directs gene expression in a particular tissue or spatiotemporal domain. The genomic regions corresponding to these DNA fragments are known as enhancers or *cis*-regulatory modules (CRMs). Enhancers are typically separated from one another by regions of DNA which cannot independently drive transcription. Enhancers typically contain clusters of binding sites for TFs, can act over many kilobases (kb) from the transcription start site (TSS), and are still functional when orientation is reversed.

Unlike the *lac* operon in *E. coli* and many other prokaryotic genes, the regulatory function of eukaryotic genes resides in multiple TF binding sites [6]. TFs can be divided into two different classes, activators and repressors, depending on their activities. Eukaryotic genes are inactive in the ground state but are activated by multiple activators bound to enhancers. The multiplicity of binding sites on enhancers allows various protein-DNA and protein-protein interactions between TFs and between TFs and transcription machinery such as a cooperative binding and greater than multiplicative synergy [7, 8, 9, 10], which enables fine regulation of transcriptional initiation. These activities are

controlled by multiple repressors. Repressors can prevent activator activities by competition, repression at a short distance of 100-150 base pairs (bp), called short-range repression, or long-range repression (called silencing) [11]. Short-range repression (also called "quenching") is of particular importance. While the silencing mechanism inactivates entire chromosomal locus, the quenching mechanism represses only the activities of nearby DNA-bound activators and does not interfere with distantly bound activators, which consequently ensures the precise control of autonomous enhancer activities [12].

Although substantial progress has been made in understanding the expression of individual enhancers, a level of understanding adequate for prediction has not yet been reached. Further, understanding individual enhancers is itself insufficient, as it is now clear that multiple enhancers act simultaneously to ensure accurate and robust gene expression [13, 14, 15]. Indeed, a complete solution to the *cis*-regulatory logic problem in metazoa requires an understanding of the control of gene expression at the level of a whole, intact genetic locus since it is the whole locus and not the enhancer which is the fundamental unit of physiological function. This study utilizes regulatory DNAs from the gene *even-skipped* in *Drosophila melanogaster* in order to understand the underlying mechanisms governing metazoan transcriptional control. In the remainder of this chapter, I will first review *Drosophila* embryo development. Next, I will outline current understanding of *even-skipped* gene regulation, the advantage of this system for studying transcriptional regulation, non-classical action of *even-skipped* regulatory sequences, and then introduce *even-skipped* enhancer fusions and their transcriptional activities—the transcription model that I have investigated in this dissertation.

1.2 *Drosophila* embryo: model system

Drosophila embryogenesis, from egg deposition to hatching of the first larval stage, takes place within 24 hours at 25°C. In the first three hours after egg deposition (AED), the embryo undergoes 13 rapid mitotic divisions during which nuclei divide almost synchronously. Dramatic rearrangements of embryonic tissue then begin with a phase called gastrulation. During gastrulation, presumptive endoderm and mesoderm tissues invaginate to establish the three germ layers, the endoderm, mesoderm, and ectoderm, of the embryo. This three-layered embryo is called a gastrula.

Prior to the onset of gastrulation, the thirteen nuclear cleavages occur without accompanying cytokinesis forming a multinuclear cell, called a syncytial or, more strictly, a coenocytic embryo. After the ninth nuclear cleavage, most nuclei migrate to the periphery of the egg, creating a hollow ellipsoid of cells, called the blastoderm [16]. Cleavage cycle n is defined as the time between the end of mitosis $n-1$ and the end of mitosis n. The part of cycle 14 which occurs before the onset of gastrulation is called cleavage cycle 14A (C14A). Cleavage cycles 10-14A (covering approximately one and half hours before the onset of gastrulation) are, therefore, called the blastoderm stage of development. Shortly after mitosis 13, nuclei elongate along their basal-apical axis. Subsequently, during the middle of cycle 14A, the cell membrane begins to invaginate between blastoderm nuclei in a process called cellularization. Note that until cellularization is complete, just prior to the onset of gastrulation, proteins can be exchanged between neighboring nuclei throughout the entire blastoderm stage. During late cycle 14A, the embryo undergoes the mid-blastula transition [17, 18], when maternal mRNA and proteins are degraded, and zygotic transcription increases many fold.

During the first three hours before the onset of gastrulation, the segments, the fourteen repeating units of the *Drosophila melanogaster* body plan, are determined by a set of genes called the segmentation genes [19]. Based on their genetic interactions, these genes can be organized into a causal hierarchy of the four levels, the maternal coordinate genes, the gap genes, the pair-rule genes, and the segment-polarity genes [20, 21]. The proteins encoded by the maternal coordinate genes, *bicoid* (*bcd*), *hunchback* (*hb*) and *caudal* (*cad*) are translated from mRNA deposited in the egg by the mother and form monotonic gradients that provide positional information for zygotic genes. The terminal maternal system acts through two genes with gap gene-like activity, *tailless* (*tll*) and *huckebein* (*hkb*). The gap genes, *hb*, *Kruppel* (*Kr*), *knirps* (*kni*) and *giant* (*gt*) are expressed in a broad overlapping domains. Together with maternal coordinate genes, gap genes provide regulatory inputs for pair-rule gene expression. Pair-rule genes, notably *even-skipped eve*, *hairy* (*h*), *runt* (*run*) and *fushi tarazu* (*ftz*), are expressed in overlapping stripes with double segment periodicity, regulating the initial expression of segment-polarity genes, for example, *engrailed* (*en*) in 14 narrow stripes. The segment-polarity genes are expressed in the germ-band after gastrulation and form the segment prepattern. This study focuses on the transcriptional regulation of the pair-rule gene *eve* in the cleavage cycles 13 and 14A of *Drosophila* development.

1.3 Transcriptional regulation of *even-skipped*

Many key parts of our current understanding of enhancers come from studies of the early control region of the *Drosophila* pair-rule gene *eve*, which directs the formation of seven transverse stripes of expression during the blastoderm stage

of embryogenesis (Figure 1.1A and B) [22]. The transcriptional regulation of these stripes is thought to be controlled by a series of separate enhancers [23, 24, 6, 25, 26] in the *eve* promoter (Figure 1.1C). A 16 kb fragment of DNA is capable of rescuing a lethal *eve* allele to viability [27]. This 16 kb fragment is thus very close to constituting the entire *eve* locus. The transcript itself is about 1.5 kb in length and the coding region is located in the middle of the 16 kb fragment. Sequences on the 5' side of the coding region control the expression of stripes 2,3 and 7, while the 3' sequences control the expression of stripes 1,4,5 and 6 [23, 24, 28].

eve is an excellent system for studying the rules of transcriptional control in an integrative manner because key features of its *cis*-regulation have been extensively studied. It is known that the 7 narrow stripes of gene expression, each about 3 nuclei wide, form by the repressive action of gap gene encoded TFs such as Hb, Kr, Kni and Gt, expressed in domains 10-15 nuclei wide [29]. Because gap gene expression domains are wider than *eve* stripes, silencing from these genes would result in a repressed region comparable in size to that of a gap domain and could not produce the observed stripes. Therefore the repressors must act primarily over short distances. *eve* stripes 2 and 3 are particularly informative. It has been shown that stripe 2 is repressed by Kr, but stripe 3 evades repression by peak levels of Kr [30]. Hb, on the other hand, represses stripe 3 while it activates stripe 2 expression [10, 26]. These observations provide stringent mechanistic constraints on transcriptional regulation.

The *eve* stripe 2 enhancer is located between –1.6 and –1.1 kb upstream of the transcriptional start site and the stripe 3 enhancer is located between –3.8 and –3.3 kb. The two enhancers are separated by about 1.5 kb of DNA. Each of those fragments is known as the smallest fragment that directs the expression

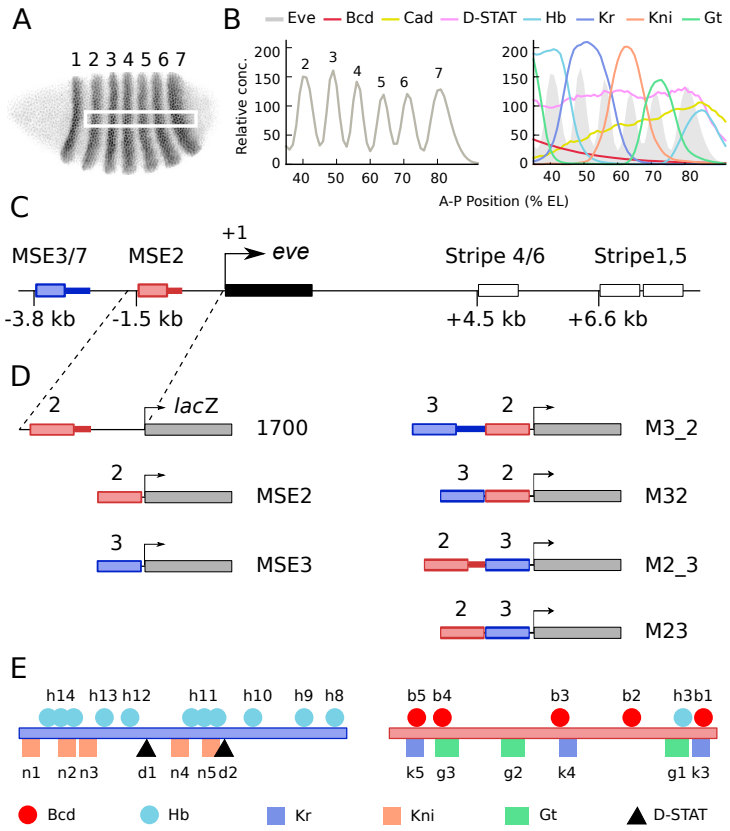


Figure 1.1: ***Even-skipped* gene and reporter constructs.** (A) The 7 striped expression pattern of *eve*, visualized with antibody staining. Here and elsewhere embryos are oriented dorsal up and anterior to the left. The white rectangle located in the middle of the embryo indicates a 10% dorso-ventral strip ranging from 35 to 92% embryo length (EL). (B) (Left) Averaged quantitative expression data for Eve protein, obtained from the area shown in the white rectangle in A. The number in the panel indicates the identity of each *eve* stripe. (Right) Averaged TF expression data. Eve expression is also shown for reference. (C) Schematic view of the *eve* gene. The transcript (black box) and early acting enhancers are shown. The distance of the 5' end of each enhancer from the TSS is specified. The colored boxes and adjacent thick lines indicate the two segments of DNA used to create various reporter constructs. (D) Key reporter constructs studied for the control of *eve* expression. Note that the orientation of the enhancers is the same in all four fusion constructs. (E) Footprinted binding sites identified in MSE3 and MSE2.

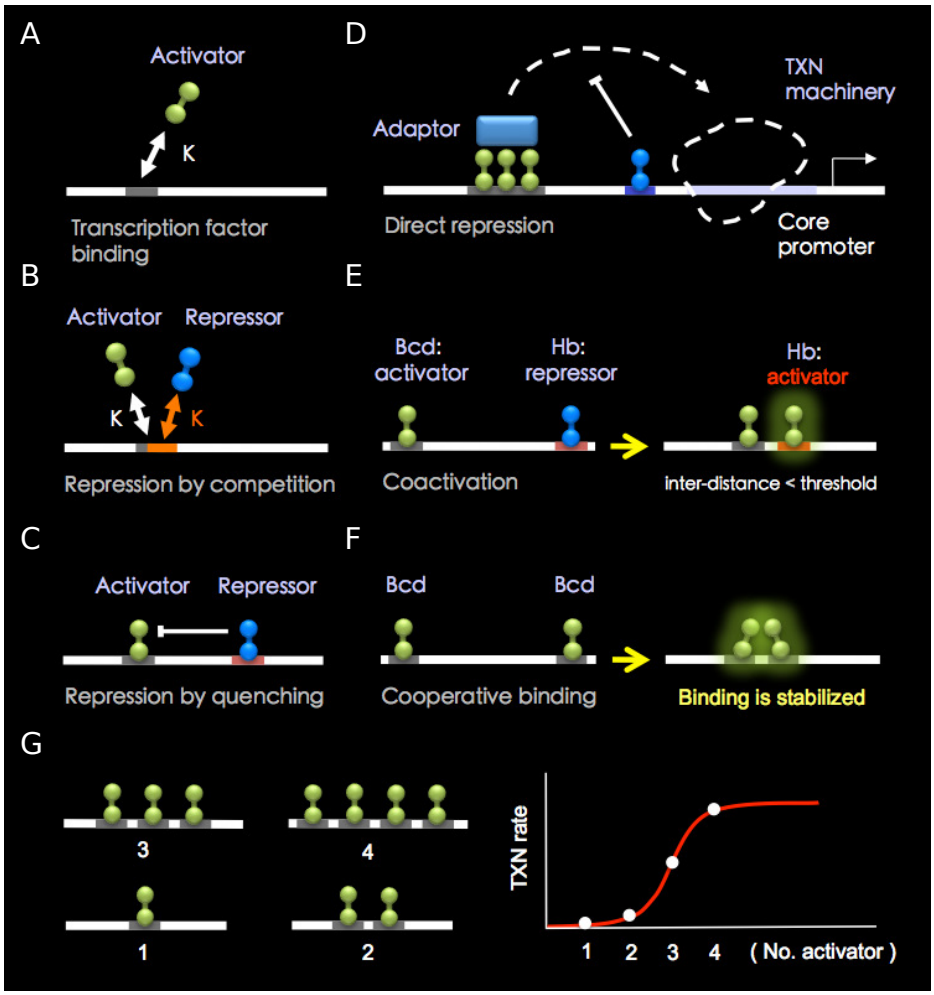


Figure 1.2: **Essential mechanisms for *eve* regulation.** (A) TF binding to DNA. K indicates a binding affinity of a site represented as a grey box. (B) Competition between TFs. Two binding sites are overlapped. (C) Short-range quenching. Repressors can act over a short distance to repress activators' activity. (D) Direct repression. Repressor bound close to TSS prevent adaptor from interacting with transcription machinery. (E) Coactivation of Hb by Bcd. Hb plays as an activator when Bcd bound nearby. (F) Pair-wise Bcd cooperativity. A Bcd binding to DNA is stabilized when two Bcd molecules bound within a certain distance. (G) Activation synergy. There is a greater than multiplicative synergy between activators for a certain range. As the number of bound activators is larger than a certain threshold, the transcription rate approaches the maximum rate.

of *lac*Z in a pattern coextensive with the native stripe 2 and 3, and hence they are called "minimal stripe element 2 and 3" (MSE2 and MSE3)[6]. Note that because MSE3 also drives weak stripe 7 expression, MSE3 is sometimes called the 3/7 enhancer. A combination of DNA binding assays [30], transient cotransfection assays [10], P-element transformation experiments [31, 6, 26, 32, 33, 34] has led to a coherent picture of stripe 2 and 3 regulation.

With respect to stripe 2, several lines of experimental evidence suggest that Bcd and Hb work synergistically to drive stripe 2 expression [10, 6, see Figure 1.2E] and that two repressors, Gt and Kr define the anterior and posterior stripe borders respectively. In addition, Slp1 binds specifically to the $(GTTT)_4$ sequence in MSE2 and MSE3 and the activities of Slp1 is involved in the repressive mechanism of *eve* stripe 2 expression in anterior region [33]. *eve* stripe 3 is thought to be formed by *Drosophila* STAT protein (D-STAT) which is a target of JAK kinase (*hopscotch*) activity [35]. Two footprinted D-STAT binding sites were identified in MSE3. When both sites are mutated, MSE3 no longer drives *eve* stripe 3 expression [36]. However, loss-of-function mutants lacking JAK-STAT components show only a partial loss of *eve* stripe 3 expression [35, 36], suggesting that other factors must be involved in enhancer activation [34]. The anterior and posterior stripe borders are formed by the Hb and Kni repressors [26, 34]. In *kni*– embryos, a complete derepression was seen between stripe 3 and 7 [26]. Furthermore, the mutation of 11 predicted Kni sites including all footprinted Kni sites in MSE3 caused similar derepression [34]. Mutations in four predicted Hb sites in MSE3 led to an anterior derepression of stripe 3 expression [34], suggesting that Hb acts as a repressor that forms the anterior boundary of eve stripe 3. In transient cotransfection assays with a CAT (Chloramphenicol AcetylTransferase) reporter containing

three Bcd and one Hb site, CAT activity increased up to 18 fold compared to background levels with increasing Bcd expression, but only a maximum of two fold with Hb alone. When Hb and Bcd were co-expressed, the CAT reporter drove more than multiplicative activation, with as much as a 44 fold stimulation in CAT activity. These results established the hypothesis that Hb plays as an activator in MSE2 having Bcd sites, but an repressor in MSE3 lacking Bcd sites. The phenomenon, in which a repressor is transformed to an activator by the binding of a coactivator nearby, is called "coactivation" in this study.

1.4 Non-classical action of *eve* stripe 2 and 3 enhancers

Since the *eve* gene was cloned in 1986 [37], investigation of the transcriptional control of *eve* has significantly extended our understanding of metazoan transcription. However, in the course of these studies, certain experimental results, difficult to understand on the basis of the conventional idea of modularized and autonomous enhancers, were obtained. An example of non-classical action of enhancers is the expression of stripe 7 driven by the 1.7 kb proximal *eve* promoter (1700, Figure 1.1D). The proximal promoter contains MSE2, 1.1 kb of proximal sequence and 100 bp additional sequence flanking the 5' end of MSE2. MSE2 drives stripe 2 expression only and the 1.1 kb proximal promoter does not drive any gene expression by itself in the blastoderm embryo. However, the whole 1.7 kb DNA drives both stripe 2 and stripe 7 expression. The generation of stripe 7 gene expression from the combined DNA of the two fragments is not compatible with the concept of additive action of enhancers—

transcriptional activities of enhancers are thought to be independent of each other, thus driving gene expression in an additive manner. Another example of non-classical enhancer action was found in the embryos bearing a homozygous deletion of S2E [2]. The S2E is defined by two short sequences conserved between *Drosophila* species [38]. It contains additional sequences at both ends of MSE2. Despite the absence of S2E, the *EVE*Δ*S2E* homozygotes drive a weak, delayed stripe 2. These findings and the transcriptional activities of *eve* enhancer fusions, which will be described in the following section, provoke fundamental questions about the transcriptional control of metazoan gene expression.

1.5 Enhancer fusions of MSE2 and MSE3

It has been shown that MSE2 and MSE3 can drive normal expression of both stripes if separated by as little as 155 bp (172 bp with polylinker) or 335 bp (360 bp with polylinker) of endogenous DNA 3' of MSE2 or MSE3 respectively, but drive abnormal expression if these DNA fragments are removed [25]. These two sequences were commonly referred to as "spacers" even though they are more than simple neutral spacing elements. In this thesis, I refer to the line bearing a fusion of MSE3 and MSE2 without the "spacer" as M32, with the "spacer" as M3_2, a reverse-order fusion without the "spacer" as M23, and reverse-order with the "spacer" as M2_3 (Figure 1.1D). The M32 fusion is of particular importance because, in the P-element reporter assay, the stripe 2 expression level increases significantly compared with M3_2. In addition to the enhanced level of stripe 2 expression, stripe 3 expression is slightly reduced and the inter-stripe region between stripes 2 and 3 is derepressed

in M32 compared with M3_2, causing a fusion of the two stripes. In M23, there is a severe reduction in stripe 2 expression relative to M2_3. It is worth mentioning that the orientation of the enhancers is the same but the juxtaposed region in each fusion creates a novel arrangement of binding sites.

Two explanations have been proposed to explain the expression phenotype of M32 [25]. First, the increased stripe 2 expression might be caused by coactivation of MSE3 bound Hb by MSE2 bound Bcd. Five footprinted Bcd sites, denoted in a 3' to 5' direction as bcd-1, bcd-2, bcd-3, bcd-4 and bcd-5 were identified on MSE2 and 11 Hb footprinted sites are widely distributed in MSE3 [6]. As described in Section 1.3, Bcd and Hb function synergistically to activate transcription within a limited range [10, 39, Figure 1.2E]. Therefore, it is likely that Hb bound to the MSE3 interact with Bcd bound to the MSE2 to augment stripe 2 expression when the two enhancers are directly coupled. It is known that the elimination of the bcd-1 site, located in the 3' end of MSE2, causes a nearly complete loss in stripe 2 expression when MSE2 alone is assayed in a reporter construct [32]. However, when MSE3 is placed directly upstream of the mutated MSE2, stripe 2 expression is restored.

Second, Kr bound to MSE2 might be able to act over short distances to repress stripe 3 expression driven by MSE3 in M32. There is independent evidence that the repressors which act in MSE2 and MSE3 are "quenchers", which, as described in Section 1.1, stop activators from functioning within a range of about 150 bp [12, 11, 40, 41, 42, 43, 44, 45, Figure 1.2C]. If the two enhancers are separated by "spacer" sequences, stripe 3 evades repression by Kr since MSE3 lacks high-affinity Kr binding sites and MSE2 bound Kr is far from it. It has been shown that a defective MSE2 lacking all three high affinity Kr-binding sites, one at the 5' end, one in the middle and one at the 3' end of

MSE2, fused to the stripe 3 enhancer restores strong stripe 3 expression [25]. This result suggests that Kr bound at the 5' end of MSE2 may act over short distances to repress stripe 3 activators. However, it is also conceivable that the observed stripe 3 restoration is due to the removal of the Kr site located at the 3' end of MSE2. It is known that when quenchers are bound within quenching range of the TSS they can prevent activators from acting at any range, a phenomenon known as direct repression [46, Figure 1.2D]. The Kr site located at the 3' end of MSE2 is within quenching range of the TSS.

In contrast to the M32 fusion, there is no clear hypothesis for the repression of stripe 2 expression in the M23 fusion. Even though three footprinted Kni binding sites, termed kni-1, kni-2 and kni-3, were identified at the 5' end of MSE3, it is not possible to repress the transcriptional activity of the adjacent MSE2 in M23 because Kni is not expressed in the stripe 2 region. Another possibility is that Hb bound to the 3' end of MSE3 represses stripe 2 expression through direct repression. With respect to M3_2 and M2_3, the "spacers" in these two fusions may permit the MSE3 and MSE2 to function independently because the short-range repressors bound to the stripe 2 enhancer work locally to block stripe 2 expression, but are unable to interfere with distantly located stripe 3 activators.

Abnormal gene expression seen in the *eve* enhancer fusions provide stringent mechanistic constraints on transcriptional regulation. This study investigates these *eve* enhancer fusions and their transcriptional activities with a quantitative model approach. In the following sections, I describe the limitations of standard experimental methods and then introduce the quantitative approach I used in this dissertation.

1.6 Limitations of experimental methods

Transcriptional control of metazoan genes has been extensively investigated in the past 30 years. For many developmentally essential TFs, their binding sites, the binding site arrangements in their target regulatory sequences and the transcriptional consequences of their activity have been characterized in deep detail. In addition, recent state of the art technologies generate lots of different types of data at scales ranging from individual base pairs to genome-wide. This technology permits us to systematically measure binding affinities of proteins to DNA, determine site occupancy of TFs *in vivo*, make 3D map of genome interaction (3C and its variants), assay methylation, chromatin marks, and even system-wide RNA levels. Having obtained such an enriched and informative dataset, the critical phenomena of metazoan transcriptional control—the precise control of gene expression—is not well understood.

What is missing from current efforts to gain an understanding of the control of transcription is not *in vivo* TF data but the fundamental understanding of the rules which determine whether a particular configuration of bound factors will activate or repress transcription and to what extent. Because gene expression is orchestrated by simultaneously operating molecular interactions taking place on the regulatory DNA between large number of regulatory proteins including TFs, it is impossible to keep track of such complex interactions by contemporary experimental approaches alone. Furthermore, because many TF concentrations are variable from nuclei to nuclei, the molecular interactions must be assayed at nuclear resolution. Therefore, in order to understand the interplay of multiple transcriptional mechanisms, I utilized a computational modeling approach with quantitative data on TFs and their transcriptional outputs at single nucleus resolution.

1.7 A quantitative model for the transcriptional regulation

To demonstrate an understanding of transcriptional control, it is necessary to be able to calculate the transcriptional response of a segment of DNA to an accuracy comparable to that observed *in vivo*. Such a calculation will involve both the DNA sequence and certain parameters determined by training on data. At the very minimum, given a set of DNA sequences and the expression patterns driven by them, the model should be able to calculate the observed expression patterns with a residual error less than or equal to the likely error of the experimental observations themselves. A statistically significant correlation of the model output with expression data is an inadequate criterion of correctness—a highly correlated pattern is typically sufficiently different from wild type that it would cause death if expressed in a real organism. Beyond this minimal level, a more stringent test is the correct prediction of expression driven by segments of DNA not used for training. Finally, understanding will be demonstrated by performing these calculations of transcriptional output on DNA segments larger than classical enhancers, ideally on an entire locus.

In 2003 Reinitz and Sharp began to address this question by proposing a model of transcriptional control which contains an explicit thermodynamic representation of the occupancies of individual binding sites as a function of the concentrations of the TFs [47]. In a study with Hou and Janssens, this model was applied to the blastoderm of *Drosophila*, a syncytium in which transcriptional control operates at an extremely precise spatial level that approaches cellular resolution. By making use of previously obtained quantitative data on TF levels [48, 29, 49], these authors were able to satisfy not only the minimum

criterion of calculating to within the margins of experimental error in measurements of quantitative gene expression, but also to extend our calculation beyond well-described enhancers to understand how expression of *Drosophila melanogaster eve* stripe 7 was driven by the sequences not present in its "classical" enhancer. While not included in this thesis, I participated in this study by experimentally configuring ectopic expression of this construct in Kni mutant embryo [50, Fig. 5e and f].

Since that time, other modeling studies have been made on certain enhancers with small numbers of binding sites [51, 52, 53, 54, 55]. At a larger scale, Segal and coworkers modeled a set of previously described enhancers in the *Drosophila* segmentation system using the the TF dataset employed in [50] together with *E. coli lac*Z reporter gene expression obtained from the literature and digitized in a binary zero/one manner [56]. A more recent study on this dataset made use of the correlation between data and model output to compare the roles of different transcriptional control mechanisms [57]. In both of these cases the calculation of transcriptional output from known sequences with trainable parameters resulted in expression patterns containing large qualitative errors that would be expected to result in *in vivo* lethality. In addition, these models were only able to fit gene expression of individual enhancers, which makes it impossible to investigate how multiple enhancers act simultaneously in the native context. Finally, these models lack strong analytic ability due to the large qualitative errors in calculation and the inability to monitor protein-DNA and protein-protein interactions taking place between individual TF binding sites.

The central contribution of this dissertation is the introduction of an *in silico* transcription model which is capable of assaying simultaneous molecu-

lar interactions taking place on the regulatory DNA, analyzing transcriptional control at single nucleus, single binding site and even single base pair resolution; and which can predict quantitative levels of gene expression directly from DNA sequence. I augmented the previously published model from our laboratory [47, 50], which represented sequence specific binding of TFs, steric competition between bound factors, activation, short-range repression, and direct repression, by including coactivation and cooperative binding of TFs to DNA. By assembling many multi-channel scanned confocal images of embryos in this embryonic stage, I was able to construct a dataset at cellular resolution in which the concentrations of TFs and the corresponding transcription rate for a given gene or reporter in each blastoderm nucleus are determined to within a relative error of less than 10% [58, 50, 29]. This enables us to treat the *Drosophila* blastoderm as an *in vivo* microarrray in which it is possible to perform many transcription assays in parallel. These assays were performed on genes in a native chromosomal context in cells with well defined concentrations of TFs that produce markedly different transcriptional outputs from relatively small changes in TF concentration, resulting in an assay system of sensitivity and reproducibility unmatched by any tissue culture system I am aware of. I then challenged this assay system with a family of seven carefully selected rearrangements of two early acting enhancers, MSE2 and MSE3, and their native flanking sequences of the *Drosophila eve* locus. Each rearrangement drives a different expression pattern, and the most informative patterns driven by the *eve* enhancer fusions, M3_2, M32, M2_3 and M23 were quantitatively compared by transforming all constructs to a common chromosomal site and quantitatively assaying reporter expression together with the levels of nine TFs—Bcd, Cad, D-STAT, Dichaete, Hb, Kr, Kni, Gt and Tll.

1.8 Dissertation overview

This dissertation presents a theoretical model that reconstitutes *eve* transcriptional control *in silico* by implementing the molecular regulatory mechanisms that are essential for the expression of *eve*. The model is then applied to *eve* enhancer fusions in order to elucidate the common rules governing the transcriptional control of *eve* as well as the *Drosophila* genome. Chapter 2 presents precise gene expression data from four *eve* enhancer fusions at single nucleus resolution. The generation of site-specific transgenic fly lines, RNA and protein imaging, image processing, quantitation, RNA expression levels of the fusion constructs and D-STAT protein expression are described. Chapter 3 provides a detailed description of the methodology employed to characterize the transcription factor binding information directly from DNA sequence and the formulation of the *in silico* transcription system. Subsequent refinement of the model and final modeling results are described. In Chapter 4, the validation of the model, including two stringent tests—model fitting and analysis of the prediction of gene expression that was not fitted in the model—is presented. Chapter 5 presents the results of functional analysis of the *eve* enhancer fusions. In Chapter 6, the mechanisms governing conserved gene expression from S2Es from four *Drosophila* species, *D. melanogaster*, *D. yakuba*, *D. erecta* and *D. pseudoobscura* and two Sepsid species, *Sepsis cynipsea* and *Themira putris* are extensively investigated using the *in silico* transcription system. Finally, Chapter 7 summarizes the findings of this dissertation and concludes with prospective avenues of further inquiry.

Chapter 2

Quantitative gene expression data at single nucleus resolution

2.1 Transgenics, imaging and image processing

Transcriptional output is a quantitative entity. In order to understand the precise quantitative changes in gene expression, the underlying mechanisms must be inferred by a quantitative transcription model against concentrations of TFs and their transcriptional outputs. A critical requirement for testing the rules is the availability of directly comparable quantitative gene expression data. The transcriptional rules must be applicable to any regulatory sequence in the genome, therefore, training the models against quantitative gene expression of multiple regulatory sequences is key to characterize the universal rules. Another requirement is that, because nuclei are the fundamental units of transcriptional processing, these data must be at nuclear resolution.

To acquire quantitative gene expression fulfilling these requirements, I placed the four fusion reporter constructs M3_2, M32, M2_3 and M23 in a

targeted integration site located in the fly genome. Then I performed fluorescence *in situ* hybridization (FISH) in order to visualize the RNA levels driven by the fusion constructs. Using confocally scanned images of the stained embryos and published image processing procedures, I measured RNA expression of these constructs at cellular resolution [58, 59, 50, 60].

2.1.1 Generating site-specific transformant lines

The P-element mediated transgenesis technique has been extensively used for measuring transcriptional output of an enhancer by integrating reporter constructs containing enhancers into the fly genome. However, this technique is not suited for comparing the quantitative gene expression of multiple enhancers. P-element recombinase integrates the reporter construct at a random position in the chromosome, hence the position effect can affect the transcriptional activity of the reporter construct. Therefore, in order to compare gene expression between the reporter constructs of interest, it is necessary to integrate the reporter constructs into a specific location in the genome, where the reporters are capable of driving gene expression.

I utilized the Cre RMCE (Recombinase Mediated Cassette Exchange) technique [61] to create transformant lines carrying the four fusion constructs, M32, M3_2, M23 and M2_3, at the same position in the genome. The reporter construct is called a 'cassette' because it lies between two recombinase recognition sites, *lox-P* and *lox2272* in the RMCE vector so that the bacterial sequence used for the molecular cloning of the vector doesn't accompany the reporter in the genome. Absence of the bacterial sequence in the integrated reporter constructs removes potential interference.

The M32, M3_2, M23, and M2_3 transformant lines were generated by

excising the *Eco*RI-*Xba*I fragments from four *eve-lac*Z pCaSpeR plasmids [25] and ligating them into the RMCE vector pBS(KS+)-*lox-white-lox2272* [61] cut with *Eco*RI and *Spe*I. Each *Eco*RI-*Xba*I fragment contained an *eve* enhancer fragment fused with the basal *eve* promoter (from –42 bp) and the intact 100 bp untranslated leader and the first 22 codons of the *eve* gene fused with *lac*Z as described [25]. The M32 *eve-lac*Z pCaSpeR plasmid contains an additional *Eco*RI site between MSE3 and MSE2. In this case, the *Eco*R1-*Xba*I fragment was first ligated into the vector, and then after transformation and amplification of the product the *Eco*RI-*Eco*RI fragment containing MSE3 was cloned into the RMCE vector after digestion with *Eco*RI. The correct orientation of MSE3 in the RMCE vector was confirmed by DNA sequencing. The pCaSpeR vectors and the RMCE vector were gifts of Stephen Small.

Transformant flies were generated by microinjection of the RMCE plasmid and Cre expression vector into the embryos of line A13 [61]. This fly line contains a landing site near 96F on chromosome III. Surviving flies were crossed to *y w* and progeny were screened for exchange events, scoring for the loss of *y* and gain of *w*. Recombination events were characterized by PCR amplification of the exchange junctions. PCR characterization of recombination events was carried out using the primers land-1 (5'-TCCGTGGGGTTTGAATTAAC-3', specific to the 5' end of landing site sequence) and cassette-1 (5'-GGCAG TTAGTTGTTGACTGTG-3', specific to the 5' end of transcript sequence in the reporter cassette) and yielded a positive product of approximately 1300 bp to 1600 bp, depending on the length of the regulatory DNA in the cassette.

2.1.2 RNA and protein imaging

To measure *lac*Z expression driven by the four fusion constructs as well as the native *eve* pattern, the transformed embryos were dechorionated—a thick egg shell, called chorion, of the embryos is removed—fixed and then hybridized with a Fluorescein (FITC)-labeled *lac*Z anti-sense RNA probe (Figure 2.1A). After hybridization, *lac*Z mRNA is visualized by sequential incubation with rabbit anti-FITC antibody, followed by fluorescently labeled anti-rabbit antibody. The embryos are simultaneously incubated with guinea pig anti-Eve (primary antibody) and a fluorescently labeled anti-guinea pig antibody (secondary antibody) to detect endogenous Eve protein (Figure 2.1A). After antibody incubation, the nuclei are stained using the dye PicoGreen.

The fluorescently stained embryo images were scanned in a confocal microscope. To generate mutually comparable confocal images between different transformant embryos, I compared all available constructs with respect to image intensity and identified a line, M32_attB, with maximal image intensity. Image intensity for each protein and *lac*Z RNA was standardized against M32_attB by setting the gain of the microscope so that the brightest pixels saturate the eight bit photon detector. The standard line was included in all the experiments done for this study to set the intensity standard. These measures ensure that the level of gene expression of M3_2, M32, M2_3 and M23 constructs is directly comparable.

2.1.3 Image processing

The confocal images were reduced to a text file containing a list of nuclei, their coordinates, and the average fluorescence levels of *eve* protein and *lac*Z mRNA

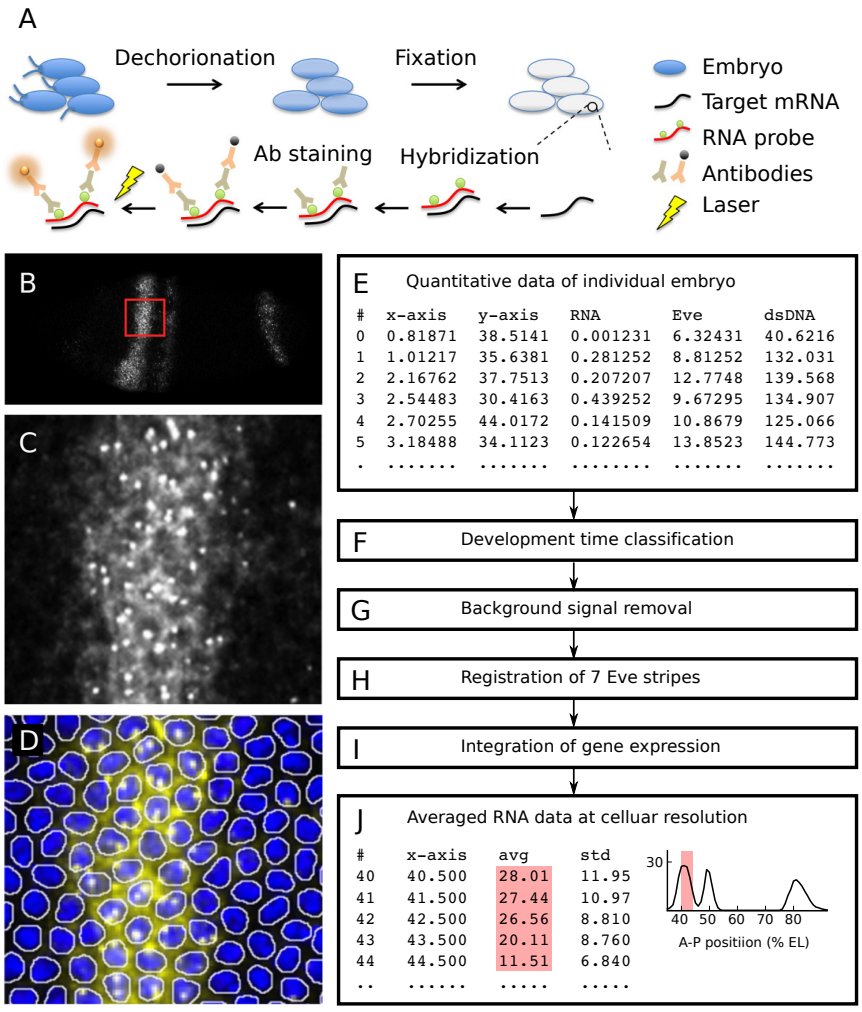


Figure 2.1: **FISH and image processing.** (A) Schematic view of fluorescence in situ hybridization procedure. The outermost shell around the embryo, called chorion, was removed before the following staining process. Black circles attached to the antibody are fluorescent molecules. (B-E) Image segmentation. (B) Reporter driven mRNA is visualized with FISH. (C) Magnified image from red square box shown in (B). (D) Segmented image with nuclear mask. Intense and punctate fluorescent spots in the nuclei are nascent transcripts. (E) Segmented data. Fluorescent intensity inside each nucleus is averaged and x and y coordinate indicate the position of the centroid of each nucleus. (F-I) Data processing. See text for details. (J) Final output of data processing. Averaged RNA data with corresponding nuclear coordinates are generated.

in each nucleus. In order to obtain the average fluorescence levels, individual nuclei must be identified. Use of morphological watershed segmentation technique permitted the detection of the nuclei in two dimensional (2D) confocal images (Figure 2.1B-D). The pixel intensities of each segmented nucleus were then averaged. (**1. Data segmentation**, Figure 2.1E). Next, background staining of RNA signal were subtracted as described [59]. The smoothed data at this step is only used for estimating background which is obtained by finding individual non-expressing nuclei (**2. Background removal**, Figure 2.1F). Then, embryos were classified temporally as belonging to either cleavage cycle 13 (C13), or one of eight time classes (T1-T8) based on the Eve protein pattern and differential interference contrast (DIC) membrane images, each about 6.5 minutes long, in cycle 14 (C14) before gastrulation. (**3. Time classification**, Figure 2.1F) [62, 29]. **4. Registration** was performed by aligning Eve expression domains of individual embryo to the Eve dataset, available in the Flyex database [49], using the wavelet method (Figure 2.1H) [62]. This step is critical for generating relevant averaged data of Eve and *lac*Z expression driven by reporter constructs because a positional variation of expression patterns between embryos makes it difficult to correctly superimpose expression patterns for averaging. Finally, all data for specific time classes is averaged by collecting intensities from individual embryos in one hundred bins according to the Anterior-Posterior (AP) position (**5. Integration**, Figure 2.1I). Data from the middle 10% of dorso-ventral positional values of each embryo was averaged for each time class. These data provide the relative expression levels of the reporter and eight TFs to 5-10% relative accuracy in each nucleus [29]. The quantitative model used in this study and all following data analysis use the 1D data from 35% to 92% AP.

2.1.4 Abolition of position effect

To confirm that the gene expression obtained from the site-specific transformants is directly comparable, I compared gene expression in two independently generated transformants for the same construct. I first quantified *lac*Z expression level in the two independent P-element lines containing MSE2, also called 1511 to indicate the –1.5 kb to –1.1 kb region of *eve*. The 1511B line contains the 1511 construct on the 3rd chromosome and the 1511C line contains the 1511 construct on the 2nd chromosome. Overall, the 1511B line shows higher expression levels than 1511C (Figure 2.2A). I then also quantified expression of the two site-specific M32A and M32B lines bearing the reporter at the same integration site on the second chromosome (Figure 2.2B). The expression levels of M32A and M32B in stripe 2 region are indistinguishable. This result demonstrates that site-specific transgenesis, in which reporter constructs are integrated at the same genomic position, permits precise comparisons between multiple transgenic constructs by eliminating position effect.

2.2 Quantitative RNA expression of fusion constructs

2.2.1 M3_2 and M32 gene expression

Quantification of gene expression of M3_2 and M32 at cellular resolution reveals novel quantitative features and dynamics. In both constructs, overall gene expression increases until T5 and stripes 2,3 and 7 begin forming at T3. In M3_2, stripes 2 and 3 are completely resolved at T6 with almost equal expression levels at the stripe peaks whereas in M32 the two stripes do not

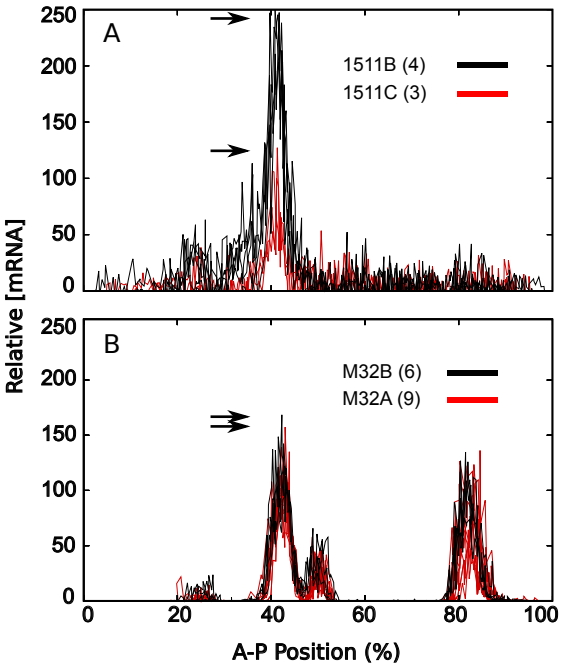


Figure 2.2: **Position effect on reporter construct expression.** Middle 10% expression data with background removed superimposed from multiple embryos bearing P-element transformed or RMCE transformed reporters. The number of embryos used to generate the expression data shown is given in parentheses in each key, and black arrows indicate the maximum expression level found in each construct. (A) Expression of MSE2 in the two independently established P-element lines, 1511B and 1511A [6]. 1511B bears a reporter construct on the second chromosome and 1511C bears the same construct on the third chromosome. (B) Expression of two M32 RMCE transformed M32A and M32B lines bearing the reporter at the same integration site on the second chromosome. The expression levels of M32A and M32B are indistinguishable.

become resolved (Figure 2.3). Stripe 2 expression in M3_2 starts declining after T5 while stripe 2 expression in M32 declines after T6. Like stripe 2, stripe 7 expression level declines in M3_2 after T5. However, Stripe 7 expression maintains its level after T6. A striking feature of M3_2 and M32 data is that stripe 2 expression shifts posteriorly from T6 to T8 in contrast to wild type Eve stripe 2, which moves anteriorly [29]. When comparing gene expression levels, four major differences are observed in M3_2 and M32. In M32, stripe 2 is upregulated by a factor of 3.5 compared with M3_2. In addition to the enhanced level of stripe 2 expression, the inter-stripe region between stripes 2 and 3 is derepressed in M32 compared with M3_2, causing a fusion of the two stripes. One interesting observation which has not been reported previously is that peak stripe 7 expression in M32 is twice that of M3_2. The positions of the peaks of stripes 2 and 7 are exactly same in M32 and M3_2. In contrast to stripes 2 and 7, M32 has 70% of M3_2 stripe 3 expression.

2.2.2 M2_3 and M23 gene expression

When the order of enhancers in the M32 construct is reversed with and without 160 bp "spacer" (see section 1.5) in between, completely different patterns in gene expression are observed (Figure 2.4). In both M2_3 and M23 transformants, stripe 3 expression is much higher than stripe 2 expression after T5. In M23, stripe 3 expression gradually increases until T7 and stabilizes thereafter. However in M2_3, stripe 3 expression declines after T7. Stripe 2 is not expressed in M23 until T4, and is expressed at low levels subsequently. Compared to M3_2 and M32, stripe 2 expression in M2_3 is weaker but has the same dynamics. One of the noticeable features of the M23 data is that stripe 3 forms at T3 while stripe 2 forms at T5. This delayed stripe 2 formation is

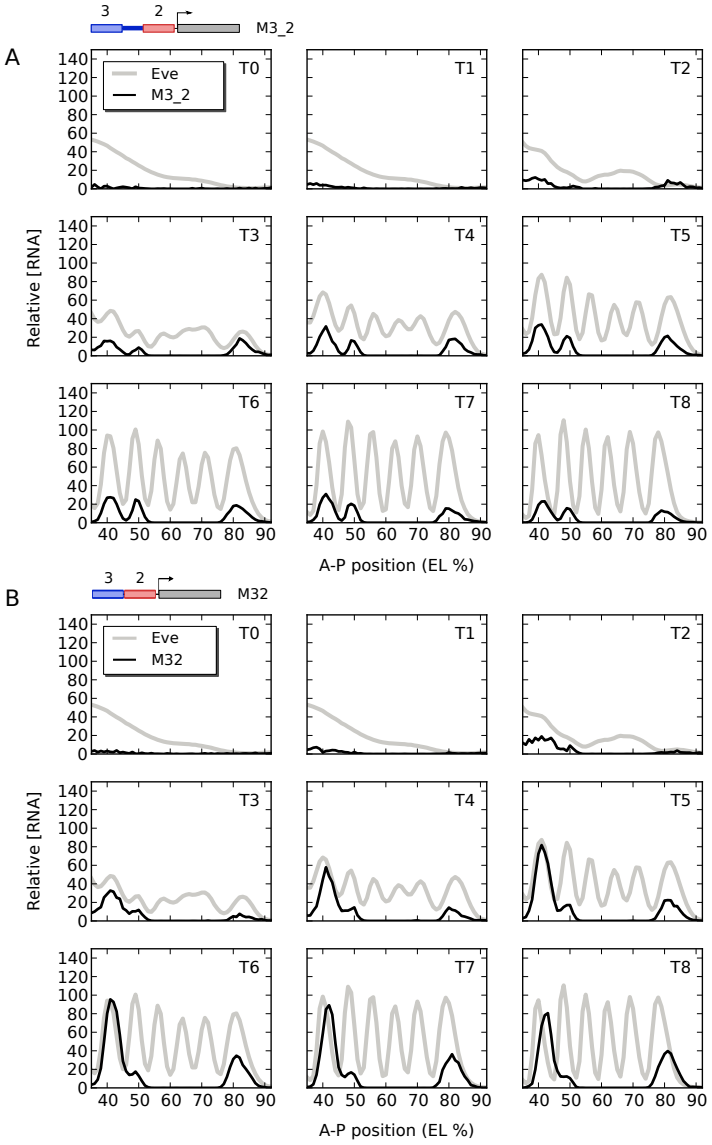


Figure 2.3: **Quantitative gene expression of M3_2 and M32.** (A) Averaged gene expression driven by the M3_2 reporter construct in *Drosophila melanogaster* embryos. (B) Averaged gene expression driven by M32. Gray and black lines indicate Eve protein and *lac*Z RNA respectively. Embryos were classified as belonging to one of cleavage cycle 13 (T0) or eight time classes (T1-T8) in cleavage cycle 14A (C14A), each except T0 about 6.5 min long.

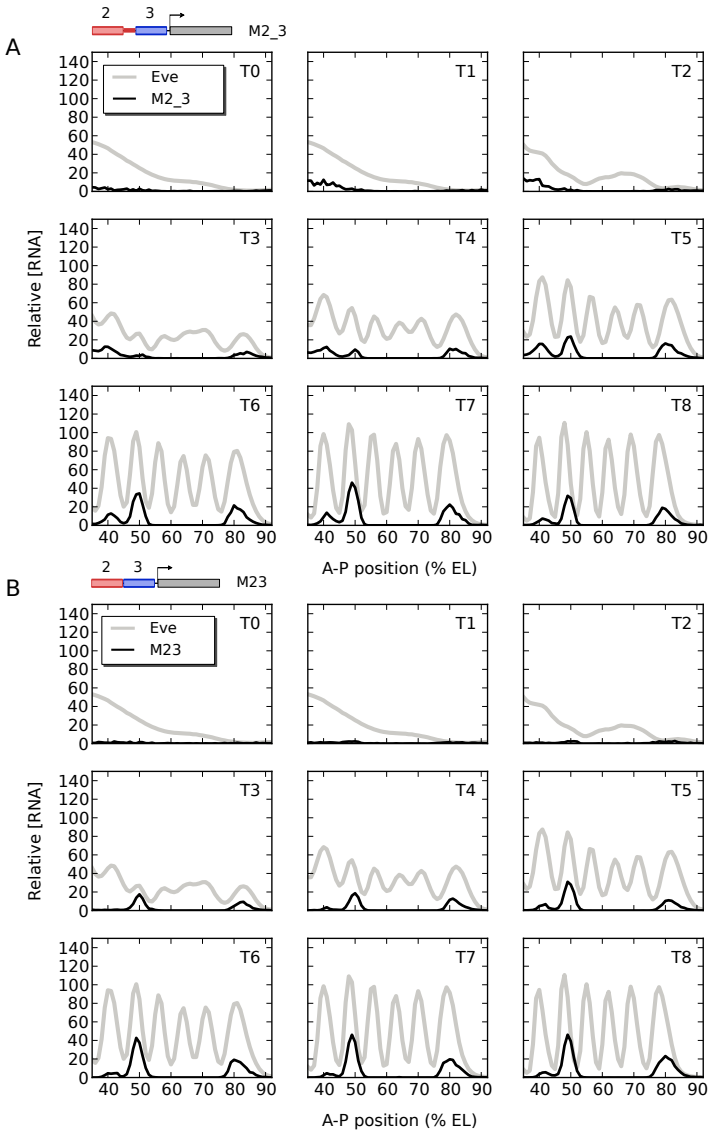


Figure 2.4: **Quantitative gene expression of M2_3 and M23.** (A) Averaged gene expression driven by M2_3 reporter construct. (B) Averaged gene expression driven by M23. Other notation is the same as Figure 2.3.

unusual because in early C14, broad stripe 2 expression is generated first in the other fusion constructs. Unlike the M3_2 and M32 fusions, the stripe 2 peak is not shifted to posterior in either M2_3 or M23. Two significant changes in gene expression are also observed between M2_3 and M23. In M23, stripe 2 expression is decreased by a factor of 0.2 at T6 and stripe 3 expression is upregulated in M23 compared with M2_3.

2.2.3 Quantitative data analysis

Rearrangements of stripes 2 and 3 enhancers create significant changes in gene expression. First, all four fusion enhancers do not drive the early broad expression seen in the native *eve* gene [29] and 1.7 kb proximal *eve* promoter (1700 promoter) [6, 50] (see T1 expression in Figure 2.5A). Second, overall expression levels of the four constructs decline after time classes 6 through 8 (see T6 and T8 in Figure 2.5A). Once the level of gene expression in each stripes' peak reaches maximum, all of them begin declining except M23. In the case of M3_2 and M2_3 at T7, a slight increase in gene expression is observed in the peak stripe 2 region, however it is obvious that the gene expression in T8 significantly drops compared with T6. Stripe 2 expression in M23 is too weak to compare to the expression of other stripes (Figure 2.5A). It is worth noting that the time when stripes 2, 3 and 7 reaches maximal expression is stripe and construct dependent. The growth curves of expression at the stripe peaks are also different between constructs (Table 2.1).

The most dramatic difference in gene expression among the four fusion constructs is observed in the stripe 2 region of M23 and M32. Simply placing the stripe 2 and 3 enhancers in reverse order without a "spacer" increases gene expression up to twenty one times at time class T6 (Table 2.1). Another

	M3_2	M32	M2_3	M23
Peak stripe 2 at T6 (v^{fl})	27.44	95.46	12.8	4.55
Relative ratio to peak the minimum	6x	21x	2.5x	1x
Peak stripe 3 at T6	25.02	17.57	33.43	42.72
Relative ratio to peak the minimum	1.4x	1x	2x	2.4x
Peak stripe 3 / stripe 2	1x	0.2x	2.7x	9.4x
Peak stripe 2 position at T6 (% EL)	41%	41%	41%	41%
Peak stripe 2 position at T7	41%	42%	41%	41%
Peak stripe 2 position at T8	42%	43%	41%	42%
highest expression level time class	T5	T6	T5	T5
Decline after reaching maximum	Yes*	Yes	Yes*	Yes*
Peak stripe 3 position at T6	49%	49%	50%	49%
Peak stripe 3 position at T7	49%	49%	49%	49%
Peak stripe 3 position at T8	49%	49%	49%	49%
highest expression level time class	T6	T6	T7	T8
Decline after reaching maximum	Yes	Yes	Yes	–
Peak stripe 7 position at T6	81%	81%	80%	80%
Peak stripe 7 position at T7	79%	81%	80%	80%
Peak stripe 7 position at T8	79%	81%	79%	80%
highest expression level time class	T5	T8	T7	T8
Decline after reaching maximum	Yes	–	Yes	–

Table 2.1: **Summary of quantitative data analysis.**

* Small fluctuations in gene expression, 10%, 5% and 15% of the maximum peak level respectively, are observed.

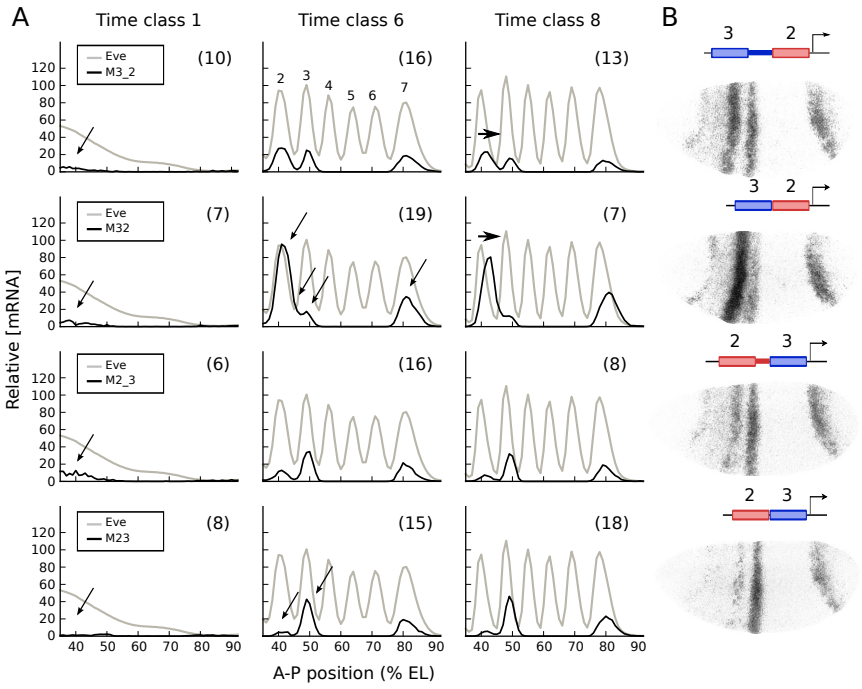


Figure 2.5: **Integrated expression data from four fusions.** (A) Quantitative expression data for Eve protein and 4 fusion constructs, obtained from the area shown in the white rectangle in B. T1, T6 and T8 data are shown here. The numbers in parentheses are the number of embryos used to generate the averaged expression profiles of each time class. Arrows indicate regions of major alteration in gene expression after spacer removal. (B) *lac*Z mRNA expression from individual embryos. 4 fusion constructs and their gene expression at T6 are shown.

dramatic change in gene expression occurs in between M3_2 and M32. Removal of the 350 bp "spacer" between MSE3 and MSE2 increases level of gene expression by a factor of 3.5 compared with M3_2. Despite the significant changes in gene expression, the peak of stripe 2 is located at exactly same position, 41% EL, at T6. This observation raises an interesting question about the robustness of the peak stripe 2 position.

Temporal analysis of gene expression reveals a dynamic shift of the peak

position in stripe 2 expression. During the time between T6 and T8, the position of the peak of stripe 2 shifts to the posterior except in the M2_3 construct while wild type Eve stripes are shifting to the anterior (Figure 2.5A). Shifting peak position is most significant in M32 (Table 2.1). On the other hand, the position of peak stripe 3 remains constant except in M2_3 and the position of the peak of stripe 7 remains constant except in M3_2 and M2_3.

2.3 Quantitative D-STAT protein expression

In previous work, quantitative data of seven *trans*-acting factors—Bcd, Cad, Hb, Kr, Kni, Gt and Tll, from our Flyex database [49]—was required to model stripe 2 and weak stripe 7 expression [50]. However, as described in Section 1.5, the model of the four fusions, M3_2, M32, M2_3 and M23 requires the concentration of an additional TF, the *Drosophila* homolog for mammalian STAT (Signal Transducer and Activator of Transcription), also called D-STAT92E (D-STAT at 92E region of chromosome 3) or just D-STAT. It is well known that the Hop-D-STAT pathway is required for optimal expression of stripe 3 [63, 26]. The minimal stripe 3 element (MSE3) contains two footprinted D-STAT binding sites, which are surrounded by five Kni and eleven Hb footprinted binding sites. Small and colleagues suggested that stripes 3 and 7 are formed by one or more ubiquitously distributed activators including D-STAT and the anterior and posterior borders of stripe 3 are established by the Hb and Kni repressors, respectively [26].

In order to obtain the D-STAT expression data, I utilized anti-D-STAT antibody (a gift from Dr. Erika Bach) to measure D-STAT expression in the blastoderm stage of *Drosophila* embryo. It has been reported that D-STAT

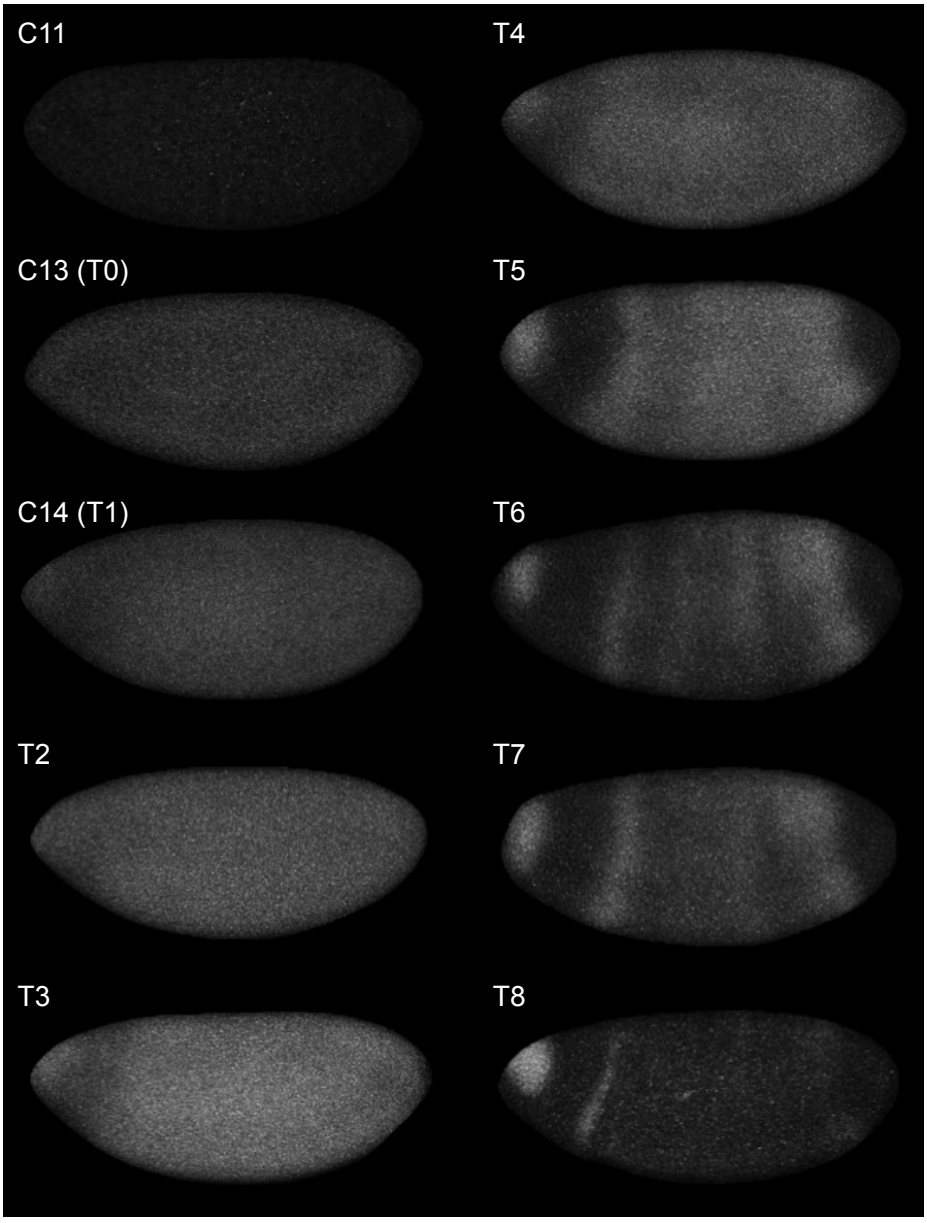


Figure 2.6: **Development of D-STAT expression.** D-STAT protein is visualized with anti-D-STAT antibody staining. Embryos shown are laterally aligned. D-STAT is not detected at C11 but is expressed ubiquitously at C13 and early C14A.

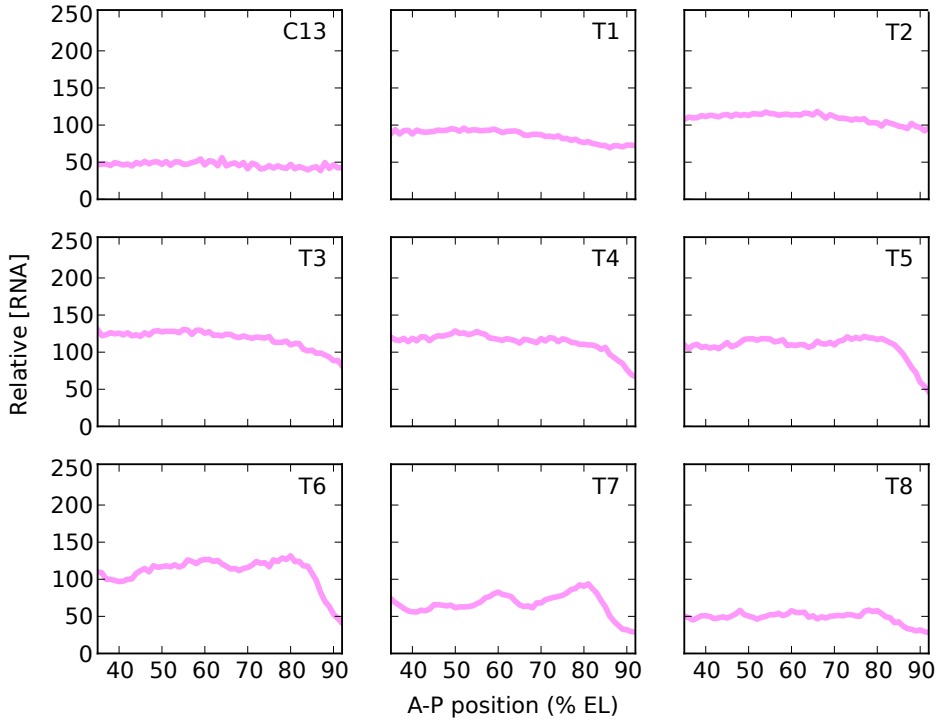


Figure 2.7: **Quantitative gene expression of D-STAT.** 6, 7, 8, 8, 7, 5, 9, 10 and 9 embryos were used to generate the averaged expression profiles of C13 and C14 from T1 to T8 respectively.

is activated by JAK mediated tyrosine phosphorylation and only activated D-STAT forms a dimer and translocates to the nucleus [64]. Even though the antiserum does not specifically binds to tyrosine phosphorylated D-STAT, I was able to obtained approximate functional D-STAT expression data because our standard segmentation method restricts the quantitative measurement of expression levels to the nuclei of a *Drosophila* embryo.

It is known that D-STAT mRNA is present in the egg at the time of fertilization and is translated early in development [35]. At the cellular blastoderm stage, a weak seven-stripe mRNA pattern is seen, with the weakest staining at

stripe 5. After gastrulation, a 14 striped segment polarity pattern is seen [36]. However, quantitative D-STAT protein data reveals that unlike its mRNA pattern, the D-STAT protein has a four striped expression pattern with no clear inter-stripes (Figure 2.6). Expression levels gradually increase from C13 until T3 and start decreasing thereafter (Figure 2.7). The level of D-STAT expression in the presumptive germ band significantly drops at T7 and is reduced below C13 levels at T8. In contrast, D-STAT expression at the anterior pole starts to increase after T4 and reaches maximal expression at T8. On the other hand, the expression in the posterior tip is strongly downregulated after T4. There is an additional domain of repression located between the anterior tip and the four striped expression. Expression in the approximate location of *eve* stripe 1 gets narrower after T6 and forms clear anterior and posterior borders. By T8, D-STAT is only expressed at the anterior pole and in a stripe around 35% EL.

Chapter 3

Construction of an *in silico* transcription system

Quantitative gene expression data of the four fusion genes, in which the expression levels are dynamically changing in several hundred nuclei, provides rich and highly informative transcriptional cues with which to characterize the relationship between transcription input and its output. I constructed an *in silico* transcription system as an assay tool to quantitatively measure the molecular interactions taking place on the regulatory DNA and the rules governing the transcriptional control. The *in silico* transcription system takes three inputs: regulatory sequences of interest, quantitative protein concentration profiles of TFs that are known as regulators and binding sequence information to predict the binding site position of the TFs and their binding affinity. Construction of the *in silico* transcription system is achieved in the following steps. (1) Formulation of a mathematical modeling framework. (2) Fitting the model to gene expression data to obtain the parameters. (3) Comparison between model-calculated gene expression and observed gene expression. (4) Parame-

ter analysis. (5) Prediction of gene expression driven by regulatory sequences that were not used in the training set. (6) If the visual inspection of the result of the model and observed RNA data or parameter analysis or prediction results indicate missing mechanism(s), add the mechanism(s) and go to step (2). (7) Functional analysis of the regulatory sequences with the resulting transcription system.

I employ a theoretical model that is intermediate between a content-based picture in which only the number of binding sites for each factor in an enhancer is significant [65], and, on the other hand, a grammar-based approach in which a precise arrangement of binding sites is required for regulatory function [66]. In our model, the physical arrangement of binding sites is quite important, but the rules are sufficiently flexible to permit many solutions, reflecting the observed variability in binding site arrangement. Four design principles guide the formulation of the model. First, a minimal set of regulatory mechanisms that are essential for the transcriptional control of the *eve* stripes 2, 3 and 7 are determined and implemented numerically. Second, I construct the model in such a way that the mechanisms operate simultaneously. Third, the mechanisms are nonetheless separable and removable so that the relative contributions of each mechanism can be visualized as can the consequences of removing a specific mechanism *in silico*. Fourth, I perform a full thermodynamics calculation to find the fractional occupancy of each binding site. Dynamic programming approaches are more computationally efficient but calculate summed fractional occupancies [56, 57]. Calculating with the the fractional occupancies of individual binding sites rather than their sum allows us to determine the contribution of each TF, binding site, and even nucleotide to gene expression.

3.1 Determination of missing essential mechanisms

The theoretical model was refined by a series of *in silico* experiments. Taking the elegant work of Janssens and colleagues [50] as a starting point, I extended the model by adding essential mechanisms for *eve* regulation to meet the minimum experimental results. In this section, I describe the determination of the newly added molecular mechanisms and their contribution to model fitting. I then describe the complete *in silico* transcription system, its fitting method and the modeling results in the following sections.

3.1.1 Coactivation of Hb by Bcd

Hb bound to MSE2 is known to behave as an activator [10, 6, 32]. The fact that mutation of the Hb binding site in MSE2 dramatically reduced stripe 2 expression and other experimental results support that idea. However, it is also known that Hb acts as a repressor when it binds to MSE3 [25, 26]. Because the four fusion constructs contain both MSE2 and MSE3, I implemented the Bcd mediated coactivation of Hb [10, 6, 25, 67] in the initial *in silico* transcription model.

After the coactivation mechanism of Hb by Bcd is added, the model was still not able to correctly calculate the observed RNA data of M32 (Figure 3.1A). When MSE3 and MSE2 are fused in M32, two footprinted Bcd binding sites located at the 5' end of MSE2 become close to the footprinted Hb binding sites at the 3' end of MSE3. As described in Section 1.5, it had been proposed that the Hb bound to the sites in MSE3 might synergistically activate transcription by MSE2 bound Bcd mediated coactivation [25]. However, it is known that

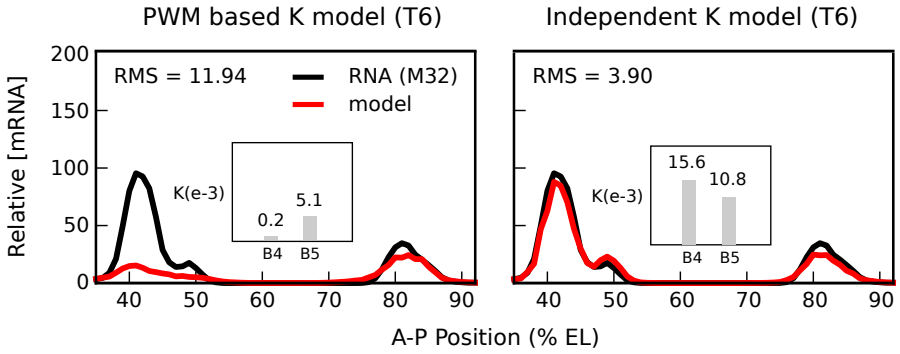


Figure 3.1: **Model suggests Bcd cooperativity.** PWM score based K model utilizes PWMs for predicting binding affinities of all TFs used in the model. Independent K model freely determines the binding affinities of Bcd sites during fitting the observed data. The boxes in the two panels show the binding affinity differences between PWM score based K and independent K models. The PWM score based K model fails to recapitulate the increased stripe 2 expression driven by M32. The independent K model prefers to have strong binding affinities for the two Bcd sites, bcd-4 and bcd-5 to recapitulate the increased stripe 2 expression accurately.

the binding affinities of the two Bcd binding sites, bcd-4 and bcd-5, are weak compared to the binding affinity of the bcd-1 site located at the 3' end of MSE2 [6]. I reasoned that Bcd might bind cooperatively on the bcd-4 and bcd-5 sites (Figure 5.4) *in vivo* so that the Bcd bound to the bcd-4 and 5 is sufficient to drive the strong synergistic activation with the Hb bound to the sites in MSE3. There are several lines of evidence supporting the distance dependent pairwise cooperativity between two Bcd proteins on DNA [6, 7, 8, 9]. I first tested the possibility of cooperative interaction by modeling the M3_2 and M32 gene expression with free Bcd binding affinity parameters. In this model, the binding affinity of Bcd sites are not calculated using the corresponding position weight matrix (PWM), a standard method of a binding affinity calculation used in this study (see Section 3.2.1 for details). Instead the model freely determined the binding affinity K for each site independently during the model training.

The independent K model was able to successfully recapitulate the increased stripe 2 expression of M32 and, as expected, a strong binding affinity for bcd-4 and bcd-5 site was observed in the model (Figure 3.1B).

3.1.2 Cooperative binding of Bcd

With this result and the experimental evidences, I implemented pairwise cooperativity of Bcd in the model with the following algorithm for finding cooperative pairs:

1. Find the strongest Bcd binding site in a given construct.
2. Find the strongest remaining Bcd sites within the cooperativity range.
3. Pair the Bcd binding sites if exists.
4. Find the strongest Bcd binding site not yet considered or paired.
5. Go to step 2 until all Bcd sites are considered or paired.

The pairing algorithm is inspired by Johnson and Burz's work [68, 9]. Johnson showed that repressors bound to the three operator sites, O_R1, O_R2 and O_R3, located in the P_{RM}/P_R promoter region in the λ phage genome, interact only in a pair-wise manner [68]. The experiment demonstrated that repressor bound to O_R2, the second strongest binding affinity site, always interacts with the repressor at O_R1, the strongest binding affinity site, instead of interacting with O_R3 under DNaseI footprint assay [68]. When the highest binding affinity site O_R1 and adjacent site O_R2 are intact, fractional occupancies of both sites were increased. However, the repressor occupancy at O_R3 which is to the left of O_R2 was exactly the same as that of DNA containing O_R3 alone. Nevertheless, if O_R1 is inactivated by a mutation, O_R2 instead coop-

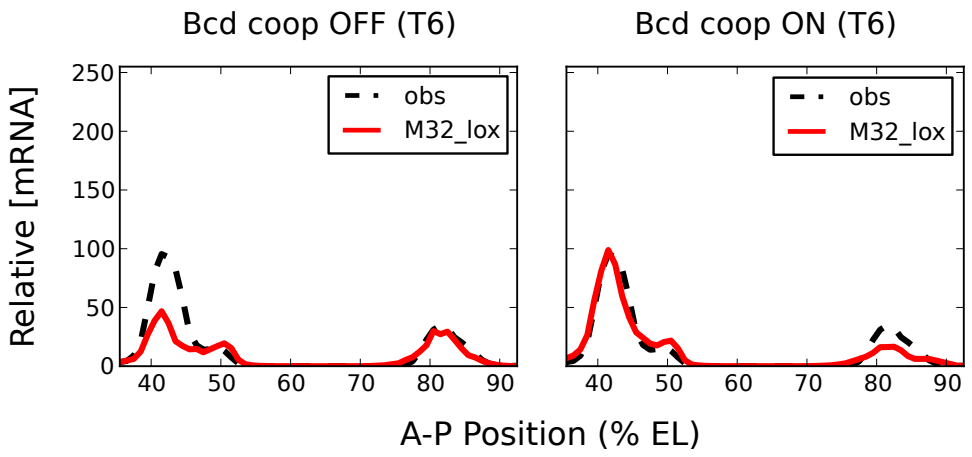


Figure 3.2: **Addition of Bcd cooperative binding.** Implementation of pair-wise Bcd cooperativity improves fitting quality of the model. Note that stripe 7 expression was not completely recapitulated when the Bcd cooperativity is added in the model.

erates with O_R3. These experiments taken together strongly support the five step algorithm given above.

In addition to λ repressor, independent experiments have indeed demonstrated pairwise cooperativity between Bcd molecules bound to nearby sites *in vitro* [8, 9]. Remarkably, the cooperative interaction has a range of at least 41 bp, the center to center distance between the A1 and X1 sites in the *hb* promoter [8]. Given the absence of a well defined upper limit for the range of cooperative interactions of Bcd, I chose a 60 bp range for the studies presented here, although a shorter range did not affect the quality of fit (Table B.1, Model 2). When I allowed the cooperative interaction of Bcd, together with Bcd-Hb coactivation, the model significantly improved the quality of the calculation of gene expression (Figure 3.2) with a small defect in stripe 7 expression.

3.1.3 Cad mediated coactivation

Unlike other stripes, it was observed that stripe 7 expression was not properly formed during early simulation (see Figure 3.1, Figure 3.2 and Figure 3.3). In order to investigate stripe 7 expression, I tried to model gene expression of MSE3 (the "3/7" enhancer) with the *in silico* transcription system containing the Bcd cooperative binding and Bcd-Hb coactivation mechanisms. MSE3, a part of the four fusion constructs, drives expression of stripes 3 and 7 [24, 25, 26]. The model trained with the newly added MSE3, in addition to the fusion constructs, failed to calculate proper stripe 7 expression (Figure 3.3A).

Inspection of the MSE3 enhancer reveals an intriguing binding site arrangement of Hb. More than 11 footprinted binding sites are tightly clustered in MSE3. The DNaseI footprint assay shows that MSE3 does not contain Bcd footprinted binding sites. In the absence of coactivating Bcd, Hb sets the anterior border of stripe 3 expression by repressing the transcriptional activity of MSE3 [26]. In the region of stripe 7, the concentration of Hb is much higher than its concentrations in the stripe 3 region (Figure 3.3), therefore MSE3 must require strong activation input to overcome the Hb mediated repression. I considered Cad mediated activation as a strong candidate for the additional activation input. Cad is a maternal and zygotic factor that gradually reaches its highest concentration around stripe 7. Cad can, therefore, provide an additional activation input to the posterior region including stripe 7. It has been suggested that *eve* stripe 7 expression is driven by different activator(s) instead of D-STAT [63, 69]. Furthermore, more direct experimental evidence supports the possibility of Cad mediated activation of stripe 7 expression [70]. However, how does the same MSE3, the activity of which is strongly repressed by Hb at the anterior border of stripe 3, can activate stripe 7 expression in the

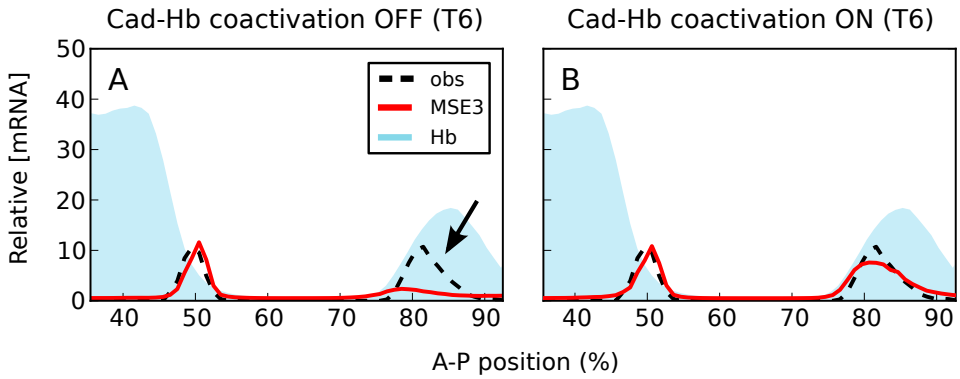


Figure 3.3: **Model suggested Cad-Hb coactivation.** (A) Cad mediated Hb coactivation is not activated in the model. MSE3 fails to drive stripe 7 expression in the model (see black arrow). (B) Cad mediated Hb coactivation is activated in the model. Stripe 7 expression is successfully recapitulated. Light blue area indicates Hb protein profile. Compare the Hb concentrations at the peaks of stripes 3 and 7 respectively.

presence of even higher Hb concentration? The fact that the stripe 2 activators cannot activate gene expression in the presence of high concentrations of repressors bound to MSE2 suggests that, in order to drive stripe 7 expression, the Hb mediated repression needs to be removed. A possible scenario is that Cad might provide the required strong activation by coactivating Hb bound to MSE3. Having coactivation of Hb by Cad enabled, the model was able to calculate the gene expression of stripe 7 in the fusions and MSE3 accurately (Figure 3.3B). This result supports the hypothesis of coactivation of Hb by Cad for *eve* stripe 7 expression.

In summary, these results suggest that three regulatory mechanisms—coactivation of Hb by Bcd, pair-wise cooperative binding of Bcd and coactivation of Hb by Cad—are essential for modeling the expression of *eve* stripes 2, 3 and 7. Hence, I extended the previous model by incorporating the these

regulatory mechanisms. In addition, in order to calculate the fractional occupancies of individual binding sites, a PWM based binding site prediction and a full thermodynamics calculation are incorporated. The *in silico* transcription system constructed in this dissertation contains representations of thermodynamic protein-DNA interactions including steric interference and cooperative binding, short range repression, direct repression, coactivation and activation synergy. I will now describe it in full technical detail.

3.2 The *in silico* transcription system

The first step for the *in silico* transcription model is to identify the position of TF binding sites (TFBSs) and calculate their relative binding affinities in a given regulatory DNA. A TF can bind specifically to various short stretches of DNA with different binding affinities. For example, Bcd, a homeodomain TF in *Drosophila melanogaster*, can bind to at least 48 different motifs. The zinc finger TF Hb is able to bind to about a hundred different motifs *in vitro* [71]. A consensus sequence approach identifies some putative TFBSs for a certain TF in regulatory sequences by comparing the sequences in a sliding window with the canonical motif or a subset of the motifs for the TF. However, the consensus sequence approach can only identify TFBSs in a qualitative manner and does not provide any information about the binding affinity of the variants of a canonical motif. An alternative to the consensus sequence approach is to utilize a PWM (position weight matrix), also called a PSSM (position specific scoring matrix) [72]. To date, the PWM is the most commonly used representation of functional binding site motifs because it makes possible to both identify TFBSs and estimate relative binding affinity. The *in silico* tran-

scription system utilizes the PWM approach in order to calculate binding site positions and their binding affinity directly from DNA sequence.

3.2.1 Generation and selection of PWMs

The idea of PWMs was first formulated by Stormo and colleague [73], then refined by Berg and von Hippel by utilizing thermodynamics to show that the PWM score was proportional to the Gibbs free energy of binding [74]. A key assumption of PWMs is that the individual base pair contributions to the free energy of a ligand binding to its site are independent and therefore, additive [72]. It was hypothesized that multiple bases in a ligand binding site could interact with the ligand and that the total Gibbs free energy ΔG of a ligand binding to its site is the sum of all the contributing interactions [72]. The essential assumption lying behind the calculation of the individual base pair contributions is that the probability of finding a given motif in the sample depends on its free energy of binding through the Boltzmann distribution $\exp(-\Delta G/kT)$. This quantity in a chemical context is simply the binding affinity K. In a statistical context, it can be interpreted as inversely proportional to the odds that a particular binding site would be identified as specific when it is in fact non-specific. For that reason, the individual base pair contributions to the free energy are calculated by taking the log-odds score of the base frequency in the motifs [74]. In this thesis I will, for clarity, always interpret the sum of all the contributions (called PWM score) as being proportional to the ΔG of binding. To obtain the PWMs for regulating TFs, three steps were taken: 1) Alignment of functional binding site sequences. 2) Generation of PWMs. 3) Selection of PWMs. Each of these steps is described in detail in the following sections.

Generation of PWMs

High quality PWMs require a fairly large sample of specific functional binding sites to provide reasonable predictive accuracy [74]. SELEX (Systematic Evolution of Ligands by EXponential enrichment) based data is suitable for fulfilling that requirement. SELEX is a method to enrich a population of bound DNAs from a random sequence pool, each of which is typically 16-24 bp in length, by successive rounds of the following steps. First, DNA-protein binding reaction. Second, PCR amplification of bound DNAs. Third, a portion of the PCR-amplified DNA from previous round of SELEX is used as the starting DNA probe pool for the next round of SELEX [75].

I obtained SELEX derived 16 bp long TF binding sequences, a courtesy of Dr. Mark Biggin, for Bcd, Cad, Hb, Kr, Kni and Gt from the BDTNP (Berkeley Drosophila Transcription Network Project) database (http://bdtnp.lbl.gov/). Then I searched binding site motifs for each TF in SELEX data using the MEME (Multiple Expectation Maximization for Motif Elicitation) software [76]. MEME is a widely used tool for searching for motifs in sets of DNA or protein sequences [76]. MEME scans each binding site sequence with a variable length scanning window and determines whether the sequence in the scanning window is a motif or background sequence using the EM (Expectation Maximization) technique [77]. The major advantages of MEME are that it adaptively estimates the best motif width in a given range and discovers motifs with high accuracy [78]. MEME generates multiple frequency matrices, which record the position-dependent frequency of each nucleotide in motifs, if more than two distinctive sets of motifs are found. I generated a family of PWMs with different widths for each of these TFs by running MEME v.3.0.4 with parameters "-evt 0.001 -dna -nmotifs 10 -minw A -maxw B -nostatus -

mod zoops -revcomp" on different selection rounds of the SELEX data, with A equal to 8 and B usually set to 12 unless the results were unsatisfactory, in which case I increased it to values up to 15. I obtained multiple frequency matrices for all of the six TFs. Each frequency matrix was converted into position weight matrix using the following equation:

$$m_{i,j} = \ln(p_a(i,j)/p_{\text{bg}}(j)), \tag{3.1}$$

where $m_{i,j}$ indicates an element at the i, j position in the PWM matrix. $p_a(i,j)$ indicates the probability of observing base j at position i of a binding site for TF a and $p_{\text{bg}}(j)$ is the genome-wide frequency in *Drosophila melanogaster*, called the background probability of base j.

Selection of PWMs

MEME adaptively estimates the motif with a variable length scanning window. It is common that more than two distinctive sets of motifs are found. Consequently, multiple PWMs are generated for each TF. For example, 7 Bcd PWMs were generated from the SELEX data and 12 PWMs for Gt. I evaluated the multiple PWMs with two criteria, the recovery rate and false positive rate. The recovery rate of a PWM is the ratio of the number of recovered footprinted sites by the PWM over the total number of footprinted sites. The false positive rate of a PWM for a TF is the ratio of the number of inert sequences, which are predicted as binding sites by the PWM, over the total number of inert sequences for the TF binding. These inert sequences were a total of fifteen segments of sequence (20 bp each) from the *eve* transcript which show no peaks on ChIP-Chip assays [79], and unprotected sequences located between known footprinted sites, which is TF dependent. Then the

best PWM for each TF was selected as follows.

With the threshold set to zero, I discarded all PWMs that failed to detect more than 70% of known footprinted sites by extending each site by 5 base pairs of genomic sequence on each side and considering the highest score of the extended site. From the remaining PWMs, I selected the one that gave the smallest number of false positives. The result, summarized in Table 3.1, led to the selection of Bcd, Hb, Kr, and Gt sites from the SELEX derived PWMs. However, the best SELEX-derived Kni and Cad matrices failed to meet the criterion. One SELEX-derived Kni matrix recovered 100% footprinted sites, however its false positive rate is also extremely high at 97%. The best Cad matrix recovered only 61% of the footprinted sites.

In the cases for which BDTNP SELEX-derived PWM were not available or if the SELEX-derived PWM failed to meet the quality criterion for a specific TF, I compared at least two independent PWMs which are publicly available [80, 81, 71, 56, 82, from Dr. Dmitri Papasenko] for each TF and chose the PWM that gave the best result. For Kni, Cad, and Dichaete, I utilized bacterial one-hybrid PWMs [82]. These matrices recovered 75%, 76%, and 100% of the footprinted binding sites respectively with a reasonably low false positive rate (see table 3.1). For D-STAT, both the one-hybrid PWM [82] and the SELEX PWM obtained from Dr. Dmitri Papasenko (http://line.bioinfolab.net/webgate/help/dxp.htm#D-stat-223) meet the requirements. I chose the SELEX PWM for this study because the Papasenko SELEX PWM predicted the relative binding affinity of two sites in MSE3 more accurately. For Tll, the PWM used was from a published source [81]. The Tll PWM shows an high recovery rate (94%) and a false positive rate that is lower than that of the footprint-derived PWM from the *Drosophila* DNase I Footprint Database [71].

	This work	Berman et al.	Segal et al.
Bcd	79% (38/48)	100% (48/48)	58% (28/48)
	8% (3/35)	42% (15/35)	14% (5/35)
Cad	76% (10/13)	100% (13/13)	69% (9/13)
	15% (4/26)	100% (26/26)	19% (5/26)
D-STAT	100% (3/3)	N/A	0% (0/3)
	17% (3/17)	N/A	17%(3/17)
Dichaete	100% (4/4)	N/A	N/A
	22% (4/18)	N/A	N/A
Hb	87% (90/103)	98% (101/103)	86% (89/103)
	3% (1/26)	19% (5/26)	0% (0/26)
Kr	75% (34/45)	86% (39/45)	86% (39/45)
	6% (1/15)	26% (4/15)	40% (6/15)
Kni	75% (25/33)	84% (28/33)	42% (14/33)
	21% (9/42)	59% (25/42)	28% (12/42)
Gt	75% (6/8)	N/A	62% (5/8)
	6% (1/15)	N/A	20% (3/15)
Tll	94% (35/37)	N/A	70% (26/37)
	63% (23/26)	N/A	52% (19/36)

Table 3.1: **Comparison between PWMs.** For each TF, the top row is the recovery rate of footprinted sites and the bottom row is the rate of false positives. In this dissertation work, SELEX data derived PWMs were used for Bcd, D-STAT, Hb, Kr and Gt. Bacterial one hybrid data derived PWMs were used for Cad, Dichaete and Kni. Gibbs sampling data derived PWM were used for Tll.

However, the false positive rate for Tll (63%) is notably higher than that of the PWMs for other TFs. All PWMs used in the model are listed in Appendix C and Table 3.1 summarizes the performance of the PWMs.

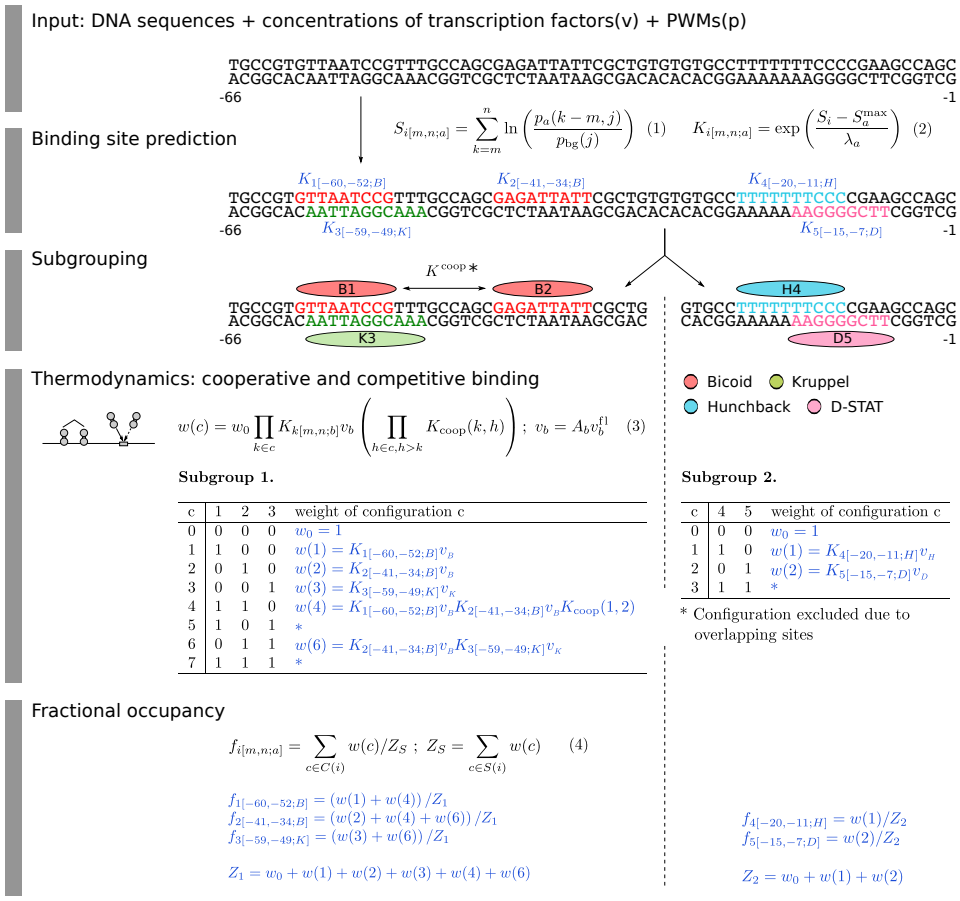
3.2.2 Implementation of the model equations

TF binding to DNA

The central players of transcriptional regulation are sequence-specific TFs that bind to DNA. The position of a TF binding site and its binding affinity are determined by a frequency matrix normalized to a position weight matrix (PWM; Figure 3.4, Eq. 1). In this equation, $p_a(k-m,j)$ is the probability of finding base j ($j \in \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$) at the kth position of a possible binding site for ligand a that extends from base m on the 5' side to base n on the 3' side, and $p_{\mathrm{bg}}(j)$ is the expected frequency of base j in *D. melanogaster*. When convolved with sequence, the score $S_{i[m,n;a]}$ of the PWM on the sequence is proportional to the free energy of binding [83], and can be exponentiated to obtain the binding affinity $K_{i[m,n;a]}$ of ligand a at site i. This is shown in Figure 3.4, Eq. (2), where $S_a^{\max}$ is the maximum possible score and λ_a is the proportionality constant to free energy. I include a binding site in the calculation when its score is above a certain threshold. This threshold can be determined with different degrees of accuracy for each TF depending on the quality of the data used to construct its PWM (See section 3.2.1).

Fractional occupancy: competition and cooperativity

In order to calculate the fractional occupancy $f_{i[m,n;a]}$ of TF a bound at a site i that extends between m and n bp from the TSS, it is useful to first determine



c	1	2	3	weight of configuration c
0	0	0	0	$w_0 = 1$
1	1	0	0	$w(1) = K_{1[-60,-52;B]} v_B$
2	0	1	0	$w(2) = K_{2[-41,-34;B]} v_B$
3	0	0	1	$w(3) = K_{3[-59,-49;K]} v_K$
4	1	1	0	$w(4) = K_{1[-60,-52;B]} v_B K_{2[-41,-34;B]} v_B K_{\mathrm{coop}}(1,2)$
5	1	0	1	*
6	0	1	1	$w(6) = K_{2[-41,-34;B]} v_B K_{3[-59,-49;K]} v_K$
7	1	1	1	*

c	4	5	weight of configuration c
0	0	0	$w_0 = 1$
1	1	0	$w(1) = K_{4[-20,-11;H]} v_H$
2	0	1	$w(2) = K_{5[-15,-7;D]} v_D$
3	1	1	*

* Configuration excluded due to overlapping sites

Figure 3.4: **Model equations: TF binding to DNA.** The model equations for binding site prediction (Eq. 1 and 2), cooperative and competitive binding (Eq. 3) and fractional occupancy calculation (Eq. 4) are shown together in a flow diagram with cartoons of each mechanism on the left and an example application in blue with 5 TF binding sites. Subgrouping process partitioning the binding sites into independent binding groups allows faster computation without losing accuracy. In the example, we set the range of quenching to 20 bp.

the effects of interacting configurations c of TFs in terms of their weights $w(c)$ (Figure 3.4, Eq. 3). These weights depend on TF concentrations v_b, which in our dataset are in units of relative fluorescence v_b^{fl} from confocal scans. To

convert to true concentration units I multiply by a free parameter A_b to obtain v_b. There are two types of interacting configurations. Some TF binding sites overlap or are closely placed. Overlapping sites lead to competitive binding by steric hindrance. I implement this phenomenon whenever sites overlap based on their averaged size determined from DNase I footprints. A second type of interaction has the opposite effect. Two adjacent sites may support cooperative binding, in which the free energy of binding of two simultaneously bound factors is greater than the sum of the free energies of them each binding separately [68, 84]. Transforming free energies to binding affinities, the nonadditive free energy term becomes a multiplicative factor $K_{\text{coop}}(k, h)$, where k and h are two interacting binding sites (Figure 3.4, Eq. 3). I implemented cooperativity only when there is independent evidence for it, which is currently the case only for Bcd [8, 9]. I allow the strongest Bcd binding site to interact cooperatively with the strongest remaining Bcd site within 60 bp (see Section 3.1.2), and repeat these assignments with the remaining sites until all pairwise cooperative interactions are assigned. With these mechanisms in hand, I use the concentration of TF a and other competing or cooperating TFs to calculate the fractional occupancy $f_{i[m,n;a]}$ (Figure 3.4, Eq. 4). We do this by summing the weights $w(c)$ for all configurations c in which site i is occupied by a. I then normalize against the sum Z_S of all weights $w(c)$ in group S, ensuring that for each site $f_{i[m,n;a]}$ is between 0 and 1. As shown in the example associated with Eqs. (3) and (4) of Figure 3.4, each interacting group can be treated independently.

Note that the quantities $f_{i[m,n;a]}$ are fully deterministic intensive thermodynamic variables akin to concentrations. Although frequently derived from statistical mechanics [84] or even the Chemical Master Equation [85], they

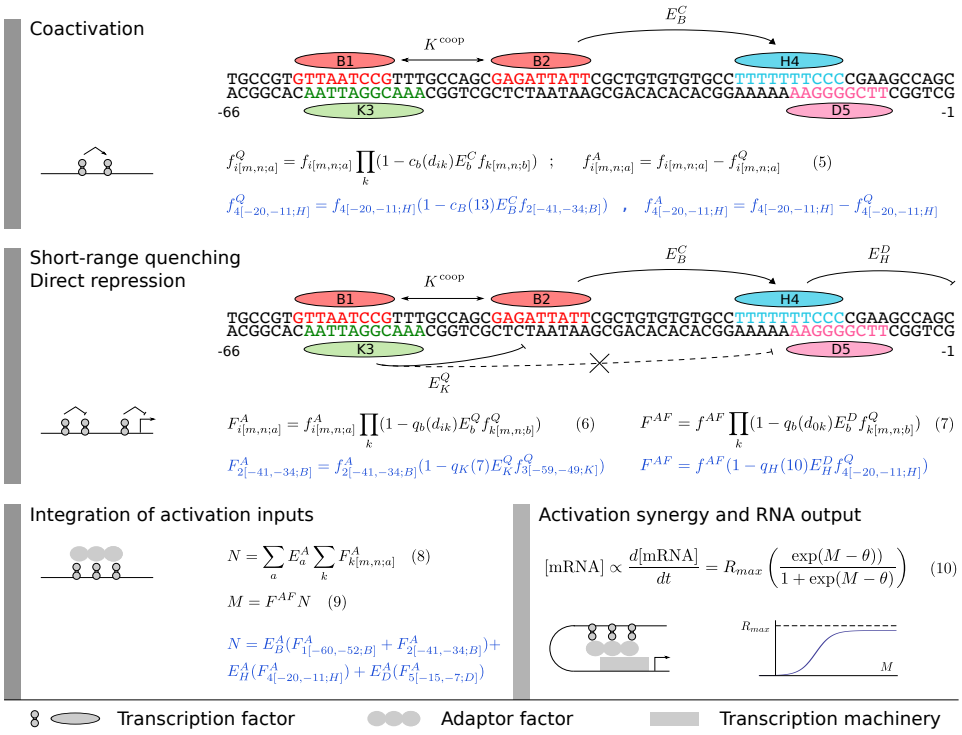


Figure 3.5: **Model equations: protein-protein interactions.** The model equations for coactivation (Eq. 5), short-range quenching (Eq. 6), direct repression (Eq. 7), adaptor factor recruitment (Eq. 8 and 9) and activation synergy (Eq. 10) are shown together in a flow diagram with cartoons of each mechanism on the left and an example application in blue with 5 TF binding sites.

can also be derived from elementary considerations of equilibrium and stoichiometry [86]. Although $f_{i[m,n;a]}$ is frequently interpreted as the probability of finding ligand a bound at site i, it is more accurate to view this quantity as the time averaged occupancy of site i by a. I thus assume that the binding states of the TFs that we explicitly consider equilibrate quickly compared to the time scale of changes in gene expression.

Coactivation

Once I have calculated f_i, I calculate the effects of protein-protein interactions (Figure 3.5). A TF b bound at site k acting on a TF bound at site i by mechanism X will be characterized by a parameter E_b^X between 0 and 1 denoting the strength of b's action and a function $0 < x_b(d_{ik}) < 1$ of the distance in bases between sites k and i which controls the range at which the mechanism acts. The equations representing each mechanism are written such that they have the property that biological function can reside in multiple binding sites. I classify TFs as repressors or activators based on independent experiments. In what follows, f_i with no superscript denotes the physical fractional occupancy of site i. I write f_i^A to denote the fractional occupancy of an activator and f_i^Q to denote the fractional occupancy of a quencher. I then allow for the possibility of coactivation, in which a repressor is transformed to an activator by the binding of a coactivator nearby. There is evidence that Bcd coactivates Hb in this manner [10, 39], as does Cad (see Section 3.1.3).

I represent coactivation as shown in Figure 3.5, Eq. (5), where E_b^C represents the coactivation efficiency of a coactivator b and the dependence of coactivation on distance is given by $c_b(d_{ik})$ that equals 1 for $d < D_1$ and 0 for $d > D_2$, with linear interpolation between these points (Figure 3.6B). I set $D_2 = 1.1D_1$ so that only one free parameter is added when coactivation distance is not fixed. For Bcd, I allow $D_1 = D_{\mathrm{B-H}}^C$ to vary within a range tightly constrained by experimental observations. If $D_{\mathrm{B-H}}^C$ were less than 150 bp, the distance between two closest sites of Bcd bound to MSE2 and Hb bound to MSE3 in the M32 construct, the Hb bound to MSE3 would repress stripe 2 (Figure 3.6C). If, on the other hand, the distance were longer than 200 bp, a "spacer" of 160 bp would not suffice to make MSE2 and MSE3 independent

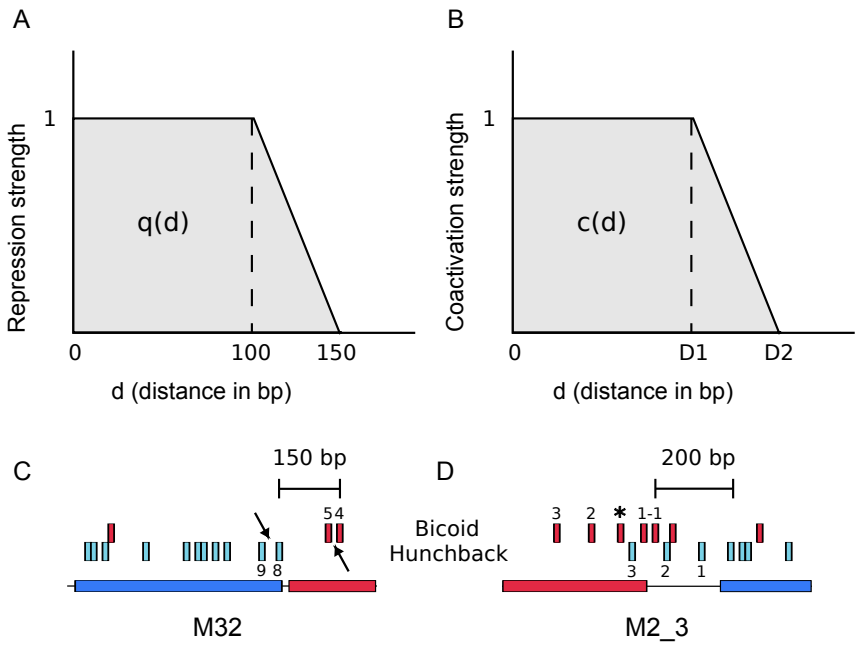


Figure 3.6: **Repression and coactivation functions.** (A) The short range repression function $q_b(d_{ik})$. (B) The coactivation function $c_b(d_{ik})$. D_1 and D_2 are indicated. Key binding sites used for establishing the coactivation range of Bcd in M32(C) and M2_3 (D) are shown. Bcd and Hb sites are in red and cyan respectively. Some sites are labeled by name. See Figure 5.4 for a diagram of all sites.

in M2_3 (Figure 3.6D). Training runs gave very constrained values of $D^C_{\mathrm{B-H}}$ that ranged from 158 to 165 bp (Table B.1). In contrast to Bcd, there were no independent constraints on the range of the parameter $D_1 = D^C_{\mathrm{B-C}}$ for Cad, so we allowed it to vary from 10 to 200 bp. I constrain the activating and repressing activity of a coactivation target to the sum of the total physical fractional occupancy.

Short-range quenching and direct repression

The gap genes are short range repressors that act when bound within 150 bp of activators [43], a fact that I represent by convolving the fractional occupancies of all activators f^A with those of quenchers f^Q as shown in Figure 3.5, Eq. (6) to obtain activator fractional occupancies F^A corrected for quenching, where E_b^Q represents the repressive strength of TF b and the function $q_b(d_{ik})$ represents its range of action (Figure 3.6C). When quenchers are bound within quenching range of the TSS they can prevent activators from acting at any range, a phenomenon known as direct repression which is represented (Figure 3.5, Eq. 7) in the same way as Eq. 6 except that d_{0k} in this equation is the distance between the repressor binding site k and TSS, and that the repressor does not act on f^A but on f^{AF}, a quantity associated with the transcription machinery that binds to TSS, as I now describe.

Adaptor recruitment and activation synergy

With respect to activation, it is now clear that in metazoa activators do not directly contact the transcription machinery [87, 88]. Instead, proteins that bind to TFs, such as Mediator [3, 4], serve as a functional bridge between TFs and the basal machinery. These proteins are referred to as "adapter factors" (AFs) here following Guarente and Tjian [89, 90, 88]. Although AFs are sometimes referred to as "corepressors" or "coactivators", I reserve that terminology in this work to TFs that bind DNA specifically. I view initiation of transcription as an enzymatic process catalyzed by AFs bound to TFs [47]. In the fly blastoderm, some AFs have been identified [91, 4, 92] and they are uniformly expressed from maternal mRNA, enabling us in this work to formulate AF action in a coarse-grained manner such that AFs are represented by a single composite

chemical species whose fractional occupancy of binding to DNA bound TFs is given by $f^{AF} = 1$ (Figure 3.5, Eq. 7). Functionally active activators a recruit the AFs with different recruiting strengths E_a^A (Figure 3.5, Eq. 8). Activators can act anywhere between the TSS and an insulator element, so here I do not need to consider d_{ik}, but simply sum the effects of the activators to obtain N, which is then corrected for the effects of direct repression to obtain M (Figure 3.5, Eqs. 7 and 9). The adapters then catalyze transcriptional initiation by decreasing the activation energy barrier of transcriptional initiation, $\Delta A = \theta$ by an increment $M = \Delta\Delta A$. We describe the effect of lowering this activation energy by a diffusion limited Arrhenius rate law (Figure 3.5, Eq.10, Figure 5.2A and Appendix A). This rate law is exponential for a certain range of M, providing the capability to represent greater than multiplicative synergy between activators [93]. As the activation energy barrier falls to zero, the transcription rate R approaches $R_{\max}$ because diffusion of new polymerase molecules to the basal complex becomes rate limiting.

3.3 Optimization of the *in silico* transcription system

3.3.1 Parameter optimization

Implementation of scaled cost function

Parameters were initially determined by minimizing the summed squared difference between the model output and the data, which consisted of 399 observations of RNA level (7 constructs $\times$ 57 RNA data from 35% to 92% EL). The summed square differences (E) was calculated using the equation

$$E_i = \sum_j ([\text{RNA}]_{i,j}^{\text{observed}} - [\text{RNA}]_{i,j}^{\text{calculated}})^2, \tag{3.2}$$

where i indicates ith construct in the model and j indicate A-P position (% EL). However, during the model training with the fusion constructs, it was found that the model fitted the M32 expression better than M3_2 expression (Figure 3.7). Because the model utilized the summed squared difference as a cost function, the model tends to fitting M32 better than M3_2 to reduce the summed square differences between the observed RNA concentration and the calculated RNA concentration. In order to fit the low expression levels of M3_2 more accurately, I rewrote the cost function

$$E_i^{\text{scaled}} = E_i \frac{W_{\max}}{W_i}, \tag{3.3}$$

where

$$W_i = \sum_j ([\text{RNA}]_{i,j}^{\text{observed}})^2. \tag{3.4}$$

If the summed square RNA concentration of a construct i (W_i) is smaller than the maximum weight ($W_{\max}$) among the given constructs, the equation forces the model to fit the gene expression of the construct i more accurately. The following result shows that the modified cost function provides better performance in fitting the gene expression driven by M3_2 (Figure 3.7).

Model parameters

With respect to the regulatory parameters, each TF a is associated with the parameter A_a that scales the observed fluorescence units v_{fl} to absolute concentration units v_a (Figure 3.4, Eq. 3) as well as the parameter λ_a that scales

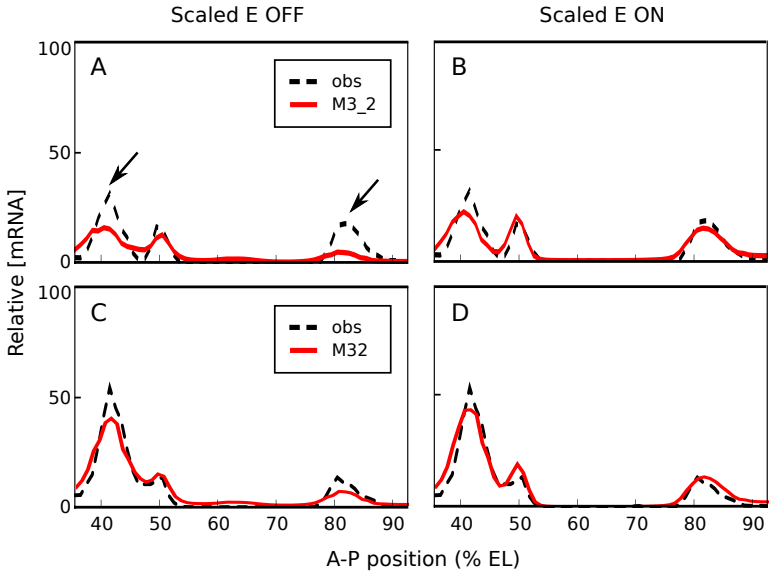


Figure 3.7: **Utilizing scaled square difference improves model fitting.** (A, C) The modeling results obtained without the scaled cost function. The model fitted better to the M32 expression better than that of M3_2 (see arrows). (B, D) The modeling results obtained with the scaled cost function. The model fitted both M32 and M3_2 expression well.

the weight matrix score to units of free energy (Figure 3.4, Eq. 2). Other parameters depend on the nature of the TF. Each activator is associated with an activation efficiency E_a^A, and each repressor to quenching and direct repression efficiencies E_a^Q and E_a^D respectively. Thus each activator and repressor is associated with three and four parameters respectively. In addition, Bcd has a free parameter $K_{\text{Bcd}}^{\text{coop}}$ representing cooperative interaction. All elements of $K_{\text{coop}}(k,h)$ from Figure 3.4, Eq. (3) are equal to $K_{\text{Bcd}}^{\text{coop}}$ or unity. Both Bcd and Cad have free parameters E_{Bcd}^C and E_{Cad}^C for the coactivation of Hb, and coactivated Hb has an activation efficiency E_{Hb}^A. The activation energy barrier of transcription, θ, was also fitted (Figure 3.5, Eq. 10). In addition

to these parameters, 10 free parameters, three positional effect scale factors S^R for the P-element constructs—1700, MSE2 and MSE3—and seven PWM threshold scores T for all TFs except for Bcd and Hb (see Section 3.5 for details). Finally, I fitted the operating distances of the Bcd and Cad mediated coactivation, but the range of the coactivation distance was set by independent experimental criteria (Section 3.2.2, Page 56). Thus, 49 free parameters are fit to 399 observations.

Optimization of these parameters was performed using the simulated annealing schedule of Lam [94, 95, 96]. Parameter search spaces were set by explicit search limits for A^a, λ^a, E_a^A,E_a^Q, E_a^D, $K_{\text{Bcd}}^{\text{coop}}$, E_{Bcd}^C, E_{Cad}^C and θ with $R_0 = 255$ and $f^{AF} = 1.0$ (Figure 3.5).

3.3.2 Code optimization

Parameter optimization of the *in silico* transcription system is not a simple task, especially for large regulatory DNA sequences because the number of configurations of bound factors increases exponentially with the number of binding sites in the sequence and consequently, the parameter optimization time or model training time also increases rapidly. In order to reduce the computation time, I employed two strategies. First, the model partitions the binding sites in a given construct into independent binding groups such that if a binding site is overlapped with other binding site(s) or a TF bound to the site cooperatively interacts with a TF on another site, they are assigned to the same group. Then the model applies the thermodynamic calculation to each group separately (Figure 3.3). This made it possible to train the model of the four fusion constructs which contains several hundred binding sites with a single high performance CPU in a month. I then further improved the calculation

speed up to six times faster (fitting is completed in three to four days) by optimizing the data structure storing all the possible binding configurations and the inner loop calculating the statistical weight of the configurations.

3.4 Restriction of modeling time class

Attempts to recapitulate gene expression at all the time classes of the fusion constructs were not successful. Because of the difficulties in modeling gene expression of the eight different time classes, I trained the model with the expression data of fewer time classes (T3-T6), wherein the reporter constructs drive distinctive expression patterns. The model reasonably fit the expression data (Figure 3.8B). Then I attempted to predict gene expression of the earlier time points (C13-T2) using the model to ask whether there is a notable discordance between the observed data and the model prediction. Interestingly, the model predicted strong activation of gene expression at C13 and T1, which is not seen in the observed data (Figure 3.8A). The model suggests that if the regulatory DNAs are fully accessible to TFs in the early time points, the reporter constructs would drive strong gene expression with the given *trans*-acting factor concentrations.

I hypothesized that the chromatin environment, which is not implemented specifically in the model, might play a role during the cleavage cycle 13 and early 14 such that the chromatin environment in the early time points might not be as favorable to transcriptional activation as in the late time points. Because the modeling of the chromatin on the regulatory DNA is beyond the capability of the current model, I restricted the modeling time classes to a single time point T6 (85.2 min after cleavage cycle 10). The fact that at T6,

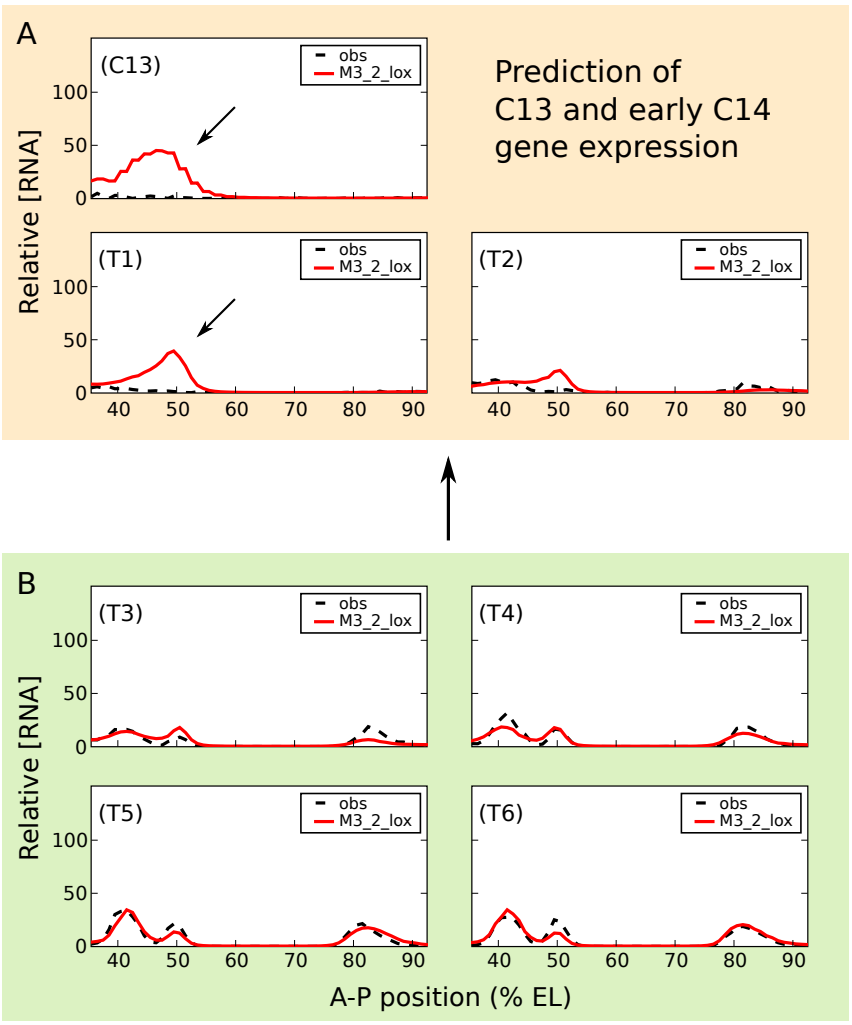


Figure 3.8: **Model suggested the repressive action of chromatin.** (A) Predicted gene expression at early time classes. (B) Model is trained with gene expression data at late time classes. The late time classes model predicts strong early activation of gene expression while observed levels of gene expression is almost zero. This results raise the possibility of repressive action of chromatin at early time classes.

all of the fusion constructs containing MSE2 and 3 are transcriptionally active at stripes 2, 3 and 7 and the gene expression reaches maximum suggests that at this time point, the effect of chromatin on transcription might be relatively smaller than the effects at other time points. This approach would provide a chance to characterize the direct effect of the protein-protein interactions and DNA-protein interaction of TFs to gene expression more accurately.

3.5 The rules of *cis*-regulation determined from expression data

After refining process of the *in silico* transcription model, I was able to obtain the transcription model used for the rest of my dissertation work (Figure 3.9) and four additional models for comparison (Figure 3.10 and 3.11). The models were generated as follows; First, the model is trained with gene expression data driven by seven constructs—M3_2, M32, M2_3, M23, 1700, MSE3 and MSE2—at single nucleus resolution. TF expression data for all proteins except Dichaete were that used [50], with the addition of new D-STAT data starting with C13, averaged from at least six embryos per each time class. Dichaete data were obtained from the t5:26-50 virtual embryo data [97]. Intensity of the gene expression from the middle 10% of dorsoventral position values was quantified by the ImageJ [98] plot profile function and was not registered to the Eve pattern. Quantitative expression data for the 1700 construct (1.7 kb proximal *eve* promoter) was previously published [50] and, in addition to M3_2, M32, M2_3 and M23, I also generated quantitative MSE2 expression data from 1511B, one of three MSE2 bearing lines that were gifts of M. Levine. (Figure 3.9 and See Figure 2.2 for a comparison of 1511B and 1511C expres-

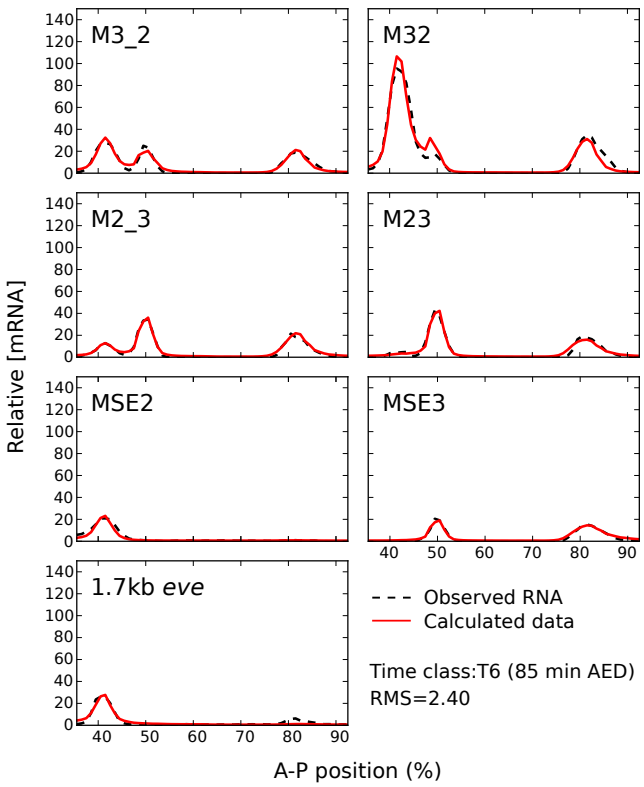


Figure 3.9: **Model training: standard 7 constructs model.** Model output is represented by the red solid lines, while the observed expression data is represented by the black dashed lines, as shown in the key; the model result trace obscures the data in regions where both are superimposed.

sion). Quantitative MSE3 expression data was obtained from M3_2 data by setting expression in stripe 2 to zero.

Second, protein expression data of a total of nine TFs, Bcd, Cad, D-STAT, Hb, Kr, Kni, Gt, Tll, Dichaete, were used as a regulatory input in the model. Third, I fitted the model described above to the gene expression driven by the four fusion constructs and three fragments of the *eve* promoters, MSE2, MSE3 and 1.7kb proximal *eve* (1700) at T6 (Figure 3.9); fits were also performed with the fusion constructs alone (Figure 3.11). Inclusion of the three addi-

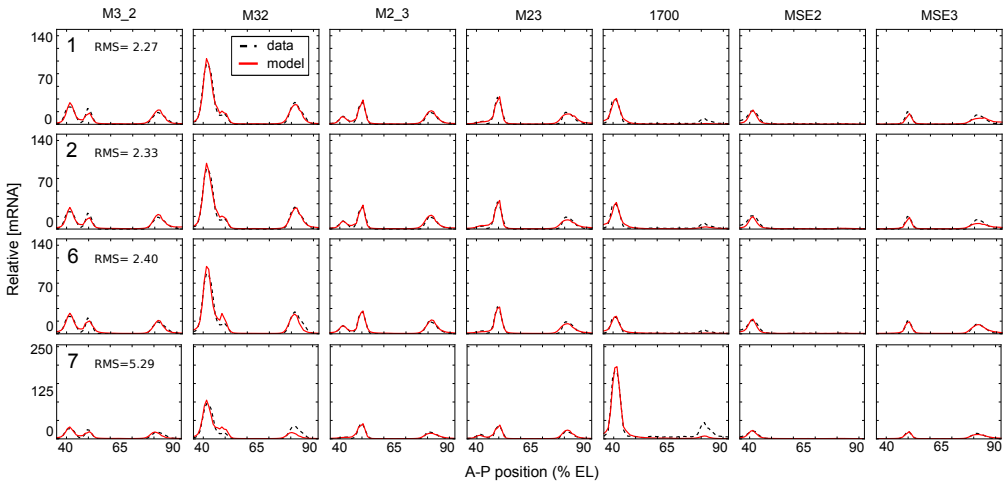


Figure 3.10: **Best four 7 constructs models.** In this study, in addition to the standard model, three other models which also had the best fit and had strong predictive power for gene expression were used in order to gain an understanding of the transcriptional control of the fusion constructs and for some predictions (see Chapter 4 for details). Model output is represented by the red solid lines, while the observed expression data is represented by the black dashed lines, as shown in the key. The behavior of models 1, 2, 6, and 7 are shown as indicated in the leftmost column, which also gives each model's rms score. Parameter sets for these four models are given in Table B.1. Note that the concentration scale for model 7 differs from the other two rows. The data is rescaled by the factor S^R, a free parameter for position effect, for the P-element constructs 1700, MSE2, and MSE3 (Table B.1).

tional P-element constructs improved the predictive power of the model at the cost of one additional free position effect scaling parameter for each P-element construct (data not shown). The sequences included in the model contains not only *eve* regulatory sequences, but also adjacent cassette sequences that were a part of the reporter constructs in the transformant line. Fourth, the PWMs for predicting the binding sites of the nine *trans*-acting factors were chosen as described in Section 3.2.1. Independent experimental data allowed us to define binding thresholds for Hb and Bcd unambiguously, but in the case of other TFs these data implied a range of values for PWM thresholds and I allowed

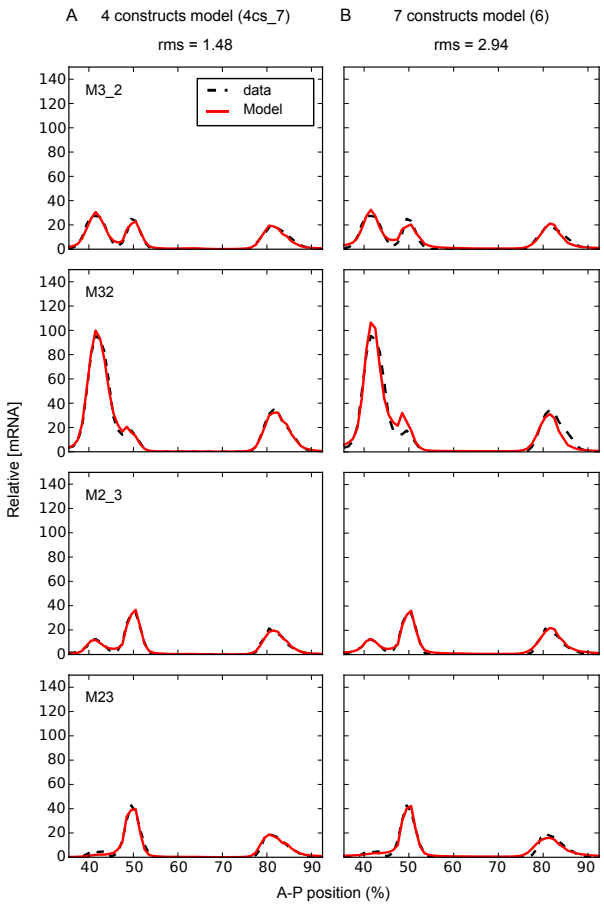


Figure 3.11: **4 constructs model vs. 7 constructs model.** One model, fit only to the four fusion constructs but fit almost completely, was also investigated to understand the changes in stripe 3 expression in M3_2 and M32. (A) The behavior of model 4cs_7 is shown with comparison to expression data, as indicated in the key. The x-axis is the percentage of A-P position and the y-axis is the relative mRNA concentration. This model was trained on expression data driven by the four constructs M3_2, M32, M2_3, and M23 only. (B) For comparison, we show the behavior of model 6, trained on seven constructs, compared to training data for the same four constructs shown in (A). The behavior of model 6 compared to its full training set is shown in Figure 4.1A1-7 and Figure 3.9. Note that model 4cs_7 fits the expression data driven by M32 better than model 6. Comparative rms scores are shown at the top. The full set of parameters for each model is given in Table B.1.

the threshold to be a free parameter within this range. Fifth, seven experimentally characterized regulatory mechanisms, all acting on *eve* regulation—TF binding to DNA, competition between TFs bound to overlapping sites, cooperative binding of Bcd, coactivation, short-range quenching, direct repression and transcription synergy is implemented. Sixth, I used gene expression of a single time point T6. Seven, model parameters were optimized in order to minimize the scaled summed square differences between the model calculation and the observed data. Runs were repeated 10 times with different random seeds for each optimization problem. The quality of the runs was judged by its root mean square (rms) score and by visual observation of the expression pattern. From these runs, I chose four models—model 1, model 2, model 6 and model 7 which had the best fit and had strong predictive power for gene expression (Figure 3.9 and 3.10) . Among the four models, Model 6, called the standard model in this study, shows the best predictive ability which will be described in the following chapter. The four models, however, show a small defect in stripe 3 expression in M3_2 or M32, hence, one model, fitted only to the four fusion constructs but fitted almost completely (Figure 3.11), was also investigated to understand the changes in stripe 3 expression in M3_2 and M32.

Chapter 4

Model validation

4.1 Quality of the model fits

Multiple fits to the training data resulted in a group of models driving essentially identical expression patterns (Figure 3.10) and having similar but not identical parameter values (Table B.1). The models resulting from the fitting procedure agree with experimental data within the limits of experimental accuracy with two very small exceptions (Figure 3.9 and Figure 3.10). First, the peak of stripe 3 in M32 is one nucleus anterior with twice the expression level in the model compared to data. Second, stripe 7 expression in the 1700 construct is almost absent in the model. It is an important validation of the model that I can numerically represent the effects of these enhancer fusions at this stringent level of precision.

4.2 Prediction of gene expression

The most stringent proof of the credibility of the theoretical model is its predictive ability on DNA sequences not used for training. The *in silico* transcription system provides a platform for in-depth biological validation of the proposed theory of the transcriptional control of the four fusion constructs and the simulated model results. In this section, I describe the prediction of various gene expression as the most important validation. Each DNA sequence tested contained one or more enhancers and the basal promoter sequence. If the basal promoter sequence for an enhancer construct was not known, the 42 bp long *eve* basal promoter sequence [6] was used. Except as noted, all predictions shown in Figure 4.1 were made from the standard model (Figure 3.10 and Table B.1) with no alterations of any parameter except the sequence itself. In Figure 4.1 and 4.2, black lines are predicted RNA expression and colored lines are quantitative protein profiles of the corresponding endogenous loci. The scale of relative fluorescence levels for RNA is shown at the left of graphs, that for proteins on the right. If a prediction from a parameter set other than model 6 is shown in Figure 4.1, the corresponding prediction from model 6 is shown in Figure 4.2. Altogether I tested 54 sequences amounting to 62 kb of DNA, and obtained good predictions for 44 sequences driven by 51 kb of DNA, as I will describe.

4.2.1 Site-directed mutagenesis

The classic literature describing the 5' regulatory region of the *eve* locus contains numerous studies of the effects of very small site-directed mutations affecting only 2 to 6 bases. Our ability to predict the effects of such mutations

is of interest not only for checking the validity of the model, but also has implications for the interpretation of single nucleotide polymorphisms (SNPs) and small indels (insertion and deletion of nucleotides). Here I consider a 3 base pair change in the bcd-1 site (Mbcd-1, M denotes mutation) in the context of both MSE2 and M32, a 5 base pair change in the bcd-3 site (Mbcd-3) in MSE2 [6], a two base pair change in each of two D-STAT sites (M2dsts) in MSE3 [36], and changes of 5, 3, and 6 base pairs respectively in the kr-3, kr-4, and kr-5 sites (MKr345) in M32 [25]. The model correctly predicts that Mbcd-1 causes a larger decrease of expression than Mbcd-3 (Figure 4.1B1-2) [6, Figure 6]. The model's prediction of greatly diminished expression in M2dsts is qualitatively correct, but experiment indicates a complete abolition of expression (Figure 4.1B3) [36]. The prediction of reduced but equivalent expression of stripes 2,3, and 7 while 2 and 3 remain fused when MKr345 is placed in M32 is completely correct (Figure 4.1B4), and the model correctly predicts the restoration of stripe 2 expression in the presence of a non-functional bcd-1 site when Mbcd-1 is placed in M32 (compare Figure 4.1B1 and 4.1B5), but the model predicts that stripe 3 is absent when in fact it is reduced [25, see Figure 4].

4.2.2 Downstream *eve* and chimeric enhancers

As an initial test of the model's predictive power on sequences with no homology to those used in training, I found that I can correctly predict expression of *eve* stripe 5 and stripes 4 and 6 from their respective enhancers (Figure 4.1D1-2 from model 2; see Figure S4.2A, B for model 6 results; see Table D.1 for sequence). I then extended this test to interspecific S2E chimeras. Altered expression patterns driven by chimeric constructs with half of the stripe 2 en-

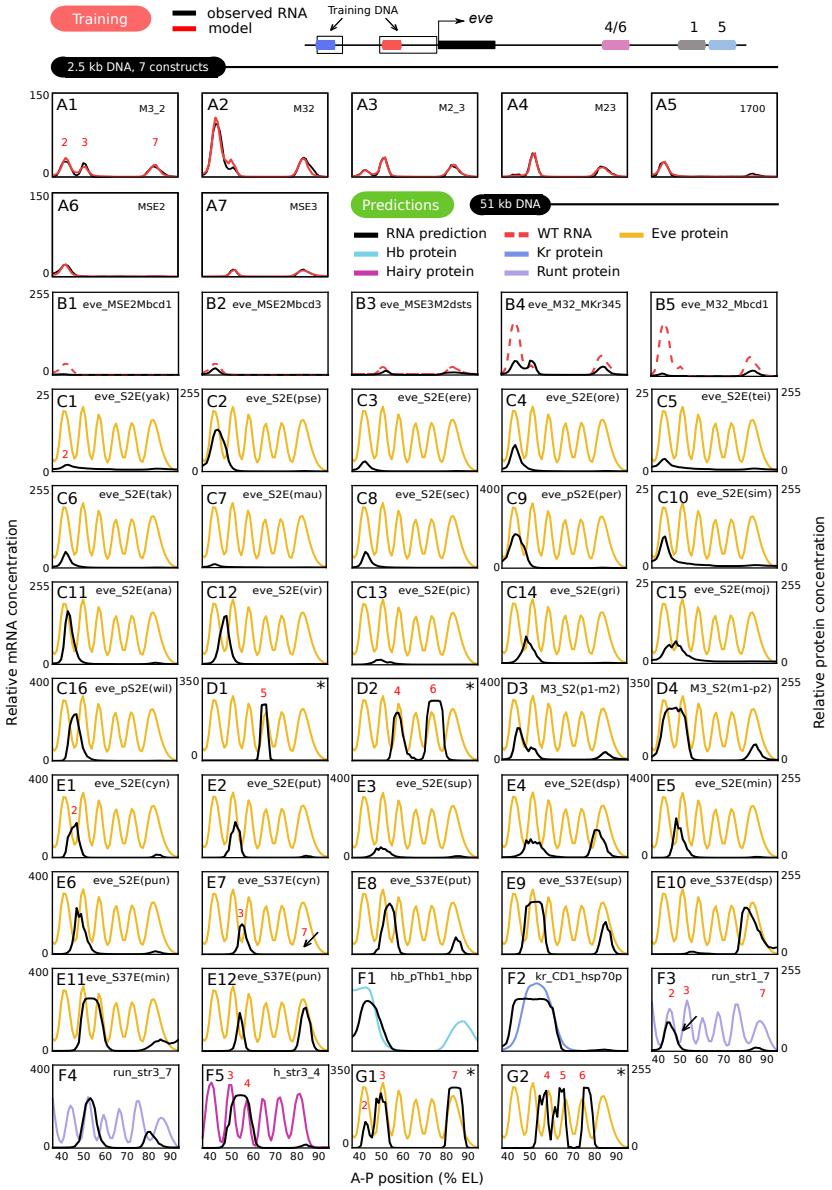


Figure 4.1: **Correct or putatively correct predictions.** The training set (A1-A7), together with predictions of gene expression driven by DNA sequences that were not used for training. Annotation is fully described in Section 4.2 and the sequences used in the predictions are described in Appendix D. All protein patterns are taken from the FlyEx database [49].

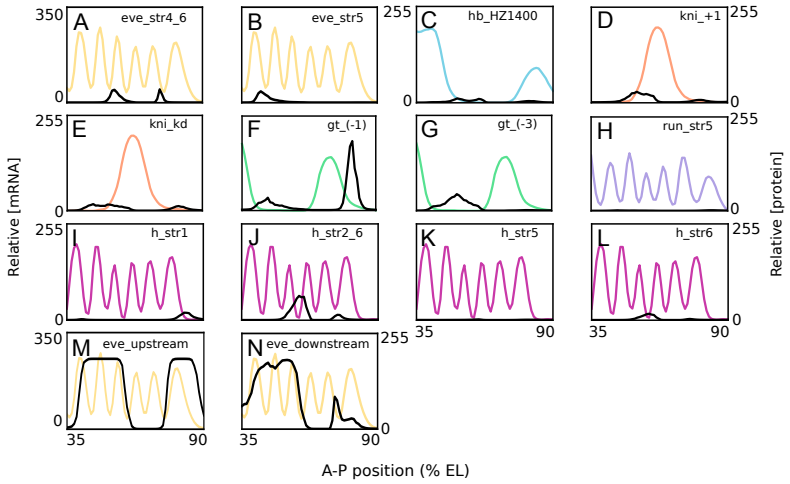


Figure 4.2: **Incorrect Predictions.** Incorrect predictions of gene expression driven by DNA sequences that were not used for training. The sequences used are fully described in Appendix D.1. All predictions in this Figure were made using the model 6 parameters (Table B.1). (A-B) Predictions for the *eve* stripe 5 (A) and 4/6 enhancers (B). Correct predictions of these enhancers from model 2 are shown in Figure 4.1D1-2. (C-L) Predicted expression driven by enhancers from the genes *hb* (C), *kni* (D-E), *gt* (F-G), *run* (H), and *h* (I-L). (M-N) Predictions for expression driven by large 5' (M) and 3' (N) *eve* regulatory DNAs that contain multiple enhancers. Correct predictions for these DNA segments from models 7 and 1 respectively are shown in Figure 4.1G1-2.

hancer from *D. pseudobscura* and and half from *D. melanogaster* have been observed by enzymatic assays [99]. With the *D. melanogaster* sequences on the 3' end, a posterior expansion of stripe 2 was described, which appears to extend to a variable fusion of the two stripes and a reduction in stripe 3 amplitude; our model predicts a partial fusion and a reduction in the amplitude of stripe 3 (Figure 4.1D3). The complementary chimeric enhancer drives a fusion of stripes 2 and 3 which is also predicted by the model (Figure 4.1D4) [99, Figure 1].

4.2.3 Evolutionarily diverged *eve* enhancers

I confronted the model with DNA sequence from the stripe 2 enhancers of 16 *Drosophila* species other than *melanogaster* (Figure 4.1C1-16), four of these enhancers, from *Drosophila persimilis, mojavensis, grimshawi*, and *willistoni* were identified in the course of this study. To do so, I used a publicly available BLAST tool [100, 101]. I used the *D. melanogaster eve* coding sequence (2R:5866746-5868284) as a query sequence and then scanned 25 kb centered on this region with the two conserved S2E sequences block-A (5'-AATATA ACCCAAT-3') and block-B (5'-TGATTATATCATCATAATAAATGTTT-3') which bracket the ends of S2E [38]. This provided sequence for S2E's from *mojavensis* and *grimshawi*. In the case of *willistoni*, block-B is found at the position 667 bp apart from the 5' end of block-A with two base pair changes. In the case of *persimilis*, it was not possible to obtain more than 753 bp of sequence 3' from block-A because the genomic database of this species lacks genomic sequence information beyond this point. We ran the model to predict gene expression from these putative enhancers and the results are shown in Figure 4.1C9 and C14-16). *Drosophila* and Sepsid species abbreviations are shown in Table 4.1.

In ten cases, stripe 2 expression was coextensive with the *D. melanogaster* stripe pattern (Figure 4.1C1-10). There is experimental evidence that *D. yakuba*, *D. pseudoobscura*, and *D. erecta* stripe 2 enhancers express coextensively with the *melanogaster* stripe 2 (Figure 4.1C1-3) [99, 2]. Our results are in substantial agreement with these findings (Figure 4.1C2-3). To our knowledge, no experimental observations have yet been made of the positions of stripe 2 driven by the remaining 13 *Drosophila* stripe 2 enhancers in *D. melanogaster*.

Abbreviation	Species	Figure index
yak	*Drosophila yakuba*	(4.1C1)
pse	*Drosophila pseudoobscura*	(4.1C2)
ere	*Drosophila erecta*	(4.1C3)
ore	*Drosophila orena*	(4.1C4)
tei	*Drosophila teissieri*	(4.1C5)
tak	*Drosophila takahashi*	(4.1C6)
mau	*Drosophila mauritiana*	(4.1C7)
sec	*Drosophila sechellia*	(4.1C8)
per	*Drosophila persimilis*	(4.1C9)
sim	*Drosophila simulans*	(4.1C10)
ana	*Drosophila ananassae*	(4.1C11)
vir	*Drosophila virilis*	(4.1C12)
pic	*Drosophila picticornis*	(4.1C13)
gri	*Drosophila grimshawi*	(4.1C14)
moj	*Drosophila mojavensis*	(4.1C15)
wil	*Drosophila willistoni*	(4.1C16)
cyn	*Sepsis cynipsea*	(4.1E1,7)
put	*Themira putris*	(4.1E2,8)
sup	*Themira superba*	(4.1E3,9)
dsp	*Dicranosepsis sp.*	(4.1E4,10)
min	*Themira minor*	(4.1E5,11)
pun	*Sepsis punctum*	(4.1E6,12)

Table 4.1: ***Drosophila* and Sepsid species abbreviations.** To distinguish *Drosophila* and Sepsid flies explicitly, I also used abbreviations having additional letters, D for *Drosophila*, S for *Sepsis* and T for *Themira*, in front of the three letter abbreviations. For example, *D. mel* indicates *Drosophila melanogaster*.

I also made predictions of expression patterns driven by regulatory sequences from the *eve* locus of six species of Sepsid flies. These species are about twice as evolutionarily distant from *D. melanogaster* as *D. melanogaster* is from the most distantly related *Drosophila* [102]. Our model, when challenged by Sepsidae DNA, predicts stripe 2, 3 and 7 expression driven by the corresponding Sepsid enhancers (Figure 4.1E1-12). Some of these predictions are confirmed (Figure 4.1E1-3 and 4.1E7-9). Stripe 2 and 3/7 enhancers from *T. cynipsea, T. putris* and *S. superba* have been tested for expression in *D. melanogaster* and have been shown to express *eve* stripes 2, 3, and 7 [102]; these are correctly predicted with the single exception of a failure to correctly predict the observed stripe 7 expression driven by the *cynipsea* 3/7 enhancer (Figure 4.1E7, arrow). The model also predicts that the Sepsid stripe 2 enhancers drive stripe 7 expression at levels which vary from species to species (Figure 4.1E1-6). It is confirmed experimentally that 78% of embryos containing the *S. cynipsea* enhancer and 55% of embryos containing the *T. putris* enhancer appear to have stripe 7 expression [102]. The model also predicts that stripe 2 expression from *S. cynipsea* and *T. putris* is shifted to the posterior (Figure 4.1E1 and 4.1E2) and that the shift is larger in *T. putris*, a point supported by published observations [102]. In this regard it is notable that our model predicts stripe 3 and 7 activity from the putative stripe 2 enhancer of *Dicranosepsis sp.* (Figure 4.1E4), and further predicts that in a *D. melanogaster* context this species' putative 3/7 enhancer drives stripe 7 expression at levels an order of magnitude greater than the maximum level of stripe 3 expression (Figure 4.1E10).

4.2.4 Gap and pair-rule enhancers

A more stringent test of the model is to predict the expression driven by the enhancers of *D. melanogaster* genes other than *eve*. Not all such reported enhancers can be tested, as some require TFs (such as pair-rule gene products) not considered in this study. I tested 15 enhancers of gap and pair-rule genes using the same TFs as were employed for the training set. Among the gap genes, I obtained correct predictions for expression driven by the pThb enhancer of *hb* (Figure 4.1F1) and the CD1 enhancer of *Kr* (Figure 4.1F2). With respect to the Runt 1_7 and 3_7 enhancers (Figure 4.1F3-4), I correctly predicted the expression of *run* stripe 3 and reduced expression of *run* stripe 7 compared to stripe 3, although in Runt 1_7 the predicted stripe 1 is coextensive with stripe 2 of the *run* protein pattern. The predicted pattern of *run* stripe 7 is shifted about 2 and 7 nuclei to the anterior of the native *run* stripe in Runt 1_7 and Runt 3_7 respectively. The predicted pattern of the h_str3_4 enhancer (Figure 4.1F5) is correct, as this enhancer drives an expression domain that does not contain the *h* 3-4 inter-stripe [103]. Ten additional enhancers from the genes *hb*, *kni*, *gt*, *run*, and *h* gave incorrect predictions (Figure 4.2C-L). In each case, expression in the correct domain was absent although in some instances small amounts of ectopic expression remained.

4.2.5 Large regulatory sequences

Our model is not limited to experimentally isolated enhancers. I attempted to predict expression driven by the approximately 4 kb of 5' and 3' noncoding DNA which respectively control stripes 2, 3, and 7 (Figure 4.1G1, parameters

from model 7; see Figure 4.2M for model 6 prediction) and stripes 4, 5, and 6 (Figure 4.1G2, parameters from model 1; see Figure 4.2N for model 6 prediction). Our initial prediction was completely incorrect, showing saturated blocks of expression without inter-stripes. When the threshold θ, the activation energy barrier of transcription initiation (Figure 3.5 Eq. 10), was increased by hand, I obtained the qualitatively correct predictions of gene expression of the 4 kb upstream and downstream of *eve* shown in Figure 4.1G1-2. Although it required manual tuning of a single parameter, I consider it highly significant that the predictive power of the model extends beyond single enhancers discovered by *in vivo* assays.

Chapter 5

Functional analysis of gene expression of four fusions

The accurate modeling of expression from the fusion constructs together with success obtained in prediction provide evidence that the model captures major elements of the underlying rules governing *eve* transcription. Given this level of predictive ability, it is also possible to use the model to understand how the interplay of multiple transcriptional mechanisms give rise to the very complex expression changes induced by removing the "spacer" DNA in the fusion constructs, M3_2 and M2_3. In this chapter, I will describe the result of model parameter analysis, a methodology of functional analysis and discuss the result of the functional analysis of the four fusion constructs. Except as noted, the standard model (Figure 3.10) is used for the rest of this dissertation study.

5.1 Model parameter analysis

Several features of the action of maternal and gap gene products were identified from inspection of the parameters of the four best models (Figure 5.1, Table B.1). First, Bcd and Cad are weak activators. In the models, multiple Bcd or Cad binding sites are required to drive gene expression. Second, Bcd shows strong pair-wise cooperativity. Even in the model having the weakest cooperativity of Bcd (model 1), two Bcd sites, for example, which fractional occupancies are 50% and 9% respectively at [Bcd]=20 (equivalent to 38% EL), can increase their fractional occupancies up to 84% and 71% respectively if they bind cooperatively. Third, Bcd has a strong capability to coactivate Hb for the synergistic activation of the transcription. Fourth, Hb plays the role of a strong activator in the presence of neighboring Bcd. Fifth, Kr, Kni and Gt appears to have considerable repressive power. For example, in model 6, a single Kr site for which the fractional occupancy is 50% can decrease the fractional occupancy f of any activator sites by a factor of two. These properties are in accord with qualitative findings in the experimental literature where it has been clearly established that multiple activators are required for gene expression [104, 105, 106], pair-wise cooperativity of Bcd occurs [8, 9], coactivation of Hb by Bcd has been observed [10, 39, 107] and that short-range repression by Kr, Kni and Gt also occurs [42, 12, 40, 43]. In addition, the model suggests the existence of coactivation of Hb by Cad. Note that D-STAT plays the role of a strong activator in all the models except model 6. The strongest Bcd cooperativity was found in model 2 and Bcd activation strength are significantly low in model 7 (Figure 5.1 and Table B.1).

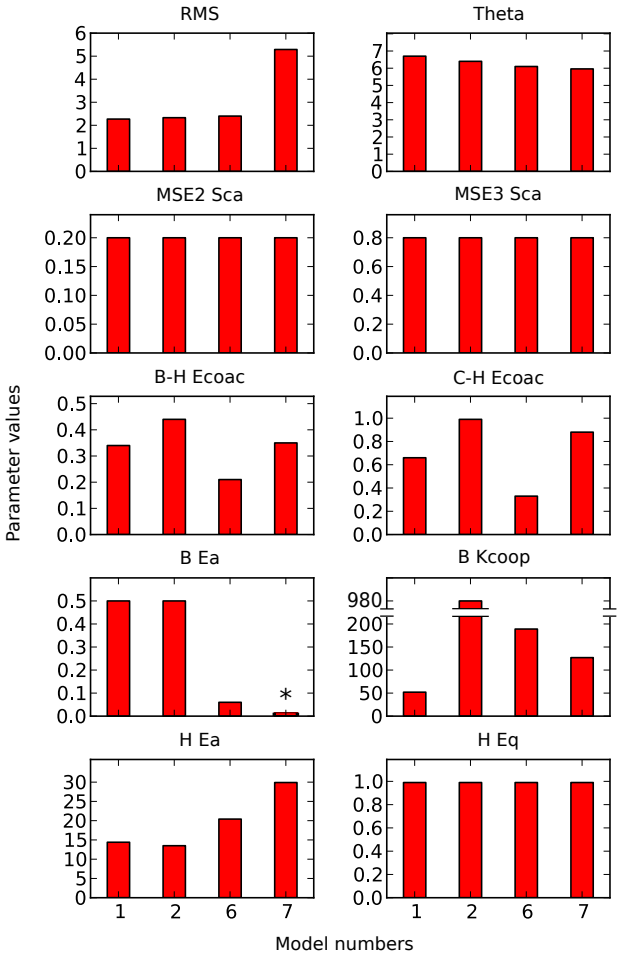


Figure 5.1: **Regulatory parameters of the four models.** Overall quality of the model fitting was evaluated first using the root mean square (rms). All of the fits have small rms score (< 6) but model 7 has relatively higher rms value than the others. The activation energy barrier theta (θ) is almost identical. Sca, Ecoac, Ea, Kcoop and Eq denote S^R, E^C, E^A, K^{coop} and E^Q respectively.

5.2 Functional analysis method

One of the key advantages of the transcription model is that it calculates the contribution of each TF, binding site, and even nucleotide to gene expression. The model utilized thermodynamics to calculate the effect of competition by steric hindrance and cooperative binding on the fractional occupancy of a TF site. Another advantage of the model is that the mechanisms are nonetheless separable and removable so that the relative contributions of each mechanism can be assayed as the consequences of removing a specific mechanism *in silico*.

Using these advantages, I devised an application, coded in python, named DyEVer (Dynamic Enhancer Viewer), that visualizes the contributions of individual binding sites on a target regulatory sequence as a 2-D map (Figure 5.2). In the DyEVer plot, the x and y axes correspond to the base pair position from TSS and the A-P axis of the embryo in % EL. The amount of the contribution of each activator binding site to gene expression is represented as a heat map of the decreased activation energy barrier ($\Delta\Delta A$). Because the map combines these two dimensions, it allows us to easily visualize which binding sites are contributing the most to initiate transcription at different spatial positions of the *Drosophila* embryo. In a macroscopic sense, it provides a new way to visualize which DNA fragments are actively involved in transcription and their contributions to the spatial patterning along the A-P axis.

Figure 5.2C shows an example of a DyEVer plot where activator contributions ($\Delta\Delta A$) are mapped at single binding site resolution for M3_2 as a function of A-P position and the base pair position relative to the TSS. $\Delta\Delta a$ for each activator binding site is shown in the central panel according to the key in Figure 5.3 B and the summed activation $\Delta\Delta A$ in the right hand bar. In the figure, three regions where the activation energy is significantly decreased

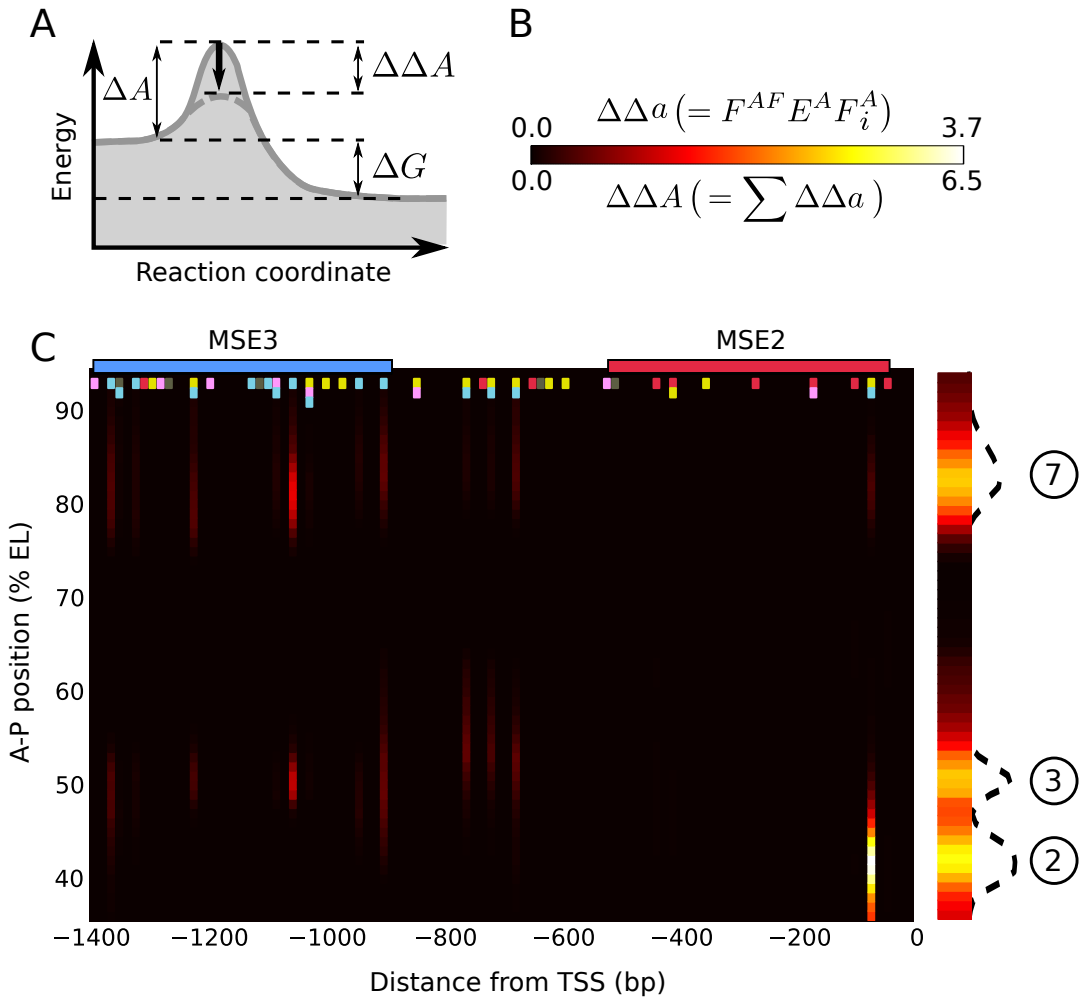


Figure 5.2: **DyEVer analysis of the M3_2 fusion.** (A) Illustration of a catalyzed reaction with free energy change ΔG and activation energy barrier ΔA. Catalysis by activators reduces the barrier by $\Delta\Delta A$. (B) A scale bar of two heatmaps used in (C) is shown. The $\Delta\Delta A$ heatmap applies to the vertical bars on the right hand side of these panels and the $\Delta\Delta a$ heatmap applies to the square panels in (C). $\Delta\Delta a = F^{AF} E^{A} F_i^{A}$; compare with Eqs. (8) and (9) in Figure 3.5. (C) Distribution of activation energy barrier changes at single binding site resolution for M3_2 as a function of A-P position on the embryo and number of basepairs 5' to the M3_2 TSS. The positions of MSE2 and MSE3 are schematically shown at the top.

are seen in the vertical bar in Figure 5.2C. The regions in the vertical bar corresponds to the multiple stripes driven by M3_2 (see circled 2, 3 and 7 in Figure 5.2C). The map shows that MSE2 and MSE3 are highly active in the stripe 2 and stripe 3 region respectively as expected from experimental observations [6, 26]. Using the DyEVer plot and the analytic capability of the model, I analyzed each region in which gene expression is significantly changed when the "spacer" between MSE2 and MSE3 is removed.

5.3 Functional analysis of fusion constructs

The fusions introduce six types of quantitative alterations in expression, each of which occurs in a small spatial region containing 2-3 nuclei, which I call a "zone" (Figure 5.3A). With respect to the M32 fusion compared to M3_2, in zone I stripe 2 expression is increased by a factor of almost four; in zone II the 2-3 inter-stripe is derepressed; in zone III stripe 3 expression is reduced; and in zone IV stripe 7 expression is increased. With respect to the M23 fusion compared to M2_3, in zone V stripe 2 expression is reduced and in zone VI stripe 3 expression is slightly increased (Figure 5.5A). I analyzed the causes of these effects by plotting the contributions to the activation $M = \Delta\Delta A$ as a function of position on the A-P axis and the regulatory sequence, where each position on the A-P axis defines a unique set of TF concentrations as shown in Figure 5.3B.

5.3.1 Control of gene expression in Zone I and II

By comparing DyEVer plots of M3_2 and M32, the binding sites responsible for driving stripe 2 expression were determined (Figure 5.3C-D). The analysis

indicates that the major source of activation of stripe 2 is from coactivated Hb bound at the hb-3 site which is coactivated by Bcd bound at the bcd-1,bcd-* and bcd-2 sites (Figure 5.3C-D and 5.4A). With respect to zone I, I found that the increase of gene expression in M32 is almost entirely the consequence of coactivation of two sites of bound Hb in MSE3 by Bcd bound to MSE2. This occurs because of the deletion of the "spacer" DNA between MSE3 and MSE2, which reduces the distance between the two Bcd sites in MSE2 and the two Hb sites in MSE3 from more than 400 bp to about 150 bp, permitting coactivation (Figure 5.3C-D, lower black arrows; Figure 5.3F, white arrow). This result confirms the previously proposed hypothesis of M3_2 and M32 [25].

These two footprinted Hb sites, hb-8 and hb-9, extend about 60 bp into MSE3, which is about 15% of its total length. These Hb sites are subject to repression by quenchers bound within 150 bp on their 5' side, including one site for Gt (Figure 5.4A). Thus, the same functional interactions characteristic of MSE2 now extend 200 bp into MSE3, about 40% of its length. These points indicate that in M32, 40% of MSE3 has been recruited to be a functional part of MSE2. This functional recruitment includes the setting of the anterior border of stripe 2 through repressive action emanating from a Gt site in MSE3 and another Gt site in MSE2. In M3_2 and M32 fusion, it is significant that, despite the novel synergistic activation of MSE3-bound Hb by MSE2-bound Bcd, the location of the anterior border of stripe 2 is unchanged in M32 compared to M3_2 based on the location of half maximum expression (Figure 3.9). Note that Bcd and Hb concentrations are essentially equivalent at the peak of the augmented stripe 2 and at its anterior border. This implies that a single predicted Gt binding site in MSE3 together with a single site

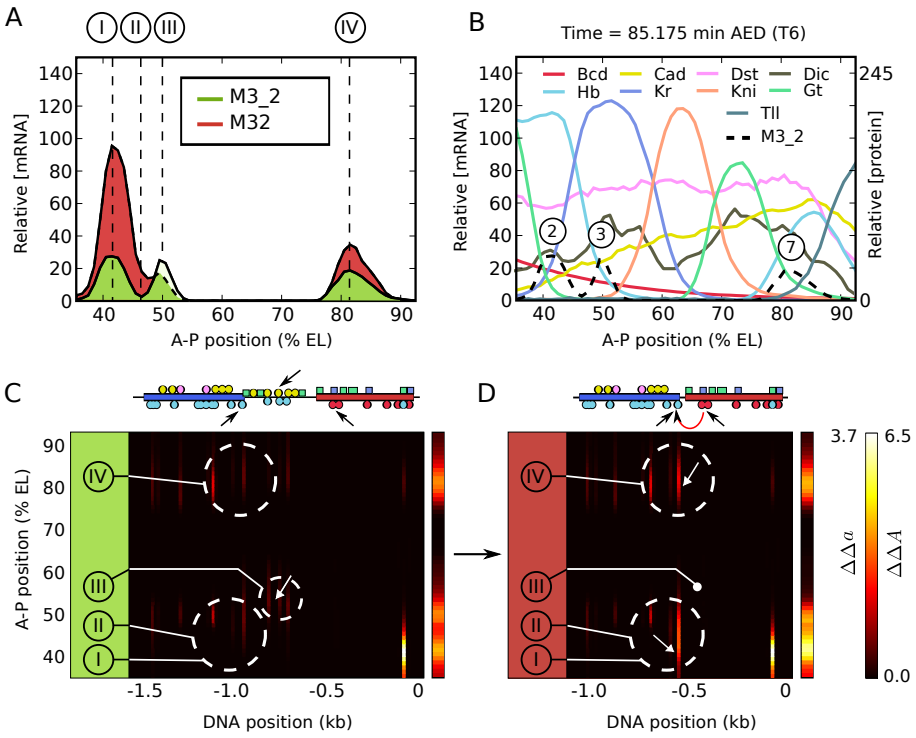


Figure 5.3: **Functional analysis of M3_2 and M32.** (A) The expression profiles driven by M3_2 and M32 are subdivided into four distinct zones I to IV for analysis as shown. Two additional zones V and VI involving expression changes between M2_3 and M23 are shown in Figure 5.5. (B) Expression levels of RNA expression driven by M3_2 together with regulating TFs at cellular resolution, as shown in the key. In the key, standard abbreviations are used except that Dst indicates D-STAT and Dic indicates Dichaete. (C) and (D) show a regulatory dissection of expression changes induced by removal of the "spacer" with activation represented. Selected binding sites for M3_2 and M32 are shown at the top of (C) and (D) respectively. The full set of binding sites is shown in Figure 5.4. The black arrows show binding sites involved in coactivation; the red arrow in (D) indicates the major coactivation interaction in M32. Circled areas indicate groups of binding sites critical for expression changes in different zones as described in the text.

in the stripe 2 enhancer are sufficient to repress anterior expression driven by the recruited portion of MSE3. Such robustness in border control would be impossible if repression were to occur only by steric competition. These results

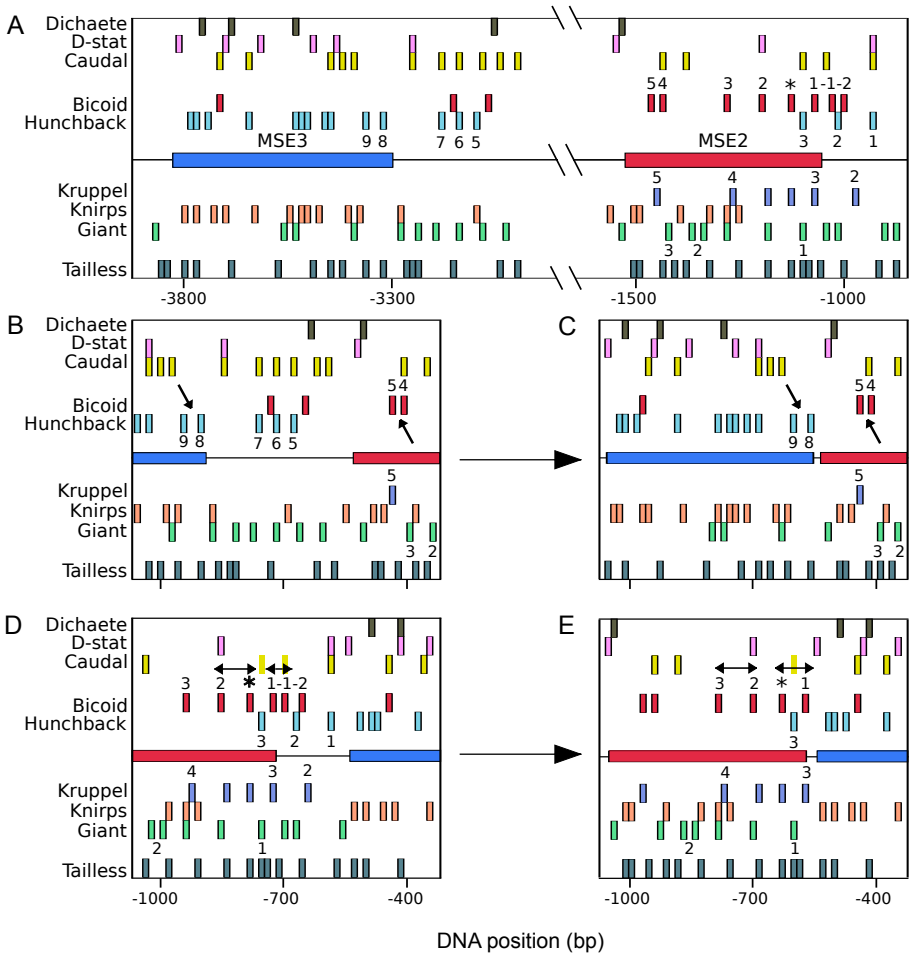


Figure 5.4: **Binding site map for model 6.** Every binding site used in model 6 is shown. Of these, all footprinted sites of the four TFs Bcd, Hb, Kr, Gt are numbered as the same way as in the original papers [30, 10]. (A) 5' upstream of *eve*. (B) M3_2 (C) M32 (D) M2_3 (E) M23. Key rearrangements of binding sites are indicated by black arrows. bcd-(–1) is a computationally identified site named in this work. bcd-* is evident on footprints [6], but was not named.

also demonstrate that the borders of enhancers are not intrinsic, but instead are determined by genomic context. In zone II, the derepression of the inter-stripe is a consequence of the fact that Kr binding sites are predominantly

distributed on the 3' end of MSE2, close to the hb-3 site (Figure 5.4). There is a single Kr binding site (kr-5) within range of the coactivated Hb bound to MSE3, and it is insufficient to provide complete repression in zone II.

5.3.2 Control of gene expression in Zone III and IV

The expression changes that occur in zones III and IV are connected with the fact the "spacer" in M3_2 is in fact a functional component of the 3/7 enhancer (Figure 5.3C, D). The reduction of stripe 3 expression levels in zone III is not recapitulated by fitting the model to the full set of seven constructs (Figure 3.9), but is found in fits made only to the four fusion constructs (Figure 3.11A). The cause of the change in expression in zone III is evident from inspection of Figure 5.3C (downward pointing arrow and white arrow), which show that the "spacer" contains Hb binding sites which are coactivated by Cad, the removal of which decreases expression. There are, in addition to activator sites, repressor sites in the "spacer" (Figure 5.4A). In zone IV, the model consistently gives a correct representation of the increase in stripe 7 expression in M32 compared to M3_2, and this is a consequence of the removal of repressor sites located in the "spacer". The effects seen in zones III and IV are critically dependent on the precise balance between activation, coactivation, and repression. This leads to residual ambiguity in how models with differing training data and parameter sets account for expression changes in these zones, but all models agree that the "spacer" plays a major functional role and is not an inert segment of DNA.

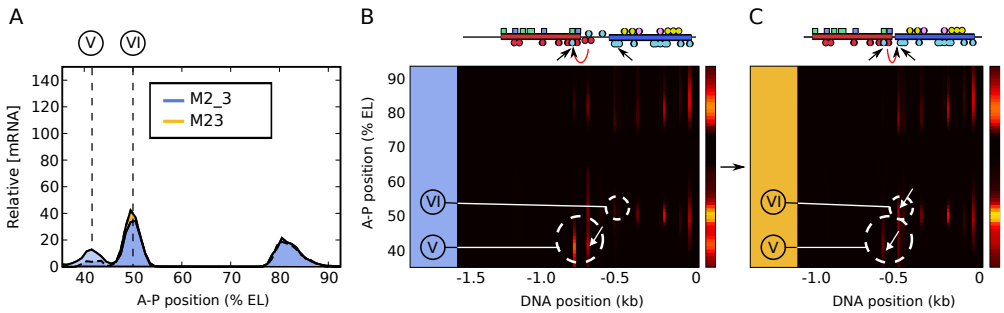


Figure 5.5: **Functional analysis of M2_3 and M23.** (A) Zones V and VI, the areas where expression changes occur between M23 and M2_3. (B-C) Distribution of activation energy barrier changes at single binding site resolution for M2_3 and M23 as a function of A-P position on the embryo and number of basepairs 5' to their TSS. In (B) and (C) the positions of MSE2 and MSE3 are schematically shown at the top. $\Delta\Delta a$ for each activator binding site is shown in the central panel according to the key in Figure 5.3B and the summed activation $\Delta\Delta A$ in the right hand bar. All footprints sites for Bcd, D-STAT, Hb, Kr and Gt are shown at the top of panels (B) and (C) except for the kr-2 site in the "spacer" (Figure S6D), which is very close to the 3' Bcd site in the "spacer". Computationally identified Cad binding sites in MSE3 and Bcd sites in the "spacer" are also shown. The black arrows in (B) and (C) indicate two Hb sites potentially subject to coactivation by Bcd. The red arrow indicates which of these sites is in fact subject to coactivation in a given construct. Circled areas highlight major changes in $\Delta\Delta A$ between M2_3 and M23, and the white arrows indicate which binding sites cause the changes seen in the circled areas. The distributions of TFs and further information about the diagrams in (B) and (C) are given in Figure 5.3B and its legend.

5.3.3 Control of gene expression in Zone V and VI

The "spacer" DNA in M2_3 is a component of the full stripe 2 enhancer S2E [31, 108], and its removal causes a severe diminution of stripe 2 expression in zone V and a much smaller increase of stripe 3 expression in zone VI, with stripe 7 unaffected (Figure 5.5A). These effects occur because the M2_3 "spacer" DNA contains two Bcd and two Hb binding sites (Figure 5.5B). The strongest Bcd site in MSE2 is bcd-1, and in M2_3 it preferentially establishes

pairwise cooperativity [9] with the next strongest site (bcd-(–1), Figure 5.4D), which is the most 5' of the two sites on the "spacer". In addition, a cooperative interaction exists between Bcd bound at the bcd-* (unnamed footprinted site; see Figure 5.4D and a DNA footprinting experiment result shown in Figure 3 in [10]) and bcd-2 sites. The net result is that in M2_3 these two pairs of cooperatively bound Bcd provide strong coactivation to two Hb sites, one of which is in the "spacer" (Figure 5.5B, zone V region and downward pointing white arrow). In M23, the absence of the "spacer" causes major rearrangements of pairwise cooperative interactions among bound Bcd molecules in MSE2 because bcd-(–1) is lost. Without the "spacer", Bcd bound at bcd-1 cooperates with Bcd bound at bcd-*, while Bcd bound at bcd-2 cooperates with Bcd bound at bcd-3 (compare Figure 5.4E and 5.4F). This configuration of cooperative interactions results in a lower fractional occupancy of Bcd compared to that seen in M2_3. Although in M23, Hb sites at the 5' end of MSE3 are recruited as a part of the stripe 2 enhancer by cooperatively bound molecules of Bcd in MSE2 (Figure 5.5C, white arrow), the net reduction in bound Bcd without the "spacer" causes a reduction of activation in zone V. These results highlight the importance of Bcd cooperativity between bcd-1 in MSE2 and bcd-(–1) in adjacent genomic sequence in providing correct levels of activation for MSE2. The contrasting small increase in expression in zone VI happens because the "spacer" also contains Kr sites (Figure 5.4D) which are heavily bound in the *Kr* expression domain which contains *eve* stripe 3 (Figure 5.3B). It is this difference in Kr levels which causes the opposite effect in zone VI compared to zone V.

Chapter 6

Functional conservation of *eve* stripe 2 enhancers

One striking characteristic of the *eve* stripe 2 enhancer is that its activity is extremely sensitive to some small, specific nucleotide changes. A change of only three nucleotides in the footprinted sites bcd-1 or bcd-2 or a change of five nucleotides in the single Hb site hb-3 cause a nearly total loss of stripe 2 expression [6, 32, see Figure 5.4 for the site names]. However, in contrast to the exceptional sensitivity to small mutations, it is also seen that the S2Es of three different *Drosophila* species, *D. yakuba*, *D. erecta* and *D. pseudoobscura*, containing various substitutions, additions and deletions of nucleotides both inside and outside the footprinted binding sites drive almost identical stripe 2 expression in the *Drosophila melanogaster* blastoderm embryo [38, 2]. Furthermore, even S2Es of distantly related Sepsid flies, *Sepsis cynipsea* and *Themira putris*, whose binding sites are almost completely rearranged relative to *D. mel*, can still produce an identical stripe 2 pattern in the *D. melanogaster* embryo [102]. These results clearly demonstrate that the stripe 2 enhancer of

eve has substantial structural flexibility in carrying out a common function.

These findings provoke several fundamental questions about the underlying logic governing transcription. How can such highly diverged S2E sequences drive an almost identical expression pattern? What are the underlying mechanisms ensuring robust gene expression given the gross perturbations in DNA sequence? There have been substantial advances towards understanding the conserved function of various *eve* stripe 2 enhancers on both sequence and gene expression levels [38, 2, 102], however the molecular mechanisms compensating for the diverged sequences remain to be elucidated. Because the mechanisms underlying the conserved function can only be inferred from the knowledge of the complex protein-DNA and protein-protein interactions that actually occur on the regulatory DNA, it wouldn't be possible to characterize the mechanisms without assaying simultaneously operating interactions. With the *in silico* transcription system, however, it is possible to approach this problem. The model is able to track the complex protein-DNA and protein-protein interactions acting in concert with clear description of their activities. I investigated conserved gene expression driven by *eve* S2Es from four *Drosophila* species, *D. melanogaster* (*D. mel*), *D. yakuba* (*D. yak*), *D. erecta* (*D. ere*) and *D. pseudoobscura* (*D. pse*) and two Sepsid species, *S. cynipsea* (*S. cyn*) and *T. putris* (*T. put*) using the standard model, which is used for both prediction of gene expression of 23 *Drosophila* and Sepsid species and the functional analysis of the four enhancer fusions, M3_2, M32, M2_3 and M23 (Section 4.2.3 and 5.3).

I considered the six different S2Es an ideal test case for a comprehensive sequence-function analysis for the several reasons. First, the fact that the differential gene expression driven by the six S2Es is entirely the result of dif-

ferences in the *cis*-regulatory sequences acting in a common *trans*-environment allows us to apply the large body of knowledge about the *trans*-acting factors, regulatory proteins and transcription machinery in *D. mel*. Second, all of the six S2Es drive gene expression at levels we can measure *in vivo* [38, 2, 102]. The expression pattern itself is the most essential and primary information that we need in order to decipher the *cis*-regulatory code driving it. Third, all of the six S2Es drive a nearly identical spatiotemporal expression pattern in the *D. mel* embryo in spite of their substantial sequence differences. The conserved expression pattern of these regulatory sequences from six different species in the *D. mel* embryo allows us to investigate the molecular mechanisms ensuring functional conservation and the structural flexibility of the regulatory sequences.

In this chapter, I focus on the enhancer structure, its function, and the molecular mechanisms connecting them. I will briefly describe the sequence differences in the S2Es of 17 *Drosophila* and 6 Sepsid species, which drive similar expression patterns in *in silico* predictions (Figure 4.1), and describe a functional binding site analysis method I devised for studying the conserved function. I then apply the method to the six experimentally tested S2Es from *Drosophila* and Sepsid species. Finally, I will propose molecular mechanisms responsible for the posterior shift of the Sepsid stripe 2 expression observed in the *D. mel* embryo.

6.1 Structural differences of *eve* stripe 2 enhancer

S2E sequences from the four *Drosophila* species, *D. mel*, *D. yak*, *D. ere* and *D. pse*, are substantially diverged. There are large insertions and deletions

in the sequences between known factor-binding sites, single nucleotide substitutions and deletions in binding sites, and complete gain or loss of binding sites [38, 2]. Furthermore, it has been reported that the Sepsid S2Es almost completely lack sequence similarity to *D. mel* S2E [102]. With this low level of similarity, it is beneficial for the sequence comparisons if there are highly conserved sequences surrounding the S2Es between different species. Fortunately, such conserved blocks were found in the four *Drosophila* species, *D. mel*, *D. yak*, *D. ere* and *D. pse*, denoted block-A (5'-AATATAACCCAAT-3') and block-B (5'-TGATTATATCATCATAATAAATGTTT-3') [38]. Furthermore, thirteen additional *Drosophila* S2Es that were used in this study also have the two conserved blocks except block-B for *D. wil* and *D. per*. In the case of *D. wil*, block-B is found 667 bp away from the 5' end of block-A with two base pair changes. For *D. per*, it was not possible to discern the block-B as the genomic database of this species lacks sequence information beyond 766 bp of sequence 5' from 5' end of block-A. In this study, I used sequences bounded by the two conserved blocks for the sequence-function analysis. In the case of *D. per*, the 766 bp fragment was used. Sepsid stripe 2 enhancers lack the two conserved blocks. In this special case, I utilized the sequences obtained from the literature [102].

A simple but effective way to quantify the structural differences between DNA sequences is to calculate minimum edit distance (MED), also called Levenshtein distance [109]. The Levenshtein distance between two sequences is defined as the minimum number of edits needed to transform one sequence into the other using three allowable edit operations: insertion, deletion, or substitution of a single nucleotide. As seen in Figure 6.1, the sister taxa *D. sec*, *D. sim* and *D. mau* have only a small MED from *D. mel*, less than 40

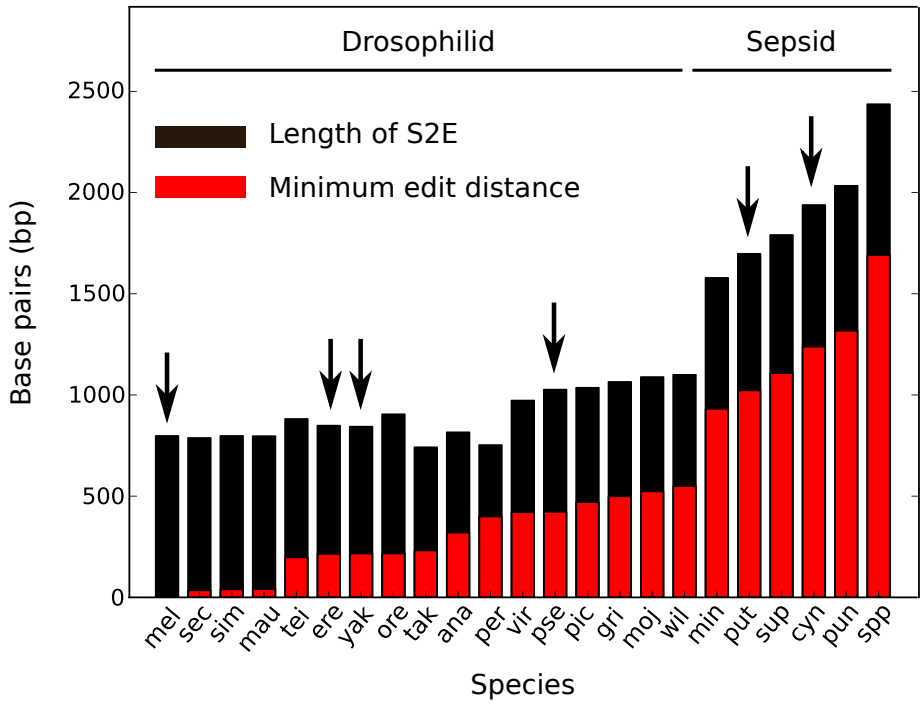


Figure 6.1: **Minimum edit distances of S2Es from 23 species.** Species are aligned according to the MED from S2E of *D. mel*. Black downward arrows indicate the experimentally characterized S2Es that drive almost identical gene expression in the *Drosophila melanogaster* blastoderm embryo.

in S2E ($<$ 5% of their sequence). However, *D. yak*, *D. ere* and *D. pse* S2Es have MEDs of 216, 214 and 422, which are equivalent to 26%, 25% and 41% of their total sequences respectively. In the case of Sepsid S2Es, their MEDs are almost triple that of *D. pse* S2E. *S. cyn* and *T. put* S2Es have 1238 and 1022 MED, respectively. Note that as Sepsid enhancers are not defined by the two conserved blocks they might contain additional sequences. Nevertheless, these MED quantifications demonstrate that significant structural differences exist in the S2Es of various species including the six S2Es that I investigate in this chapter.

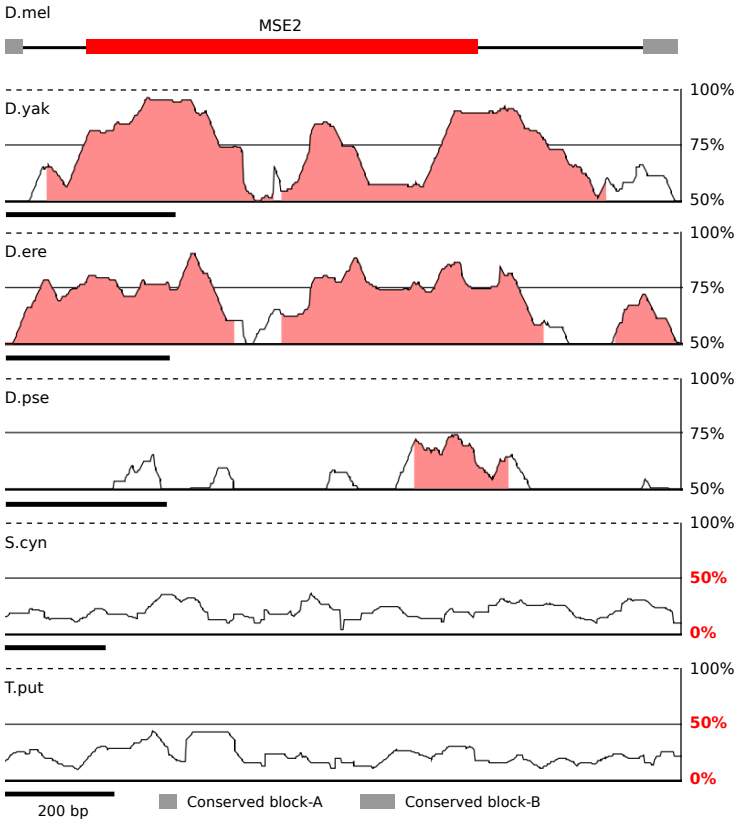


Figure 6.2: **Sequence conservation in six S2Es.** The conserved sequences are shown in pink. *D. mel* S2E is used as a reference sequence. Note that the scale of the genomic intervals plotted differs between panels (black bar = 200 bp).

The divergence among the six S2Es is also shown when the sequences are aligned using LAGAN, a global pair-wise alignment tool [110]. The percent identity of the S2Es in a 100 bp window at each base pair is calculated and visualized in Figure 6.2. In this plot, a conserved segment (pink colored area) is defined to be a region in which every contiguous subsegment of 100 bp is at least 70% identical to its paired *D. mel* sequence. These segments are merged to define the conserved region. In the case of the *D. yak* and *D. ere* S2Es, MSE2 is highly conserved at both its 5' and 3' ends, while *D. pse* contains

only a narrow conserved region corresponding to around 3' end region of *D. mel* MSE2. For the Sepsid S2Es, both *S. cyn* and *T. put* completely lack statistically significant sequence conservation (Figure 6.2). Compared to *D. mel*, the total conserved regions of *D. yak*, *D. ere*, *D. pse*, *S. cyn* and *T. put* are 81%, 79%, 14%, 0% and 0%, respectively. Note that the only conserved region shared among four species *D. mel*, *D. yak*, *D. ere* and *D. pse* corresponds to 3' end of *D. mel* MSE2 and its neighboring genomic fragment located between the 3' end of MSE2 and the conserved block-B (Figure 6.2). This may indicate the functional importance of the conserved region. However, these results, together with the previous works [38, 102], clearly show that the functional conservation of stripe 2 expression does not arise from sequence conservation of S2E.

In spite of these sequence differences in various S2Es, the conserved S2E driven spatiotemporal expression is correctly predicted by the model (Figure 6.3). The model recapitulates conserved gene expression driven by the four *Drosophila* S2Es in the same position of native *eve* stripe 2. This is in agreement with experimental results showing that RNA expression driven by S2Es of the four species, *D. mel*, *D. yak*, *D. ere* and *D. pse* are co-localized with native *eve* stripe 2 [38]. In regard to Sepsid S2Es, the model correctly predicts that the stripe 2 expression driven by *S. cyn* and *T. put* S2Es is shifted posteriorly with the degree of shift larger in *T. put* than *S. cyn* in the *D. mel* embryo [102]. Given this level of predictive ability and concordance between the model and the experimental results, I was able to investigate with confidence the conserved function between the six S2Es and the functional differences of the Sepsid S2Es compared to *Drosophila* S2Es as described in the following sections.

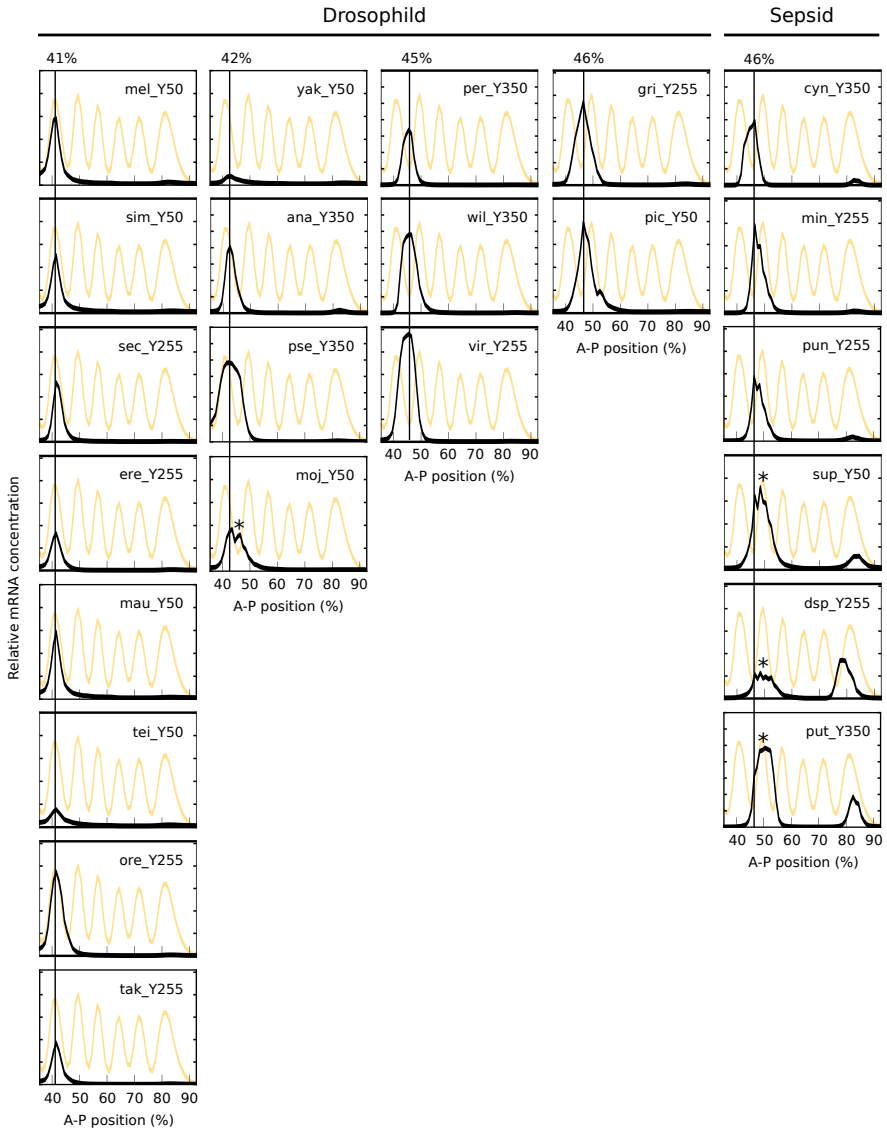


Figure 6.3: **Prediction of *eve* stripe 2 expression of 22 species.** Each panel has three letter abbreviation for the corresponding species (see Table 4.1 for details). The following numbers after the abbreviations are the height of the y axis.

6.2 Functional binding site analysis method

The fundamental functional unit of a regulatory sequence is a TF binding site. In S2E of *D. mel*, the contributions of the TF binding sites to gene expression are quite diverse. A disruption of bcd-1 by substituting three nucleotides in its binding site sequence, for example, causes a severe reduction in stripe 2 expression while a five nucleotide substitution in bcd-3 has little effect on stripe 2 expression [6]. Disruption of both gt-1 and gt-3 by nucleotide deletions slightly affects the anterior border of stripe 2. However, a deletion of thirteen contiguous base pairs in gt-2 causes noticeable derepression at the anterior border of stripe 2 [32]. Furthermore, many TF sites in the *D. mel* S2E are relevant only in a specific region of the embryo because their TFs such as Gt and Kr are only expressed locally. Therefore, in order to understand the sequence-function relationship of the different S2Es, we must determine the functional binding sites essential to stripe 2 expression at each A-P position and then compare the functional binding sites, their configurations and molecular interactions on the functional binding sites between species.

Using the *in silico* transcription system, I first tried to identify the highly active functional binding sites in *D. mel* S2E at three different positions—the anterior, peak and posterior border of *eve* stripe 2 on the A-P axis at T6. I defined a highly active functional site as a TF binding site essential for the full levels of stripe 2 expression driven by S2E. I reasoned that as the altered gene expression is due to the change in activator activity on the regulatory DNA and because the repressors, coactivators, and cooperatively interacting factors bound close to or overlapping the activator binding sites regulate the activator activity, I expected that the functional binding sites will be found in clusters. Because S2E contains multiple activator sites, S2E should have

multiple functional clusters and all or part of these clusters should be critical for gene expression.

To identify the highly active binding sites in *eve* S2E at a given A-P position, I applied the following criteria. First, I determine the top contributing activator binding sites such that the sum of their contributions, quantified as the decrease in the activation energy barrier (θ) of transcription initiation (Figure 3.5, Eg. 10 and Figure 5.2A) to gene expression is equal to or more than 80% of the total decreased energy at a given A-P position. Second, I identify the regulator binding sites repressing, coactivating, cooperatively interacting with and overlapping the top contributor sites. Third, I consider the top 80% of contributors and their regulators as the highly active functional binding sites. Fourth, any regulator sites whose net fractional occupancies are equal to or lower than 5% are discarded. I then inspect the arrangement of the top contributors and their regulator sites and investigate their protein-DNA and protein-protein interactions. This method, termed HOT (Highly active sites Of TFBSs) analysis, allows us to isolate and compare the functional binding sites between different S2Es at single nucleus resolution. In this study, I performed HOT analysis of S2Es at three positions on the A-P axis, 38%, 41% and 44% EL, which represent the positions of the anterior border, the peak and the posterior border of stripe 2 respectively.

Figure 6.4 shows an example of a HOT plot where the highly active functional binding sites on *D. mel* S2E are visualized. Activators and repressors are displayed above and below the middle horizontal axis, respectively. In the case of Hb sites, both activating and repressing sites are displayed (Figure 6.4A). Figure 6.4B shows which binding sites are highly active at a given A-P position and their spatial arrangement. This plot reveals that, among a hundred

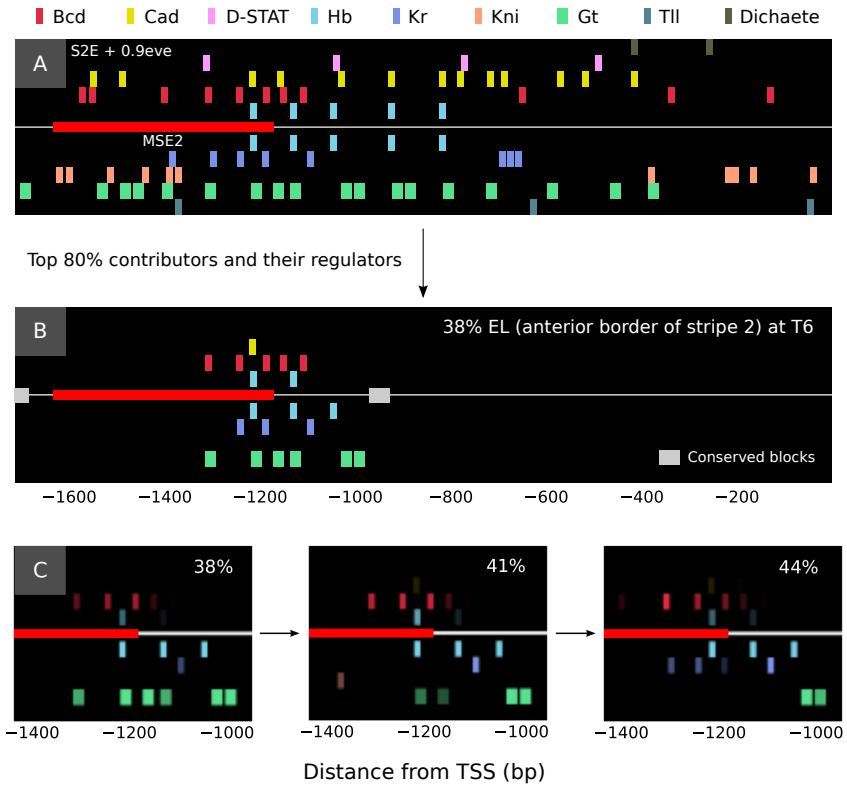


Figure 6.4: **Functionally active binding site analysis.** (A) All of the predicted binding sites in the 1.6 kb *eve* upstream regulatory DNA are shown. The MSE2 region and two conserved blocks are marked. (B) Top 80% contributors and their regulators (coactivator and repressor) are shown. The highly active binding sites driving stripe 2 expression are clustered at the 3' end of MSE2. Other sites, including bcd-4 and bcd-5, provide the rest of the contribution to stripe 2 expression but they are not shown. (C) TFs bind to a different set of binding sites at different A-P positions. 38%, 41% and 44% A-P position correspond to the anterior, peak and posterior border of stripe 2 respectively. In order to highlight the dynamic binding of TFs, levels of their fractional occupancy are represented by adjusting transparency such that the transparency of the binding sites is inversely proportional to the fractional occupancy. Bcd and Hb binding sites are highly occupied at the peak position (41% EL) while Gt and Kr bind strongly to their binding sites at the anterior and posterior border respectively.

footprinted and computationally identified binding sites, only a small subset are highly involved in transcriptional control at a given A-P position (Compare Figure 6.4A and B). For example, stripe 2 expression is almost entirely driven by 7 activator sites, bcd-2, bcd-*, bcd-1, bcd-(–1), bcd-(–2), hb-3 and hb-2 (Figure 5.4 and Figure 6.4). They are tightly clustered in the 3' side of S2E, from –1.3 kb to –1.0 kb. The region drives gene expression by synergistic coactivation of Bcd and Hb. In addition, multiple functionally active Gt and Kr binding sites are located in the region (Figure 6.4B). The net fractional occupancies of the highly active binding sites change dynamically according to their positions on A-P axis (Figure 6.4C). At the anterior border of stripe 2, Gt binds strongly to multiple sites in the region and antagonizes activator binding while, at its posterior border, Kr binds strongly and represses gene expression. On the other hand, at the peak stripe 2, both Gt and Kr binding is relatively weak such that the synergistic activation of Bcd and Hb is maximized. These results strongly support the previous work on *eve* stripe 2 regulation [10, 6].

I initially predicted that the functional binding sites would be found in clusters, in which functional activators are surrounded by their regulators. HOT analysis reveals two regions of highly active clusters (Figure 6.5B, C). The two clusters have very similar structure in that each of them contains footprinted Hb sites surrounded by the multiple Bcd, Gt and Kr binding sites. Notably, the clusters share many of their regulator sites, for example five Bcd sites, bcd-2, bcd-*, bcd-1, bcd-(–1) and bcd-(–2) and three Gt sites, including the footprinted site gt-1, are located in the effective range capable of coactivating or quenching the two Hb sites, hb-3 and hb-2 (Figure 6.5A). I denote these two clusters as the hb-3 and hb-2 cluster hereafter.

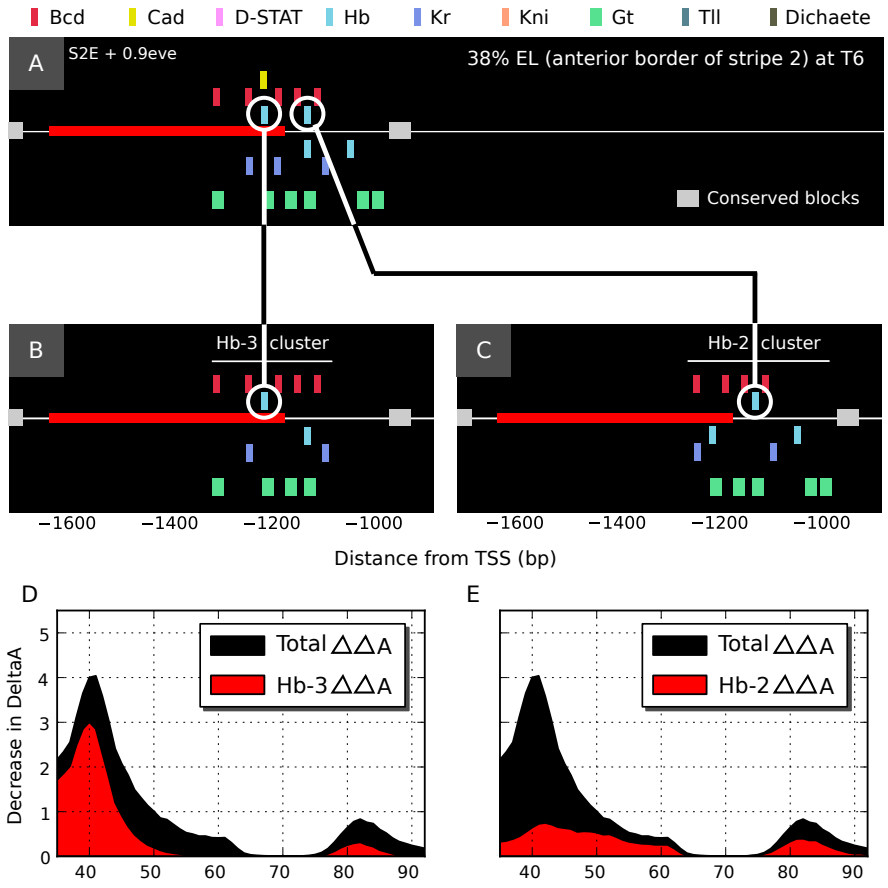


Figure 6.5: **Two functionally active clusters in S2E.** (A) Highly active functional binding sites are clustered in *eve* upstream region between –1.4 kb and –1.0 kb. (B, C) The highly active binding sites can be divided into two functionally independent clusters, the hb-3 cluster (B) and the hb-2 cluster (C). They share some functional binding sites but the model predicts that the contributions of the two functional clusters to gene expression are significantly different. (D) Decreased activation energy barrier driven by the hb-3 cluster are shown along A-P axis. The hb-3 clusters are highly active in stripe 2 region. (E) Decreased activation energy barrier driven by hb-2 cluster is shown. The hb-2 cluster behaves differently than the hb-3 cluster. It is activated in a broad anterior region between 35% and 65% EL and the posterior region between 75% and 90% EL. Note that both the hb-3 and hb-2 clusters are activated in the stripe 7 region.

The model indicates that the transcriptional activity of the hb-3 cluster is quite distinct from the action of hb-2 cluster despite the similar binding site structure (Figure 6.5D, E). The hb-3 cluster alone provides more than 70% of the total contribution to stripe 2 expression while hb-2 cluster is involved in decreasing the activation energy barrier in a broad anterior domain covering the region of *eve* stripes 2, 3 and 4. This can be interpreted as the 200 bp fragment containing the hb-3 cluster behaving as a "mini stripe 2 enhancer" such that the 3' side fragment of S2E drives transcription in the stripe 2 region. This result shows that MSE2 is not completely minimal in the sense that an even smaller region can drive stripe 2 expression. This is in agreement with a previous work showing that smaller regions of DNA within MSE2, containing a part of the hb-3 cluster, can drive weak and variable stripe 2 expression [10, 6]. Furthermore, this analysis shows that the region outside of MSE2 containing hb-2 cluster, from –1.1 kb to –1.0 kb, is necessary for normal stripe 2 expression driven by S2E. Interestingly, in the model, the hb-3 and hb-2 clusters are weakly activated in the posterior domain where *eve* stripe 7 is formed. This result supports the experimental result that wild-type levels of stripe 7 expression require an additional fragment covering the 3' side of S2E [26] and suggests that the small regions in S2E might be actively involved in maintaining wild-type levels of stripe 7 expression *in vivo*.

Hereafter I restrict the functional binding site analysis of S2Es to the top 80% of contributors and their regulator sites. These approximate all essential functional binding sites necessary for complete stripe 2 expression *in silico*. I use a center-to-center distance when I measure the distance between two binding sites. All of the functional binding sites and their affinities were calculated by the *in silico* transcription system. For the Bcd and Hb sites, the

key contributor sites of stripe 2 expression, their binding affinities are further investigated using empirical data derived from state-of-the-art microfluidics experiments (MITOMI) [111]. The MITOMI device measured the binding energy change of observed mutations in Bcd and Hb binding sites, therefore, the MITOMI based PWMs (a courtesy of Bin He) are ideal for estimating actual binding affinity of Bcd and Hb sites.

6.3 Structure–function analysis of *D. mel* S2E

In order to understand the functional conservation of S2Es of four different species, I first determined the relationship between structure and function of *D. mel* S2E. Inspection of the occupancies of functional clusters of *D. mel* S2E at three positions, the anterior border (38% EL), the peak (41%) and the posterior border (44%) of *eve* stripe 2 area reveals the precise spatial dynamics of protein-DNA and protein-protein interactions (Figure 6.4, Figure 6.5 and Figure 6.6). The local arrangements of the two functional clusters are highlighted in Figure 6.6B. There are 5 functional Bcd sites in the effective coactivation range of the top contributors (see red dashed line in Figure 6.6). At the anterior border, Gt is competing with Hb in the hb-3 and hb-2 clusters but at the posterior border, it no longer tightly binds to the overlapping site while Kr sites are highly occupied.

I investigated the mechanism of repression by Gt at the anterior border where Gt tightly binds to multiple site in both the hb-3 and hb-2 clusters (Figure 6.5B). The footprinted site gt-1 overlaps hb-3 and a computationally identified site, gt-(–1), overlaps hb-2 [6, Figure 5.4]. If the gt-1 site is completely removed *in silico*, while the binding affinity of the overlapping hb-3

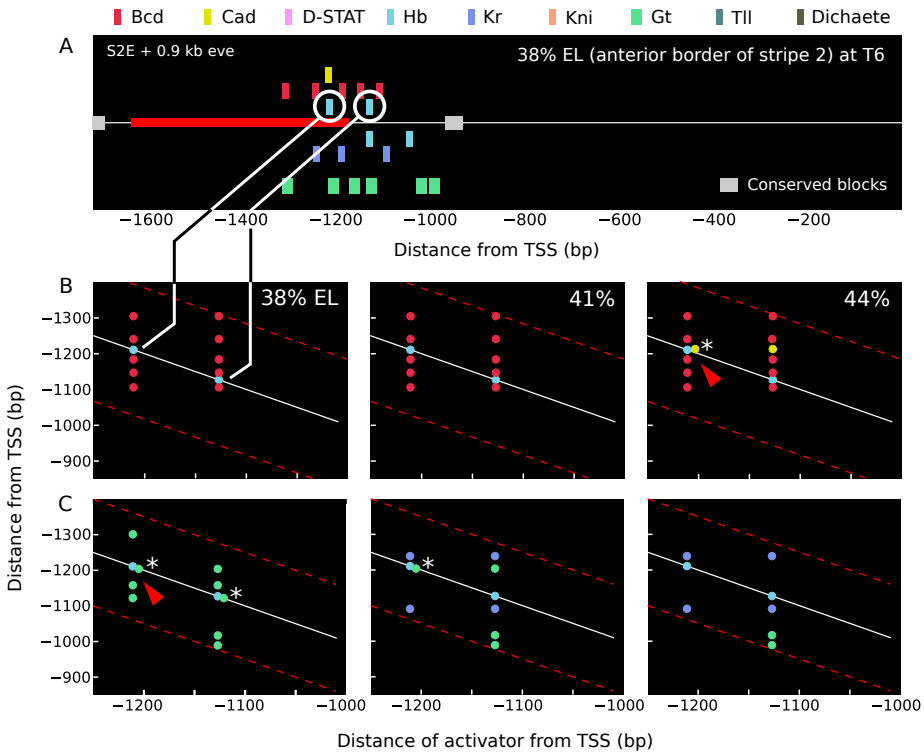


Figure 6.6: **Functional cluster analysis of *D. mel* S2E.** (A) Top 80% contributors and their regulators are shown. The highly active functional binding sites are clustered. (B) Functional hb-3 and hb-2 clusters are visualized at three different A-P positions. The top activators are placed on x-y coordinates according to their positions from TSS. However, their regulators are aligned vertically on top or below their target activator such that the plot allows us to easily determine the differences of the functional binding site arrangements in the clusters. Distance values on x-axis applies only to the top activators. Regulators at the same x-axis indicate that they belong to the same functional cluster. Regulators on the same y-axis indicate that they are identical binding sites. The red arrows indicate the competitive binding, showing the interaction is not static. Note that coactivators and repressors form homotypic clusters.

is left intact, noticeable derepression in the 1-2 inter-stripe region is observed (Figure 6.7B) and the overall levels of stripe 2 expression are increased. Nevertheless, the anterior border of stripe 2 is still maintained. Except for the

derepression in the inter-stripe, this result agrees the previous experimental work suggests that the levels of gene expression are increased but the anterior border is still present when a fourteen contiguous nucleotide deletion was created in gt-1 [32]. This result shows that the competition between the Gt and Hb binding sites gt-1 and hb-3 is a component of the anterior border formation but there must be an additional repression mechanism to maintain the anterior border.

In order to access the contribution of Gt mediated short-range quenching to the anterior border, I ran the *in silico* transcription system while specifically disabling the short-range quenching ability of Gt. Although the competition between Bcd and Gt is still operating on the overlapping Bcd and Gt sites, the anterior border is abolished in the inter-stripe region of *eve* stripes 1 and 2 (Figure 6.7C). This result shows the importance of the short-range quenching mechanism for the anterior border formation. Further *in silico* experiments show that the robust anterior border formation depends on short-range quenching by multiple Gt binding sites. The deletion of a single Gt binding sites has no significant effect while the deletion of multiple Gt sites abolish the anterior border of stripe 2 (Figure 6.8C).

At the posterior border, a completely different configuration of repressor occupancies is observed. The overlapping Gt binding sites are no longer populated by Gt. However, the three Kr binding sites are highly occupied. Interestingly, none of them overlap the hb-3 and hb-2 sites. Instead, the Kr sites overlap bcd-*, bcd-1 and bcd-(–2) so that they indirectly repress the Hb sites by competing on the overlapped binding sites and then repressing the Hb and Bcd sites through short-range quenching. If the model allows Kr mediated competition but disallows short-range quenching interactions, the

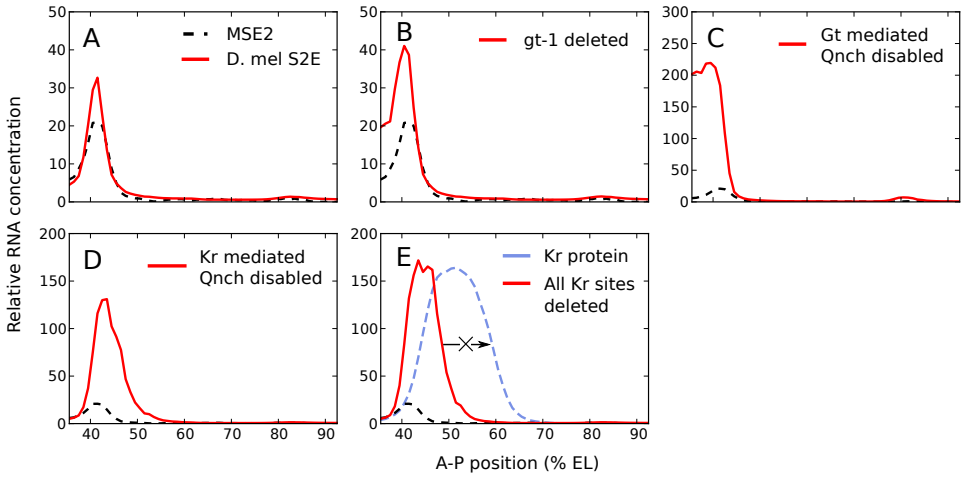


Figure 6.7: **Multi-tier mechanisms of repression.** (A) Predicted gene expression driven by *D. mel* S2E. The model predicts higher levels of gene expression than that of the observed MSE2. (B) Footprinted site gt-1 is deleted *in silico*. The anterior border of stripe 2 is still maintained. (C) Gt mediated short-range quenching is disabled while Gt mediated competition is still operating *in silico*. The anterior stripe 2 border is abolished in the inter-stripe region of *eve* stripes 1 and 2. (D) Kr mediated short-range quenching is disabled *in silico*. The posterior border expansion is seen but the border is not abolished. (E) The posterior border of stripe 2 is not abolished even when the all Kr sites in S2E are deleted *in silico* (see an black arrow). The Kr protein profile is also shown for a reference.

posterior border is extended (Figure 6.7D). This indicates the importance of Kr-mediated short-range quenching for the establishment of the posterior border. Unlike the anterior border however, the model shows that the posterior extension is strongly constrained by diminishing levels of Bcd and Hb concentration. Although shifted, the posterior border of stripe 2 is still maintained even when the all Kr sites are completely deleted (Figure 6.8D), indicating the multi-tier mechanisms of the posterior border formation. This result agrees with the finding that there is no loss of the posterior border when the wild type MSE *lac*Z fusion gene is expressed in Kr– embryos as visualized by antibody

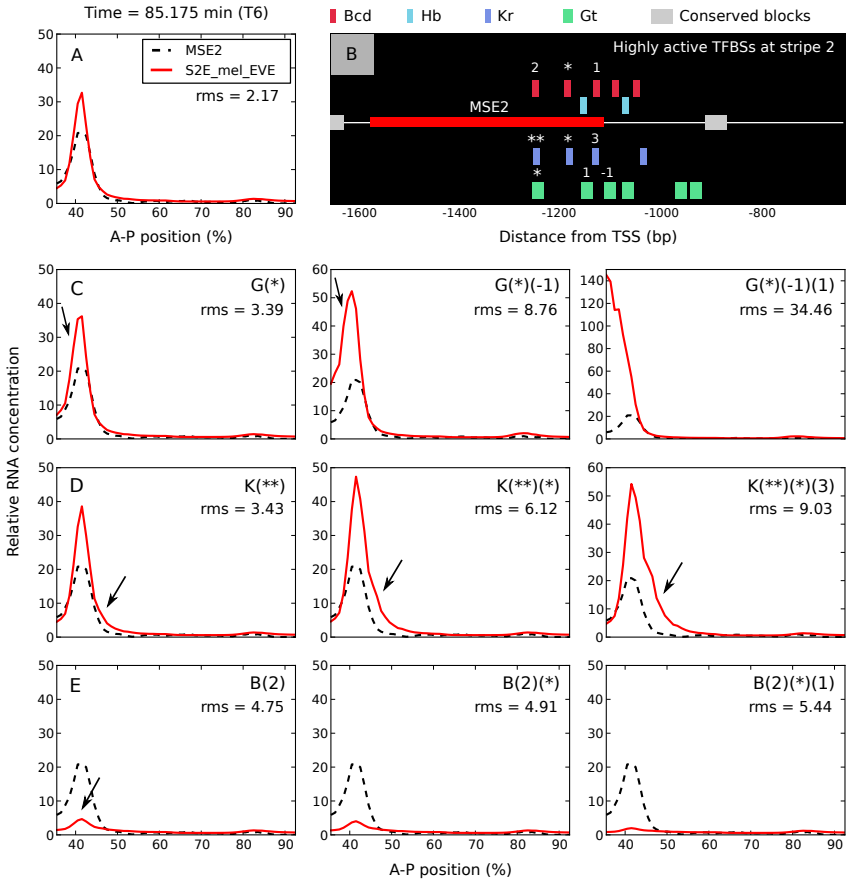


Figure 6.8: **Robust stripe 2 border formation.** (A) Predicted gene expression driven by *D. mel* S2E. The model predicts higher levels of gene expression than that of the observed MSE2 (shorter than S2E). (B) Top 80% contributors and their regulators are shown except for Cad, Kni and Tll. (C) Functionally active Gt binding sites are removed one by one from the left except gt-1. gt-1 overlaps the hb-3 site and it is removed last. Until three Gt sites are removed, the anterior border of stripe 2 is not abolished (see black arrows). (D) Functionally active Kr sites are removed one by one from the left. The posterior border of stripe 2 is shifted posteriorly but the border is not abolished even when the three Kr sites are all deleted (see black arrows). (E) Functionally active Bcd sites are deleted one by one from the left. Deletion of a functional activator site causes more dramatic change in gene expression than repressor sites. A single binding site deletion is enough to cause near total loss of gene expression (see black arrows).

staining [6]. In conclusion, these results shows that S2E of *D. mel* has an intrinsic tolerance to some mutations in carrying out its essential function—the establishment of *eve* stripe 2 in the *D. mel* embryo. Furthermore, these results demonstrate that the structural flexibility of S2E arises, in part, from the multi-tiered repressive mechanisms by multiple Gt and Kr molecules bound in the clusters and diminishing levels of Bcd and Hb concentration.

6.4 Structure–function analysis of four *Drosophila* S2Es

In the previous sections, I investigated the gross sequence differences in S2Es found in four *Drosophila* species, *D. mel*, *D. yak*, *D. ere* and *D. pse* and two Sepsid species, *S. cyn* and *T. put*. I then identified and characterized two highly active clusters of functional binding sites, the hb-3 and hb-2 clusters, in *D. mel* S2E. In this section, I identify the highly active clusters in the *D. yak*, *D. ere* and *D. pse* S2Es and compare the structures of the functional clusters to that of *D. mel* S2E in detail. I then investigate whether there are molecular mechanisms that compensate for the structural differences in the four different S2Es.

HOT analysis of the three *Drosophila* S2Es reveals two and sometimes three highly active clusters. The top two functional clusters in these species share many functional binding sites with the *D. mel* S2E, although their configurations are different from each other (compare the clusters in each panel of Figures 6.9 and 6.10). In the S2Es of all four species, each cluster contains one Hb site coactivated by multiple Bcd sites (Figure 6.9) as well as multiple Gt and Kr binding sites in the effective range of quenching the Hb and Bcd

sites (Figure 6.10). Because the top two clusters in the S2Es of *D. yak*, *D. ere* and *D. pse* contain the hb-3 and hb-2 sites respectively, I also denoted them as hb-3 and hb-2 clusters.

6.4.1 Structure of functional clusters in S2Es

I examined the structure of hb-3 and hb-2 clusters in four *Drosophila* S2Es to determine whether the conserved transcriptional activities arises from the conservation of functional binding site sequences and their configurations. As the hb-3 and hb-2 clusters are conserved across the four species, I was able to align the two clusters based on the hb-3, the top contributor site and one of the most conserved binding sites in the hb-3 cluster. As seen in Figure 6.11A-C, the configurations and the sequences of the functional binding sites have differences (Figure 6.11D for the sequence differences). First of all, the DNA sequences of the top contributors, hb-3 and hb-2, and their coactivator sites, bcd-2, bcd-*, bcd-1, bcd-(–1) and bcd-(–2) are not completely conserved among the four *Drosophila* species. First, although the footprinted bcd-1 site of *D. mel* is also found in *D. yak*, *D. ere* and *D. pse*, their 14 bp motifs defined by the SELEX PWM have one, three and three nucleotide substitutions respectively (Figure 6.11D). Second, the distances between the functional activators and the conserved block-B, defining the 3' border of S2E, are variable. In *D. pse*, the distances of the two Hb sites hb-3 and hb-2 from the conserved block are 305 bp and 210 bp respectively, while the same Hb sites in *D. ere* are 260 bp and 169 bp (Figure 6.11A). Third, the number of functionally active Bcd sites is different in each species. At the peak of stripe 2, *D. mel* S2E contains five functional Bcd sites whereas *D. pse* S2E has six functional Bcd sites (Figure 6.9). Fourth, the distances between the Bcd and Hb sites are

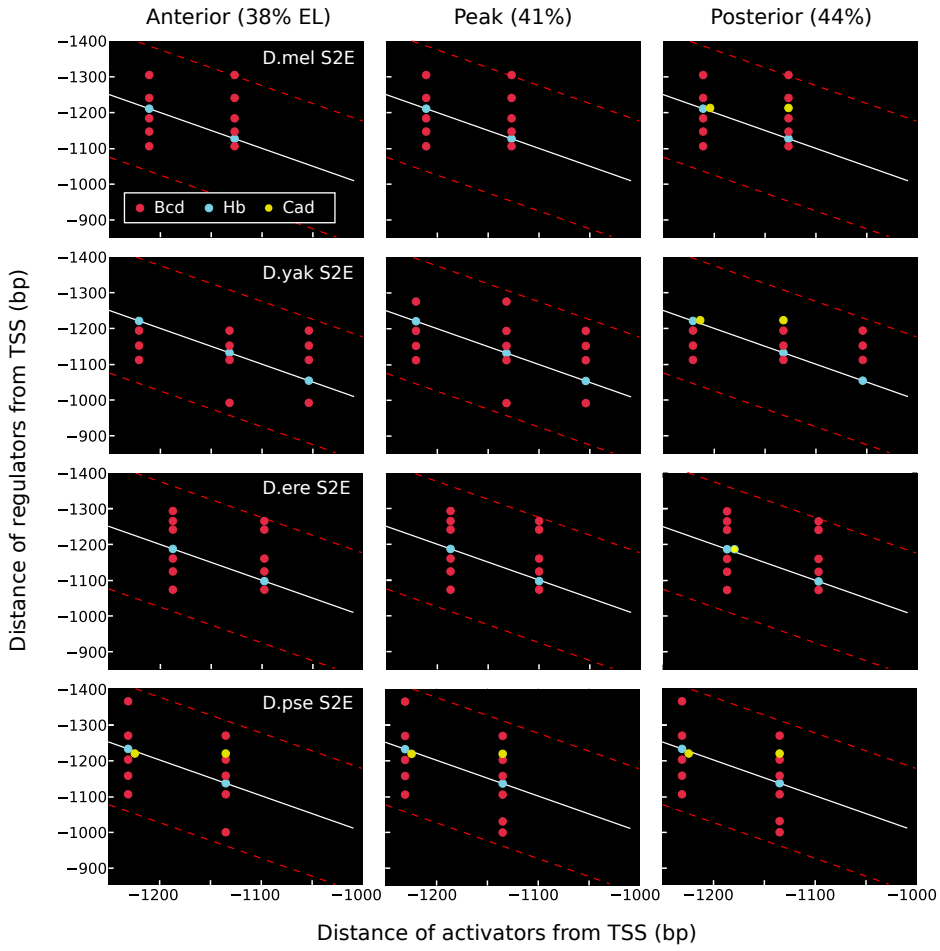


Figure 6.9: **Arrangements of functional activators in S2Es.** Each cluster maintains the same number of functionally occupied Hb and Bcd binding sites at three different A-P positions, the anterior border, peak and posterior border of stripe 2, except *D. yak*. Even though the distances of the Hb and Bcd sites from the TSS and distances between the activator and coactivator sites are variable, multiple functionally conserved Bcd sites are positioned in the range capable of coactivating hb-3 and hb-2 sites. Red dashed lines indicate the ranges capable of coactivating each Hb site. At the posterior border, Cad is competing with Hb over the hb-3 sites in the four S2Es, but its effect is negligible at the three A-P positions. Only functional binding sites having more than 5% fractional occupancies are displayed.

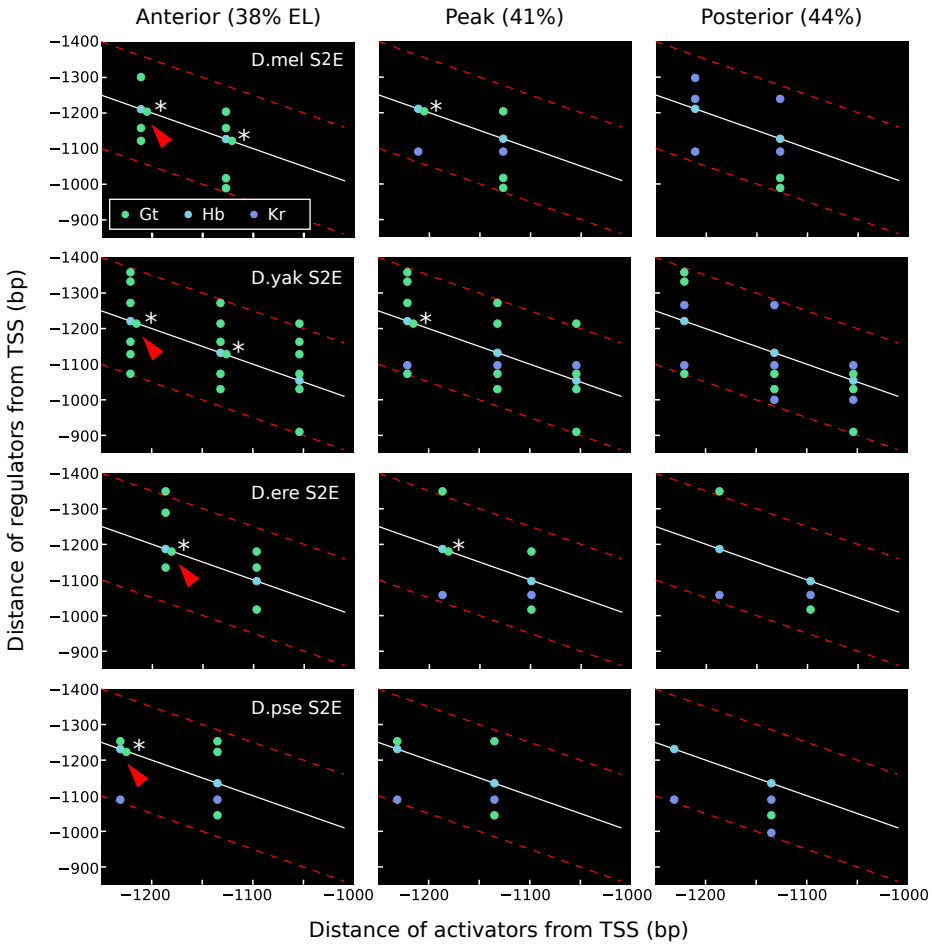


Figure 6.10: **Arrangements of functional repressors in S2Es.** The occupancies of functional repressor sites are significantly different at the anterior border, peak and posterior border of stripe 2. At the anterior border, Gt binding sites are strongly occupied and Gt and Hb are competing over hb-3 and hb-2 sites (see red arrow). At the posterior border, the majority of functional binding sites are Kr sites. Red dashed lines indicate the ranges capable of quenching each Hb sites. The positions of Gt and Kr sites in the clusters are variable, but always located in the effective quenching range of the hb-3 and hb-2 sites. Asterisks indicate the competition between two functional binding sites. Red arrows shows that the competition between hb-3 and gt-1 is conserved in all four species. Only functional binding sites having more than 5% fractional occupancies are displayed.

different. Among the hb-3 clusters in the four *Drosophila* species, for example, the distance between the hb-3 site and its closest upstream Bcd sites is 31 bp in *D. mel* and 56 bp in *D. yak*. The distance between the hb-3 site and the second closest Bcd sites is 96 bp in *D. mel* and 135 bp in *D. pse* (Figure 6.11A).

Between species, there is significant structural divergence of both the distances from the functional repressor sites to their target activator sites and the number of functional repressor sites in the hb-3 and hb-2 clusters. For example, in *D. yak*, the distance between gt-* and hb-3 is 52 bp while the distance between the equivalent binding sites in *D. ere* is 103 bp (Figure 6.11B). The number of functional Gt and Kr sites is also variable. Each cluster contains from two (*D. pse*) to six (*D. yak*) functional Gt sites at the 38% EL (the anterior border of stripe 2) (Figure 6.10). Especially, *D. pse* S2E completely lacks the two functional Gt sites, gt-* and gt-(–1) and the two functional Kr sites, kr-* and kr-** as shown in Figure 6.11B, C. HOT analysis indicates that there are non-functional sequences in *D. pse* S2E which are similar to the functional sites in *D. mel* but contain various substitution mutations (Figure 6.11D).

The analysis of these functional binding site alignments clearly shows substantial differences in both the sequences and configurations of the functionally active binding sites. I conclude, therefore, that the observed conserved function of S2Es is not maintained through simple conservation of functional binding site sequences or configurations in the four S2Es.

6.4.2 Conserved molecular interactions in S2Es

As sequence conservation for functional binding sites is an inadequate explanation for the conserved expression pattern, I then asked whether there are conserved features at the protein-DNA and protein-protein interaction levels.

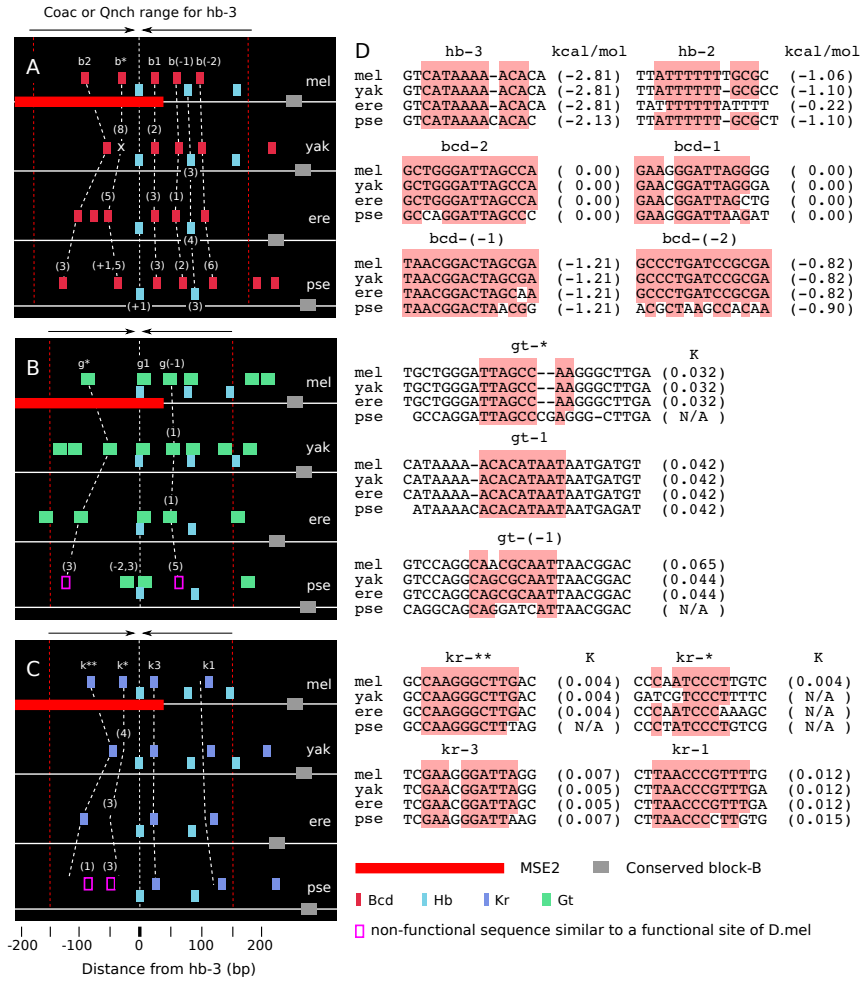


Figure 6.11: **Conserved functional clusters in four S2Es.** (A-C) Functionally active clusters at stripe 2 are aligned by the top contributor hb-3 site. A number in parenthesis indicate the number of a nucleotide change. Plus, minus and no sign indicate a base insertion, deletion substitution respectively. Conserved Bcd and Hb sites are connected through white dashed line. Some repressors are co-localized with the Bcd and Hb sites. (D) Sequences of functional binding sites in the functional clusters. Conserved PWM motifs are highlighted if 3 or more of 4 nucleotides in each position are identical. White dashed lines in panels A to C indicates the shifts of functional Bcd and Hb sites between species.

I examined DNA-protein interactions first and then compared protein-protein interactions on the S2Es in the four different *Drosophila* species. I then asked whether the conserved interactions in the four S2Es are sufficient to drive the conserved stripe 2 expression.

Conserved DNA-protein interactions

Experimental data has shown that a mutation of just a few nucleotides in a functional activator site can alter the binding affinity of a target TF, and in some cases, induce significant change in gene expression [6, 32]. I inspected binding affinities and fractional occupancies of functional binding sites in the different S2Es and found some conservation of binding interactions between TFs and their binding sequences despite differences from their *D. mel* counterparts. TF binding to hb-3 and hb-2 sites and four Bcd sites, bcd-2, bcd-1, bcd-(–1) and bcd-(–2) in *D. mel* S2E is conserved in the other species (Figure 6.11A). The hb-3 sequence in *D. pse* contains an additional nucleotide in the middle of its motif, which moves four contiguous nucleotides in the motif towards the 3' side (Figure 6.11D). However, the MITOMI data and the *in silico* model confirm that the binding affinity is not reduced by this sequence alteration (compare the Hb binding energy in Figure 6.11D). Furthermore, as in *D. mel* S2E, the *D. pse* hb-3 site is occupied by Hb at stripe 2 and serves as a top contributor site *in silico*. The sequence of hb-2 in *D. mel* S2E is not well conserved in the other species, although its binding affinity remains almost identical, or in the case of *D. ere*, even increased. This modeling result further shows that, as in *D. mel*, Hb remains bound to the hb-2 sites in the three species after competition and short-range quenching interactions in the stripe 2 region. Like hb-3, the bcd-(–2) sequence is completely conserved in *D. mel*,

D. yak and *D. ere* but not *D. pse*. The sequence of the *D. pse* bcd-(–2) differs by 6 bp of 14 bp from *D. mel*, but both MITOMI data and the model indicates that the binding affinity remains almost the same (Figure 6.11D). These results reveal that there are conserved DNA-protein interactions despite the sequence differences in the four S2Es, however the levels of the fractional occupancies at the equivalent binding sites in the four S2Es are not conserved *in silico*. This demonstrates that the binding affinities of the conserved binding sites are not a sole determinant of the fractional occupancies of the sites, and consequent S2E expression.

Conserved synergistic coactivation

HOT analysis reveals that the synergistic coactivation of Bcd and Hb for transcription initiation is a universal mode within the four S2Es (Figure 6.9 and 6.11A). Although the total number of functional Bcd sites in each cluster differs from species to species, four Bcd sites, bcd-2, bcd-1, bcd-(–1) and bcd-(–2), are functionally conserved. Furthermore, even though the distances between the Bcd sites and Hb sites are variable in the equivalent clusters between four S2Es, they are all located within the effective coactivation range of hb-3 and actively involved in coactivation of Hb bound to the hb-3 site. In *D. mel*, the hb-3 cluster provides more than 70% of the total contribution to gene expression at the native stripe 2 position (Figure 6.12). The activity of the hb-3 cluster reaches its maximum at around the stripe 2 peak and strongly reduced at the anterior and posterior borders of *eve* stripe 2. The hb-2 cluster is activated in broad anterior domain and provides additional contribution to initiate transcription in the stripe 2 region. This result shows that the non-MSE2 sequence, containing the core of hb-2 cluster, assists MSE2 to ensure

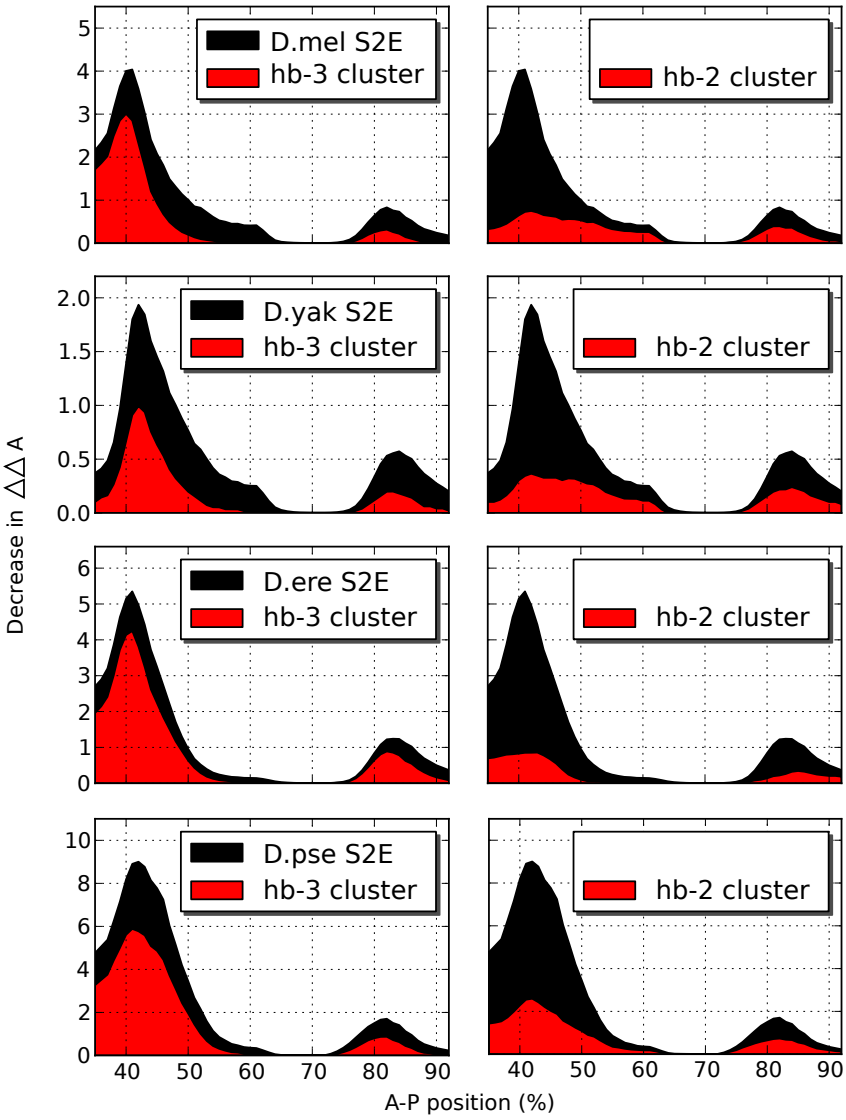


Figure 6.12: **Transcriptional activities of the hb-3 and hb-2 clusters.** Decreases in activation energy barrier ($\Delta\Delta A$) are almost entirely driven by synergistic coactivation of Bcd and Hb. In all four *Drosophila* species, hb-3 cluster provides the major contribution to stripe 2 expression, while the hb-2 cluster assists the hb-3 cluster to achieve the full levels of gene expression.

full levels of stripe 2 expression driven by S2E.

Like *D. mel*, the other species, *D. yak*, *D. ere* and *D. pse*, also show a similar pattern (Figure 6.12). The hb-3 clusters provide a considerable amount of contribution to stripe 2 expression and the hb-2 clusters assist the transcriptional activity of hb-3 cluster. Note that in the four species the hb-3 and hb-2 clusters are also activated in the stripe 7 region (Figure 6.12). This result raises the possibility that the stripe 2 enhancers of *D. yak*, *D. ere* and *D. pse* species might also be actively involved in *eve* stripe 7 formation in their native context.

Conserved competitive interactions between TFs

I then examined the protein-protein interactions between functional activators and repressors in the four S2Es. I observed a striking conservation of pattern when I superimpose the configurations of the functional repressor sites upon the configurations of the functional activator and coactivator sites (see white dashed lines in Figure 6.11A-C). For the functional Kr sites in the four species, all sites of each cluster are co-localized with one of the four Bcd sites, bcd-2, bcd-*, bcd-1 and bcd-(–2), despite the changing positions of the Bcd sites within the clusters (Figure 6.11C). For example, the two footprinted sites kr-3 and kr-1 follow the bcd-1 and bcd-(-2) sites in all the four S2Es. Interrogating these sequences in the model further confirms that competition interactions between Bcd and Kr occur at overlapping sites, kr-3 over bcd-1 and kr-1 over bcd-(–2), at the posterior border of stripe 2. The removal of the Kr sites in the four S2Es causes a more posterior expansion than simply disabling the short-range quenching capability of Kr *in silico*. This result suggests that the competitive interactions between the Bcd and Kr sites are a relevant component of posterior border formation. This interaction is completely conserved

between the four species. In addition, the three Gt binding sites, gt-*, gt-1 and gt-(–1), also track with the bcd-2, hb-3 and bcd-(–1) respectively (Figure 6.11B). Notably, the overlap of gt-1 and hb-3 is completely conserved in the four species while gt-* and gt-(–1) overlap over bcd-2 and bcd-(–1) in all species except *D. pse*. *In silico* experiments show that the competition between Bcd and Gt or Hb and Gt over these sites is critical for the anterior border formation. These results show that three competitive binding interactions, between Bcd and Kr, Bcd and Gt and Hb and Gt, are functionally conserved in a subset of each clusters despite sequence differences in the overlapping sites.

Conserved short-range quenching interactions

I analyzed the short-range quenching interactions between functional sites in each cluster. As seen in Figure 6.11B and C, the positions of the repressor Gt and Kr sites found in the *D. mel* stripe 2 clusters vary in different S2Es, however, all of them, when present, are located in the range capable of quenching the conserved activator sites bcd-2, bcd-1, bcd-(–1), bcd-(–2), hb-3 and hb-2. Furthermore, even though the strength of their repressive action on equivalent sites differs between different species, the *in silico* model indicates that these Gt and Kr sites are actively involved in the anterior or posterior border formations. These results suggest that many Gt and Kr mediated quenching interactions found in *D. mel* S2E are functionally conserved in the other species and the distances between the repressor sites and their target activator sites in different S2Es could be constrained by their physical quenching range *in vivo* [42, 12, 40, 43].

6.4.3 Insufficiency of conserved interactions between S2Es

Having identified the specific conserved molecular interactions needed for stripe 2 expression, I then asked whether these interactions are sufficient to drive the conserved spatiotemporal expression of stripe 2 in the *in silico* transcription system. S2Es from the three species exhibit large differences in their expression levels and rescue abilities in the *D. mel* embryo [2]. As only *D. yak* and *D. pse* S2Es drive similar levels of *D. mel* S2E expression in *D. mel* blastoderm embryo (Dr. Martinez's unpublished data) and rescue the embryonic lethal $EVE\Delta S2E$ deletion [2], I restrict all further functional analysis to these two species hereafter. The *in silico* transcription system successfully recapitulates the *in vivo* findings that the *D. yak* S2E drives expression levels similar to that of *D. mel* S2E. While the *D. pse* S2E recapitulates a similar spatiotemporal pattern of expression *in silico*, the level of S2E expression at the peak of stripe 2 is about four times higher than *D. mel*. Therefore, for *D. pse* S2E, I limited my investigation of *D. pse* S2E to anterior and posterior border formations.

In order to measure the net transcriptional activities of the conserved molecular interactions in the functional clusters of the *D. yak* and *D. pse* S2Es, I only allowed interactions occurring on conserved functional binding sites based on following criteria. First, among the top functional activator, coactivator and repressor sites identified by HOT analysis for each S2E (see Section 6.2), only these binding sites which are conserved between the *D. mel* and a target S2E are used. Second, any coactivation, competition and short-range quenching interactions taking place on the conserved functional sites are allowed. Third, because the Bcd-Bcd cooperative interactions operate in a pair-wise manner among the Bcd sites in the clusters, I only allowed the cooperative interactions seen in the *D. mel* S2E between bcd-2 and bcd-* and

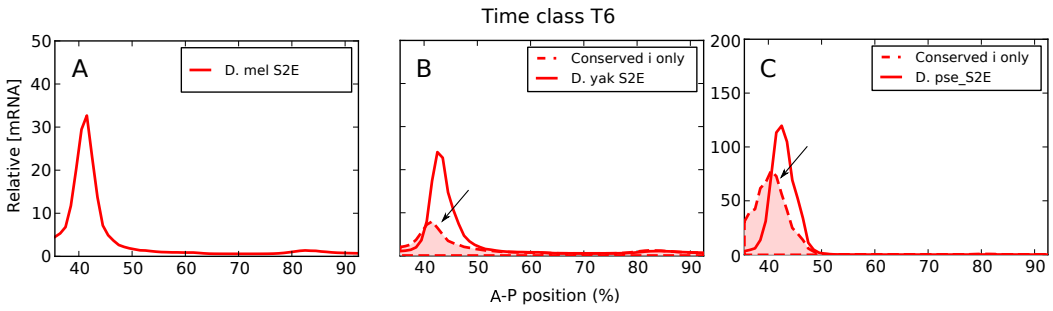


Figure 6.13: **S2E expression driven by conserved interactions.** Protein-DNA and protein-protein interactions occurring only on the conserved binding sites are allowed with criteria described in the section 6.4.3. (A) *D. mel* S2E expression *in silico*. (B, C) Conserved interactions between *D. mel* and *D. yak* (B) and between *D. mel* and *D. pse* (C) are not sufficient to generate conserved expression. Arrows indicate the conserved interaction-driven S2E expression.

between bcd-1 and bcd-(–1) when available in a target S2E. Finally, I left all binding sites outside of the active clusters and consequent molecular interactions intact in this *in silico* experiment.

As seen in Figure 6.13, conserved interactions between species were not able to generate full levels of conserved gene expression. Furthermore, in the case of *D. pse*, substantial derepression in the 1-2 inter-stripe region is observed. In both S2Es, their expression is shifted anteriorly. This result demonstrates that the conserved interactions alone are not sufficient for the establishment of conserved gene expression and suggests the existence of novel molecular mechanisms assuring the conserved expression *in vivo*.

6.4.4 Novel molecular interactions in *D. yak* and *D. pse* S2Es

The *D. yak* and *D. pse* S2Es do not contain five TF sites found in *D. mel* (Figure 6.11B, C). Because they are functional components of *D. mel* S2E function, the conserved interactions remaining in *D. yak* or *D. pse* S2Es alone were not be able to recapitulate conserved stripe 2 expression *in silico* (Figure 6.13B, C). In the *D. yak* S2E, the cooperativity of Bcd between bcd-2 and bcd-* is lost due to the absence of the bcd-* site while in *D. pse* S2E many repressive interactions are missing due to the lack of the four repressor sites gt-*, gt-(–1), kr-** and kr-*. Despite the loss of these molecular interactions critical for *D. mel* S2E, almost identical stripe 2 patterns were observed when the *D. yak* and *D. pse* S2Es are assayed *in vivo* [2, Dr. Martinez's unpublished data] and *in silico* (see solid red lines in Figure 6.13B, C).

I found that, despite the absence of bcd-*, *D. yak* bcd-2 is able to maintain strong cooperative interactions through reconfiguration of its cooperative pairing (Figure 6.14A). In the *D. mel* S2E, cooperative interactions are established between bcd-2 and bcd-* and between bcd-1 and bcd-(–1) and these interactions are essential for full levels of stripe 2 expression *in silico*. In *D. yak*'s S2E, however, novel cooperative interactions between bcd-2 and bcd-1 and between bcd-(–1) and bcd-(–2) compensate for the absence of bcd-* (Figure 6.14E). These reconfigurable Bcd-Bcd cooperative interactions are possible because Bcd executes its cooperative action over a certain distance (see Section 3.1.2). Previously, I constrained the cooperative interaction range to 60 bp (see Section 3.1.2). With this restriction, the restoration of stripe 2 expression seen in Figure 6.14E was not observed (see Figure 4.1C1). However, a 10 bp increase of the cooperative range results in the similar levels of stripe 2

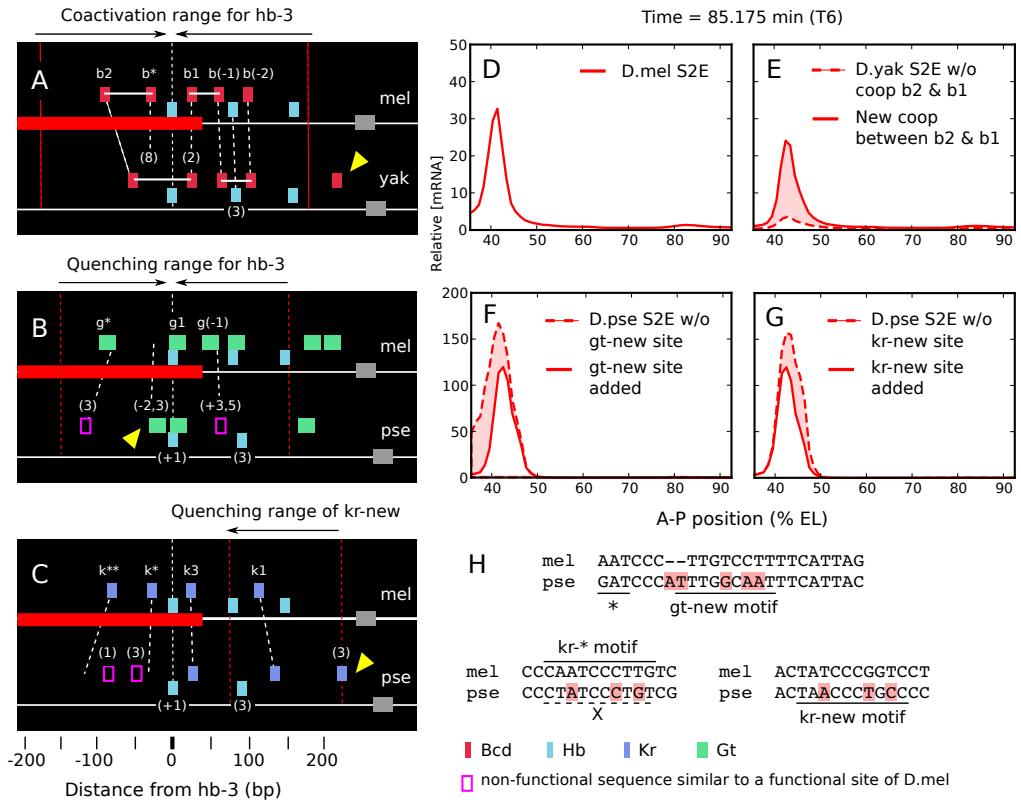


Figure 6.14: **Novel molecular mechanisms for conserved expression.** Novel cooperative and repressive interactions are necessary for normal stripe 2 pattern. (A) Adaptive cooperative interactions. Despite the absence of bcd-*, *D. yak* bcd-2 is able to maintain the strong cooperative interactions through reconfiguration of the cooperative pairing. White horizontal lines between Bcd sites indicate cooperative interactions. (B) A new Gt site, indicated by a yellow arrow, becomes a functional component of *D. pse* S2E. (C) A new Kr site, indicated by a yellow arrow, becomes a functional component of *D. pse* S2E. (D) *D. mel* S2E driven expression *in silico*. (E-G) Contributions of *de novo* molecular interactions to conserved stripe 2 expression. (H) Putative mutations in *D. pse* S2E create new functional sites. '*' indicates the motif of overlapping bcd-* site. 'x' indicates the loss of kr-* site in *D. pse* S2E.

expression compared to *D. mel* S2E (Figure 6.14E). This result suggests that the reconfigured cooperative interactions might be a major component ensur-

ing the conserved levels of stripe 2 expression observed *in vivo*. In addition, this further suggests that S2E acquires an additional intrinsic capability that tolerates a certain degree of structural alterations through a Bcd-Bcd pairwise cooperative interaction.

In case of *D. pse* S2E, two novel repressor sites, not found in *D. mel* S2E, play an essential role for the conserved stripe 2 expression border (see yellow arrows in Figure 6.14B and C). As seen in Figure 6.14B and C, *D. pse* lacks the *D. mel* Gt sites gt-* and gt-(–1) and two Kr sites kr-** and kr-*. However, for Gt, a new site appears in the *in silico* model. This site has a strong binding affinity in *D. pse* and its sequence differs by four nucleotides from the corresponding *D. mel* sequence (Figure 6.14H). The new Gt site, termed gt-new, is located close to the top contributor hb-3 site and, without gt-new, severe derepression is observed in the *eve* 1-2 inter-stripe region *in silico* (Figure 6.14F). This results suggests that the gt-new site becomes a functional component of *D. pse* S2E. For Kr, a new Kr site, kr-new, is located close to the conserved block-B (6.14C). This site contains three nucleotide substitutions in its sequence compared to *D. mel* sequence (Figure 6.14H). Deletion of kr-new causes a noticeable posterior border expansion *in silico* (Figure 6.14G). This result establishes that the kr-new site, repressing the activities of activators at posterior border, might be a functional component of the *D. pse* S2E *in vivo*. In conclusion, these results suggest that novel molecular interactions found in the altered structures of S2Es are required for maintaining the conserved biological function of *D. yak* and *D. pse* S2E in the *D. mel* embryo. Given their strong contributions to gene expression, it is likely they are also highly functional in their native context.

6.5 Structure–function analysis of Sepsid stripe 2 enhancers

Two of the most extreme examples of structural flexibility are the *eve* stripe 2 enhancers from two Sepsid species *S. cyn* and *T. put*. These enhancers have a completely different arrangement of binding sites compared to their well-characterized *D. mel* counterpart yet produce an almost identical stripe 2 expression domain in the *D. mel* embryo [102]. It has been reported that only 5% of *D. mel* binding sites are conserved in pairwise comparisons with Sepsid species and only a few short 20-30 bp sequences are shared in stripes 2, 3 and MHE enhancer between *Drosophila* and Sepsid species [102].

In the *eve* stripe 2 enhancer, Hare *et al.* were able to find one highly conserved Bcd-Kr binding site pair, through a pair-wise sequence alignment across the entire six *Drosophila* and six Sepsid species in the 5' side of their stripe 2 enhancers. Among the six Sepsid species, one conserved Bcd-Kr site pair and two Slp binding sites were identified. Because of these very limited sequence similarities, it was almost impossible to align the Sepsid enhancers to their *Drosophila* orthologs based on the sequence similarities [102]. Consequently, it is impossible to infer the transcriptional control of gene expression of the Sepsid S2Es based on the conserved binding sites between *Drosophila* and Sepsid. Furthermore, it has been reported that stripe 2 expression driven by the *S. cyn* and *T. put* S2Es are shifted posteriorly compared to the *D. mel* S2E and the degree of the shift of *T. put* is larger than that of *S. cyn* [102]. Functional dissection of such changes in gene expression at the molecular level has been a major research challenge.

In this section, I sought to elucidate fundamental mechanisms that allow

Sepsid S2Es to generate a remarkably similar stripe 2 expression pattern in the *Drosophila* embryo using the *in silico* transcription system. I then investigated the experimentally reported functional differences between *Drosophila* and Sepsid S2Es. The peak positions of *S. cyn* and *T. put* S2E expression are 46% and 50% EL respectively *in silico*, hence I performed the HOT analysis for the two Sepsid S2Es at 46% and 50%, and also at 41% EL as a reference, the peak position of *D. mel* S2E (Figure 6.15). In this study, 41%, 46% EL, the peak positions of *D. mel* and *S. cyn*, represent the anterior border positions of *S. cyn* and *T. put eve* stripe 2 and 50% and 53% EL represent their posterior border positions respectively.

6.5.1 Differential features of *S. cyn* and *T. put* S2Es

The functional binding site analysis reveals that more than three functional clusters are actively involved in stripe 2 formation in the Sepsid S2Es. Because of the substantial sequence differences between *Drosophila* and Sepsid S2Es, almost no similarity in functionally active binding sites relative to the *D. mel* S2E clusters was found except for two Hb sites. The two Hb sites in *S. cyn* and *T. put* are different from the hb-3 and hb-2 sites in *D. mel* S2E by one or two nucleotides. The *in silico* model and MITOMI data confirm that the two Hb sites in *S. cyn* and *T. put* S2E have similar or higher binding affinities compared to that of *D. mel* S2E.

Given the limited similarities of the functional clusters between *Drosophila* and Sepsid S2Es, it is hard to determine whether the two Hb sites are equivalent to the hb-3 and hb-2 sites in *Drosophila* S2Es. However, like the hb-3 and hb-2 sites in the *Drosophila* S2Es, they are one of the top 80% contributors for stripe 2 expression and located right next to each other. Hence, in

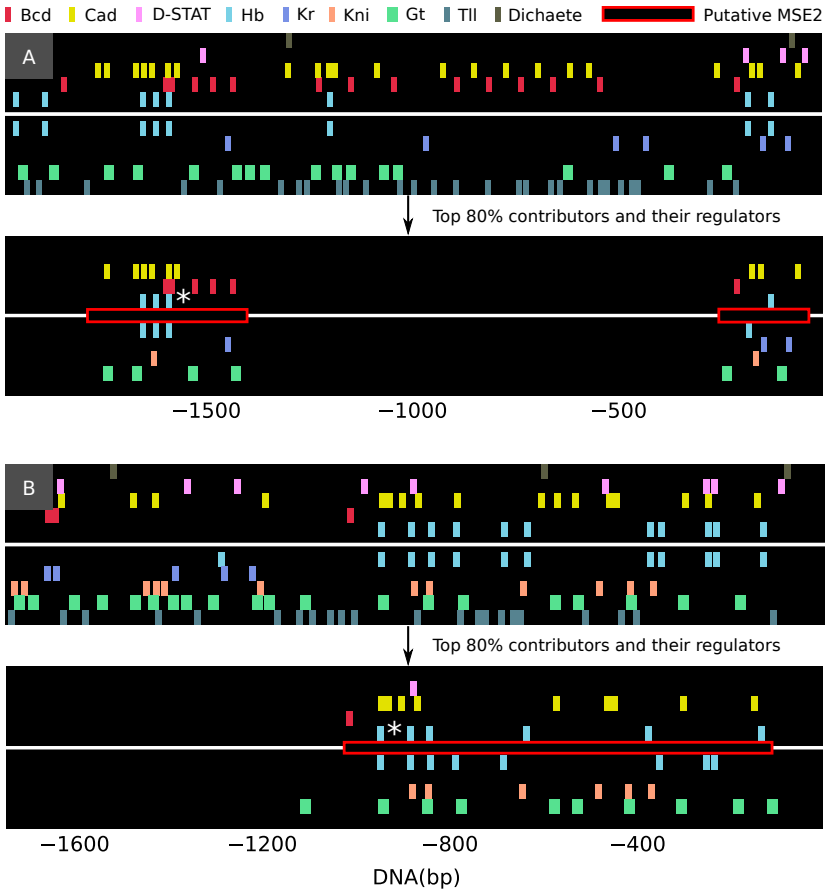


Figure 6.15: **Highly active sites in *S. cyn* and *T. put* S2Es.** Top 80% contributors and their regulators at the anterior, peak and posterior positions of stripe 2 were identified and visualized together using *in silico* system. (A) *S. cyn* S2E contains two highly active regions. Bigger fragment contains two Hb sites similar to hb-3 and hb-2. (B) Top contributors are dispersed in about 1 kb fragment in *T. put* S2E. Note that only one Bcd site is functionally active. Putative hb-3 site is indicated by white asterisk.

this study, I denoted them as hb-3 and hb-2 respectively and, based on the hb-3 site, I aligned the functional clusters identified in the Sepsid S2Es with the top clusters in *Drosophila* S2Es (Figure 6.16). The functional binding site alignment shows that there are significant differences in the composition and

configuration of TF binding sites in the clusters compared to that of *Drosophila* species. One significant feature of Sepsid S2Es is that the S2Es contain multiple strong Cad binding sites (Figure 6.16A). The *in silico* model confirms that these sites are occupied at the peak positions and even at 41% EL, the peak position of *D. mel* S2E, Cad is present on the binding sites. In contrast, although the *D. mel* S2E also contains computationally identified Cad binding sites, they are not occupied at 41% EL and only one site is occupied at 44% EL (Figure 6.9).

Another noticeable feature of the Sepsid S2Es is that, in their functional clusters, a limited number of functional Bcd sites is found. In *S. cyn* S2E, five functional Bcd sites are found in the clusters, but MITOMI data and the model confirmed that none of their binding affinities are stronger than or similar to bcd-1 or bcd-2 in the *D. mel* S2E. Furthermore, only three of them are actively involved in coactivation of hb-3 *in silico* because two of them overlap hb-3 site (Figure 6.15A). The model confirmed that competitive interactions between Bcd and Hb occur on the overlapping sites in the region of stripe 2 expression. Among the six *Drosophila* and Sepsid species, the competitive interactions between Bcd and Hb are only seen on these particular sites in *S. cyn* S2E. With respect to *T. put* S2E, only one functional Bcd site is found in the hb-3 and hb-2 clusters (Figure 6.15B) and its binding affinity is also lower than the footprinted Bcd sites, bcd-1 or bcd-2 in *D. mel* S2E. This result suggests that Sepsid S2Es might require the *trans*-acting factor Cad to drive stripe 2 expression in a *D. mel* embryo.

With respect to functional repressor sites in Sepsid S2Es, two distinctive features were found. First, functional Kr sites are almost entirely absent in the Sepsid S2Es. As seen in Figure 6.15, *T. put* S2E does not contain any func-

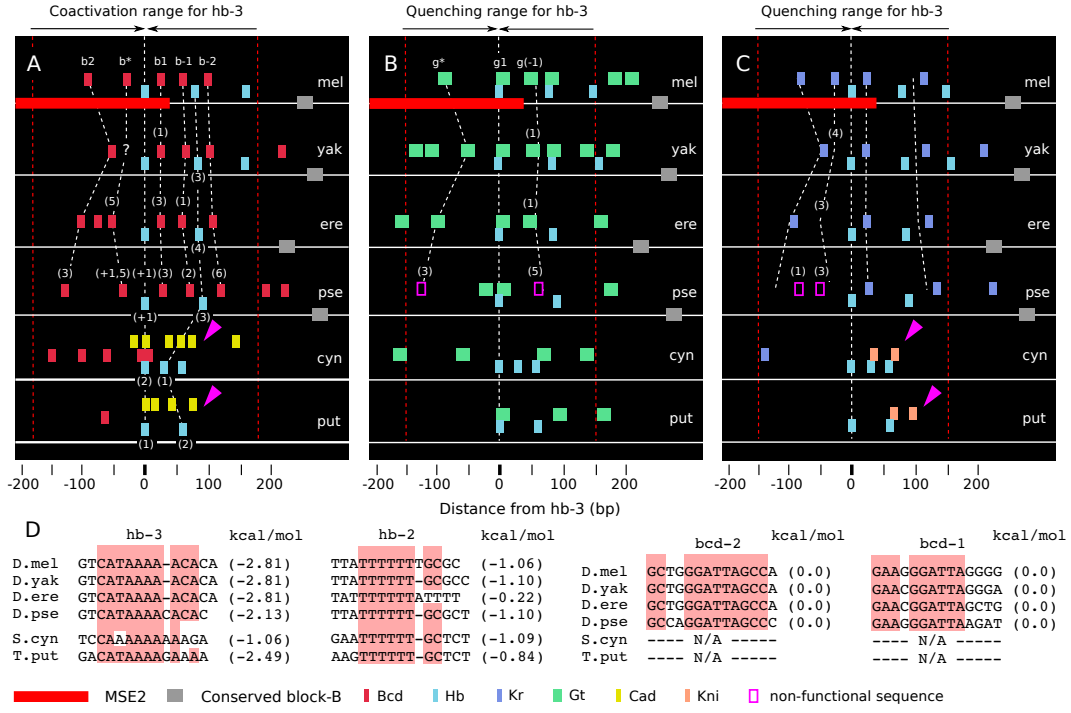


Figure 6.16: **Comparison of functional clusters in six S2Es.** Substantially different configurations drive similar function. (A) Sepsid S2Es utilize Cad sites for gene expression. Two Hb sites resemble hb-3 and hb-2 sites of *D. mel* S2E. Magenta arrows indicate functionally active Cad sites. (B) Sepsid S2Es contain functional Gt binding sites. As gt-1 overlaps over hb-3 in *D. mel* S2E, one of Gt sites in each Sepsid S2E overlaps over a functional Hb site. (C) At the posterior border of Sepsid S2E expression, completely different configurations of TF occupancies are observed. Unlike *Drosophila* S2Es, Sepsid S2Es almost entirely lack functional Kr sites in their active clusters. Instead, Sepsid S2Es utilize Kni sites for the posterior border formation. Magenta arrows indicate functionally active Kni sites. Both *S. cyn* and *T. put* have similar Kni configuration. (D) DNA sequences of two highly active Hb sites in Sepsid S2Es are similar to hb-3 and hb-2 sequences in *D. mel* S2E. However, the DNA sequences of surrounding TF binding sites such as coactivator Bcd sites and spacers between TF sites are substantially different from the sequences of *D. mel* S2E. Conserved PWM motifs are highlighted if 4 or more of 6 nucleotides in each position are identical.

tional Kr sites in the top 80% clusters. In the case of *S. cyn* S2E, it contains three functional Kr sites, but only one site is located in the hb-3 and hb-2 clusters (Figure 6.15A and 6.16C). Second, unlike *Drosophila* S2Es, the *T. put* S2E contains multiple functional Kni sites in their top 80% clusters (Figure 6.15B). In the case of *D. mel* and *S. cyn* S2E, only one Kni site is functional at 50% EL, which is 9% and 4% posterior from their peak positions. However, *T. put* S2E have four functional Kni sites at 50% EL, the peak position and six at 53% EL, the posterior border position of the *T. put* S2E expression. Having identified these differential configurations of TF occupancies relative to *D. mel* S2E, I then investigated the relationship between the configurations and their function in the following sections.

6.5.2 Utilization of *trans*-acting factor Cad for *eve* stripe 2

I asked whether the highly occupied Cad sites are essential for the Sepsid S2E expression. In order to test the Cad dependency, I ran the *in silico* transcription system without Cad binding sites for three enhancers, *S. cyn*, *T. put* S2Es, and *D. mel* S2E as a control. Binding affinities of any overlapping TF sites are left intact. As seen in Figure 6.17, deletion of all Cad sites in *D. mel* S2E does not cause any significant changes in gene expression *in silico*. However, gene expression driven by *S. cyn* or *T. put* S2Es is completely abolished when the Cad binding sites are deleted (Figure 6.17). This result suggests that Cad is essential for the Sepsid S2Es to drive stripe 2 expression in *D. mel* embryo. I then asked whether Cad initiates transcription through coactivation of Hb in the Sepsid S2Es *in silico*. Note that, in Section 3.1.3, the model suggested coactivation of Hb by Cad for proper stripe 7 expression. When I disabled

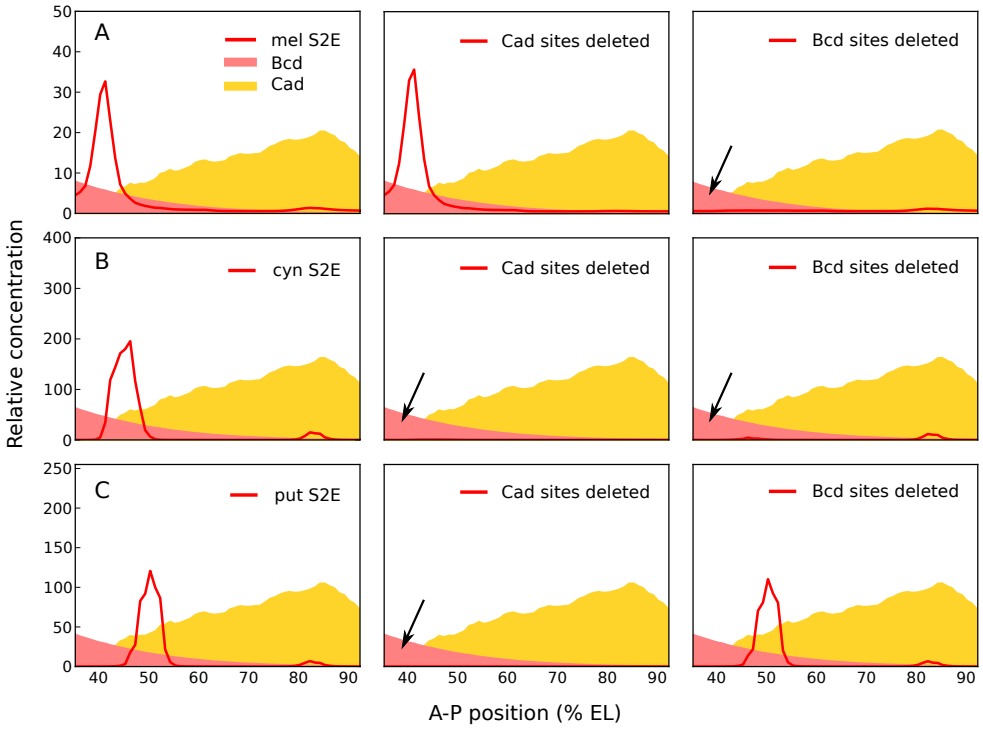


Figure 6.17: ***In silico* prediction of maternal input dependency.** In order to test Cad dependency of Sepsid S2Es, two Sepsid *S. cyn* and *T. put* S2Es were tested and also the *D. mel* S2E for comparison. Two *in silico* experiments were performed for each S2E; Deletion of entire Cad sites or Bcd sites respectively. Binding affinities of TF binding sites overlapping the Cad or Bcd sites are left intact. (A) *D. mel* S2E does not depend on Cad as an activation input. However, the transcriptional activity of *D. mel* S2E is completely abolished in the absence of Bcd binding. (B) *S. cyn* S2E requires both Bcd and Cad input for its expression. Gene expression of *S. cyn* S2E is completely abolished without Cad binding sites or Bcd sites. (C) In the *T. put* S2E, coactivation of Hb by Cad is necessary and sufficient for the activation of stripe 2 expression. Deletion of Bcd sites slightly reduced the levels of expression. Black arrows indicate the loss of gene expression.

coactivation of Hb by Cad *in silico*, even though many functional Cad sites are occupied, *S. cyn* and *T. put* S2E did not generate gene expression (data not shown). This suggests that the Cad mediated Hb coactivation is essential for the Sepsid S2E-driven gene expression in the *D. mel* embryo.

Further *in silico* experiments were performed to ask whether the Sepsid S2Es are able to drive the conserved expression pattern without Bcd sites. As described in the previous section, *S. cyn* contains a few weak Bcd sites in the hb-3 and hb-2 clusters and the *T. put* S2E contains only one Bcd site in the clusters (Figure 6.15). Deletion of all Bcd sites completely abolishes the *S. cyn* S2E expression while the *T. put* S2E drives almost identical expression in the absence of the one Bcd binding site *in silico*. This result suggests that *S. cyn* S2E requires both Bcd and Cad input for its expression while *T. put* S2E maintains its expression by compensating for the lack of functional Bcd sites through coactivation of Hb by Cad.

6.5.3 Posterior shift of stripe 2 expression of Sepsid S2Es

When the two Sepsid S2Es are driven in the *D. mel* embryo, a noticeable difference is observed in their anterior and posterior border positions. This is in contrast to *Drosophila* S2Es, as double staining experiments showed that the S2Es from *D. yak*, *D. ere* and *D. pse* produce a pattern of *lac*Z expression that is coincident with the *D. mel eve* stripe 2 without any consistent shift or expansion of stripe 2 transgene expression [38]. The border positions of the Sepsid S2E-driven *lac*Z expression are shifted posteriorly compared to *D. mel* MSE2-driven gene expression [102]. Hare *et al.* reported that the posterior borders of gene expression driven by *S. cyn* and *T. put* S2Es are shifted 3% and 4% EL posterior respectively. The *in silico* model predicts the peak positions of *S. cyn* and *T. put* S2E expression shifted 5% and 9% EL posteriorly (Figure 6.3E1,2). Due to the differences of the data quantification method, genomic position of the reporter constructs, and the lack of knowledge about the definition of the border position used in the previous study, it is impossible

to precisely evaluate the *in silico* predictions made in this study. However, the qualitative features were well captured (compare Figure 6.17A, B and C). In order to investigate the posterior shift, I aligned the top clusters of the S2Es vertically (Figure 6.18) and then examined molecular interactions taking place on the configurations at five different A-P positions representing the anterior, peak and posterior border of S2E expression of four *Drosophila* and two Sepsid species.

Loss of Bcd and gain of Cad sites cause anterior border shift

At 41% EL, the peak position of *D. mel* stripe 2, multiple functional Bcd sites in the *Drosophila* S2Es are highly occupied around the top contributor sites hb-3 and hb-2 and drive the maximum level of expression. In Sepsid S2Es, in contrast, only a limited number of Bcd sites are shown in the top clusters at 41% EL. The lack of strong Bcd sites together with multiple Gt molecules bound to the Sepsid S2Es, causes loss of gene expression at the peak of native *eve* stripe 2 *in silico* (Figure 6.18A, B). However, at the peak positions of *S. cyn* and *T. put* S2E expression, multiple Cad sites are occupied in the Sepsid S2Es and compensate for the lack of Bcd sites through Cad mediated Hb coactivation (compare the clusters in Figure 6.18A). Because the *T. put* S2E contain fewer Bcd sites than *S. cyn* S2E, it requires more activation input from Cad for expression, and consequently, the anterior border of *T. put* S2E expression is shifted more posteriorly.

Loss of Kr and gain of Cad sites cause posterior border shift

At 46% EL, the posterior border of *D. mel* stripe 2, gene expression driven by the *D .mel* S2E is strongly repressed by Kr mediated short-range quenching

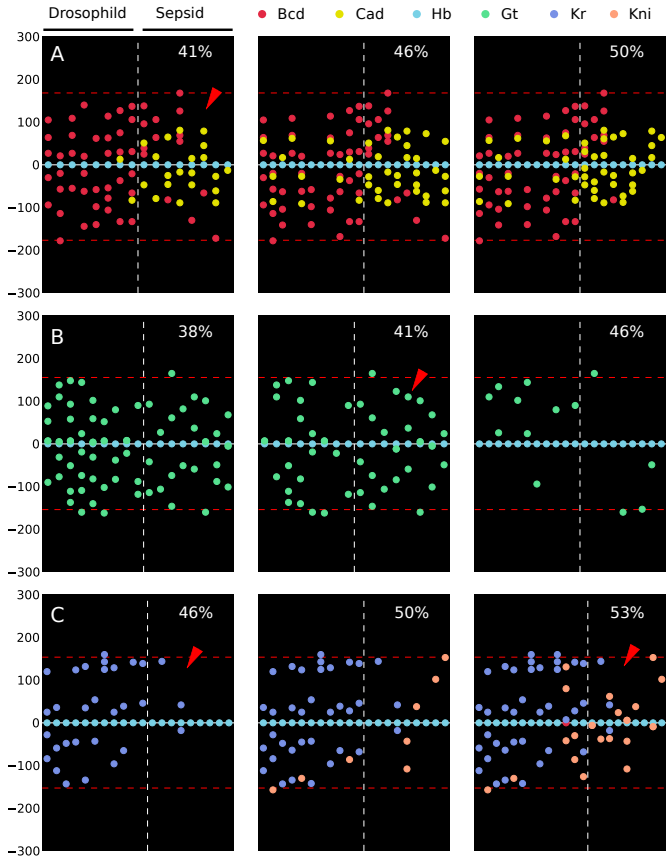


Figure 6.18: **Analysis of posterior shift of Sepsid stripe 2.** Top 80% clusters of S2Es from four *Drosophila* species, *D. mel*, *D. yak*, *D. ere* and *D. pse* and two Sepsid species, *S. cyn* and *T. put* are aligned vertically. Red dashed lines indicate the ranges of coactivation or short-range quenching of top activators located in the middle horizontal line. (A) Multiple functional Bcd and Hb sites are clustered in the *Drosophila* S2Es. While only few functional Bcd sites are observed (see a red arrow), Sepsid S2Es contain multiple functional Cad sites. But even at 50% EL, only few Cad sites are occupied in the *Drosophila* S2Es. (B) Multiple Gt sites are occupied and functional in both *Drosophila* and Sepsid S2Es. At 41% EL, the peak position of *D. mel* S2E expression, the number of functional Gt sites are significantly reduced in *Drosophila* S2Es compared to at 38% EL, but not in Sepsid (see a red arrow). (C) At 46% EL, functional Kr sites are almost entirely absent in Sepsid S2Es (see a red arrow). At 53% EL, multiple functional Kni sites are observed in Sepsid S2Es, but not in *Drosophila* S2Es (see a red arrow).

and competition *in silico* (Figure 6.8D). However, functional Kr sites are almost entirely absent in the Sepsid S2Es (Figure 6.18C), which allows posterior expansion of Sepsid S2E expression, helped by increasing Cad concentration. Note that *S. cyn* S2E contains three functional Kr sites in the clusters but, unlike *D. mel*, only one Kr site is located in the range of quenching the hb-3 cluster and none of the Kr sites are seen in the hb-2 cluster (Figure 6.16C). In the case of *T. put* S2E, it does not contain any functional Kr sites even at the middle of Kr domain (see 50% EL panel in Figure 6.18C). The complete deficiency of functional Kr sites in the *T. put* S2E and its independence from Bcd mediated activation further expands the posterior border.

In conclusion, I have established through *in silico* analysis that Sepsid S2Es maintain the conserved expression pattern by utilizing Cad, which is a novel form of activation input for the *eve* stripe 2 pattern. I have also shown that the posterior shift of Sepsid S2E expression arises mainly from the absence of functional Bcd and Kr sites. Despite the differences, these results clearly show the possibility that conserved expression patterns can be driven by recruiting completely different *trans*-acting factors *in vivo*.

Chapter 7

Conclusions

There are six major results reported in this dissertation. First, I have generated highly precise gene expression data of the four *eve* enhancer fusions, M3_2, M32, M2_3 and M23, permitting direct comparisons between the constructs at single nucleus resolution. Second, I have demonstrated an *in silico* transcription system having a strong predictive and analytic ability. Third, significantly altered gene expression of the four *eve* enhancer fusions is numerically recapitulated within the limits of experimental accuracy. Fourth, the model demonstrated that *cis*-regulatory "elements" are not elementary objects. The analysis of the modeling result has shown that the highly elevated expression driven by a fusion of MSE2 and MSE3 is a consequence of the recruitment of a portion of MSE3 for stripe 2 function. Fifth, I demonstrated that conserved molecular interactions taking place on the conserved functional TF binding sites between different S2Es are not sufficient to maintain conserved stripe 2 expression seen in the *in vivo* experiment [2]. Using the *in silico* transcription system, I then identified novel molecular interactions required for the conserved expression, but not seen in the *D. mel.* Sixth, the model has predicted

gene expression of *eve* stripes 2, 3 and 7 or 4, 5 and 6, driven by about 4 kb long *eve* regulatory DNAs which contain multiple enhancers. This result opens the door to an understanding of the control of gene expression at the level of a whole, intact genetic locus.

The first section of this chapter discusses the quantitative gene expression data and the *in silico* transcription system. In the second section, the implications of the results for *cis*-regulatory logic are discussed. The third section highlights the mechanism of conserved enhancer function inspired by this work. Fourth section discusses the limitations of this study and finally future prospects is described in the last section.

7.1 Predictive and analytic transcription model

The work described here represents a substantive advance in modeling of transcriptional control. I have overcome the limit of modeling only individual experimentally identified enhancers, and have done so at a level of resolution comparable to that required for organismal survival. The expression data generated during my dissertation work not only involved two enhancers, but more importantly considered a situation in which the function of these enhancers was critically altered by juxtaposing them. These rearrangements provided a powerful constraint on the possible rules of transcriptional control, as demonstrated by the success in prediction of expression patterns reported here (Figure 4.1). Furthermore, the model can be used as an analytic tool with which to understand how multiple transcriptional mechanisms operate simultaneously to produce observed patterns of expression (Figure 5.3, 5.5 and 4.1).

7.1.1 Importance of high quality expression data

Highly precise experimental data made this study possible, and their importance cannot be overemphasized. Transcriptional machinery is inherently precise, and fundamental understanding of its functioning requires data at a cellular level of precision. The dataset of the four fusions, M3_2, M32, M2_3 and M23, has that level of precision because I performed simultaneous staining of reporter-driven *lac*Z expression and native Eve protein, allowing us to register the reporter data with our full TF dataset [50].

The intrinsic variability of gene expression prevents such registration by measurements of the position of reporter expression alone. This point illuminates a problem regarding the current unbalanced state of technology in genomics, where sequence can be obtained readily and cheaply, but the inability to monitor gene expression at cellular resolution in a high throughput manner together with a lack of understanding of the code for regulatory logic has in general limited genomic level investigations of regulatory DNA to statistical association studies [112, 113].

7.1.2 Predictive ability of the model

The quality of fit to the training data indicates that the model is reasonably complete for the stripe 2 and 3 *eve* enhancers at the developmental time assayed. Further support for the current model is afforded by its predictive ability, which is a qualitative advance on previous efforts. In *D. melanogaster*, high quality predictions for the stripe 5 and 4_6 enhancers were obtained. I was also able to correctly predict the effects of site-directed mutations affecting only 2-6 base pairs. This result indicates that the model might ultimately

have utility in predicting the effects of SNPs, a point with implications for both medicine and evolutionary biology.

With respect to stripes 2, 3, and 7 in non-*D.melanogaster* species, the model correctly predicts stripe 2 expression in all enhancers assayed [38, 2, 102]. This is a strong test of the model since these diverged enhancers have considerable turnover in binding site composition [38, 99] among the Drosophilids and no homology except for short sequences involving overlapping binding sites in Sepsids [102]. The ability of the model to predict gene expression of these sequences indicates that it has captured major elements of the fundamental rules of transcription.

With respect to predictions of the expression of other *Drosophila* genes, I obtained good results for the *h* 3_4 and *run* 3_7 enhancers (Figure 4.1E5, E4). The predicted *run* 1_7 enhancer pattern is in perfect alignment with *run* stripe 2 rather than stripe 1. This last prediction is probably but not certainly erroneous. Although I am aware of no published co-staining data of the *run* 1_7 enhancer with native *run* protein or RNA, such data exists for a larger segment of DNA which drives *run* stripes 1, 3, and 5 and contains *run* 3_7 [114]. With respect to gap genes, the model has good agreement of predicted patterns for the *hb* pThb1 and *Kr* CD1 enhancers, but is not able to predict the other *Kr* and *hb* enhancers, *kni* and *gt* correctly. In the case of *gt*, the lack of expression in the native domain is a consequence of the presence of numerous Gt binding sites. There are indications that Gt has autoactivation activity [115]. It is possible that Gt has a coactivator binding to its own promoter that was not included in this study.

7.1.3 Analytic ability of the *in silico* transcription system

The ability to infer the *cis*-regulatory mechanisms underlying significantly altered gene expression of the four fusion constructs, M3_2, M32, M2_3 and M23 and highly conserved stripe 2 expression of six *Drosophila* and Sepsid S2Es by functional analysis demonstrates the analytic power of the model. This power stems from the fact that the model keeps track of the fractional occupancy of each individual binding site. This level of resolution combined with the capability of removing a specific mechanism *in silico* allows us to assay the relative contributions of the multiple mechanisms of transcriptional control that operate simultaneously. Moreover, fractional occupancy in turn depends on affinity and hence DNA sequence, affording us a way to precisely characterize regulatory changes introduced at the level of individual base pairs. The analytic ability of the model allowed me to investigate fundamental questions about *cis*-regulatory biology, as discussed in the following sections.

7.2 A dynamic view of enhancer and regulatory logic

7.2.1 Are enhancers elementary objects?

Although enhancers are frequently referred to as *cis*-regulatory "elements", they are not elementary or fundamental objects. They are not elementary because they do not have well-defined boundaries. I demonstrated the context-dependent border of MSE2 in this study by showing that the increased level of stripe 2 expression in M32 was a consequence of the recruitment of 40% of MSE3 to become a functional component of MSE2 (Figure 5.3D). Moreover,

MSE2 and S2E both drive stripe 2 and can rescue lethality [108], and MSE2 is not completely minimal in the sense that smaller regions of DNA within it can drive weak and variable stripe 2 expression [31, 6].

The analysis of *D.mel eve* S2E (see Section 7.3) identified two functional binding site clusters, the hb-3 and hb-2 clusters (Figure 6.4 and Figure 6.5). Each of the clusters contain a Hb site and multiple Bcd, Gt and Kr sites around the Hb site. The model indicates that the hb-3 cluster, about 150 bp sequence at the 3' end of stripe 2 enhancer, has a complete function that integrates transcriptional input and drives gene expression specifically at stripe 2. Note that the smaller region of DNA within MSE2, called proximal cluster [10], driving weak stripe 2 expression is the core region of the hb-3 cluster. However, the model also demonstrates that the full levels of the observed stripe 2 expression driven by S2E requires additional activation input from hb-2 cluster (Figure 6.5) and two bcd sites, bcd-4 and 5 at the 5' end (Figure 5.4).

I have also found widely distributed many small regions that are highly active at stripe 7 (Figure 5.2). They exist in MSE3, the genomic sequences between MSE3 and MSE2, MSE2 and TSS and even in MSE2. They are only weakly activated so that individual regions wouldn't be identified experimentally. However, the model suggests that such small regions would be absolutely necessary for the complete functionality of *eve* stripes. For example, the model demonstrated that hb-2 cluster, which is located outside of MSE2 and only weakly activated at stripe 2, helps hb-3 cluster to maintain the full levels of S2E-driven stripe 2 expression, but also adjusts the peak position to be at the peak stripe 2 of *eve* (Figure 6.5 in Section 7.3).

Widespread small stripe 7 regions found in *eve* upstream DNA suggest that enhancers might not be functionally fundamental objects. In addition,

most enhancers drive expression domains which are similar to but not identical with those driven by the intact locus. With respect to *eve* stripe 3, this point has been evident for some time in mutant genotypes, although the additional sequences required are as yet unidentified (compare [26, Figure 4B] with [116, Figure 5A] and [26, Figure 5B]). In the case of *hb*, the lack of fidelity is evident in wild type and complete fidelity is restored by a shadow enhancer [15].

7.2.2 Two requirements for the independent enhancer action

Our ability to model expression of the fusion constructs and to predict expression of stripes 2, 3, and 7 driven by large 5' upstream sequence demonstrated that short-range coactivation and short-range repression are essential for the independent action of multiple enhancers. Previous theoretical models failed to recapitulate *eve* stripes 2 and 3 simultaneously [56, 57], most probably because of a failure to incorporate short-range coactivation of Hb by Bcd. Because MSE2 and MSE3 both contain Hb binding sites, if Hb is a dedicated activator, the anterior border of stripe 3 wouldn't be established. On the other hand, if Hb is a committed repressor, stripe 2 would never be formed, suggesting the importance of coactivation mechanism for the independent action of the two enhancers. Furthermore, if such coactivation acts over long distance, stripe 3 expression would never be made even though coactivation of Hb by Bcd is incorporated in the model. All of the modelling results I investigated in this work consistently indicate that short-range coactivation is required for proper expression of stripe 2 and 3 in the M3_2.

It is evident that modelling large regulatory sequence having both MSE2 and MSE3 requires short-range quenching mechanism. If repressors act over

long distance without a limit, Hb bound to MSE3 would repress MSE2 because the two enhancers co-exist in the regulatory DNA, contradictory to the observed independent stripes. Indeed, lines of evidence from both experiment [31, 6, 26, 28] and theory [96, 50] indicate that *eve* stripes are generated by repression from gap genes. Because gap gene expression domains are wider than *eve* stripes, silencing from these genes would result in a repressed region comparable in size to that of a gap domain and could not produce the observed stripes.

7.3 Mechanisms governing functional conservation

Striking functional conservation of S2Es from four different species, *D.mel*, *D.yak*, *D.pse* and *D.ere* was systemically investigated by Ludwig et al. (1998) and (2005). Despite the sequence differences in both the footprinted binding sites and the sequences between the binding sites, spatiotemporal expression pattern driven by the four S2Es is indistinguishable in *D.mel* embryo [38, 2]. It was also reported that S2Es of Sepsid species whose binding sites have been almost completely rearranged can still produce identical outputs in *Drosophila* embryo [102]. These results clearly demonstrate the substantial flexibility of the enhancer structure.

The fact that the model correctly predicts the spatiotemporal control of gene expression of the four S2Es (Figure 4.1C1-C3) allowed me to investigate the rules permitting the conserved function. Among hundreds of computationally identified binding sites in each S2E, the functional binding sites that are highly involved in activating and regulating gene expression were found

in the 3' end of S2E. At stripe 2, gene expression is almost entirely driven by 7 activators binding to bcd-2, bcd-*, bcd-1, bcd-(–1), bcd-(–2), hb-3 and hb-2 in the region (Figure 5.4). The fact that any of experimentally mutated sites in the cluster cause almost total loss of stripe 2 expression [6, 32] and smaller region having bcd-1 and hb-3 can drive weak stripe 2 expression [6] supports this result. Note that five of them, bcd-2, bcd-1, bcd-*, hb-3 and hb-2 are footprint sites [30, 10]. Other two sites are computationally identified but SELEX and MITOMI data for Bcd both indicate that Bcd binds to the sites *in vitro*. In addition, footprint site kr-1 and predicted site gt-(–1) overlap bcd-(–2) and bcd-(–1) respectively (Figure 5.4A). These result suggest that bcd-(–1) and bcd-(–2) might be functional and be regulated by Kr and Gt *in vivo*.

Functional analysis of the TF binding to DNA and protein-protein interactions taking place on the binding sites reveals intriguing shared features (Figure 6.11). I have found that all of the four S2Es heavily depend on synergistic coactivation of Bcd and Hb to initiate transcription. In addition, Gt and Kr actively binds to S2E at anterior and posterior border respectively. Furthermore, it is also conserved in all the four S2Es that Gt and Kr repress the activating action of Bcd and Hb by the two tier of repression mechanism, competition and short-range quenching. Because the conserved coactivation and short-range quenching mechanisms act over a distance of 100 to 150 bp, they might allow considerable flexibility in binding site arrangement, so that binding site rearrangements up to 150 bp wouldn't seriously alter the net regulatory function of the top contributing clusters. Indeed, even though the observed sequences differ, the top contributing binding sites, bcd-2, bcd-1, bcd-(–1), bcd-(–2), hb-3 and hb-2 found in *D.mel* are functionally conserved

in *D.yak*, *D.ere* and *D.pse* and their binding affinities are almost identical or higher than that of *D.mel* (Figure 6.11). Furthermore, kr-3, kr-1 and gt-1 are functionally conserved and located in the range regulating the Bcd and Hb which are bound to the top contributing sites in the four different S2Es (Figure 6.11).

However, I demonstrated that the conserved molecular interactions are not sufficient for the conserved stripe 2 expression. As seen in Figure 6.13, in *D. yak* and *D. pse* S2Es, levels of stripe 2 expression are significantly reduced and their expression is shifted anteriorly when only the conserved interactions between species are allowed. In the case of *D. pse*, substantial derepression in the 1-2 inter-stripe region is observed. Using the *in silico* transcription model, I identified novel cooperative interactions between bcd-2 and bcd-1 in *D. yak* and two novel functional repressor sites in *D. pse*, which compensates for the structural differences *in silico*. In addition, the model prediction result suggests that the conserved stripe 2 expression driven by Sepsid S2Es is maintained by recruiting novel *trans*-acting input Cad as an activator. Future experiments will allow us to test whether these novel interactions play a major role in maintaining conserved stripe 2 expression *in vivo*. If so, this would be the first documented case demonstrating compensatory adaptation of *eve* S2E function through the novel molecular interactions not seen in the *D. mel*.

7.4 Limitations of the current *in silico* transcription system

Although a qualitative improvement over previous efforts, the work presented here does not constitute a complete solution to the problem of understanding

cis-regulatory logic. In assessing what is required for a complete solution to that problem, it is important to distinguish between limitations on available data and limitations inherent in the model. With respect to the predictions reported here, it is significant that I was able to predict the expression of highly rearranged Sepsid enhancers up to the resolution of available data, while the results for gap and pair-rule enhancers other than *eve* in *melanogaster* were more mixed. I believe that this is a consequence of the fact that some of these enhancers utilize TFs and perhaps interactions among the TFs that are not important for driving *eve* stripes 2, 3, and 7. A possible example of a missing interaction is the spurious auto-repression of *gt* in its own expression domain (Figure 4.2F and G).

In addition, even though the model predicts the expression pattern of S2Es from the four *Drosophila* species, *D.mel*, *D.yak*, *D.ere* and *D.pse* accurately, the predicted levels of gene expression were incorrect (Figure 4.1C1-C3) except for *D. yak* when the cooperative interaction range of Bcd is set to 70 bp (Figure 6.14E). The *D.ere* S2E-driven *eve* expression is much weaker than the other species in *D.mel* embryo [2], but the model predicts that *D.ere* S2E drives higher levels of gene expression than *D.mel* S2E. Given that the expression training set used in this study was driven by only 2.5 kb of DNA from a single locus, it is likely that the use of a more diverse training set would result in greatly improved predictions.

Our predictions of expression driven by large DNA segments are less clean than those of single enhancers in the sense that they required hand tuning of the threshold θ, the activation energy barrier of transcription initiation (Figure 3.5 Eq. 10), to prevent completely saturated expression domains comprising stripes 2-3 and 4-6 respectively. This saturation appears to involve a lack

of balance between activators and repressors as the length of modeled DNA increases, but it is not possible at this time to distinguish between problems with the model and the training data. With respect to the model, this lack of balance may stem from the unlimited range of activators and the limited range of quenchers.

In order to know whether this model property is biologically correct or incorrect, it is necessary to quantitatively determine how the amplitude of a given stripe changes as it is driven by larger DNA fragments. This point is not captured in our training data because only the four fusion constructs, all of similar total length, were transformed to a targeted site. Shorter and longer DNA fragments used in the model training (Figure 3.9) were not targeted transformants and hence required a free parameter scaling the amplitude to account for position effect. The quantitative characterization of expression driven by fragments of varying size transformed to a common chromosomal site is an important experimental task for future work. It will also be important to generate rescue constructs containing both native and *lac*Z message in order to standardize between observed levels of native and reporter transcripts. I believe that these results, while imperfect, demonstrate the feasibility of constructing a precise, quantitative, and predictive model of an entire locus that would also account for its enhancer structure.

Limitations exist not only for the data but also for the model. The model itself is clearly incomplete in the sense that it does not contain a complete set of regulatory mechanisms. I incorporated a representation of a regulatory mechanism into the model only when there is specific evidence that it acts in the experimental system under consideration. This means that some mechanisms that are known to occur and are easy to represent mathematically,

such as corepression [117, 118] and cooperative binding by heterologous pairs of proteins [119], were not incorporated in this study because there is no evidence that they occur in that portion of the *eve* control region used for the training set. With respect to cooperative binding to DNA, there is a pressing need for high-throughput quantitative data. Microfluidic methods provide a feasible way to address this problem [120].

7.5 Future prospects

A more fundamental issue concerns the role of chromatin structure, an area where new theoretical ideas are required. Silencing is thought to involve changes in chromatin structure. This phenomenon cannot be modeled simply by modifying the distance function $q(d)$ for short range repression because such a modification cannot account for radical changes in the range of silencing observed when the number of silencer binding sites is altered [121]. It is likely that the way forward involves spreading inactivation models of the type proposed by Sengupta [122]. A critical unsolved problem is the incorporation of regulators into such models, and the study of so-called chromatin marks [123, 124] may be useful in this regard.

The *eve* locus itself may prove a useful system in which to pursue such studies. The proximal 1.7 kb of 5' noncoding DNA from the *eve* gene drives a pattern of expression in cleavage cycle 13 and the first 6 minutes of cleavage cycle 14A that closely resembles that of the entire locus [50, 29]. In contrast, the fusion constructs considered here do not express at these early stages (Figure 2.5A), nor does MSE2 (data not shown). Moreover, changes of expression occur after T6 that suggest early signs of the midblastula transition. These

changes take the form of decreases of expression in stripes 3 and 7 by T8, together with a loss of registration with the native *eve* pattern caused by the fact that reporter expression does not follow the anterior shifts observed in expression driven by the native locus [29]. It is possible that these changes of chromatin state can be probed in a manner that will suggest new theoretical ideas by conducting ChIP-seq or hypersensitivity studies on embryos prepared with extremely high temporal and/or spatial resolution [125].

One of the putative chromatin regulator for *eve* might be a maternal zinc-finger protein Zelda (Zld). Recently, it has been shown that Zld is required for activation of *eve* stripes 2, 3 and 7 [34]. Zld is a large protein of $\tilde{1}$80 kD and has four zinc-fingers near the C-terminus that are involved in DNA binding [126]. Zld is detected in nuclei as early as C2 (cleavage cycle 2), which is earlier than any known maternal factors and is present ubiquitously until gastrulation in the *Drosophila melanogaster* embryo. Later, Zld becomes restricted to the nervous system and specific head regions. Zld is known to bind specifically to several TAG containing DNA sequences, called TAGteam sites, and is required for activation of a large number of genes early in development [127].

It has been shown that Zld binds to non-canonical sites, which are different from the previously identified TAGteam sites, in *eve* MSE2 and MSE3 *in vitro*, and, in the case of MSE3, deletion of a non-canonical Zld site (5'-CAGGCAA-3') caused a strong reduction in expression levels of both stripe 3 and 7 *in vivo* [34]. Furthermore, in an embryo lacking maternal expression of Zld, expression driven by MSE2 and MSE3 was completely abolished. Note that the *eve* stripe 2 and 3 enhancers are located in high-ranking Zld-bound regions where Zld binding was observed in an *in vivo* ChIP-seq experiment [126].

Several lines of evidence suggest that Zld binding might increase transcrip-

tional activity by facilitating the access of TFs [128, 126], but acting locally [34]. *In vivo* ChIP-seq experiment shows that Zld is bound during C8-9 to a large fraction of the transcriptional enhancers that control gene expression at C14A [128]. Early Zld binding is strongly associated with open chromatin and transcription factor binding at C14A. Furthermore, in many cases, loss of Zld does not completely abolish gene expression but results in delayed transcriptional activation [126]. Thus, rather than being specifically involved in the control of gene expression, Zld binding at earlier stage might facilitate the subsequent binding of TFs that drive gene expression at C14A. Incorporation of Zld as a chromatin regulator might extend the current *in silico* model to fit to the dynamic expression patterns observed in the entire C14, which will provide scientific insights for understanding the role of chromatin in regulating the dynamics of metazoan gene expression during embryo development.

Perhaps one of the biggest advance in understanding transcriptional control will come with the construction of entire locus model. In higher eukaryotes, gene expression is regulated by large *cis*-regulatory sequence that contains multiple enhancers and neighboring genomic sequences. Several lines of evidences indicate that there is a serious functional difference between an isolated enhancer and an entire regulatory sequence of a gene [13, 14, 129, 130]. Currently however, the majority of works on the functional analysis of the transcriptional control is limited to short enhancer-reporter constructs. I believe that characterizing the simultaneous action of the multiple enhancers and understanding how the enhancers and their neighboring genomic sequences controls gene expression and to what extent would be a substantial advance for the deep understanding of the large contiguous regulatory sequence, and ultimately the genome.

Bibliography

[1] Maston GA, Evans SK, Green MR (2006) Transcriptional Regulatory Elements in the Human Genome. Annual Review of Genomics and Human Genetics 7: 29-59.

[2] Ludwig MZ, Palsson A, Alekseeva E, Bergman CM, Nathan J, et al. (2005) Functional evolution of a *cis*-regulatory module. PLoS Biology 3: e93.

[3] Naar AM, Lemon BD, Tjian R (2001) Transcriptional coactivator complexes. Annual Reviews of Biochemistry 70: 475-501.

[4] Park JM, Gim BS, Kim JM, Yoon JH, Kim HS, et al. (2001) *Drosophila* mediator complex is broadly utilized by diverse gene-specific transcription factors at different types of core promoters. Molecular and Cellular Biology 21: 2312-2323.

[5] Ong CT, Corces VG (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature Reviews Genetics 12: 283-293.

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

[7] Hanes SD, Riddihough G, Ish-Horowicz D, Brent R (1994) Specific DNA recognition and intersite spacing are critical for action of the bicoid morphogen. Molecular and Cellular Biology 14: 3364-3375.

[8] Ma X, Yuan D, Diepold K, Scarborough T, Ma J (1996) The *Drosophila* morphogenetic protein Bicoid binds DNA cooperatively. Development 112: 1195-1206.

[9] Burz DS, Rivera-Pomar R, Jaeckle H, Hanes SD (1998) Cooperative DNA-binding by Bicoid provides a mechanism for threshold-dependent

gene activation in the *Drosophila* embryo. The EMBO journal 17: 5998-6009.

[10] Small S, Kraut R, Hoey T, Warrior R, Levine M (1991) Transcriptional regulation of a pair-rule stripe in *Drosophila*. Genes and Development 5: 827-839.

[11] Gray S, Cai H, Barolo S, Levine M (1995) Transcriptional repression in the *Drosophila* embryo. Philosophical Transactions of the Royal Society 349: 257-262.

[12] Gray S, Szymanski P, Levine M (1994) Short-range repression permits multiple enhancers to function autonomously within a complex promoter. Genes and Development 8: 1829-1838.

[13] Hong JW, Hendrix D, Levine M (2008) Shadow enhancers as a source of evolutionary novelty. Science 321: 1314.

[14] Perry M, Boettiger AN, Bothma JP, Levine M (2010) Shadow enhancers foster robustness of *Drosophila* gastrulation. Current Biology 20: 1562-1567.

[15] Perry M, Boettiger AN, Levine M (2011) Multiple enhancers ensure precision of gap gene-expression patterns in the *Drosophila* embryo. Proceedings of the National Academy of Sciences of the United States of America 108: 13570-13575.

[16] Gilbert SF (2003) Developmental Biology. Sunderland, MA: Sinauer Associates, seventh edition.

[17] Foe VE, Alberts BM (1983) Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in *Drosophila* embryogenesis. The Journal of Cell Science 61: 31-70.

[18] Renzis S, Elemento O, Wieschaus STE (2007) Unmasking activation of the zygotic genome using chromosomal deletions in the *Drosophila* embryo. PLoS Biology 5: e117.

[19] Simcox AA, Sang JH (1983) When does determination occur in *Drosophila* embryos? Developmental Biology 97: 212-221.

[20] Akam M (1987) The molecular basis for metameric pattern in the *Drosophila* embryo. Development 101: 1-22.

[21] Ingham PW (1988) The molecular genetics of embryonic pattern formation in *Drosophila*. Nature 335: 25-34.

[22] Frasch M, Hoey T, Rushlow C, Doyle HJ, Levine M (1987) Characterization and localization of the even-skipped protein of *Drosophila*. The EMBO Journal 6: 749-759.

[23] Harding K, Hoey T, Warrior R, Levine M (1989) Autoregulatory and gap gene response elements of the *even-skipped* promoter of *Drosophila*. The EMBO Journal 8: 1205-1212.

[24] Goto T, MacDonald P, Maniatis T (1989) Early and late periodic patterns of *even-skipped* expression are controlled by distinct regulatory elements that respond to different spatial cues. Cell 57: 413-422.

[25] Small S, Arnosti DN, Levine M (1993) Spacing ensures autonomous expression of different stripe enhancers in the *even-skipped* promoter. Development 119: 767-772.

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

[27] Fujioka M, Jaynes JB, Goto T (1995) Early *even-skipped* stripes act as morphogenetic gradients at the single cell level to establish *engrailed* expression. Development 121: 4371-4382.

[28] Fujioka M, Emi-Sarker Y, Yusibova GL, Goto T, Jaynes JB (1999) Analysis of an *even-skipped* rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients. Development 126: 2527-2538.

[29] Surkova S, Kosman D, Kozlov K, Manu, Myasnikova E, et al. (2008) Characterization of the *Drosophila* segment determination morphome. Developmental Biology 313: 844-862.

[30] Stanojevic D, Hoey T, Levine M (1989) Sequence-specific DNA-binding activities of the gap proteins encoded by *hunchback* and *Krüppel* in *Drosophila*. Nature 341: 331-335.

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

[32] Arnosti DN, Barolo S, Levine M, Small S (1996) The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development 122: 205-214.

[33] Andrioli LPM, Vasisht V, Theodosopoulou E, Oberstein A, Small S (2002) Anterior repression of a *Drosophila* stripe enhancer requires three position-specific mechanisms. Development 129: 4931-4940.

[34] Struffi P, Corado M, Kaplan L, Yu D, Rushlow C, et al. (2011) Combinatorial activation and concentration-dependent repression of the *Drosophila even skipped* stripe 3+7 enhancer. Development 138: 4291-4299.

[35] Hou XS, Melnick MB, Perrimon N (1996) *marelle* Acts Downstream of the *Drosophila* HOP/JAK Kinase and Encodes a Protein Similar to the Mammalian STATs. Cell 84: 411-419.

[36] Yan R, Small S, Desplan C, Dearolf CR, Jr JED (1996) Identification of a *stat* gene that functions in *Drosophila* development. Cell 84: 421-430.

[37] Macdonald PM, Ingham P, Struhl G (1986) Isolation, structure, and expression of *even-skipped*: a second pair-rule gene of *Drosophila* containing a homeo box. Cell 47: 721-734.

[38] Ludwig MZ, Patel NH, Kreitman M (1998) Functional analysis of eve stripe 2 enhancer evolution in *drosophila*: rules governing conservation and change. Development 125: 949-958.

[39] Simpson-Brose M, Treisman J, Desplan C (1994) Synergy between the Hunchback and Bicoid morphogens is required for anterior patterning in *Drosophila*. Cell 78: 855-865.

[40] Arnosti D, Gray S, Barolo S, Zhou J, Levine M (1996) The gap protein Knirps mediates both quenching and direct repression in the *Drosophila* embryo. The EMBO Journal 15: 3659-3666.

[41] Nibu Y, Zhang H, Levine M (1998) Interaction of short-range repressors with *Drosophila* CtBP in the embryo. Science 280: 101-104.

[42] Nibu Y, Zhang H, Bajor E, Barolo S, Small S, et al. (1998) dCtBP mediates transcriptional repression by Knirps, Krüppel and Snail in the *Drosophila* embryo. The EMBO Journal 17: 7009-7020.

[43] Hewitt GF, Strunk B, Margulies C, Priputin T, Wang XD, et al. (1999) Transcriptional repression by the *Drosophila* Giant protein: Cis element positioning provides an alternative means of interpreting an effector gradient. Development 126: 1201-1210.

[44] Nibu Y, Zhang H, Levine M (2001) Local action of long-range repressors in the *Drosophila* embryo. The EMBO Journal 20: 2246-2253.

[45] Nibu Y, Levine M (2001) CtBP-dependent activities of the short-range Giant repressor in the *Drosophila* embryo. Proceedings of the National Academy of Sciences USA 98: 6204-6208.

[46] Gray S, Levine M (1996) Short-range transcriptional repressors mediate both quenching and direct repression within complex loci in *Drosophila*. Genes and Development 10: 700-710.

[47] Reinitz J, Hou S, Sharp DH (2003) Transcriptional control in *Drosophila*. ComPlexUs 1: 54-64.

[48] Poustelnikova E, Pisarev A, Blagov M, Samsonova M, Reinitz J (2004) A database for management of gene expression data in situ. Bioinformatics 20: 2212-2221.

[49] Pisarev A, Poustelnikova E, Samsonova M, Reinitz J (2008) FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. Nucleic Acids Research 37: D560-D566.

[50] Janssens H, Hou S, Jaeger J, Kim AR, Myasnikova E, et al. (2006) Quantitative and predictive model of transcriptional control of the *Drosophila melanogaster even skipped* gene. Nature Genetics 38: 1159-1165.

[51] Louis M, Holm L, Sanchez L, Kaufman M (2003) Theoretical model for the regulation of *Sex-lethal*, a gene that controls sex determination and dosage compensation in *Drosophila melanogaster*. Genetics 165: 1355-1384.

[52] Zinzen RP, Senger K, Levine M, Papatsenko D (2006) Computational models for neurogenic gene expression in the *Drosophila embryo*. Current Biology 16: 1358-1365.

[53] Zinzen RP, Papatsenko D (2007) Enhancer responses to similarly distributed antagonistic gradients in development. PLoS Computationial Biology 3: e84.

[54] Papatsenko D, Levine M (2008) Dual regulation by the hunchback gradient in the drosophila embryo. Proceedings of the National Academy of Sciences of the United States of America 105: 2901-2906.

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

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

[57] He X, Samee MAH, Blatti C, Sinha S (2010) Thermodynamics-based models of transcriptional regulation by enhancers: The roles of synergistic activation, cooperative binding and short-range repression. PLoS Computational Biology 6: e1000935.

[58] Janssens H, Kosman D, Vanario-Alonso CE, Jaeger J, Samsonova M, et al. (2005) A high-throughput method for quantifying gene expression data from early *Drosophila* embryos. Development, Genes and Evolution 215: 374-381.

[59] Myasnikova E, Samsonova M, Kosman D, Reinitz J (2005) Removal of background signal from *in situ* data on the expression of segmentation genes in *Drosophila*. Development, Genes and Evolution 215: 320-326.

[60] Surkova S, Myasnikova E, Janssens H, Kozlov KN, Samsonova A, et al. (2008) Pipeline for acquisition of quantitative data on segmentation gene expression from confocal images. Fly 2: 58-66.

[61] Oberstein A, Pare A, Kaplan L, Small S (2005) Site-specific transgenesis by cre-mediated recombination in *Drosophila*. Nature Methods 2: 583-585.

[62] Myasnikova E, Samsonova A, Kozlov K, Samsonova M, Reinitz J (2001) Registration of the expression patterns of *Drosophila* segmentation genes by two independent methods. Bioinformatics 17: 3-12.

[63] Binary R, Perrimon N (1994) Stripe-specific regulation of pair-rule genes by hopscotch, a putative Jak family tyrosine kinase in *Drosophila*. Genes and Development 8: 300-312.

[64] Turkson J, Jove R (2000) STAT proteins: novel molecular targets for cancer drug discovery. Oncogene 19: 6613-6626.

[65] Lusk RW, Eisen MB (2010) Evolutionary mirages: Selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers. PLoS Genetics 6: e1000829.

[66] Erives A, Levine M (2004) Coordinate enhancers share common organizational features in the *Drosophila* genome. Proceedings of the National Academy of Sciences USA 101: 3851-3856.

[67] Sauer F, Hansen SK, Tjian R (1995) Multiple TAFIIs Directing Synergistic Activation of Transcription. Science 270: 1783-1788.

[68] Johnson AD, Meyer BJ, Ptashne M (1979) Interactions between DNA-bound repressors govern regulation by the λ phage repressor. Proceedings of the National Academy of Sciences USA 76: 5061-5065.

[69] Zeidler MP, Bach EA, Perrimon N (2000) The roles of the *Drosophila* JAK/STAT pathway. Oncogene 19: 2598-2606.

[70] Olesnicky EC, Brent AE, Tonnes L, Walker M, Pultz MA, et al. (2006) A caudal mRNA gradient controls posterior development in the wasp nasonia. Development 133: 3973-3982.

[71] Bergman CM, Carlson JW, Celniker SE (2005) *Drosophila* DNase I footprint database: A systematic genome annotation of transcription factor binding sites in the fruitfly, *d. melanogaster*. Bioinformatics 21: 1747-1749.

[72] Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16: 16-23.

[73] Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *e. coli*. Nucleic Acids Research 10: 2997-3011.

[74] Berg G, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. Journal of molecular biology 193: 723-50.

[75] Orgawa N, Biggin MD (2012) High-Throughput SELEX Determination of DNA Sequences Bound by Transcription Factors In Vitro. Methods in Molecular Biology 786: 51-63.

[76] Bailey TL, Williams N, Misleh C, Li WW (2006) Meme: discovering and analyzing dna and protein sequence motifs. Nucleic acids research 34: 369-373.

[77] Bailey T, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the International Conference on Intelligent Systems for Molecular Biology 2: 28-36.

[78] Hu J, Li B, Kihara D (2005) Limitations and potentials of current motif discovery algorithms. Nucleic Acids Research 33: 4899-4913.

[79] Li X, MacArthur S, Bourgon R, Nix D, Pollard DA, et al. (2008) Transcription factors bind thousands of active and inactive regions in the *Drosophila*blastoderm. PLoS Biology 6: e27.

[80] Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, et al. (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. Proceedings of the National Academy of Sciences USA 99: 757-762.

[81] Rajewsky N, Vergassola M, Gaul U, Siggia ED (2002) Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. BMC Bioinformatics 3: 30.

[82] Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, et al. (2008) A systematic characterization of factors that regulate *drosophila* segmentation via a bacterial one-hybrid system. Nucleic Acids Research : 1-14.

[83] Berg OG, Hippel PH (1988) Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. Journal of Molecular Biology 200: 709-723.

[84] Ackers GK, Johnson AD, Shea MA (1982) Quantitative model for gene-regulation by lambda-phage repressor. Proceedings of the National Academy of Sciences USA 79: 1129-1133.

[85] Louis M (2003) Sex determination in *Drosophila melanogaster*: a theoretical model for the regulation of the *Sex-lethal* gene. Ph.D. thesis, University of Cambridge.

[86] Reinitz J, Vaisnys JR (1990) Theoretical and experimental analysis of the phage lambda genetic switch implies missing levels of cooperativity. The Journal of Theoretical Biology 145: 295-318.

[87] Ptashne M, Gann A (1997) Transcriptional activation by recruitment. Nature 386: 569-577.

[88] Lemon B, Tijian R (2000) Orchestrated response: a symphony of transcription factors for gene control. Genes and Development 14: 2551-2569.

[89] Berger SL, TriezenbergÂă WDCACSJ, Guarente L (1990) Selective inhibition of activated but not basal transcription by the acidic activation domain of VP16: Evidence for transcriptional adaptors. Cell 61: 1199-1208.

[90] Berger SL, Pina B, Silverman N, Marcus GA, Agapite J, et al. (1990) Genetic isolation of ADA2: A potential transcriptional adaptor required for function of certain acidic activation domains. Cell 70: 251-265.

[91] Tamkun JW, Deuring R, Scott MP, Kissinger M, Pattatucci AM, et al. (1992) Brahma: a regulator of *Drosophila* homeotic genes structurally related to the yeast transcription activator SNF2/SWI2. Cell 68: 561-572.

[92] Saurin AJ, Shao Z, Erdjument-Bromage H, Tempst P, Kingston RE (2001) A *Drosophila* Polycomb group complex includes Zeste and dTAFII proteins. Nature 412: 655-660.

[93] Han K, Levine M, Manley JL (1989) Synergistic activation and repression of transcription by *Drosophila* homeobox proteins. Cell 56: 573-583.

[94] Lam J, Delosme JM (1988) An efficient simulated annealing schedule: Derivation. Technical Report 8816, Yale Electrical Engineering Department, New Haven, CT.

[95] Lam J, Delosme JM (1988) An efficient simulated annealing schedule: Implementation and evaluation. Technical Report 8817, Yale Electrical Engineering Department, New Haven, CT.

[96] Reinitz J, Sharp DH (1995) Mechanism of *eve* stripe formation. Mechanisms of Development 49: 133-158.

[97] Rubel O, Weber GH, Keranen SVE, Fowlkes CC, Hendriks LCL, et al. (2006) PointCloudXplore: Visual Analysis of 3D Gene Expression Data Using Physical Views and Parallel Coordinates. In: Eurographics/IEEE-VGTC Symposium on Visualization. pp. 203-210.

[98] Abramoff MD, Magalhaes P, Ram S (2004) Image Processing with ImageJ. Biophotonics International 11: 36-42.

[99] Ludwig MZ, Bergman CM, Patel NH, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403: 564-567.

[100] Gilbert DG (2005). DroSpeGe, a public database of *Drosophila* species genomes. Http://insects.eugenes.org/DroSpeGe/.

[101] Gilbert DG (2007) DroSpeGe: rapid access database for new *Drosophila* species genomes. Nucleic Acids Research 35: D480-D485.

[102] Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB (2008) Sepsid *even-skipped* enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. PLoS Genetics 4: e1000106.

[103] Howard KR, Struhl G (1990) Decoding positional information: regulation of the pair-rule gene *hairy*. Development 110: 1223-1231.

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

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

[106] Dearolf CR, Topol J, Parker CS (1989) The *caudal* gene product is a direct activator of *fushi tarazu* transcription during *Drosophila* embryogenesis. Nature 341: 340-343.

[107] Sauer F, Hansen SK, Tjian R (1995) DNA Template and Activator-Coactivator Requirements for Transcriptional Synergism by *Drosophila* Bicoid. Science 270: 1825-1828.

[108] Ludwig MZ, Manu, Kittler R, White KP, Kreitman M (2011) Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. PLoS Genetics .

[109] Levenshtein VI (1966) Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady 10: 707.

[110] Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, et al. (2003) LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA. Genome Research 13: 721-731.

[111] He BZ, Holloway AK, Maerkl SJ, Kreitman M (2011) Does Positive Selection Drive Transcription Factor Binding Site Turnover? A Test with Drosophila Cis-Regulatory Modules. PLoS Genetics 7: e1002053.

[112] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102: 15545-15550.

[113] Manolio TA (2010) Genomewide association studies and assessment of the risk of disease. The New England journal of medicine 363: 166-176.

[114] Klingler M, Soong J, Butler B, Gergen JP (1996) Disperse versus compact elements for the regulation of *runt* stripes in *Drosophila*. Developmental Biology 177: 73-84.

[115] Jaeger J, Blagov M, Kosman D, Kozlov KN, Manu, et al. (2004) Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster*. Genetics 167: 1721-1737.

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

[117] Kirov N, Zhelnin L, Shah J, Rushlow C (1993) Conversion of a silencer into an enhancer: evidence for a co-repressor in dorsal-mediated repression in *Drosophila*. The EMBO Journal 12: 3193-3199.

[118] Kirov N, Lieberman P, Rushlow C (1996) The transcriptional corepressor DSP1 inhibits activated transcription by disrupting TFIIA-TBP complex formation. The EMBO Journal 15: 7079-7087.

[119] Kerppola TK, Curran T (1991) Fos-Jun heterodimers and Jun homodimers bend DNA in opposite orientations: Implications for transcription factor cooperativity. Cell 66: 317-326.

[120] Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. Science 315: 233-237.

[121] Barolo S, Levine M (1997) Hairy mediates dominant repression in the *Drosophila* embryo. The EMBO Journal 16: 2883-2891.

[122] Sedighi M, Sengupta AM (2007) Epigenetic chromatin silencing: bistability and front propagation. Physical Biology 4: 246-255.

[123] Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, et al. (2006) A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell 125: 315-326.

[124] Hon G, Hawkins RD, Ren B (2009) Predictive chromatin signatures in the mammalian genome. Human Molecular Genetics 18: R195-201.

[125] Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, et al. (2012) Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nature Genetics 44: 148-156.

[126] Nien CY, Liang HL, Butcher S, Sun Y, Fu S, et al. (2011) Temporal Coordination of Gene Networks by Zelda in the Early Drosophila Embryo. PLoS Genetics 7: e1002339.

[127] Liang HL, Nien CY, Liu HY, Metzstein MM, Rushlow NKC (2008) The zinc-finger protein Zelda is a key activator of the early zygotic genome in *Drosophila*. Nature 456: 400-404.

[128] Harrison MM, Li XY, Kaplan T, Botchan MR, Eisen MB (2011) Zelda Binding in the Early *drosophila melanogaster* Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition. PLoS Genetics 7: e1002266.

[129] Frankel N, Davis GK, Vargas D, Wang S, Payre F, et al. (2010) Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature 466: 490-493.

[130] Dunipace L, Ozdemir A, Stathopoulos A (2011) Complex interactions-between cis-regulatory modules in native conformation are critical for *Drosophila* snail expression. Development 138: 4075âĂŞ4084.

[131] Hawley DK, McClure WR (1982) Mechanism of activation of transcription initiation from the lambda-PRM promoter. The Journal of Molecular Biology 157: 493-525.

[132] Ludwig MZ, Kreitman M (1995) Evolutionary dynamics of the enhancer region of *even-skipped* in *Drosophila*. Molecular Biology and Evolution 12: 1002-1011.

[133] Sackerson C (1995) Patterns of conservation and divergence at the *even-skipped* locus of *Drosophila*. Mechanisms of Development 51: 199-215.

[134] Margolis JS, Borowsky ML, Steingrimsson E, Shim CW, Lengyel JA, et al. (1995) Posterior stripe expression of *hunchback* is driven from two promoters by a common enhancer element. Development 121: 3067-3077.

[135] Hoch M, Schröder C, Seifert E, Jäckle H (1990) *Cis*-acting control elements for *Krüppel* expression in the *Drosophila* embryo. The EMBO Journal 9: 2587-2595.

[136] Schroeder MD, Pearce M, Fak J, Fan HQ, Unnerstall U, et al. (2004) Transcriptional control in the segmentation gene network of *Drosophila*. PLoS Biology 2: e271.

[137] Pankratz MJ, Busch M, Hoch M, Seifert E, Jäckle H (1992) Spatial control of the gap gene *knirps* in the *Drosophila* embryo by posterior morphogen system. Science 255: 986-989.

[138] Riddihough G, Ish-Horowicz D (1991) Individual stripe regulatory elements in the *Drosophila hairy* promoter respond to maternal, gap, and pair-rule genes. Genes and Development 5: 840-854.

[139] Langeland JA, Carroll SB (1993) Conservation of regulatory elements controlling *hairy* pair-rule stripe formation. Development 117: 585-596.

[140] Häder T, Rosée AL, Ziebold U, Busch M, Taubert H, et al. (1998) Activation of posterior pair-rule stripe expression in response to maternal *caudal* and zygotic *knirps* activities. Mechanisms of Development 71: 177-186.

Appendix A

Diffusion limited Arrhenius rate law

The diffusion limited Arrhenius rate law (Figure 3.5, Eq. 10) was not formulated in the course of this dissertation. It was derived by Dr. Alexandre Ramos from a stochastic three state Markov process model, derived from a minimal model of diffusion-limited transcription initiation [131]. We imagine that the system can have the following three states, in which 1) there is no PolII bound to the basal promoter; 2) there is a PolII bound to the basal promoter, but the PolII is stalled; 3) there is a PolII bound to the basal promoter and transcription is initiated, but a new PolII cannot yet bind. Transitions can occur between states 1 and 2 in either direction, but state 3 can only be reached from state 2 and can only change to state 1. Every time the system enters state 3, one new transcript is initiated.

The probabilities P_1, P_2, and P_3 of finding the system in states 1, 2, and 3

respectively are governed by

$$\frac{dP_1}{dt} = -k_1 P_1 + k_{-1} P_2 + k_3 P_3,$$
$$\frac{dP_2}{dt} = k_1 P_1 - (k_{-1} + k_2) P_2,$$
$$\frac{dP_3}{dt} = k_2 P_2 - k_3 P_3,$$

where the k_i are first order rate constants. We wish to calculate the steady state probabilities $\overline{P_i}$ in terms of the kinetic rate constants k_i. In a steady state the derivatives vanish and we make use of the fact that probabilities add up to one, allowing us to write

$$\overline{P}_1 = \frac{k_3(k_{-1} + k_2)}{k_3(k_{-1} + k_2) + k_1(k_2 + k_3)},$$
$$\overline{P}_2 = \frac{k_1 k_3}{k_3(k_{-1} + k_2) + k_1(k_2 + k_3)},$$
$$\overline{P}_3 = \frac{k_1 k_2}{k_3(k_{-1} + k_2) + k_1(k_2 + k_3)}.$$

k_2 is the rate-limiting Arrhenius term used in previous non-diffusion limited versions of this model [47, 50], given by

$$k_2 = \exp\left[-(\Theta - M)\right].$$

The rate of transcription will be the probability of finding the system in state 3, given by

$$\overline{P}_3 = \frac{A \exp\left[-(\Theta - QM)\right]}{B + C \exp\left[-(\Theta - QM)\right]}.$$

In the absence of detailed kinetic information, we take $A = B = C = 1$ to obtain Eq. 10 in Figure 3.5.

Appendix B

Estimated parameter values

Model parameters used in this study are inferred from the observed expression patterns by fitting transcription models to quantitative data. Daggers indicate parameters held fixed during the training process. $S^R_{\text{construct}}$ is the positional effect scale factor for each reporter construct. $R_{\max}$ is the maximum rate of transcription. S^P_{ligand} is the scale factor for protein concentration. Other parameters are described in Section 3.2.2.

Table B.1: **Parameters of 5 best models.**

Parameter	Run number					Parameter	Run number				
	1	2	6	7	4cs_7		1	2	6	7	4cs_7
RMS	2.27	2.33	2.40	5.29	1.48	A^P_{Bcd}	1.36	0.69	0.03	0.08	3.99
$S^R_{\mathrm{M3\_2}}\dagger$	1	1	1	1	1	A^P_{Cad}	0.02	0.01	0.05	0.03	0.05
$S^R_{\mathrm{M32}}\dagger$	1	1	1	1	1	$A^P_{\mathrm{D-STAT}}$	3.98	3.98	1.09	3.99	3.98
$S^R_{\mathrm{M2\_3}}\dagger$	1	1	1	1	1	A^P_{Dichaete}	3.85	0.41	3.89	0.18	0.09
$S^R_{\mathrm{M23}}\dagger$	1	1	1	1	1	A^P_{Hb}	0.82	2.73	2.72	0.04	0.09
$S^R_{\mathrm{1.7kb}}$	0.3	0.3	0.2	1.41	N/A	A^P_{Kr}	0.07	0.13	0.03	2.95	0.1
S^R_{MSE2}	0.2	0.2	0.2	0.2	N/A	A^P_{Kni}	2.58	3.99	2.23	0.27	2.8
S^R_{MSE3}	0.8	0.8	0.8	0.8	N/A	A^P_{Gt}	3.99	2.07	2.53	0.04	0.02
$R_{\mathrm{max}}\dagger$	255	255	255	255	255	A^P_{Tll}	2.88	3.96	0.02	1.95	0.004
Θ	6.7	6.4	6.1	5.96	10.6	λ_{Bcd}	1.53	1.99	4.99	2.16	1.68
$E^C_{\mathrm{B-H}}$	0.34	0.44	0.21	0.35	0.35	λ_{Cad}	4.98	4.98	4.97	3.18	4.99
$E^C_{\mathrm{C-H}}$	0.66	0.99	0.33	0.88	0.88	$\lambda_{\mathrm{D-STAT}}$	1.62	1.98	2.58	0.69	0.89
$K^{\mathrm{coop}}_{\mathrm{B-B}}$	52	982	189	127	86	$\lambda_{\mathrm{Dichaete}}$	0.91	2.33	1.98	4.54	3.49
$D^C_{\mathrm{B-H}}$	165	161	158	150	150	λ_{Hb}	1.93	1.5	1.83	4.99	4.25
$D^C_{\mathrm{C-H}}$	57	58	70	22	28	λ_{Kr}	3.08	2.31	4.04	0.98	4.99
$D^{\mathrm{coop}}_{\mathrm{B-B}}\dagger$	60	20	60	60	60	λ_{Kni}	1.6	1.17	2.48	1.56	1.82
E^A_{Bcd}	0.5	0.5	0.06	0.0001	0.001	λ_{Gt}	1.25	1.47	1.71	4.99	4.63
E^A_{Cad}	0.0001	0.0001	0.0001	0.0001	0.39	λ_{Tll}	0.87	1.26	4.98	0.96	4.99
$E^A_{\mathrm{D-STAT}}$	19.9	19.9	0.0001	19.9	16.6	$T_{\mathrm{Bcd}}\dagger$	1.71	1.71	1.71	1.71	1.71
E^A_{Dichaete}	0.0001	0.0001	0.0001	0.45	0.004	T_{Cad}	2.53	2.22	3.06	2.06	3.0
E^A_{Hb}	14.4	13.5	20.41	29.9	19.1	$T_{\mathrm{D-STAT}}$	2.21	2.19	2.83	3.63	2.83
E^Q_{Hb}	0.99	0.99	0.99	0.99	0.99	T_{Dichaete}	2.22	4.92	4.79	2.96	2.08
E^Q_{Kr}	0.99	0.99	0.9	0.99	0.51	$T_{\mathrm{Hb}}\dagger$	0.63	0.63	0.63	0.63	0.63
E^Q_{Kni}	0.54	0.75	0.06	0.99	0.26	T_{Kr}	0.009	0.02	2.11	0.07	2.06
E^Q_{Gt}	0.75	0.43	0.72	0.74	0.99	T_{Kni}	2.2	2.48	2.23	4.85	2.46
E^Q_{Tll}	0.99	0.14	0.99	0.99	0.81	T_{Gt}	0.6	0.59	0.59	0.50	0.71
E^D_{Hb}	0.53	0.31	0.32	0.58	0.37	T_{Tll}	1.83	1.82	1.97	1.97	1.97
E^D_{Kr}	0.99	0.73	0.99	0.99	0.6						
E^D_{Kni}	0.24	0.14	0.12	0.99	0.05						
E^D_{Gt}	0.87	0.99	0.17	0.99	0.51						
E^D_{Tll}	0.0001	0.0001	0.0002	0.0001	0.98						
$D^Q_{\mathrm{all}}\dagger$	100	100	100	100	100						
$D^D_{\mathrm{all}}\dagger$	100	100	100	100	100						

Appendix C

Alignment matrices used in the model

Bicoid

A	83	74	108	48	6	381	379	5	0	6	72	61	65	68
C	114	159	127	149	1	0	0	0	383	340	136	174	166	158
G	106	72	114	11	0	2	4	4	0	3	132	60	52	49
T	80	78	34	175	376	0	0	374	0	34	43	88	100	108

Caudal

A	9	12	3	4	12	38	0	4	22	1
C	10	6	3	0	0	0	0	0	0	8
G	4	4	3	0	2	0	0	7	15	10
T	11	16	29	34	24	0	38	27	1	1

D-STAT

A	1	1	2	1	1	5	3	0	24	28	27	5
C	0	0	1	27	20	16	3	2	3	0	1	8
G	0	1	0	1	6	8	22	27	1	1	0	6
T	29	28	27	1	3	1	2	1	2	1	2	1

Dichaete

A	1	0	0	20	0	0	2	0	1	4	6
C	8	25	17	0	0	0	0	0	2	10	1
G	7	0	0	0	0	3	27	0	4	6	1
T	13	4	12	9	29	26	0	29	22	9	21

Hunchback

A	53	2	0	2	0	0	0	281	31	20
C	6	6	2	0	2	3	2	0	43	100
G	224	3	0	0	0	0	0	3	78	109
T	7	279	288	288	288	287	288	6	138	61

Kruppel

A	17	187	158	0	1	0	8	0	2	44
C	73	5	39	194	194	197	22	2	34	109
G	6	0	0	1	0	0	6	0	2	15
T	101	5	0	2	2	0	161	195	159	29

Knirps

A	19	25	16	5	0	21	0	17	1	0	25	5
C	1	1	0	9	4	0	0	0	3	26	0	12
G	2	0	0	6	1	5	26	8	18	0	1	7
T	4	0	10	6	21	0	0	1	4	0	0	2

Giant

A	86	12	776	8	83	0	1020	1106	15
C	62	108	25	762	19	556	88	0	378
G	19	359	275	65	996	0	1	0	85
T	942	630	33	274	11	553	0	3	631

Tailless

A	12	1	1	5	2	11	1	0	0
C	8	2	2	1	3	1	17	2	3
G	0	2	1	0	15	5	0	1	2
T	0	15	16	14	0	3	2	17	15

Appendix D

Regulatory sequences used for predictions

All DNA sequences used in this work are listed here. Index indicates the figure panel where the results of the prediction are shown. Name indicates the sequence designator used in that panel. DNA source gives the source of the sequence itself, and Reference where it was first described. We give the genomic position if known. Asterisks in the second column indicate that there were small differences between the regulatory sequences we utilized and the corresponding sequences available in FlyBase (http://www.flybase.org). The REDfly database is at http://redfly.ccr.buffalo.edu.

Table D.1: **Regulatory sequences used for predictions.**

Index	Name	Length (bp)	DNA source	Reference	Genomic position (bp)
3B1	eve_MSE2Mbcd1	489	[6]	[6]	
3B2	eve_MSE2Mbcd3	489	[6]	[6]	
3B3	eve_MSE3M2dsts	502	[36]	[36]	
3B4	eve_M32_MKr345	1016	[25]	[25]	
3B5	eve_M32_Mbcd1	1016	[25]	[25]	
3C1	S2E(yak)*	844	[132]	[132]	C2L:18492244,18493087
3C2	S2E(pse)*	1027	[132]	[132]	C3:10905710,10906728
3C3	S2E(ere)*	849	[132]	[132]	S4929:8503125,8503973
3C4	S2E(ore)	905	Text S2	This work	
3C5	S2E(tei)	882	Text S2	This work	
3C6	S2E(tak)	742	Text S2	This work	
3C7	S2E(mau)	797	Text S2	This work	
3C8	S2E(sec)*	788	Text S2	This work	S359:11527,12271
3C9	S2E(per)	753	Text S2	This work	S4:6229414,6230166
3C10	S2E(sim)*	798	[132]	[132]	C2R:4497397,4498185
3C11	S2E(ana)*	816	Text S2	This work	S13266:15364458,15365264
3C12	S2E(vir)*	973	Text S2	This work	S12875:1336908,1337886
3C13	S2E(pic)	1036	[133]	[133]	
3C14	S2E(gri)	1065	Text S2	This work	S15245:9655365,9656429
3C15	S2E(moj)	1089	Text S2	This work	S6496:4429248,4430336
3C16	S2E(wil)	1100	Text S2	This work	S180700:33743,34842
3D1	MSE5	804	Redfly	[28]	C2R:5498538,5499341
3D2	MSE4_6	800	Redfly	[28]	C2R:5495712,5496511
3D3	M3_S2(p1-m2)	1544	[99]	[99]	
3D4	M3_S2(m1-p2)	1433	[99]	[99]	
3E1	S2E(cyn)	1939	[102]	[102]	
3E2	S2E(put)	1698	[102]	[102]	
3E3	S2E(sup)	1791	[102]	[102]	
3E4	S2E(dsp)	2437	[102]	[102]	
3E5	S2E(min)	1579	[102]	[102]	
3E6	S2E(pun)	2034	[102]	[102]	
3E7	S37E(cyn)	2044	[102]	[102]	
3E8	S37E(put)	1682	[102]	[102]	
3E9	S37E(sup)	1887	[102]	[102]	
3E10	S37E(dsp)	1540	[102]	[102]	
3E11	S37E(min)	1575	[102]	[102]	
3E12	S37E(pun)	2120	[102]	[102]	
3F1	hb_pThb1_hbp	298	Redfly	[134]	C3R:4520323,4520620
3F2	Kr_CD1_hsp70p	1159	Redfly	[135]	C2R:20730219,20731377
3F3	run_str1_7	1611	Redfly	[114]	CX:20490688,20492298
3F4	run_str3_7	2404	Redfly	[114]	CX:20493864,20496267
3F5	h_str3_4	1745	Redfly	[103]	C3L:8637477,8639221
3G1	eve_ups	3942	FlyBase	[31]	C2R:5487187,5491128
3G2	eve_downs	3500	FlyBase	[28]	C2R:5496129, 5499628

Index	Name	Length (bp)	DNA source	Reference	Genomic position (bp)
S5C	hbHZ1400_hsp70p	1421	Redfly	[134]	C3R:4526522,4527942
S5D	kni_+1	1479	Redfly	[136]	C3L:20533736,20629274
S5E	kni_kd	875	Redfly	[137]	C3L:20630383,20631257
S5F	gt_(-1)	1239	Redfly	[136]	CX:2285171,2286409
S5G	gt_(-3)	1209	Redfly	[136]	CX:2286417,2287625
S5H	run_5	1340	Redfly	[114]	CX:20492298,20493637
S5I	h_str1	876	Redfly	[138]	C3L:8644872,8645747
S5J	h_str2_6	1081	Redfly	[103]	C3L:8640258,8641338
S5K	h_str5	564	Redfly	[139]	C3L:8644027,8644590
S5L	h_str6	547	Redfly	[140]	C3L:8640797,8641343

Appendix E

Full S2E sequences first identified in this dissertation

S2E sequences first identified in this dissertation are listed in FASTA format. All of these sequences drive gene expression at *eve* stripe 2 position in the *in silico* transcription system. Three letter codes in the parentheses indicate species abbreviations. See Table 4.1 for the full names and see Section 4.2.3 for the identification method.

>eve_S2E(per)

AATATAACCCAATAATTTTAACTAACTCGCAATGGACAGGGCAGTAGAGCAGTAGAGCATTG
CAGGAAGGATGCATTACTCGGGAATGGAATGCATAACAATGGGCAAGGACCAGGGTTCCGTT
TCGCGAGATGAGGTTCTTTGACGGTTCCTTGACGGTTCCCTGTGTGCTCTCTGCTCTGTGTT
AATCCGTTTGCCATCAGCAAGATTATTAGTCAATTTTCATATTTCCAGTCGAGTCGCAGTTTT
GGTTTCACTTTCCTCCTTTGCCACTTCTTGCCTTGCCTCATGTGGATGCCGATGCCGATGCCG
TTGCCGTTGCCGTTGCCGACCGACGAGTTAGATTTTATTGCAGCATCTTGAACAATCAACTG
GAATTTGGTAACATGCTGCGCGGCCTAACCCTGGAGATTGCTCTACTTTCGCCTCAATTGAAT
CGGAGTTAGGCGGAAGACGGCGGACCCTTGCAACCAAGGGTTGTCTCCTGGCCTCAGGAGTT
TCCACAGTCAACGCTTTCGCTGGTTTGTTTATTTGTTTGTTTGTTTTAGCCAGGATTAGCCCG
AGGGCTTGACTTGGAACCCGACCAAAGCCAAGGGCTTTAGGGCATGCTCAAGAGATCCCTAT
ATCCCTATCCCTGTCGCGATCCCTAAACCGATCCCATTTAGCAATTTCATTAGAAAGTCATAA
AACACACATAATAATGAGATGTCGAAGGGATTAAGATTAAGGGACGCACACACAGGCAGCAG
GATC

>eve_S2E(gri)

AATATAACCCAATAATTTGAACTAACTCACAGCAACAACAACTGGGAGAGTTACTTAGTAAT
GCATAACAATAAGTTGAGGCTGAAATTGAGACTGAAATGCTGTTTGCCGAAGGTTTTCAGCC
ACAACGTTTTTCCAAGGGTTCTCATCGGCATTGACTGGTTCAGAATCCTGTGCGTTAATCCGT
TTTGCCATCAGCGACATTATTACTCTATTTTCCATTTCTCTCTAAAATTTGAACATTTTCTCA
ACCGTTTGCATATCCATTTCCATTTTCATTTTCCATTTTCACTTTCGCCTGCGGATACGAGTT
AGATTTTATTGCAGCATCTTGAACAATCGCCTCAACTCGAACTCGAACTCGAACGCGAACTCG
GAATTGGATGTGCAGTTTTGGAACCATGCTGGGTTTTGTTTGCTGTCATTGCTCTAGTTTTG
CTTTCATTTCCTCACTCTTAGCTGGTGATTTTTAGGCAGAATTCCGCTGTCCTGGCATTGTCA
TTGAATCGCCGACCGGGTTACCCTCAAACTAGGTTTGAGTTTAACTTTCAACTTTAACATTAC
CAAAACCGACTTCAACTCCATTTTCGACTTTGCTGGCGGAGTTTCCACATGCCTCCGTTTTTG
TTTATTTGTTTGTTTTGCGGATTTAGAGATTAGAGAAAAGGGGCCAATGGCTTTAGACTGAT
GCCTGATCTGCTCGCCTTTTCATTAGAAAGTCACACAAACGCATAATGATGACAAGGGGGAT
TAAGCTCACATACCTACACATAACCTAATTAGCGGATTTACAAATTGGTTTATTTTTTCCCTT
TTTTTTATTGTAGTTGGTCTGCCCCAGCTTAAACCCAAGCCACTGCATCAGGCCAATCCAAAA
AACCCGAGCAGGTATCAACTTACGCAAGCCAAAAAAAAAAAGACAAAAGACAAAGAAAATATC
ATAATAACATTTCAGAGTTGGCAGCAACTCAGTTTCAGCGCCAGTAACTGCTCCCGGCCAGT
AACTGCTCTCTGGGTACAGTTGCGCATTTTGGGCAACATGATTATATCATCATAATAAATGTT
T

>eve_S2E(moj)

AATATAACCCAATAATTTTAACTAACTCACAACGACAACAACAACAGCAACAACAACAACAAC
AACAACAACTGGTTGAGTTACTTAGTAATGCATAACAATGAGAGCGAGAGACAGTGAAACCG
AAATTGAAATTGAAATTGAGCGAGATCTGTAGGTTGAAGGTTTCCTTCATCCAGCCATCCAT
CCATCCAGCCATGGGTTTGGCATTGAGCTGCTTGTGCGTTAATCCGTTTGCCATCAGCGACA
TTATTAGTCGATTTTACAAAGATTTTGAGCAAACAGTTTTCACTTTCGAGTTAGATTTTATTG
CAGCATCTTGAACAATCGCTGCGTCAGAGACAAACTAGGAAATTGGATGTGGATTTTGGACA

CACGCTGTGTCTCTCTACTCTCAGTCTGCACTCAGTCGCCGTGCTCTAGTTTCGCTCGACTAA
GGTGATTTAGTTGGAATTGATGTCATGTCATTGTGTTTGACTCCGTGTTTGACCGGGTTACC
CTCAGGCAGGCGACTTTAACTTTCAGTTTAAGCGGCGACTTTAACTCCCTTTTCGACTTTAGC
TGCTGGCATTTGACGGGCCTTTCCGCAAAGGCCGCAACATCTTTGTTTGTTTGTTTATTTAGT
GGATTAGACAGCAGAGAGCGAGGGAGAGGCAGAGAGAGAGGGCTGAGCAGAGCGGGGCTTG
ACTTGGTTGAACTTTTGGCGAACGGCTTTAGCCGCGCTTGATCCATAAGTCATAAACAGACA
TAATGATGACAAGGATTAGCTAAACACACACACACACGCATATATATATGGCATAATTAGCG
TTTTCTGAAATTTGGTTTATTTTTGCCTTTTCTCCCTGAACAGCTTAAGCCTAAGCCAAAACC
CTAGCCCAAAAACCCGCGCAGGTATCAAACACGCTTACGCAAGCAAAAAAAAAAAAGGAAAA
TAAAACAGCAAATACAACAAAAAACAAAAACTGGCAGCACTTTAGCTCGCTTTGCCTCCAGAC
GCGCCCCTCTCAATTGGCGCCCCCTTGTAACTGCGCTCTGGGTACAGTTGCGCCTTTAGGCAA
CATGATTATATCATCATAATAAATGTTT

>eve_S2E(wil)

AATATAACCCAATAATTTTAACTAACTCATGGAATGGGCAAACTAGAGCAGGAAGGATGCCT
TAGTTACTTGGGAACTATGCTGAGGTAACAATAGGCCCTCTGGGCAATGGTTGAAGGTTACG
TTAATCCGTTTGCCATCAGCAAGATTATTAGTCAATTTCAGATGAGTTTTTCACTTTTCCTCA
TCGTTGTTGCCTTTCGCTTTCCGCGGCGTTCCGACGAGTTAGATTTTATTGCAGCATCTTGAA
CAATCAACTTGGATTTGGTAACATGCTGCGCGATGTGCTCTCAATTTTTCCTTTCAATCCATA
TAGATGTATCCTTTGCATTTATAGAGAATTTTACCTTGACAAAGCCAAGCCAAAAGCAAACTA
AATGTAACCAAGAGGCAATCGACTGACCGGGTTGCCTCTGGAGTTTCCACAACTTTGTTTAT
TTGTTATTTGTTTGCGGGATTAACTAGTCCTAGGGGCTTGACTTGAACTCTCTATCCCTGATC
CCATTTGGCAACTTCATTAGAAAGTCATAAAAATGCATAATGATGTCGATGGGATTAGATGG
GAATGGGAAGCGGGATGGGTCAGGTAGAGTAACCCCATCCAAACCGTTGGGCACTTGCTCCA
TCTTTAGCTGAAAGTACAGTTGCCACCACATCAAGACGCACATGATTGTATCATCATAATAAA
TGCTTTTCCTGAAACGGAAACTCTTCCCCCCGACTCCTCCCATCTCTCTTTCCCACAAAAACC
AAACAAAATGGAAACTTTTCAAATTAAGTTTTCTAGCTACCCAAAAAAAAAAACATAAAAGCA
AAGCCAAGTAAATATTTCATTATAATGGACATACACAAAAATGGTTACATTTTGGTGGGGGG
AGGGGGGTTCGAAAACTTTGGGTTCATCCCTGGGACTGCGACTTCATCAAGTGTGAGTCTGT
TAAACGTGCGGAATATTAAGACTTCATAAAAGGCGCAAATAATTTGGTTCACTGACGATCGA
TACTCTCAACCCAAACCCAGACCCATACCCAGACCAAGACCAACCACCAGCTATGACTTTGAC
TCTGGCAGAGAAGGCATTATGCCATATCATTCTTGAT